QUALITY ENGINEERING IN PRODUCTION SYSTEMS

GENICHI TAGUCHI

ELSAYED A. ELSAYED

THOMAS HSIANG

QUALITY ENGINEERING IN PRODUCTION SYSTEMS

Genichi Taguchi
International Consultant

Elsayed A. Elsayed
Professor and Chairman,
Department of Industrial Engineering,
Rutgers University

Thomas C. Hsiang
Director of Statistical Services
Universal Foods Corporation

McGraw-Hill Publishing Company

New York St. Louis San Francisco Auckland Bogotá Caracas
Hamburg Lisbon London Madrid Mexico Milan
Montreal New Delhi Oklahoma City Paris San Juan
São Paulo Singapore Sydney Tokyo Toronto

This book was set in Times Roman by Publication Services.
The editors were John Corrigan and John M. Morriss;
the cover was designed by John Hite;
the production supervisor was Friederich W. Schulte.
Project supervision was done by Publication Services.
R. R. Donnelley & Sons Company was printer and binder.

QUALITY ENGINEERING IN PRODUCTION SYSTEMS

3 4 5 6 7 8 9 0 DOC DOC 9 3 2 1 0

ISBN 0-07-062830-0

Library of Congress Cataloging-in-Publication Data

Taguchi, Genichi, (date).
Quality engineering in production systems.

(McGraw-Hill series in industrial engineering and management science)
Bibliography: p.
Includes index.
1. Quality control—Statistical methods. 2. Production management—Quality control. I. Elsayed, Elsayed A. II. Hsiang, Thomas C. III. Title. IV. Series.
TS156.T33 1989 658.5'62 88-8882
ISBN 0-07-062830-0

ABOUT THE AUTHORS

Dr. Genichi Taguchi is an international consultant in the field of quality control and assurance. He has served as Director of the Japanese Academy of Quality from 1978 to 1982. He was awarded the Deming Prize in 1960 in recognition of his contributions in the development of various techniques for industrial optimization while he was on staff at Electrical Communication Laboratories of Nippon Telegraph and Telephone Public Corporation. From 1964 to 1981, he served as a professor at Aoyamagokuin University. Since 1981, he has served as a full time consultant to various companies such as AT&T, Bell Communications Research, Ford, Xerox, and many companies in the United States, Japan, Taiwan, and the People's Republic of China. Throughout the many years of his career, Dr. Taguchi has developed methods for on-line and off-line quality control which form the basis of his approach to a total quality control and assurance in a product's development cycle. Dr. Taguchi received his Doctor of Science from Kyushu University in 1962.

Dr. Elsayed A. Elsayed is Professor and Chairman of the Department of Industrial Engineering, Rutgers University. His research interests are in the areas of quality and reliability engineering, production planning and control, and automated manufacturing systems. Dr. Elsayed held teaching and research positions at Cairo University, University of Windsor, and University of Utah. In 1987–1988, he spent his sabbatical with AT&T Bell Laboratories conducting research on reliability modeling and analysis of semiconductors. He served as a consultant to Bell Communications Research, Sea-Land, Personal Products, AT&T, and other manufacturing companies. He is a coauthor of "Analysis and Control of Production Systems," Prentice-Hall, 1985. He is a senior member of IIE, ASME, SME, and ASEE and is listed in Who's Who in the East, and Who's Who in Engineering. Dr. Elsayed received his Ph.D. in industrial engineering from University of Windsor in 1976.

Dr. Thomas Hsiang is the Director — Statistical Services of Universal Foods Corporation, Milwaukee, Wisconsin, where he plays an integral role in training, consulting, and developing the implementation of all needed elements of the Total Quality Effort throughout the company. He previously held positions in Quality and Statistics with various divisions in Bell Canada, AT&T, Bell Laboratories, and Bell Communications Research, where he served as a manager, teacher, and consultant in using statistics for the improvement of quality and productivity. He also taught at the Department of Statistics, Rutgers University, as an adjunct professor. He has had numerous publications in quality and statistics. He was co-winner of the Best Paper Award of the 1982 American Society for Quality Control Congress. A native of Szechuan, China, Tom was educated in Taiwan and Canada, earning his B.S. (1965) in Chemical Engineering from Tunghai University; M. Eng. (1967) in Chemical Engineering from McMaster University; and M. Math (1969) and Ph.D. (1971) in Statistics from the University of Waterloo. Tom is a member of the American Statistical Association (ASA) and the American Society for Quality Control (ASQC), and he is a registered Professional Engineer.

CONTENTS

PREFACE

The traditional role of quality control is basically to eliminate from production lines those parts that do not conform to specifications, and to inspect and test–finish products for defects. Given this definition, quality control is almost limited to inspecting and testing on a detailing or sampling basis.

The increased emphasis on "higher quality" products at lower costs, combined with the competition from overseas for U.S. markets, has magnified the importance of quality control. Consequently, quality control activities have been redefined to ensure the quality of the product during every phase of its life cycle. This book describes quality control from an engineering standpoint. It is intended to amplify the concept of "on-line" quality control as applied during production.

Chapter 1 provides a comprehensive discussion on the value of product quality and the relationship between product quality and price. The chapter also introduces the concept of an Overall Quality Control System, including its components and applications.

Chapter 2 covers a "loss function" approach as a measure of quality, and its use in determining product specification, target values of product characteristics, and desired tolerance(s) relevant to each target value.

The aim of Chapter 3 is to discuss tolerance design as well tolerancing for each of three basic types of variable characteristics.

Chapter 4 presents "on-line feedback quality control" and also illustrates the use of the "loss function" for the design and evaluation of a feedback control system.

Approaches for "on-line process parameter control" for variable characteristics are examined in Chapter 5.

Chapter 6 provides approaches for an on-line quality control system for attribute characteristics, including techniques for determining the optimal diagnosis interval and its effect on the quality loss per unit of production.

Chapter 7 considers different methods for improving the parameters of the production, diagnosis, and adjustment processes to minimize the total quality loss.

Chapter 8 discusses the use of preventive maintenance as a means of improving the parameters of the production process to reduce the quality loss of production.

Examples are provided throughout this book to reinforce and illustrate the concepts, methods, and approaches involved. The end of each chapter contains problems to facilitate practice.

This book is intended primarily for those involved in the research, development, and manufacturing phases of a product's life-cycle, including design engineers, production and manufacturing engineers, quality engineers, and applied statisticians. It is also intended for senior undergraduate students and first year graduate students in industrial-, production-, manufacturing-, and systems-engineering disciplines. Students should have a familiarity with basic statistical quality control.

Although we have included a few results of our own research, this textbook is not solely the work of the three authors, and the many references throughout the book reflect our indebtedness to others. We have attempted to give appropriate credit to everyone involved. We would like particularly to acknowledge the Japanese Standards Association for permission to use certain material in Chapters 6 and 7.

The interest in this book was originated by Thomas Hsiang when he worked for Bell Communications Research, Inc. (Bellcore) as the District Manager—Quality and Reliability Engineering at the Quality Assurance Operations Center. During his tenure, Genichi Taguchi and Elsayed A. Elsayed served as consultants to the Quality Assurance Operations Center. We are grateful to Bellcore for the continued editorial, technical, typing, and drafting support. In particular, we wish to express our thanks to Pete Pence and Norm Sherer for their support during the initiation of this work; Jen Tang for his technical review; Loinel Howard, Jr., and Jim Falk for their assistance and coordination in making the publication of this book possible; Donald Rector and Pamela Richardson for their editorial work and Christine DeHanes for her managerial efforts in the professional typing, editing, formatting, and drafting of this work.

Additional thanks are due to Professor Yu-In Wu for technical discussions, to John J. Gordon (formerly with AT&T Technologies, Inc.) for editorial efforts and to Wayne B. Clark and Universal Foods Corporation for interest and support of this endeavor.

We would also like to thank Anne Duffy and John Corrigan of McGraw-Hill for their efforts in facilitating the review of the book and bringing the project to fruition. Our great appreciation to the following reviewers who provided many useful comments and suggestions: Jeya Chandra, The Pennsylvania State University; Frank Kaminsky, University of Massachusetts; J. Bert Keats, Arizona State University; Hau Lee, Stanford University; Joseph Mize, Oklahoma State University; Ronald Snee, Dupont Company; and Dennis B. Webster, The University of Alabama.

Finally, we reserve the greatest thanks for our families for their encouragement and support.

Genichi Taguchi
Elsayed A. Elsayed
Thomas Hsiang

QUALITY ENGINEERING IN PRODUCTION SYSTEMS

CHAPTER 1

QUALITY VALUE AND ENGINEERING

Prior to World War II, the quality of Japanese-made products was poor. Prices were low, and it was difficult, if not impossible, to secure repeat sales. A simple comparison between the quality of products made in the United States and Japan at that time can be illustrated by the following example: Assume that a product made in Japan was one-half the price of a similar product made in the United States. If losses incurred by the customer through use of the Japanese-made product were nine times its purchase price, and if losses incurred by the customer through use of the U.S.-made product were equal to the price of the U.S.-made product, then the total loss to the customer because of the purchase of the Japanese-made product would be $2\frac{1}{2}$ times the loss caused by the purchase of the U.S.-made product. This would suggest that the U.S.-made product was $2\frac{1}{2}$ times superior in quality to the Japanese-made product, according to the following calculations:

	U.S.-made product	Japanese-made product
Purchase price	P	$0.5P$
Losses due to product use	P	$9(0.5P)$
Total cost to customer	$2P$	$5P$

The trade-off between quality and price is an important subject. Product quality is affected by its tolerance design; we therefore present a detailed methodology of tolerance design in Chap. 3.

To further illustrate the importance of trade-off between product quality and price, we quote an expression used by many executives, "The quality of our product is excellent, but the price is too high." As indicated earlier, there must be a balance between quality loss and product price. The price represents the loss to the customer at the time of purchase, and poor quality represents an additional loss to the customer during the use of the product. A goal of quality engineering should be to reduce the total loss to the customer.

Applying the quality-price trade-off, we need to predict quality loss at the product design, process design, and production phases. A product is usually priced at several times the unit manufacturing cost (UMC). Therefore, UMC becomes a very important factor in the trade-off analysis. In terms of quality loss, a monetary prediction is necessary. Many companies still use percent defective as a measure of quality level. Defective products, however, are usually not shipped. Only the products that are shipped cause quality problems to consumers. For this reason, the loss caused by unshipped defective products should be considered as a cost, not a quality loss.

In this book, quality loss is defined as the loss a product costs society from the time the product is released for shipment. Losses such as failure of function, harmful effects, pollution, operating costs, and maintenance costs are all included. Chapter 2 explains how to evaluate quality loss in terms of monetary units, assuming product tolerances are correct. The concept of loss function is also introduced. Chapter 3 discusses a method for specifying tolerances; it is intended especially for product designers. However, tolerance is so important that production engineers and production personnel also need a good understanding of its meaning. Chapters 4 through 8 deal with various on-line quality control and engineering aspects in production, including preventive maintenance.

It is important to note that company-wide activities are needed to improve quality and productivity. These activities are

1. Product planning
2. Product design
3. Process design
4. Production
5. Service after purchase

The remainder of this chapter provides an overview of the various activities that constitute an overall quality system.

1.1 AN OVERALL QUALITY SYSTEM

The following sections describe a company-wide quality control system. The overall system involves the quality concept and quality cost through all phases of

a product's life cycle. The life cycle begins with product planning and continues through the phases of product design, production process design, on-line production process control, market development, and packaging, as well as maintenance and product service. This section discusses the concept underlying an overall quality control system, that is, *what* constitutes an overall quality control system, and *how* it affects the quality of the products produced.

From the standpoint of value received, product quality is determined by the economic losses imposed upon society from the time a product is released for shipment. A typical example is loss caused by *functional variation:* the deviation of one of a product's principal functional characteristics from the specified "nominal (target) value" of the product design specification. If process design and quality control engineering are not capable of sufficiently reducing deviation by process adjustments, then inspection may be an economically useful alternative.

Product design engineering may in itself operate as a primary factor in major losses by including tolerances that result in assembly misfits or special production processes that are unnecessary. Similar examples can be found in other phases of the product life cycle, such as market development, packaging, or product maintenance.

Ideally, an integrated system of overall quality control, in which all activities interact to produce products with minimum deviations from target values, will minimize quality costs and make the most economic use of human and other company resources. This system, which aims to achieve controlled production of products with superior quality, can be called an *overall quality system*.

As discussed earlier, quality loss, in broad terms, is defined as the losses imposed on society from the time a product is released for shipment. The discussion in this book is specifically concerned with the losses caused by deviation in a product's functional characteristics from their specified nominal values (desired target values).

The two types of undesirable and uncontrollable factors that can cause deviation from target values in a product's functional characteristics are known as external and internal noise factors. Operating environment variables (such as temperature or humidity) are examples of external noise factors. There are two categories of internal noise factors. They are

> deterioration, such as the wearing out of parts caused by friction and the loss of spring resilience, and
>
> manufacturing-process imperfections, such as variations in machine setting

The broad purpose of the overall quality system is to produce a product that is *robust* with respect to all noise factors. Robustness implies that the product's functional characteristics are not sensitive to variation caused by noise factors.

In order to achieve robustness, quality control efforts must begin in the product design phase and be continued through production engineering and production operation phases. During the product design and production engineering phases, these three steps must be followed:

1. **System design**
   This step depends on the product phase during a life cycle. For example, during the research and development phase, system design involves the development of a prototype design and determination of materials, parts, components, and assembly system. In the production engineering phase, the determination of the manufacturing process is involved.
2. **Parameter design**
   In this step, the levels (values) of controllable factors (design parameters) are selected to minimize the effect of noise factors on the functional characteristics of the product.
3. **Tolerance design**
   This step applies if the reduction in variation of the functional characteristic achieved by parameter design as described above is insufficient. Narrow tolerances are then specified for the deviations of design parameters in relation to the levels determined by the parameter design.

These three steps have been widely used in many industries for improving the quality of products. Quality control activities at the product planning, design, and production engineering phases will be referred to as *off-line* quality control or quality engineering, whereas the quality control activities during actual production will be referred to as *on-line* quality control.

Taguchi and Wu (1979) summarize these activities for a typical manufacturing facility in Table 1.1. An asterisk (*) as in the fourth column of the table indicates that the external noise can be controlled at that step in the product's life cycle. A plus (+) indicates that it is not preferable to control the external noise

TABLE 1.1
Functional quality engineering activities

Quality control activities	Product phase	Steps	External noise	Internal noise	Variation among units (tolerances)
Off-line	Product design	System design	*	*	*
		Parameter design	*	*	*
		Tolerance design	+	*	*
	Production engineering	System design	+ +	+ +	*
		Parameter design	+ +	+ +	*
		Tolerance design	+ +	+ +	*
On-line	Production operation	Process control	+ +	+ +	*
		Feedback	+ +	+ +	*
		Inspection, etc.	+ +	+ +	*

Key:
* = possible
+ = not preferable
+ + = impossible

at that step. A double plus (+ +) indicates the impossibility of controlling the noise at a step.

1.2 QUALITY ENGINEERING IN PRODUCT DESIGN

As indicated earlier, product design has the greatest impact on product quality. It is essential to consider all aspects of the design (including factors built into the product) that affect the deviation of functional characteristics of the product from target (nominal) values. It is also necessary to consider methods to reduce the undesirable and uncontrollable factors (such as noise) that cause functional deviations.

The three steps—system design, parameter design, and tolerance design—are applied to a design of a product as described below.

1.2.1 System Design

System design denotes the development of a basic prototype design that performs the desired and required functions of the product with minimum deviation from target performance values. It includes the selection of materials, parts, components, and the assembly system. For example, the design of an electrical circuit for a television set that converts an input of 100-V alternating current to 115-V direct current requires a search for the technically best circuit that is specifically relevant to this design. An automatic control system might be included in the design of the circuit so that a target value of the desired voltage (115 V) is set, and then continuous measurements of the output power of the circuit are taken. If there are deviations between the measurements and the target value, the automatic control system should change the relevant parameter in the circuit. For instance, it may change the resistance value of a rheostat so that the difference between the target value and the measured output voltage is reduced to zero.

In brief, the system design of such an electric circuit requires an appropriate circuit design.

1.2.2 Parameter Design

Once the system design is established, the next step is to ascertain the optimal levels for the parameters of each element in the system so that the functional deviations of the product are minimized. As an illustration of parameter design, consider the design example of the electric power circuit for a television set with the capacity to convert an input of 100 V AC to an output of 115 V DC. After the selection of a prototype circuit (system design), it is necessary to determine the optimal levels of the circuit parameters.

Consider an example where 100 V is supplied to the prototype circuit but an output of only 80 V is obtained. To reduce the gap of 35 V (difference between the target voltage and that actually measured in the prototype circuit), the parameter h_{FE} (transistor gain) of a transistor used in the circuit is set at a different level.

The effect of h_{FE} on the output voltage is shown in Fig. 1-1. Since the value of h_{FE} varies considerably during its life in the circuit, choosing its level to be A' in order to reduce the gap will result in a circuit with significant exposure to output voltage deviation from the target value. As shown in Fig. 1-1, when a low-grade transistor is used ($h_{FE} = A'$), the value of h_{FE} would be likely to vary as much as 30 percent of the mean value, resulting in variations in the output voltage as large as the interval $Y_1 - Y_0$.

Therefore, the level of h_{FE} should be chosen at A_0, because variations about A_0 will have minimal or no effect on the output voltage. This choice will reduce the difference between actual and nominal (target) voltage to 20 V, which must be eliminated by changing another circuit parameter. Assume a resistor used in the circuit has a linear effect on the output voltage and that an increase of 1 kΩ in resistance decreases the voltage by 5 V. The gap will be diminished by choosing a value of the resistor 4 kΩ larger than that currently present in the circuit.

Determination of the optimal levels of parameters is an off-line process and is usually accomplished by using an experimental design approach. The end result of this design step is to determine the optimal combination of levels of parameters and all components of the prototype, that minimizes or diminishes the effects of various noises while keeping performance as close as possible to its nominal (target) value.

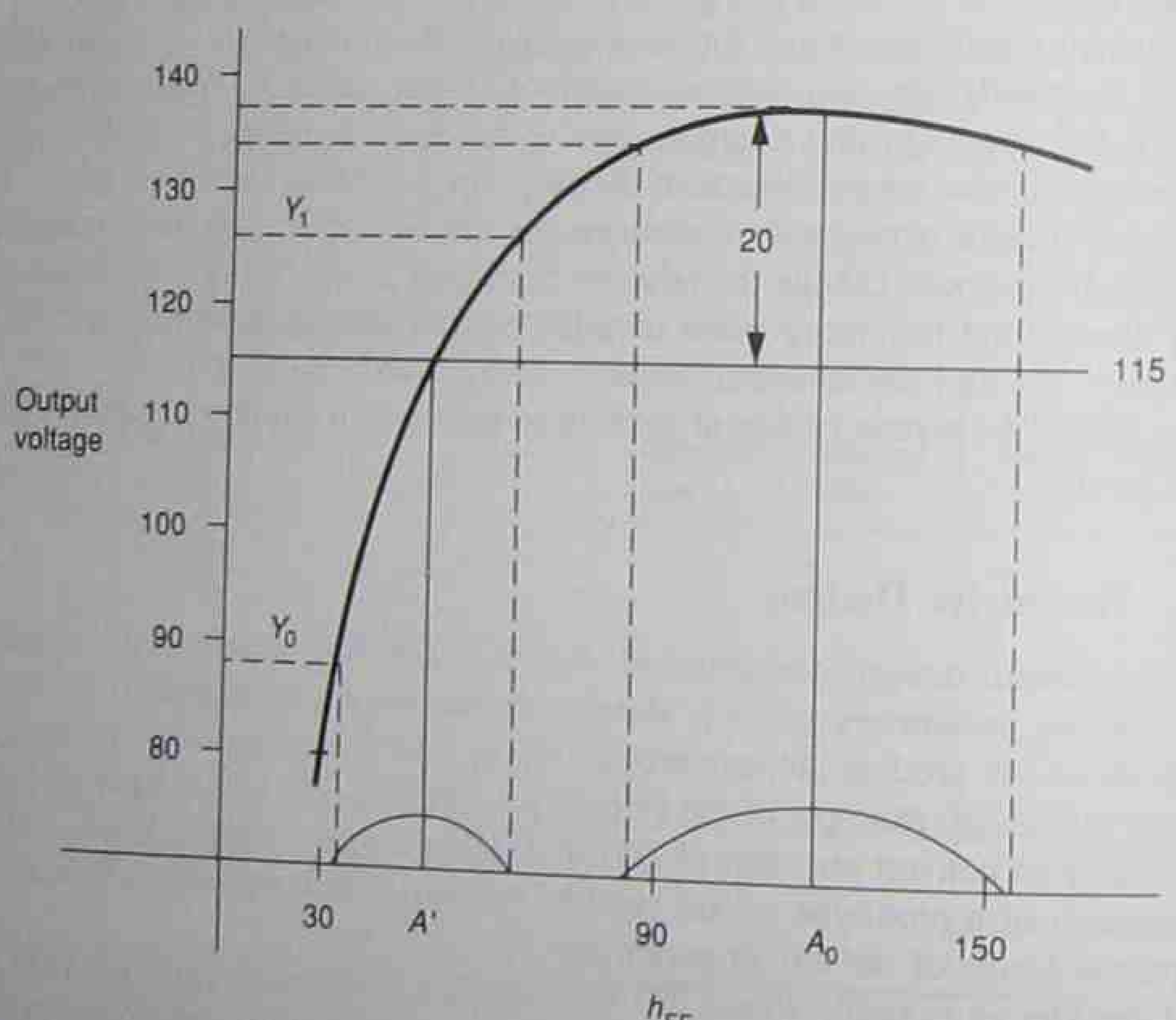


FIGURE 1-1
Effect of h_{FE} on the output voltage.

1.2.3 Tolerance Design

Once the system design is completed and the optimal values of the parameters of the elements (components) are obtained, the next step is to determine the tolerance of each individual parameter (factor) by trading off quality loss and cost. In effect, it is necessary to define allowable ranges for deviation in the parameter value. Obviously, the narrower the range of deviation, the more costly the product becomes, as a result of increased manufacturing cost. On the other hand, the wider the range of deviation, the larger the deviation in product function from a specific target value. The tolerance design step determines the most economical tolerances: those that minimize product cost for given tolerable deviation from target values. In Chap. 3, approaches are examined for optimal tolerance design, taking into consideration losses caused by deviation from nominal values given by the parameter design as well as costs of different grades of components.

1.3 QUALITY ENGINEERING IN DESIGN OF PRODUCTION PROCESSES

When the design and fabrication of the prototype with optimally determined tolerances and specifications are completed, product manufacturing proceeds with three design steps. These are similar to the three steps of the product design phase, namely system design, parameter design, and tolerance design. These steps are now explained in terms of how they are applied to the design of production processes.

1.3.1 System Design

System design (as related to production processes) determines the manufacturing processes required to move a workpiece from partial completion to a more advanced stage of completion. During the manufacturing processes, energy is added to the workpiece in order to change its shape, remove material from it, or change its physical properties or function. There are generally many manufacturing processes that can perform the same function on a workpiece. For example, metal removal can be performed by using turning operations, milling operations, or shaping operations.

The main objective of system design is to determine the manufacturing processes that can produce the product within the specified limits and tolerances at the lowest cost. This system design function is usually performed by manufacturing, production, and industrial engineering.

1.3.2 Parameter Design

Parameter design in production process design determines the operating levels of the manufacturing processes so that variation in product parameters is minimized. Typical examples of variations in the operating levels of the manufacturing processes include temperature variation, raw-material variation, input-voltage variation, and tool-condition variation. These variations, as well as several unidentified noise factors, can cause nonuniformity in the production processes, resulting in out-of-specification products or nonuniform production output. The nonuniformity of the production processes can be minimized by determining optimal levels for the parameters of the processes. Operating levels can then be shifted to points where the effects of process variations on the product are minimal.

The parameters that affect the performance of the production process are established during the "test setup" run. Consequently, production process design is classified as an off-line quality control process, and experimental design approach is used to determine the optimal levels of the parameters for the process.

1.3.3 Tolerance Design

Once the optimal operating conditions for each element of the production process have been determined, the allowance ranges for changes in operating conditions and other variables are determined. As mentioned earlier, the narrower the range of the operating conditions, the smaller the variation or nonuniformity of the product, with a necessary increase in production-process cost. Thus, the objective is to find optimal ranges of the operating conditions that minimize the sum of variation cost and cost of the product. This is the on-line feedback control system design problem explained in this book.

1.4 QUALITY ENGINEERING IN PRODUCTION

On-line quality control as explained in this book deals with the daily activities to control process conditions by observing either quality characteristics of products or process parameters. The methods used in such activities are extensions of engineering methodologies called feedback control, feed-forward control, and calibration. It is known that all processes will drift if control is not applied. Therefore, the purpose of on-line quality control is to produce uniform products by adjusting processes according to information about the process and/or the product produced. Based on this information, a solution to minimize quality loss or cost can be devised.

One observation is usually enough to control the process for each period, even in attribute cases. A typical example of using one observation to control a process is the error control of a watch. The error control of a watch can be accomplished by periodically checking the error with the standard time signal and correcting it when the magnitude of error exceeds some control limits.

The error of the watch is a function of the controlling methods. One does not need to take a sample of the error of time, but it is necessary to predict the error for the controlling method devised. Consider an example where there are two watches, and the quality of one is good, while that of the other is poor. The good-quality watch has an error of 30 seconds a year, while the poor-quality watch has an error of 30 seconds a day. If the poor-quality watch is calibrated daily, and

the good-quality watch is calibrated yearly, there would be no difference in the quality of the two watches. Quality is a function of not only design, but of the control system.

Without controlling the process, it is not possible to control a product's quality. How often should we observe the process or product, and what are the optimal control limits? To answer this question we have to start with the prediction of quality level, which becomes a most important step in on-line quality control. Chapters 4 through 8 discuss the quality prediction and control issues in more detail.

1.5 QUALITY ENGINEERING IN CUSTOMER SERVICE

Despite tight controls applied in the design and production steps, some defective products find their way to customers. Such defective products may create problems in subsequent processes or may result in liability claims by consumers once the product reaches the market. Appropriate postmanufacturing service must be provided for cases where consumer claims are justified.

The sales department of the manufacturer has an obligation to provide adequate service to customers with justifiable claims. This service should take the form of repairing or replacing the defective products and compensating for damages that the customer may have incurred.

1.6 SUMMARY

This chapter introduced a definition of quality as the loss incurred due to deviations of product characteristics from their target values, and also introduced the concept of the overall quality system. Three basic steps were discussed to achieve robustness of the product and the production processes, namely system design, parameter design, and tolerance design. These steps constitute what we refer to as off-line quality control.

We conclude this chapter by emphasizing that quality of products or performance of processes must continuously be improved so that the deviations of the product characteristics from the target values are minimized.

PROBLEMS

1.1. Quality is viewed differently by different people. How would you view it if you were
(*a*) a customer?
(*b*) an engineer who is responsible for a product line?
(*c*) an applied statistician in a manufacturing company?
(*d*) the president of a manufacturing company?

1.2. Propose a new definition of quality and an appropriate measure for it.

1.3. Describe the elements of a total quality system in a manufacturing company. What would the differences be in a service company?

1.4. There are three basic steps to be applied in each phase of the product life cycle; these steps are system design, parameter design, and tolerance design. Show how these steps can be applied for the manufacturing processes of a product.

1.5. What are the main differences between off-line and on-line quality engineering systems? Can some of the activities of these two systems overlap? If so, can you determine the areas where each system must be restricted?

1.6. The receiving department of a manufacturer inspects incoming parts. The inspection is limited to measuring the principal dimensions of the parts. If the dimensions of a measured part do not meet the required specifications, the part is either discarded or returned to the supplier; if the requirements are met, the part is accepted. The inaccuracy of the measurement process may cause the acceptance of parts that do not meet the specifications or the rejection of good parts. What are the parameters of the measurement process? Design off-line and on-line quality control systems that ensure the quality of the process.

1.7. Assume that the inspection department in Prob. 1.6 uses GO and NO-GO gauges to check the dimensions of the incoming parts. Define the parameters of the measurement process under this environment. How do the off-line and on-line quality control systems differ from those you developed for Prob. 1.6?

1.8. Variation in product quality is usually attributed to three factors: (i) the fabrication equipment, (ii) the material, and (iii) the operator. Describe examples for each of the above factors that may result in variations of product quality.

REFERENCES

Burgam, P. M. "Design of Experiments—the Taguchi Way." *Manufacturing Engineering*, vol. 94, no. 5 (May 1985), pp. 44–47.

Crosby, P. B. *Quality is Free*. New York: McGraw-Hill, 1979.

Feigenbaum, A. V. *Total Quality Control*. New York: McGraw-Hill, 1983.

Garvin, D. A. "What Does 'Product Quality' Really Mean?" *Sloan Management Review*, vol. 26, no. 1, pp. 25–43.

Gingerich, J. Z., and J. W. Peery. "CAD/CAM and Quality." *Quality Progress*, 1984.

Juran, J. M. *Quality Control Handbook*. New York: McGraw-Hill, vol. 17, no. 6 (1979), pp. 8–9.

Shewhart, W. A. *Economic Control of Quality of Manufactured Product*. New York: Van Nostrand, 1931.

Shewhart, W. A. "Statistical Methods from the Viewpoint of Quality Control." Department of Agriculture, Washington, D.C., 1939.

Taguchi, G. "Off-line and On-line Quality Control Systems." *Proceedings of the International Conference on Quality Control*, vol. B4 (1978), pp. 1–5.

Taguchi, G. "Quality Engineering in Japan." *Proceedings of CAD/CAM, Robotics and Automation International Conference*, February 1985, pp. 9–12.

Taguchi, G. *Introduction to Quality Engineering*. Tokyo: Asian Productivity Center, 1986.

Taguchi, G., and Y. Wu. *Introduction to Off-line Quality Control*. Tokyo: Central Japan Quality Control Association, 1979. (Available from American Supplies Institute, 6 Parklane Boulevard, Suite 511, Dearborn, MI, 48126.)

Wyckoff, D. D. "Dynamic Aspects of Quality Management." Working paper, Harvard Business School, 1984.

CHAPTER 2

LOSS FUNCTION AND QUALITY LEVEL

This chapter focuses on evaluating the quality level of manufactured products. Chapter 3 discusses methods for specifying tolerance. On-line quality control activities during production are discussed in Chaps. 4 through 8.

The concept of percent defective has been widely used as a measure of quality level. Usually, however, the percentage of defective products in shipped goods is small. Also, there are not many manufacturers who ship defective products. Even in developing countries a manufacturer usually screens products to ensure that only nondefective units are shipped. When defective product units are not shipped, the consumer is not directly affected except in increased cost. Therefore, the incidence of defective product units should not be considered a quality problem, but a cost problem. How to evaluate the quality level of products shipped to consumers is the problem of concern.

In the past, percent defective, process capability index (C_p index, defined as the tolerance interval divided by 6 times the square root of the mean squared deviation from the target—see Eq. 2.1), and warranty cost have been used as measures of quality level for shipped products. There are manufacturers in the United States and Japan who require suppliers to produce items with a process capability index (C_p) of more than 1.00. However, one major weakness of the capability index (C_p) is that there is no apparent immediate basis for specifying process capability index is that there is no apparent immediate basis for specifying the optimal value of C_p. This index is a poor measure of quality level because management and engineers cannot comprehend the actual significance of its values; for example, what is the actual improvement when C_p changes from 0.9 to 1.2? Percent defectives or warranty costs are understandable because they are monetary-related measures. Still, warranty cost is useless for taking action in production processes because of time lag, and it cannot measure the loss in market share. We need appropriate methods for predicting quality before shipping or during production. Therefore, we introduce a monetary evaluation of the quality of products, assuming that tolerances are correct.

2.1 THE LOSS FUNCTION

Consider a comparison between the quality of color television sets produced by two factories belonging to the same manufacturing company. One factory (A) is located in America, and the other factory (B) in Japan. Suppose the comparison was based on color concentration, which relates to the color balance of the television sets. Although both factories used the same design, the television sets produced in the American factory had lower quality, and consumers consequently preferred products made in Japan.

Figure 2-1 shows the differences in quality characteristic (i.e., color concentration) distributions. The figure shows that the quality distribution of the Japanese-made television sets (shown by the solid curve) is approximately a normal distribution with a target value at the center; its standard deviation is about 1/6 of the tolerance, which in this case equals 10.

In quality control, the index of tolerance divided by 6 standard deviations is called the process capability index, denoted by C_p.

$$C_p = \frac{\text{tolerance}}{6 \times (\text{standard deviation})} \tag{2.1}$$

The process capability of the Japanese-made television sets is therefore 1.

On the other hand, the quality distribution of the American-made television sets (shown in Fig. 2-1 by a dotted curve) has less out-of-specification products than the Japanese-made products and is quite similar to the uniform distribution for those products that are within the tolerance limit. Since the standard deviation of the uniform distribution is given by $1/\sqrt{12}$ of the tolerance, the process capability index for these sets is given by

$$C_p = \frac{\text{tolerance}}{6 \times (\text{tolerance} / \sqrt{12})} = 0.577 \tag{2.2}$$

and is worse than that for the sets made in Japan.

Loss is always incurred when a product's functional quality characteristic (denoted by y) deviates from its target (nominal) value (denoted by m), regardless of how small the deviation is. Figure 2-2 shows a simplified relationship between quality loss and the amount of deviation from the target value. As shown in this figure, quality loss caused by deviation equals zero when $y = m$; the loss increases when the value of the functional characteristic moves in either an upward or downward direction from m. When the value of the functional characteristic

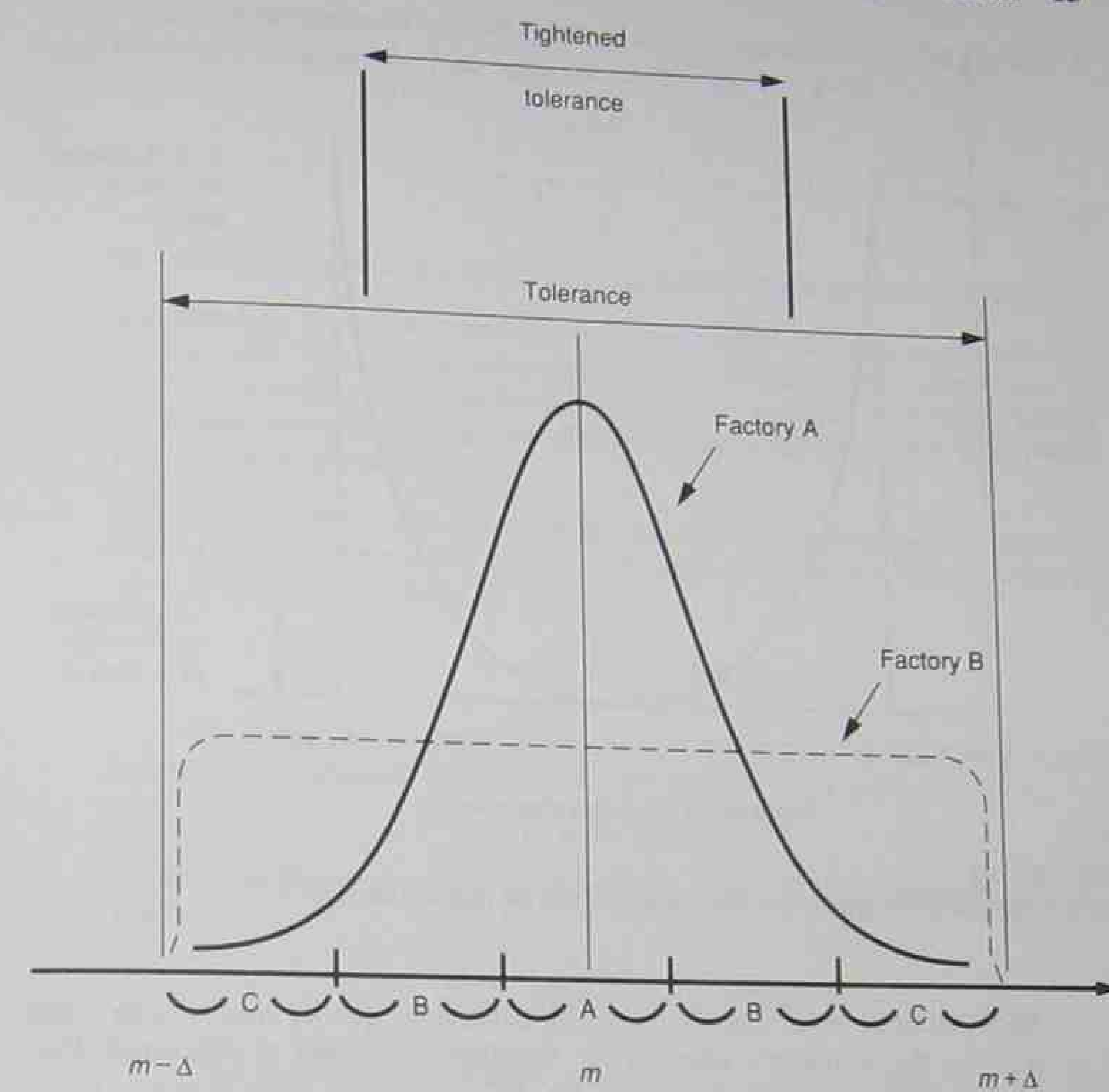


FIGURE 2-1
Distribution of color concentration.

exceeds either one of the limits $m + \Delta$ or $m - \Delta$ (where Δ is defined as the tolerance and 2Δ is the tolerance limit), the quality loss is equal to the cost of the product's disposal or manufacturing.

2.1.1 Derivation of the Loss Function

Assume the loss due to a defective part (because of discarding, repairing, or downgrading) is A. Then denote the loss function by $L(y)$ and expand it in a Taylor series about the target value m:

$$L(y) = L(m + y - m)$$

or

$$L(y) = L(m) + \frac{L'(m)}{1!}(y - m) + \frac{L''(m)}{2!}(y - m)^2 + \cdots \tag{2.3}$$

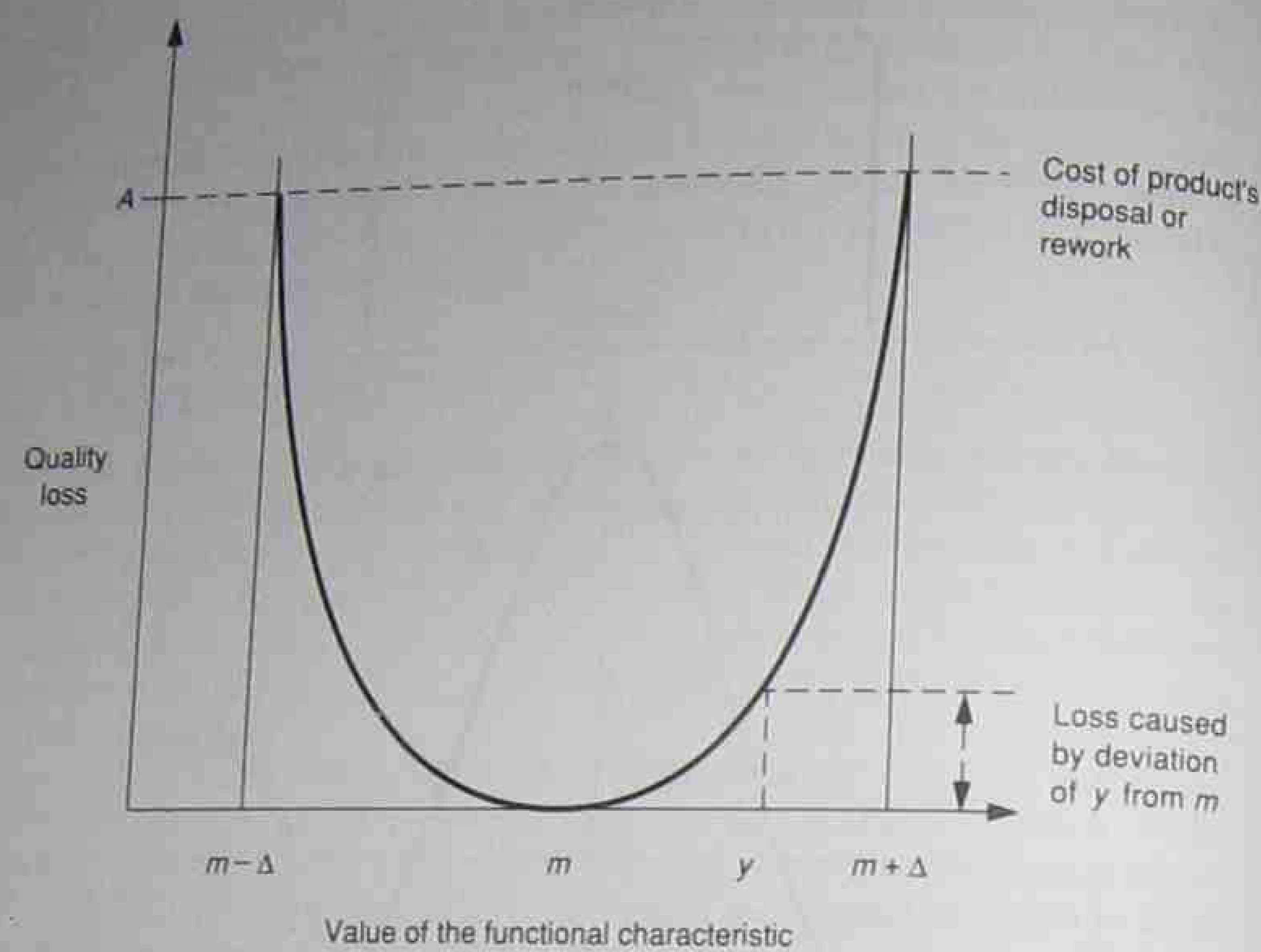


FIGURE 2-2
Relationship between quality loss and deviation from the target value (m).

Because $L(y) = 0$ when $y = m$ (by definition, quality loss is zero when $y = m$), and the minimum value of the function is attained at this point (Fig. 2-2), its first derivative with respect to m, $L'(m)$, is zero. The first two terms of Eq. 2.3, then, are equal to zero. When we neglect terms with powers higher than 2, Eq. 2.3 reduces to

$$L(y) = \frac{L''(m)}{2!}(y - m)^2 \tag{2.4}$$

or

$$L(y) = k(y - m)^2 \tag{2.5}$$

where k is a proportionality constant.

The use of quadratic loss function, minimum mean squared error, and minimum mean squared deviation criteria has been fairly extensive in the statistical and control theory literature (e.g., Athan and Falb, 1966, Box and Jenkins, 1976). More recently, Jessup (1985) discussed the rationales behind the quadratic loss function in relation to the value of continuing improvement. He used several practical examples to show the losses caused by variation in process output and the effects of different patterns of variation on the long-term economic performance, equivalent to the "loss to society" approach as advocated and explained in this book.

When the deviation of a product's functional characteristic is an amount Δ from the target value m, the loss equals A. Then Eq. 2.5 gives

$$A = k\Delta^2$$

or

$$k = \frac{A}{\Delta^2} \tag{2.6}$$

In the following, the loss function is applied to the television set example.

Example 2.1. Assume that the cost of repairing a failed television set in the factory is \$2 per unit. Compare the losses caused by deviations from the target value for two television sets, one produced in Factory A and the other produced in Factory B, as described earlier in Fig. 2-1. Recall that the tolerance interval ranges from $m - \Delta$ to $m + \Delta$, where $\Delta = 5$.

Solution. In order to calculate the losses caused by deviation from the target, we need to determine the constant k of Eq. 2.5. Since $\Delta = 5$ and $A = \$2$, the following result is obtained from Eq. 2.6:

$$k = \frac{A}{\Delta^2} = \frac{2.0}{(5)^2} = 0.08$$

The expected loss caused by deviation in the production of the television set in Factory A is obtained by taking the expectation of $L(y)$ in Eq. 2.5. Thus,

$$L = k\nu^2(\$/\text{unit}) \tag{2.7}$$

where ν^2 is the mean squared deviation from target m. For Factory A, $\nu = 10/6$, and

$$L = 0.08\left(\frac{10}{6}\right)^2 = \$0.222/\text{unit}$$

The loss caused by deviation in the production of the television set in Factory B is also

$$L = k\nu^2$$

but now $\nu = 10/\sqrt{12}$ and

$$L = 0.08\left(\frac{10}{\sqrt{12}}\right)^2 = \$0.667/\text{unit}$$

A summary of losses caused by deviation for television sets produced in both factories is given in Table 2.1. The table indicates that the losses caused by deviation for television sets produced in Factory B are three times those for the same type of television set produced in Factory A, despite Factory B's zero fraction defective.

2.1.2 Uses of the Loss Function

The loss function approach can be used in evaluating the effect of quality improvement. For example, assume that Factory A has improved the process

TABLE 2.1
Quality comparison between two manufacturers of television sets

Manufacturer	Target	Mean squared deviation ν^2	Expected loss per unit, L	Number defective
Factory A	m	$\frac{100}{36}$	\$0.222	0.27%*
Factory B	m	$\frac{100}{12}$	\$0.667	0.00%

*This is obtained by using the standard normal distribution (see Appendix A).

so that a new standard deviation from target of $10/8$ is attained. What would be the losses caused by deviations from the target value?

With the use of Eq. 2.7 and $k = 0.08$, as obtained in Example 2.1, the following is obtained:

$$L = k\nu^2$$

$$L = 0.08\left(\frac{10}{8}\right)^2 = \$0.125$$

The loss per unit of production would decrease from \$0.222 (current process) to \$0.125, resulting in \$0.097 savings per unit.

Let us see how the loss function is related to the process capability index C_p. For the current process and the improved process discussed above, Eq. 2.7 gave

$$L_1 = k\nu_1^2 \quad \text{(loss with current process)} \tag{2.8}$$

$$L_2 = k\nu_2^2 \quad \text{(loss with improved process)} \tag{2.9}$$

Divide Eq. 2.8 by Eq. 2.9 to obtain

$$\frac{L_1}{L_2} = \frac{\nu_1^2}{\nu_2^2} \tag{2.10}$$

But

$$C_{p_1} = \frac{\text{tolerance}}{6\nu_1} \tag{2.11}$$

and

$$C_{p_2} = \frac{\text{tolerance}}{6\nu_2} \tag{2.12}$$

Then substitute Eqs. 2.11 and 2.12 into Eq. 2.10, to obtain the following:

$$\frac{L_1}{L_2} = \frac{C_{p_2}^2}{C_{p_1}^2} \tag{2.13}$$

Equation 2.13 implies that the losses caused by deviation are reciprocally

proportional to the squares of the C_p indices. If the loss and the C_p index at the beginning of a process are known, then the loss after changes in the process capability can be determined thereafter by using the C_p index at the desired production period and substituting it into Eq. 2.13.

2.1.3 Economic Consequences of Tightening Tolerances as a Means to Improve Quality

As illustrated in the following example, the loss function approach can be used to determine the economic impact of tightening the tolerance to improve product quality. In order to reduce the difference in quality and process capability indices between television sets produced in Factories A and B (Example 2.1), the management of Factory B tightened the tolerance from $m \pm 5$ to $m \pm 5 \times \frac{2}{3}$. The cost of repairing an out-of-specification unit is still \$2. What is the economic impact of tightening the tolerance?

With the original tolerance, the expected loss is $L = k\nu^2 = \$0.667$, as shown in Table 2.1. The expected loss after tightening the tolerance is

$$L = k\nu^2 = 0.08\left(\frac{2}{3} \times \frac{10}{\sqrt{12}}\right)^2 = \$0.296/\text{unit}$$

If improvement of the process was obtained by repairing the failed units (units outside the new tolerance $m \pm 5 \times \frac{2}{3}$) at a cost of \$2 per unit, then the average cost of repair is as follows:

Average cost of repair per unit = percent of production that needs repair to meet the tightened tolerance × repair cost per unit

= 0.333 × 2 = \$0.667

A summary of the results of this example is shown in Table 2.2. In this case, tightening tolerance is an uneconomical alternative because the expected total loss of tightening tolerance and repair (0.667 + 0.296 = \$0.963) is greater than the expected loss using the original tolerance (\$0.667).

TABLE 2.2
Quality improvement by tightening tolerances

	Tolerance	Percent defective	ν	k	ν^2	Expected quality loss per unit
Original	$m \pm 5$	0.00	$\frac{10}{\sqrt{12}}$	0.08	$\frac{100}{12}$	\$0.667
Tightened	$m \pm 5 \times \frac{2}{3}$	0.00	$\frac{10}{\sqrt{12}} \times \frac{2}{3}$	0.08	$\frac{44.44}{12}$	\$0.296

2.1.4 The Loss Function for Similar Products (or for a System with Independent Components)

Products having similar functions can be collectively evaluated, regardless of their sizes and specifications, by using the loss function approach. For example, suppose that resistors of k different types are supplied to an assembly plant. The quality of the supplied resistors can be measured by

$$\text{Expected quality losses } L = \left(\frac{A_1}{\Delta_1^2}\nu_1^2 + \frac{A_2}{\Delta_2^2}\nu_2^2 + \cdots + \frac{A_k}{\Delta_k^2}\nu_k^2\right) \tag{2.14}$$

where $2\Delta_i$ $(i = 1, 2, \ldots, k)$ = tolerance limit of type i resistor
A_i $(i = 1, 2, \ldots, k)$ = price (loss) per unit of type i resistor
ν_i^2 $(i = 1, 2, \ldots, k)$ = mean squared deviation of the resistance of type i resistor

When a product has several measurable functional quality characteristics, the total losses caused by deviations can be estimated by using Eq. 2.15 following, where k represents, in this case, the number of measurable functional quality characteristics of the product.

$$\text{Losses due to deviations } L = \sum_{i=1}^{k} \frac{A_i}{\Delta_i^2}\nu_i^2 \tag{2.15}$$

Equation 2.15 provides an index of the total quality of the product because it takes into consideration all the functional quality characteristics of the product.

Example 2.2. A manufacturer of gauge blocks requires that the blocks meet certain flatness (surface roughness) standards for the measuring ends as well as having a specified length between the ends of each block. The loss caused by unacceptable flatness is \$50, and the loss caused by unacceptable length is \$20. These losses represent the cost of repairing (if possible) the defective blocks. The specifications of a 1-in gauge block follow:

Length	1.00000 ± 0.00010 in
Surface roughness	0.00020 μm or less

The following length measurements were taken:

1.000010	1.000020	0.999990	0.999995	1.000010
1.000005	1.000020	1.000000	0.999998	0.999990

Also, the following surface roughness measurements were taken:

0.00010	0.00020	0.00015	0.00005	0.00003
0.00010	0.00006	0.00018	0.00010	0.00020

What are the expected total losses caused by deviations?

Solution. Expected loss caused by length deviations (L_1):

$$L_1 = \frac{A_1}{\Delta_1^2}\nu_1^2$$

$$A_1 = \$20$$

$$\Delta_1 = 0.00010$$

$$\hat{\nu}_1^2 = \frac{1}{10}[(1.000010 - 1.000000)^2 + (1.000020 - 1.000000)^2 + \cdots + (0.999990 - 1.000000)^2] = 1.2 \times 10^{-10}$$

where $\hat{\nu}_1$ is the estimate of the mean squared deviation.

$$L_1 = \frac{20}{(0.00010)^2} \times 1.2 \times 10^{-10} = \$2.4$$

Expected loss caused by surface roughness deviations (L_2):

$$L_2 = \frac{A_2}{\Delta_2^2}\nu_2^2$$

$$A_2 = \$50$$

$$\Delta_2 = 0.00020$$

$$\hat{\nu}_2^2 = \frac{1}{10}[(0.00010)^2 + (0.00020)^2 + \cdots + (0.00020)^2] = 1.7 \times 10^{-8}$$

$$L_2 = \frac{50}{(0.00020)^2} \times 1.7 \times 10^{-8} = \$21.2/\text{unit}$$

Expected total loss $= L_1 + L_2 = \$23.6$/unit

2.1.5 The Loss Function and Justification of Improvements

The loss function can also be used to justify improvements of the process, as illustrated in the following example.

Example 2.3. Assume that Factory A wishes to improve the quality of its television sets by reducing deviations from the target value so that the new standard deviation will be 10/8. This improvement can be technologically achieved at an additional cost of \$0.05 per unit of production. Should the factory improve its process? (Assume that no inspection is performed.)

Solution. Total loss per unit of the current process (see Table 2.1):

$$L = k\nu^2$$

$$= 0.08\left(\frac{10}{6}\right)^2 = \$0.222 \tag{2.16}$$

Total loss per unit after improving the process:

$$L = 0.08\left(\frac{10}{8}\right)^2 = \$0.125$$

$$\text{Additional cost of improvement} = \$0.05/\text{unit}$$

$$\text{Additional cost plus loss per unit} = 0.05 + 0.125$$

$$= \$0.175 \tag{2.17}$$

The net gain resulting from improvement in the process capability is obtained by subtracting Eq. 2.17 from Eq. 2.16, and equals \$0.047 per unit of production. If the production rate of this factory is 100,000 units per month, then the expected savings will be \$4700 per month, or \$56,400 annually.

2.1.6 The Loss Function and Inspection

The loss function approach can be used effectively to determine whether 100-percent inspection can be justified or not. It should be noted that the objective of inspection is to screen or repair defective products that cannot meet the given specifications. Therefore, inspection cannot be used to improve the quality of items within the specifications. The improvement of the process can only be accomplished through improved manufacturing techniques or product design, not through screening or 100-percent inspection.

Example 2.4. Consider the case where the diameter of a stainless-steel bar is $m \pm 5\mu m$. The cost of repairing a defective bar is \$6, and the cost of inspection is \$0.03 per unit. Would a 100-percent inspection of items be justified? The estimated standard deviation of the process is 10/6.

Solution. The expected loss without inspection is

$$L = kv^2$$

where

$$k = \frac{A}{\Delta^2} = \frac{\$6.00}{5^2} = \$0.24$$

Therefore

$$L = 0.24\left(\frac{10}{6}\right)^2 = \$0.667/\text{unit}$$

Assuming that the characteristic of the product follows a normal distribution, the proportion of the products falling outside the specification $m \pm 5$ is 0.27 percent, as indicated in Table 2.3. The variance after screening defective products by using 100-percent inspection (v_{out}^2) is obtained using the procedure shown below.

After the total inspection, the out-of-specification products shown by the hatched area in Fig. 2-3 are removed. The probability density function of those items that have passed the screening (acceptable items) is given by dividing the probability density function of the normal distribution by Q, the proportion of acceptable items.

Let $f(y)$ be the density function of the normal distribution, which is given by

TABLE 2.3
Distribution of quality characteristic and its loss†
Specification: $m \pm 5\mu m$, loss by unit defective (A) = \$6, loss function $L = 0.24v^2$ dollars

Case	Mean	Standard deviation	Screening	Variance v_{out}^2	Fraction defective discovered	L(dollars)	Outgoing fraction defective (%)
1	m	$\frac{10}{2}$	no	$\left(\frac{10}{2}\right)^2$	0.00	6.000	31.73*
2	m	$\frac{10}{2}$	yes	$0.539^2 \times \left(\frac{10}{2}\right)^{2*}$	31.73*	1.743*	0.0*
3	m	$\frac{10}{4}$	no	$\left(\frac{10}{4}\right)^2$	0.00	1.500	4.55*
4	m	$\frac{10}{4}$	yes	$0.880^2 \times \left(\frac{10}{4}\right)^{2*}$	4.55*	1.162*	0.00*
5	m	$\frac{10}{6}$	no	$\left(\frac{10}{6}\right)^2$	0.00	0.667	0.27*
6	m	$\frac{10}{6}$	yes	$0.986^2 \times \left(\frac{10}{6}\right)^{2*}$	0.27*	0.648*	0.00*
7	m	$\frac{10}{8}$	no	$\left(\frac{10}{8}\right)^2$	0.00	0.375	0.01*
8	m	$\frac{10}{16}$	no	$\left(\frac{10}{16}\right)^2$	0.00	0.094	0.00*
9	m	$\frac{10}{\sqrt{12}}$	no	$\left(\frac{10}{\sqrt{12}}\right)^2$	0.00	2.000	0.00**
10	$m - 2.5$	$\frac{10}{6}$	no	$2.5^2 + \left(\frac{10}{6}\right)^2$	0.00	2.167	6.68*
11	$m - 2.5$	$\frac{10}{12}$	no	$2.5^2 + \left(\frac{10}{12}\right)^2$	0.00	1.667	0.14*
12	$m - 2.5$	$\frac{10}{16}$	no	$2.5^2 + \left(\frac{10}{16}\right)^2$	0.00	1.594	0.00*
13	$m - 2.5$	0	no	2.5^2	0.00	1.500	0.00
14	$m - 5.0$	0	no	5.0^2	0.00	6.000	0.00

† Based on Taguchi (1981)
* Normal distribution
** Uniform distribution
(no mark) Applicable to any type of distribution

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(y-\mu)/\sigma]^2} \qquad -\infty < y < \infty$$

where μ and σ are the mean and standard deviation of the normal distribution, respectively. The proportion of acceptable items, Q, is the area under the normal curve bounded by $m - 5$ and $m + 5$.

$$Q = \int_{m-5}^{m+5} \frac{1}{\sqrt{2\pi}} \times \frac{6}{10} e^{-(1/2)(6/10)^2(y-m)^2} dy$$

$$= 0.9973$$

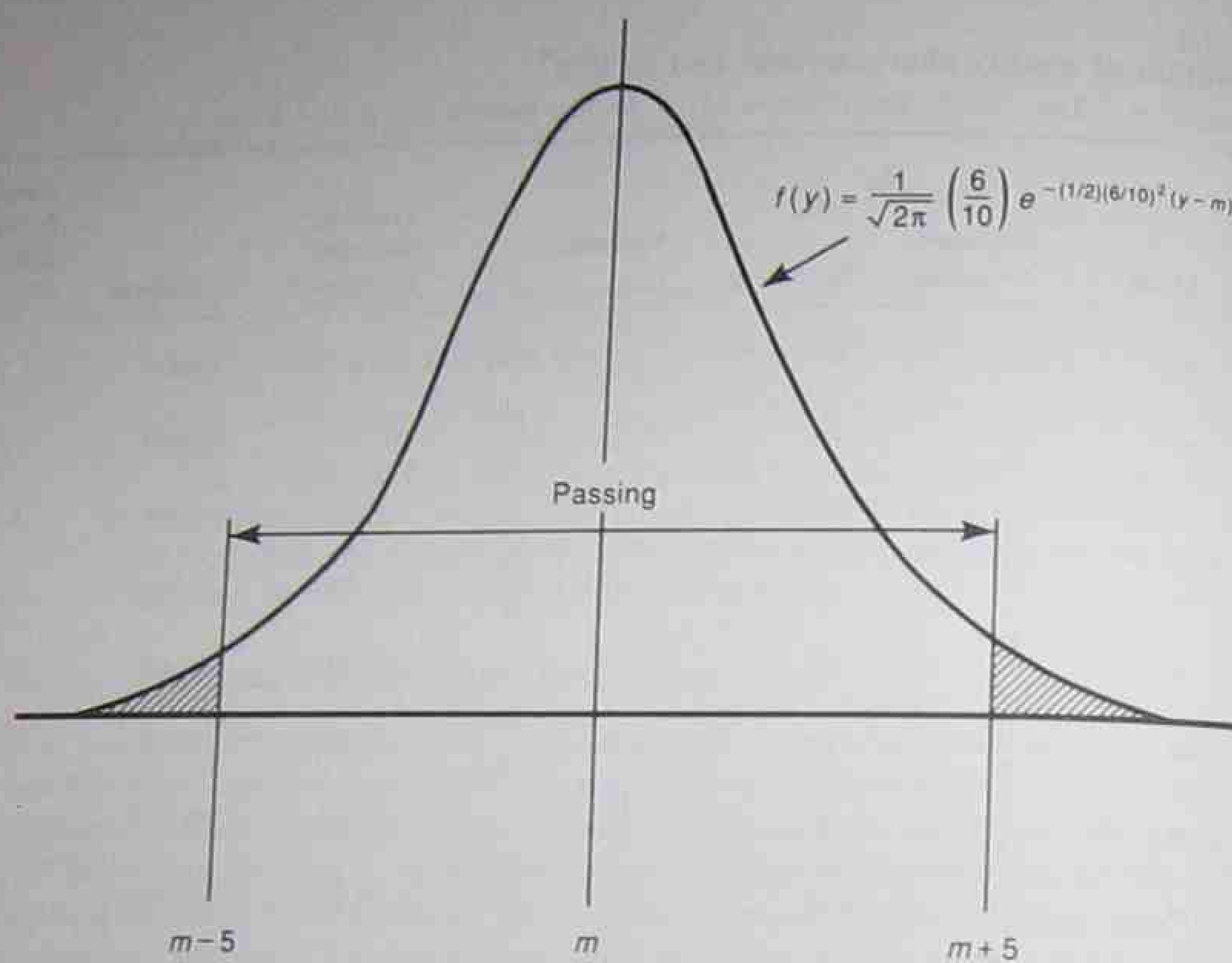


FIGURE 2-3
Normal distribution and specification.

Therefore, the variance of the passed items, ν_{out}^2, is

$$\nu_{out}^2 = \frac{1}{0.9973}\int_{m-5}^{m+5} \frac{1}{\sqrt{2\pi}} \times \frac{6}{10}(y-m)^2 e^{-(1/2)(6/10)^2(y-m)^2}\,dy$$

Using integration by parts, ν_{out}^2 is obtained as

$$\nu_{out}^2 = \left(\frac{10}{6}\right)^2 \times (0.986)^2$$

The expected total loss in the case of 100-percent inspection, L, is therefore

$$\begin{aligned} L &= \text{inspection cost per item} + (\text{loss of a defective found in the inspection} \times \text{fraction defective}) + k\nu_{out}^2 \\ &= 0.03 + 6.00 \times 0.0027 + 0.24 \times (0.986)^2 \times \left(\frac{10}{6}\right)^2 \\ &= \$0.694/\text{unit} \end{aligned}$$

Since the loss in the case of 100-percent inspection is higher than the loss when no inspection is performed (0.667/unit), 100-percent inspection is not justified.

One might conclude that in this case 100-percent inspection is useless in improving quality, because the fraction defective is only 0.27 percent. It is different, of course, when the purpose of 100-percent inspection is to find serious defectives.

In the case of a normal distribution with a standard deviation that is $\frac{1}{4}$ of the tolerance, the loss without inspection, L, is

$$L = 0.24 \times \left(\frac{10}{4}\right)^2 = \$1.50/\text{unit}$$

The proportion of the product falling outside the specification is 4.55 percent (see Table 2.3), and the variance of the outgoing items is $(0.88)^2$ times that of the original value. The total loss in the case of 100-percent inspection when σ equals 10/4 becomes

$$\begin{aligned} L &= 0.03 + 6.00 \times 0.0455 + 0.24 \times (0.88)^2 \times \left(\frac{10}{4}\right)^2 \\ &= 0.03 + 0.273 + 1.162 \\ &= \$1.465/\text{unit} \end{aligned}$$

This result is an improvement of \$0.035 per item. If there are 200,000 items produced each month, the amount of improvement is \$7,000 each month.

Assuming that the standard deviation is $\frac{1}{2}$ the tolerance and the production output still follows a normal distribution, the portion of the product falling outside the specification is 31.7 percent. Even if all the products are inspected and the defective ones screened out, the standard deviation of the outgoing quality is reduced to only 53.9 percent of the original value (σ = tolerance/2). Therefore, the loss caused by variation (using the values given above) is

$$L = 0.24\left(0.539 \times \frac{10}{2}\right)^2 = \$1.743$$

Not only is this worse than the loss of \$0.667 for $\sigma = 10/6$ and no inspection, but it is also worse than the loss of \$1.50 for $\sigma = 10/4$ and no inspection, with 4.55 percent defective products. Thus, the solution to the quality problem is, in this case, through improvement of the process and not through 100-percent inspection.

Table 2.3 shows a summary of the expected losses caused by variation for different probability distributions. These expected losses do not include the cost of inspection or loss caused by defective products found by inspection. Cases 1 through 6 demonstrate how screening reduces total losses for the given parameters. A detailed analysis of Case 2 follows, in order to illustrate how the results of the table are obtained.

Since $$L = k\nu^2$$

$$6.00 = k\left(\frac{10}{2}\right)^2$$

then

$$k = 0.24$$

and

The probability density function of all units produced is

$$f(y) = \frac{1}{\sqrt{2\pi}}\left(\frac{2}{10}\right)e^{-(1/2)(2/10)^2(y-m)^2}$$

The percentage Q of units falling within the specification limits $m \pm 5$ is obtained as follows:

$$Q = \int_{m-5}^{m+5} \frac{1}{\sqrt{2\pi}}\left(\frac{2}{10}\right)e^{-(1/2)(2/10)^2(y-m)^2}dy$$

One can substitute $t = \frac{2}{10}(y - m)$ in the above equation to obtain

$$Q = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-(t^2/2)}dt$$

$$= 0.6827$$

The fraction defective is $1 - 0.6827 = 0.3173$.

When inspection is performed, the conditional mean squared deviation of the passed (i.e., acceptable) items, ν^2_{out}, is

$$\nu^2_{\text{out}} = \frac{1}{0.6827}\int_{m-5}^{m+5} \frac{1}{\sqrt{2\pi}} \times \frac{2}{10}(y-m)^2 e^{-(1/2)(2/10)^2(y-m)^2}dy$$

$$= \left(\frac{10}{2}\right)^2(0.539)^2$$

The expected loss is

$$L = 0.24 \times \left(\frac{10}{2}\right)^2(0.539)^2 = \$1.743$$

2.2 QUALITY EVALUATIONS AND TYPES OF TOLERANCES

This section illustrates the evaluation of the quality level of products by using the loss function approach for three types of tolerances. The three types are listed below:

1. The-Nominal-The-Best (N type)

2. The-Smaller-The-Better (S type)
3. The-Larger-The-Better (L type)

2.2.1 The-Nominal-The-Best (N Type)

This type of tolerance is required for many products, parts, elements, and components when a nominal size (or characteristic) is preferred. Dimensions, clearance, and viscosity are typical examples of this type of characteristic. Also, the basic size of a screw thread or the diameter of a gear are the nominal (target) sizes from which large variations are undesirable. Bilateral tolerances are usually used in this type of application. The term *bilateral tolerance* means that the tolerance, as related to a basic dimension, is given in two directions—plus and minus.

Quality evaluations for bilateral tolerances are presented for situations in which the plus and minus tolerances are of equal amount, as well as when the plus and minus tolerances are unequal. Examples follow.

2.2.2 N-Type Tolerance When the Plus and Minus Tolerances Are of Equal Amount

As mentioned above, the bilateral system of tolerances allows variation in both directions from the target value. If one desires to specify an equal variation in both directions, then a combined plus-or-minus symbol ($\pm$) is used with a single value, as shown in Fig. 2-4.

Under an N-type tolerance, the manufacturer should aim for the target value for production, and the variation should be reduced to a minimum. The following example illustrates how the quality level of a product is evaluated for an N-type tolerance with equal variation around the target value.

Example 2.5. A manufacturer of ball bearings used in gas turbines requires that tolerances of the diameter and hardness of each ball be as follows:

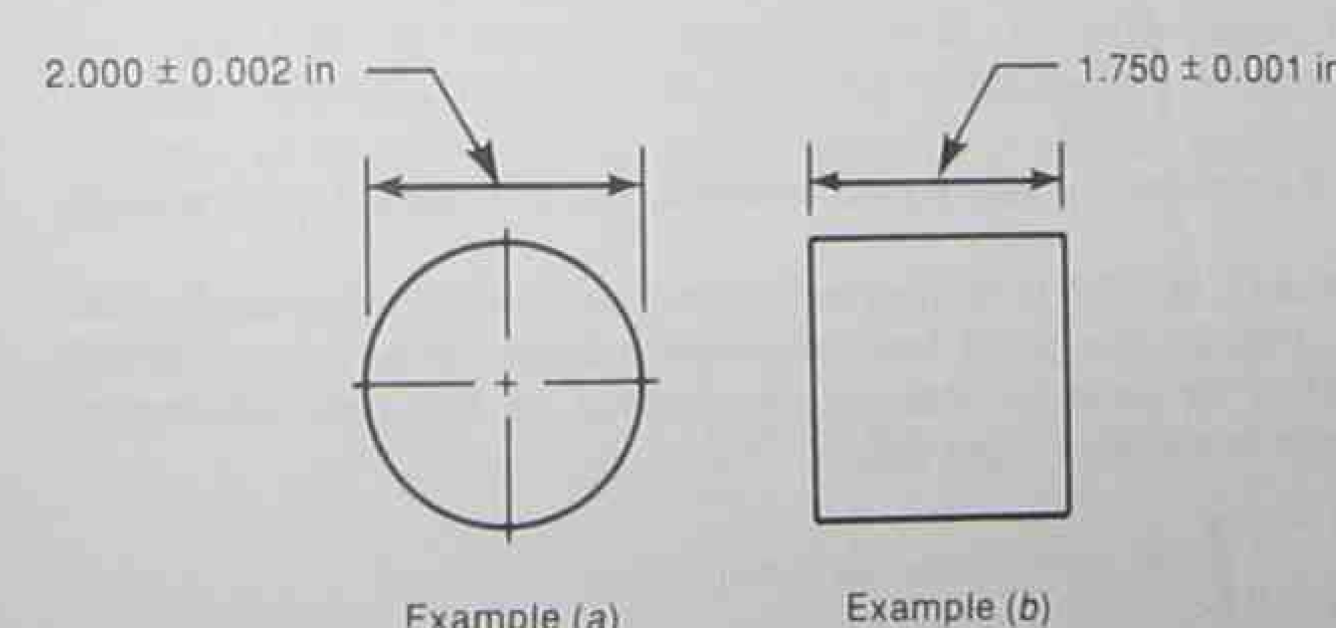


FIGURE 2-4
Examples of bilateral tolerances with equal variations.

Tolerance of diameter $m_1 \pm 0.6\ \mu$m
Tolerance of hardness $m_2 \pm 2.0$ (Brinell hardness)

where m_1 and m_2 are the target values of the diameter and the hardness, respectively. The production rate is 80,000 balls per day at a cost of 30¢ per ball. Defective balls cannot be reworked and are scrapped. The following deviations from the diameter and the hardness target values were recorded.

Deviations from the target diameter:

0.3	0.0	−0.1	0.0	0.3	0.2	0.1	−0.2	0.6	0.4
−0.2	0.1	0.0	−0.4	0.5	0.4	−0.2	0.0	0.0	0.2

Deviations from the target hardness:

−1.0	−1.6	−0.4	−1.0	0.6	0.4	−1.2	−1.3	−0.2	−0.4
−0.4	0.5	−0.3	0.6	0.6	−0.9	−0.7	−0.9	−0.7	−1.3

Based on the diameter and hardness measurements recorded above, determine the quality levels of the production process for the diameter and hardness attributes of the balls.

Solution. By the analysis in this chapter, the loss caused by deviation from the target value is estimated as

$$L = \frac{A}{\Delta^2} \nu^2 \tag{2.18}$$

where A = cost of a defective product
2Δ = tolerance limit
ν^2 = mean squared deviation of the produced values (y's) from the target value (m)

The quantity ν^2 is usually estimated based on available data. The usual estimate of ν^2 has been referred to as $\hat{\nu}^2$ and is obtained as follows:

$$\hat{\nu}^2 = \frac{1}{n}\left[(y_1 - m)^2 + (y_2 - m)^2 + \cdots + (y_n - m)^2\right] \tag{2.19}$$

where n is the number of measurements available, and y_i is the value of measurement i.

Equations 2.18 and 2.19 are used to determine the quality level of the diameter of the balls. Actually, this quality level (expressed in dollars) is the loss attributable to deviations of the particular characteristic involved (in this case, ball-bearing diameter) from its specified target value.

$$A = 30¢$$

$$\Delta = 0.6$$

$$\hat{\nu}^2 = \frac{1}{20}\left[(0.3)^2 + (0.0)^2 + (-0.1)^2 + \cdots + (0.2)^2\right] = 0.075$$

The quality level of the diameter of the balls, as measured by the loss function, is

$$L = \frac{30}{(0.6)^2} \times 0.075 = 6.25¢$$

Similarly, the quality level of the hardness of the balls, as measured by the loss function, can be obtained as follows:

$$A = 30¢$$

$$\Delta = 2.0$$

$$\hat{\nu}^2 = \frac{1}{20}[(-1.0)^2 + (-1.6)^2 + (-0.4)^2 + \cdots + (-1.3)^2] = 0.704$$

$$L = \frac{30}{(2.0)^2} \times 0.704 = 5.28¢$$

The difference between the quality levels relating to the required hardness and diameter of the balls becomes evident when the yearly production rate is considered, assuming 250 working days per year.

$$\text{Difference in quality levels} = (0.0625 - 0.0528) \times 250 \times 80{,}000$$
$$= \$194{,}000 \text{ per year}$$

As shown above, the loss function approach can be used to compare the quality levels of the various processes underlying the different attributes.

It should be noted that the mean squared deviation of y from the target value m, as given by Eq. 2.20 below, is decreased by reducing either the variance of y or the term $[E(Y) - m]^2$. The reduction of variance can be achieved by using both off-line and on-line quality control approaches.

$$\text{Mean squared deviation} = E(Y - m)^2 = \text{var}(Y) + [E(Y) - m]^2 \tag{2.20}$$

where $E(Y)$ is the expectation of Y, and var(Y) is the variance of Y.

Let $\bar{y}$ be the estimated expectation of Y. No adjustments of the process are needed when predicted $\bar{y} = m$; however, if predicted $\bar{y} \neq m$ then an amount of adjustment equal to $\bar{y} - m$ can be made on y. If the amount of adjustment is denoted as $\bar{e}$, then

$$\bar{e} = (\bar{y} - m) = \frac{1}{n}\sum_{i=1}^{n}(y_i - m) = \text{predicted deviation from target} \tag{2.21}$$

In other words, if predicted $\bar{y}$ does not equal m, as shown in Figure 2-5, the process should be adjusted (if possible) so that predicted $\bar{y}$ coincides with m, to obtain a significant decrease in the mean squared deviation.

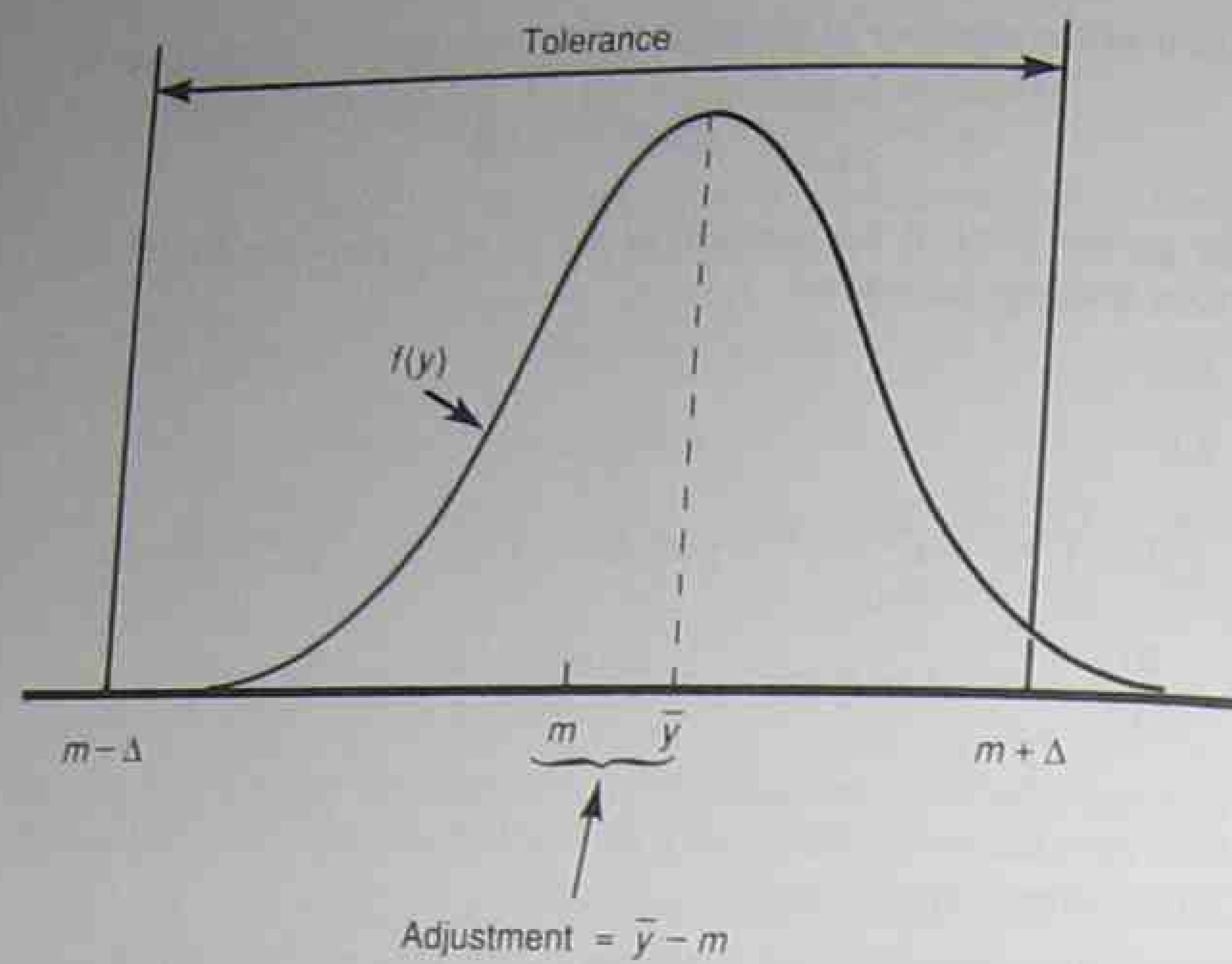


FIGURE 2-5
A process without adjustment.

Example 2.6. Consider the ball-bearing manufacturer in Example 2.5. Examination of the diameter data shows more positive deviations than negative ones, whereas the hardness data show more negative deviations than positive ones. Assume that the manufacturer can shift the means of the data to the target values. What are the quality levels of the diameter and the hardness after the adjustments?

Solution. The process should first be adjusted so that the value of every diameter is adjusted by an amount $\bar{e}$, the predicted deviation of the diameter from the target value. The new deviation (after adjustment) from the target value is

$$\text{Deviation after adjustment} = y - m - (\bar{y} - m)$$
$$= (y - \bar{y}) \tag{2.22}$$

Let the sum of squared deviation after adjustments of the diameter of the balls be denoted as $\hat{\nu}_d^2$. Then

$$\hat{\nu}_d^2(\text{after adjustment}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] \tag{2.23}$$

$$\hat{\nu}_d^2 = \frac{1}{19}\left[0.3^2 + 0.0^2 + \cdots + 0.2^2 - \frac{2^2}{20}\right]$$
$$= 0.0684$$

The denominator $n - 1$ is used instead of n in the above equation because

one degree of freedom is used to estimate the mean. The quality level of the diameter after the adjustment is

$$L = \frac{30}{(0.6)^2} \times 0.0684 = 5.70¢$$

Thus, the annual improvement is

$$(0.0625 - 0.0570) \times 80{,}000 \times 250 = \$110{,}000$$

Similarly, by using Eq. 2.23 and denoting the squared deviation after adjusting the hardness of the balls by $\hat{\nu}_h^2$:

$$\hat{\nu}_h^2(\text{after adjustment}) = \frac{1}{19}\left[(-1.0)^2 + (-1.6)^2 + \cdots + (-1.3)^2 - \frac{(-9.6)^2}{20}\right]$$
$$\hat{\nu}_h^2 = 0.499$$

The quality level of the hardness after the adjustment is

$$L = \frac{30}{(2.0)^2} \times 0.499 = 3.74¢$$

The annual improvement resulting from this adjustment is

$$(0.0528 - 0.0374) \times 80{,}000 \times 250 = \$308{,}000$$

Also, a comparison of quality levels between the diameter and the hardness after the adjustment indicates that the quality level of the hardness is 5.70/3.74, or 1.5 times better than that of the diameter.

The above procedure can also be used to evaluate the quality levels of the same product when provided by different suppliers.

Example 2.7. An automobile manufacturer requires that the steering knuckle for disk brakes be made of two separate parts, the knuckle and the spindle. They must be assembled by shrink-fitting the spindle into the knuckle, thereby achieving a more desirable stress distribution. The specification of the diameter of the spindle is $m \pm 20\mu\text{m}$. The loss caused by a defective spindle is \$24. The manufacturer observed the deviations shown in Table 2.4 from three different suppliers. What are the quality levels of their spindles? If adjustments could be made, what would the quality levels be after adjustments?

Solution. By using Eq. 2.18 we obtain the suppliers' losses before adjustments are made as follows:

$$k = \frac{A}{\Delta^2} = \frac{24}{(20)^2} = 0.06$$

and
$$L = 0.06\hat{\nu}^2$$

where
$$\hat{\nu}^2 = \frac{1}{n}\left[(y_1 - m)^2 + (y_2 - m)^2 + \cdots + (y_n - m)^2\right]$$

A summary of the suppliers' losses before adjustments is given in Table 2.5.

TABLE 2.4
Suppliers' observed data

Supplier	Deviation from target										Price per unit
1	−5	8	5	−4	3	−2	5	4	0	1	$36
	−2	0	3	8	−4	−6	2	5	−3	0	
2	−6	−3	−5	−6	−7	−5	−3	−7	−7	−6	$36
	−5	−4	−8	−6	−3	−8	−6	−5	−4	−9	
3	−7	18	0	15	−16	−7	−10	−9	−17	−3	$32
	−8	−10	12	−10	−6	−9	10	4	16	−13	

Adjustments could be made so that $\bar{y}_1 = \bar{y}_2 = \bar{y}_3 = m$. The mean squared deviations after adjustment for Suppliers 1, 2, and 3 are obtained using Eq. 2.23 as follows:

$$\hat{\nu}_j^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] \qquad j = 1, 2, 3$$

Thus,

$$\hat{\nu}_1^2 = \frac{1}{19}\left[(-5)^2 + (8)^2 + \cdots + (0)^2 - \frac{(18)^2}{20}\right] = 17.67$$

$$\hat{\nu}_2^2 = 2.98$$

$$\hat{\nu}_3^2 = 122.26$$

The suppliers' losses, if adjustments could be made, are given in Table 2.6. The above data suggest that the manufacturer should choose Supplier 2 as the source for the spindles if Supplier 2 is able to keep the distribution of spindle diameters centered at the target value m.

2.2.3 N-Type Tolerance When the Plus and Minus Tolerances Are Not Equal

The most general case of the bilateral tolerance system is where the tolerance limits are set at unequal distances from the target value m. This is usually written

TABLE 2.5
Suppliers' losses before adjustments

Supplier	$\hat{\nu}^2$	$L = 0.06\hat{\nu}^2$
1	17.60	$1.06
2	34.75	$2.08
3	122.40	$7.34

TABLE 2.6
Suppliers' losses after adjustments

Supplier	$\hat{\nu}^2$	$L = 0.06\hat{\nu}^2$
1	17.67	$1.06
2	2.98	$0.18
3	122.26	$7.34

as $m^{-\Delta_1}_{+\Delta_2}$, where Δ_1 and Δ_2 are the lower and the upper limits of the tolerance, respectively. The loss caused by deviation of a data point y from the target value is shown in Fig. 2-6 and is expressed as

$$L(y) = \begin{cases} \dfrac{A_1}{\Delta_1^2}(y-m)^2 & \text{if } y \le m \\ \dfrac{A_2}{\Delta_2^2}(y-m)^2 & \text{if } y > m \end{cases} \qquad (2.24)$$

where A_1 is the loss caused by y being below the lower limit of tolerance, and A_2 is the loss caused by y being above the upper limit of tolerance.

When n observations are taken, the expected loss L is obtained as

$$L = \frac{1}{n}\left[\frac{A_1}{\Delta_1^2}\sum{}' (y-m)^2 + \frac{A_2}{\Delta_2^2}\sum{}'' (y-m)^2\right] \qquad (2.25)$$

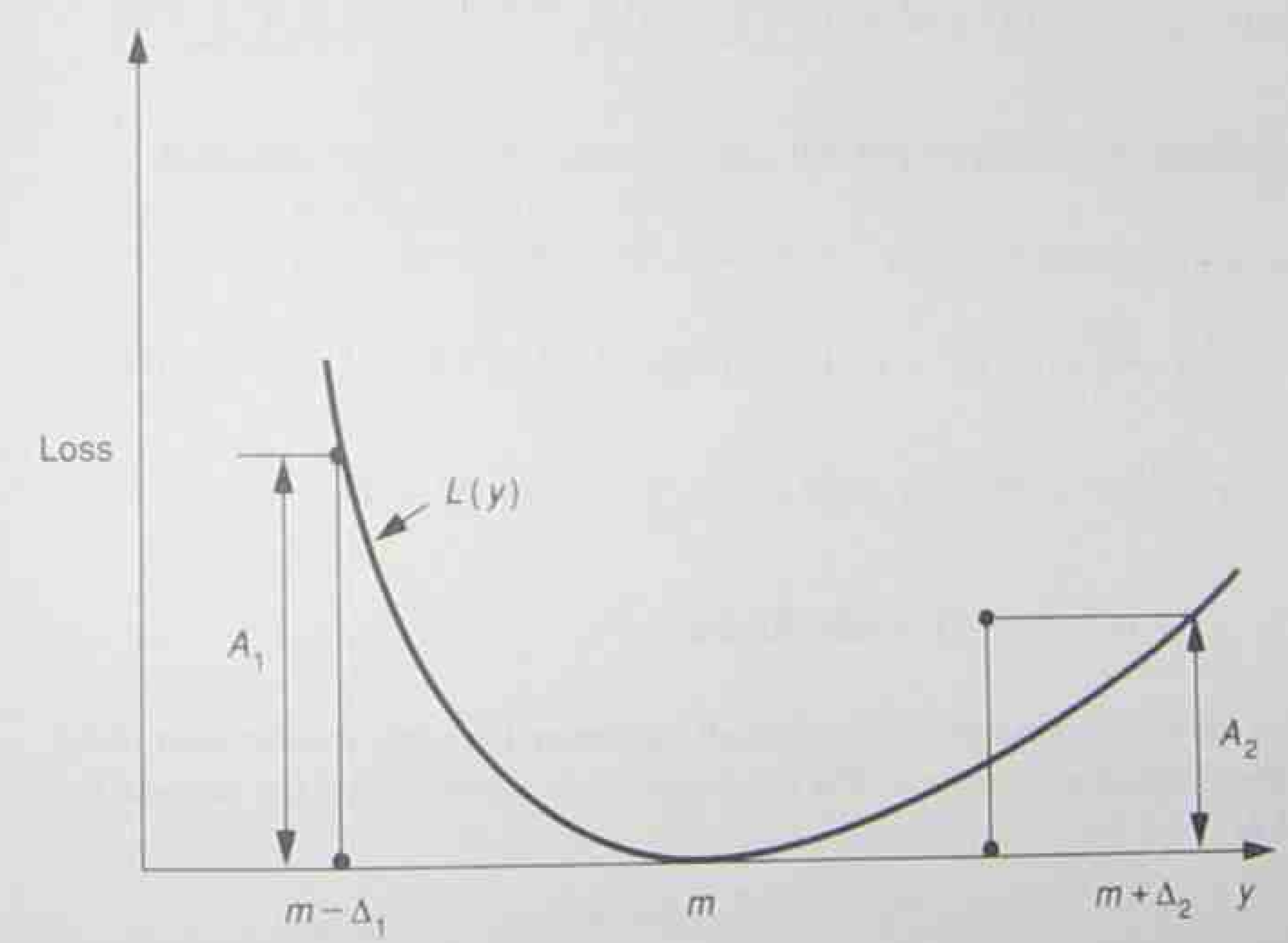


FIGURE 2-6
Loss due to deviations from m.

where $\sum'(y-m)^2$ is the sum of the squared deviation for y_i smaller than m, and $\sum''(y-m)^2$ is the sum of the squared deviation for y_i larger than m.

Example 2.8. An automobile manufacturer requires that the clearance between the cylinder and the piston of a six-cylinder engine be 3^{-2}_{+7} μm. Defect loss for each cylinder and piston assembly is \$200, and the monthly production is 50,000 units. Data showing deviation from the target value for the first two months of production are shown below. What are the quality levels during these two months? What is the improvement, if any, of the quality level?

Month	Deviations											
1	−2	3	0	4	5	−2	3	−2	0	−1	−1	−3
	0	4	3	−2	0	1	0	5	6	2	−1	
2	3	2	0	1	−1	−1	0	−2	3	0		
	6	−2	4	3	0	−2	0	−1	2	4		

Solution. The quality level of the production during Month 1 is determined by using Eq. 2.25:

$A_1 = A_2 = \$200 \qquad \Delta_1 = -2 \qquad \Delta_2 = 7$

$$L_1 = \frac{1}{23}\left\{\frac{200}{(-2)^2}[(-2)^2 + (-2)^2 + (-2)^2 + (-1)^2 + (-1)^2 + (-3)^2 + (-2)^2 + (-1)^2] + \frac{200}{7^2}[3^2 + 4^2 + 5^2 + 3^2 + 4^2 + 3^2 + 1^2 + 5^2 + 6^2 + 2^2]\right\}$$

$$= \frac{1}{23}\{1400 + 612.24\} = \$87.50$$

The loss during Month 2, L_2, is

$$L_2 = \frac{1}{20}\left\{\frac{200}{(-2)^2}[(-1)^2 + (-1)^2 + (-2)^2 + (-2)^2 + (-2)^2 + (-1)^2] + \frac{200}{7^2}[3^2 + 2^2 + 1^2 + 3^2 + 6^2 + 4^2 + 3^2 + 2^2 + 4^2]\right\}$$

$$= \frac{1}{20}\{750 + 424.49\} = \$58.70/\text{unit}$$

The losses resulting from variation on the lower limit are greater than those of the upper limit in each month. The improvement of quality level per unit during Month 2 is

$$87.5 - 58.7 = \$28.80$$

which results in an improvement of

$$28.8 \times 50{,}000 = \$1{,}440{,}000 \text{ per month}$$

2.2.4 The-Smaller-The-Better (S Type)

A The-Smaller-The-Better type tolerance involves a nonnegative characteristic, whose ideal value is zero. A typical example of such a characteristic is impurity. Wear, shrinkage, deterioration, and noise level are also examples of this type.

Under The-Smaller-The-Better (S-type) tolerance, the characteristic value is $y \geq 0$, the target value is $m = 0$, and the upper tolerance limit is Δ. The quality level (loss function) is estimated as

$$L = \frac{A}{\Delta^2}\nu^2 \qquad (2.26)$$

where A is the loss caused by exceeding the upper tolerance limit.

Example 2.9. A manufacturer of gauge blocks requires that the flatness of the surface of each block be within 12 μm. Obviously, the smaller the deviation in the flatness of the block, the better the block. The loss caused by out-of-tolerance conditions is \$80. Two machine tools, M_1 and M_2, are used in manufacturing the blocks. The flatness data given below were obtained by measuring two blocks each day for two successive weeks. Compare the quality levels of the two machine tools.

Machine	Flatness data (μm)									
M_1	0	5	4	2	3	1	7	6	8	4
	6	0	3	10	4	5	3	2	0	7
M_2	5	4	0	4	2	1	0	2	5	3
	2	1	3	0	2	4	1	6	2	1

Solution. This is a The-Smaller-The-Better (S-type) case with an upper tolerance limit $\Delta = 12$ μm and $A = \$80$.

The quality level of machine M_1 is

$$L_1 = \frac{A}{\Delta^2}\hat{\nu}_1^2$$

where
$$\hat{\nu}_1^2 = \frac{1}{20}(y_1^2 + y_2^2 + y_3^2 + \cdots + y_{20}^2)$$

$$= \frac{1}{20}(0 + 5^2 + 4^2 + \cdots + 7^2)$$

$$= 23.4\ \mu\text{m}^2$$

Hence,
$$L_1 = \frac{80}{12^2} \times 23.4 = \$13$$

The quality level of machine M_2 is

$$L_2 = \frac{A}{\Delta^2}\hat{\nu}_2^2$$

where
$$\hat{\nu}_2^2 = \frac{1}{20}(5^2 + 4^2 + \cdots + 1^2)$$
$$= 8.8\ \mu\text{m}^2$$

giving
$$L_2 = \frac{80}{12^2} \times 8.8 = \$4.90$$

Therefore, the quality level of machine tool M_2 is better than that of M_1 by \$8.10. A daily production of 2000 pieces will result in a difference of \$4,050,000 per year (250 working days).

2.2.5 The-Larger-The-Better (L Type)

There are cases where The-Larger-The-Better is applicable to characteristics such as the strength of materials and fuel efficiency. In these cases, there are no predetermined target values, and the larger the value of the characteristic, the better it is.

Under this type of tolerance, the characteristic value is $y \geq 0$, the lower tolerance limit is Δ, and the target (or ideal) value is $m = +\infty$. A is the loss caused by falling below the lower tolerance limit (i.e., $y < \Delta$). The quality level of this type of characteristic is obtained by transforming the L-type tolerance to an S-type tolerance as follows:

Let

$$z = \frac{1}{y} \tag{2.27}$$

The characteristic $z \geq 0$ has an S-type tolerance with the target value $m = 0$ and the upper specification limit $1/\Delta$. The loss function of the characterisitic z is

$$L(z) = \frac{A}{(1/\Delta)^2} z^2 \tag{2.28}$$

Equation 2.28 is put in terms of the desired characteristic, y, by substituting Eq. 2.27 in Eq. 2.28. Thus,

$$L(y) = \frac{A\Delta^2}{y^2} \tag{2.29}$$

or

$$L = A\Delta^2 \hat{\nu}^2 \tag{2.30}$$

where

$$\hat{\nu}^2 = \frac{1}{n}\left(\frac{1}{y_1^2} + \frac{1}{y_2^2} + \cdots + \frac{1}{y_n^2}\right) \tag{2.31}$$

In effect, the loss function (quality level) associated with a The-Larger-The-Better type of tolerance is obtained by taking the reciprocals of the measurements and treating them as The-Smaller-The-Better type.

Example 2.10. The strength of an adhesive is usually determined by the kilograms force (kgf) needed to break apart specimens joined by the adhesive. Two types of adhesives, S_1 and S_2, which cost \$50 and \$60 per unit weight, respectively, are to be compared. The lower specification limit Δ is 5 kgf for the breaking force. The out-of-specification units are discarded, resulting in a loss of \$70 per unit. The annual production rate is 120,000 units. Sixteen units were tested for each type of adhesive, and the following data for the breaking force were obtained:

Type of adhesive	Breaking force (kgf)							
S_1	10.2	5.8	4.9	16.1	15.0	9.4	4.8	10.1
	14.6	19.7	5.0	4.7	16.8	4.5	4.0	16.5
S_2	7.6	13.7	7.0	12.8	11.8	13.7	14.8	10.4
	7.0	10.1	6.8	10.0	8.6	11.2	8.3	10.6

Compare the quality levels of S_1 and S_2.

Solution. In this case, the larger the force, the better the adhesive. The quality level of S_1 is obtained by using Eq. 2.30.

$$L_1 = A\Delta^2\hat{\nu}_1^2$$
$$A = \$70$$
$$\Delta = 5\text{ kgf}$$
$$\hat{\nu}_1^2 = \frac{1}{16}\left(\frac{1}{10.2^2} + \frac{1}{5.8^2} + \cdots + \frac{1}{16.5^2}\right)$$
$$= 0.0228$$
$$L_1 = 70 \times 5^2 \times 0.0228$$
$$= \$39.90$$

Similarly, the quality level of S_2 is:

$$L_2 = A\Delta^2\hat{\nu}_2^2$$
$$\hat{\nu}_2^2 = \frac{1}{16}\left(\frac{1}{7.6^2} + \frac{1}{13.7^2} + \cdots + \frac{1}{10.6^2}\right)$$
$$= 0.01139$$
$$L_2 = 70 \times 5^2 \times 0.01139$$
$$= \$19.90$$

Considering the cost per unit of S_1 (\$50) and S_2 (\$60), the savings that result from the use of S_2 instead of S_1 are

$$(50 + 39.90) - (60 + 19.90) = \$10 \text{ per unit}$$

and the annual savings are \$1,200,000.

The loss function can be used to evaluate the quality level of production during different intervals of time. This evaluation might suggest recommendations for quality improvement or changes in the production process.

Example 2.11. The tensile strength data, in kilograms per millimeter squared, of thread samples collected every day for two consecutive months of production are as follows:

Month	Tensile strength data (kg/mm^2)							
First month (23 work days)	15.2	10.3	12.4	13.6	14.5	12.8	13.2	11.4
	14.6	12.5	13.8	14.0	12.5	14.0	13.6	14.2
	14.6	12.3	15.4	13.7	13.0	12.7	15.0	
Second month (22 work days)	13.6	15.0	14.3	12.6	14.3	15.0	14.6	13.2
	11.5	12.6	13.6	12.0	13.1	12.1	15.1	13.7
	14.2	13.6	14.1	12.8	11.7	10.2		

Compare the quality levels of the two months assuming the lower specification limit is 10.0 kg/mm^2. Defect loss is \$0.5 per linear meter, and the monthly production is 200 million meters.

Solution. This case involves The-Larger-The-Better (L-type) tolerance. The quality levels are determined by using Eq. 2.30 and 2.31.

The quality level during the first month of production, L_1, is estimated as

$$L_1 = A\Delta^2 \hat{\nu}_1^2$$

where

$$A = \$0.5$$

$$\Delta = 10.0$$

$$\hat{\nu}_1^2 = \frac{1}{23}\left(\frac{1}{15.2^2} + \frac{1}{10.3^2} + \cdots + \frac{1}{15.0^2}\right)$$

$$\hat{\nu}_1^2 = 0.005682$$

Therefore,

$$L_1 = 0.5 \times 10^2 \times 0.005682 = \$0.2841$$

$$\text{Monthly loss} = 0.2841 \times 200 \times 10^6 = \$56{,}820{,}000$$

The quality level during the second month of production is found similarly:

$$L_2 = A\Delta^2 \hat{\nu}_2^2$$

$$\hat{\nu}_2^2 = \frac{1}{22}\left(\frac{1}{13.6^2} + \frac{1}{15.0^2} + \cdots + \frac{1}{10.2^2}\right)$$

$$\hat{\nu}_2^2 = 0.005809$$

$$L_2 = \$0.2905$$

$$\text{Monthly loss} = 0.2905 \times 200 \times 10^6 = \$58{,}100{,}000$$

The quality level of the first month is better (L_1 is lower) than that of the second month of production by an amount

$$58{,}100{,}000 - 56{,}820{,}000 = \$1{,}280{,}000$$

As demonstrated above, the quality level should be evaluated periodically. It should be used as an indicator of the performance of the production line and considered an important index.

Quality-level calculations can be used to evaluate suppliers or products having similar functions, or to evaluate the quality level of a product having several characteristics of the same or of different tolerance types. Suppose that k types of products are produced. The quality level of the production line is given by the sum of the losses that correspond to the specified tolerances. For example:

$$L = \frac{A_1}{\Delta_1^2}\hat{\nu}_1^2 + \frac{A_2}{\Delta_2^2}\hat{\nu}_2^2 + \cdots + \frac{A_k}{\Delta_k^2}\hat{\nu}_k^2 \qquad \text{(for N and S types)}$$

$$L = A_1\Delta_1^2\hat{\nu}_1^2 + A_2\Delta_2^2\hat{\nu}_2^2 + \ldots + A_k\Delta_k^2\hat{\nu}_k^2 \qquad \text{(for L type)}$$

where A_i = loss caused by a defective component of type $i (i = 1, 2, \ldots, k)$
$\hat{\nu}_i^2$ = mean squared deviation of type i component
Δ_i = tolerance limit of an individual component of type i

2.3 DETERMINATIONS OF TOLERANCES

This section explains a method for determining tolerances of the quality characteristics. The determination of tolerance is illustrated in the following example.

Example 2.12. Consider the production of high-voltage transformers. During the life of this kind of transformer, output voltage might change because of the deterioration of transistors in the power circuit. Assume that a transformer is not suitable for its intended function when its output voltage exceeds the tolerance limits of ± 25V. Exceeding the limits results in a loss (denoted by A) of \$300. Before shipping to a customer, the manufacturer can adjust the voltage in the plant by changing a resistor at a cost of \$1. What should the manufacturer's specifications be?

Solution. The loss caused by product variation from the target value, $L(y)$, is

$$L(y) = k(y - m)^2 \qquad (2.32)$$

where m is the target value (115 V in this case) and k is the proportionality constant. Therefore,

$$k = \frac{A}{\Delta^2} = \frac{300}{(25)^2} = 0.48$$

The loss function is

$$L(y) = 0.48(y - 115)^2 \qquad (2.33)$$

It is assumed that the allowable varying range of the output voltage for

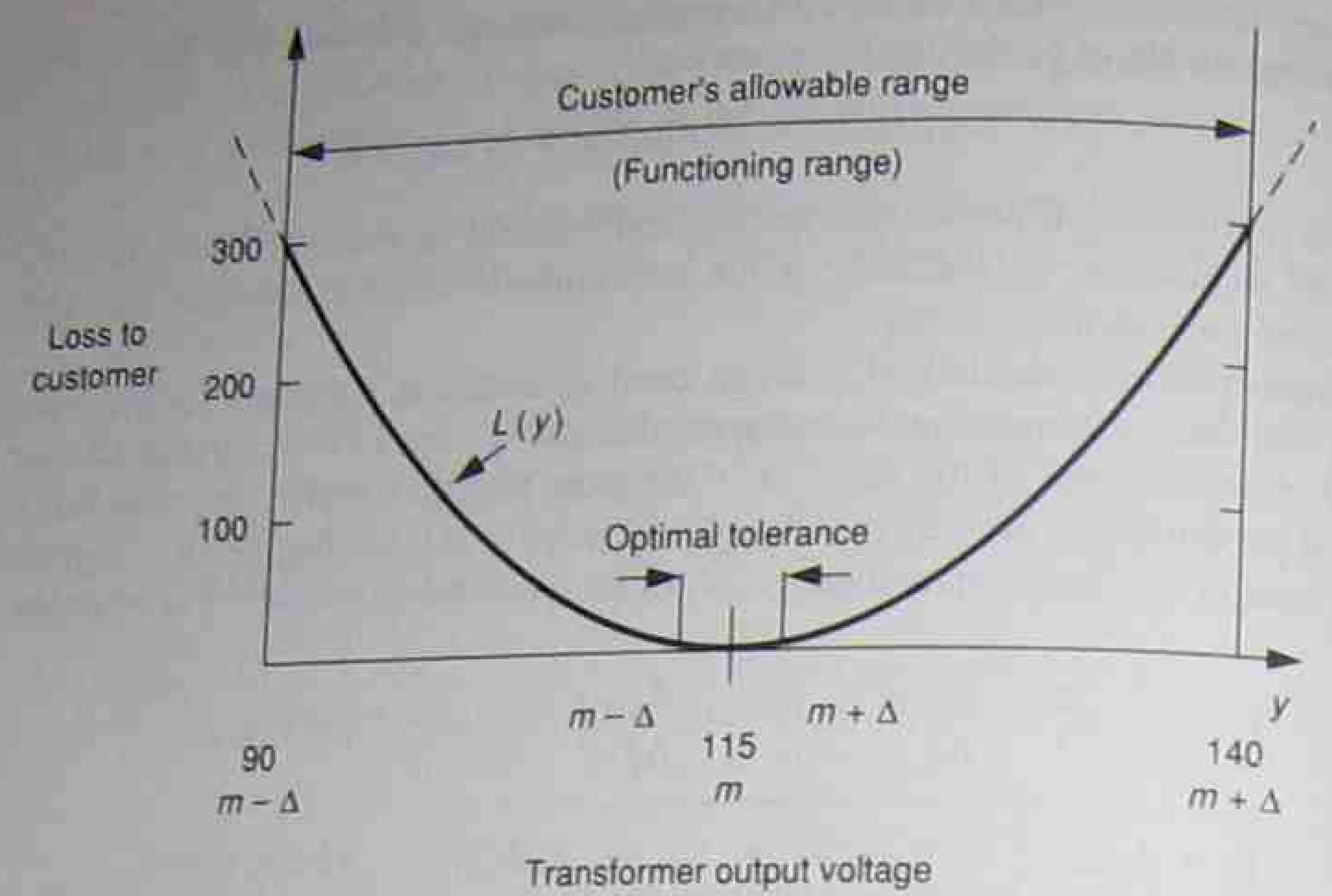


FIGURE 2-7
Tolerance determination for Example 2.12.

the customer is 115 ± 25 V. The allowable varying range in the plant will be different, because it is easy to adjust the voltage to the target value by changing a resistor in the circuit. The loss or cost of adjustment to the manufacturer is $1. Substitution of this value in Eq. 2.33 yields

$$1.0 = 0.48(y - 115)^2$$

and

$$y = 115 \pm \sqrt{\frac{1}{0.48}}$$

$$y = 115 \pm 1.4 \text{ V}$$

Although the varying range for the customer is 115 ± 25 V, the varying range in production should be 115 ± 1.4 V, and units whose output voltage differs from 115 by more than ± 1.4 V should be considered defective by the manufacturer, as shown in Fig. 2-7. It is beneficial to the customer for a transformer to be adjusted in the plant if its output voltage falls outside the range 115 ± 1.4 V.

As shown above, the manufacturer should adjust (repair) the characteristic y if

$$\frac{A}{\Delta^2}(y - m)^2 \le \frac{A_0}{\Delta_0^2}(y - m)^2 \qquad (2.34)$$

and the manufacturer's allowance is

$$\Delta = \sqrt{\frac{A}{A_0}}\Delta_0 \qquad (2.35)$$

where 2Δ = manufacturer tolerance limit
$2\Delta_0$ = customer tolerance limit or functional limit
A = manufacturer loss caused by a defective product
A_0 = customer loss caused by a defective product

2.4 SUMMARY

This chapter discussed the use of the loss function concept for evaluating the quality level of a product or process by quantifying the deviations from the target value. The loss function was used to evaluate the effect of process improvement, the impact of tightening tolerances, and the effect of 100-percent inspection.

The chapter also covered the use of the loss function in the quality evaluations of three types of tolerances: The-Nominal-The-Best, The-Smaller-The-Better, and The-Larger-The-Better. An application of the loss function in the determination of tolerances was also explored.

PROBLEMS

2.1. A manufacturer of enhanced graphics adapters (EGAs) for personal computers requires that the distance between any two adjacent pins of a memory chip be 1.5 ± 0.001 mm. The loss due to a defective chip is $3.

The manufacturer measured the distance between two randomly selected adjacent pins on 20 chips and recorded the following observations:

1.500	1.501	1.499	1.501	1.500
1.500	1.500	1.499	1.500	1.501
1.501	1.500	1.499	1.499	1.500
1.501	1.499	1.501	1.501	1.500

What is the quality loss? If the mean is adjusted on target, how much quality improvement can be expected? Develop a system for inspecting the distance between pins.

2.2. A prototype of a robot monitoring system is built to evaluate the performance of the system before the beginning of the high-volume production. The purpose of the monitoring system is to increase productivity by tracking a robot's repeatability, trajectory, and accuracy. Measurements are compared to user-specified target values. The system consists of LEDs (light-emitting diodes) attached to selected points on the robot, two cameras to detect light emitted from the LEDs, a microprocessor, and a central processing unit. The robot should reach the target value (coordinates of a specified point) within a tolerance range of +0.0005 and −0.0003 in. The following deviations were recorded:

0.0001	0.0003	0.0002	−0.0002	−0.0001
0.0003	0.0004	−0.0003	−0.0002	0.0002
0.0001	−0.0003	−0.0002	0.0003	0.0004
0.0005	0.0004	−0.0003	0.0001	−0.0002

0.0004	−0.0001	−0.0002	0.0003	0.0001
−0.0003	0.0002	0.0004	0.0005	0.0001
0.0003	0.0003	−0.0002	−0.0003	−0.0001

When the loss per failure is \$30, what is the quality loss? If the process engineer replaced the current tolerance range by +0.0004 and −0.0004, what would the quality loss be? Perform sensitivity analyses to find the effect of a change in tolerance range on the quality loss.

2.3. A robot is used in a seam welding process. The robot is programmed so that it tracks the product on an overhead conveyor and welds the seam. The tolerance of the deviations of the weld from the center of the seam is ±0.005 in. The following observations of deviations were taken:

0.003	0.002	0.005	−0.004	−0.003
−0.002	0.003	0.004	−0.003	−0.005
−0.004	0.003	0.005	−0.004	0.003
0.004	0.005	−0.003	−0.003	0.004
−0.003	−0.002	0.005	0.004	0.005

A robot guidance system is introduced to increase the quality of the welded seams. The guidance system, essentially a machine vision system, guides the robot along the seam rather than using previously defined programmed points along the seam. The following deviations were observed after the introduction of the guidance system:

0.001	0.002	−0.001	−0.002	0.003
−0.002	0.003	−0.001	−0.003	0.002
−0.004	−0.002	0.001	0.002	0.001
0.003	0.002	0.001	−0.003	−0.001
0.000	−0.002	−0.003	0.002	0.001

(*a*) What is the effect of the introduction of the robot guidance?
(*b*) If the cost of a defective welded product is \$150, what would the savings be (if any) after the introduction of the guidance system?

2.4. A manufacturer of heat exchangers requires the plate spacing to be 0.25 ± 0.01 in. The quality control engineer sampled 25 heat exchangers and randomly measured the spacing between two plates on each exchanger. The following are the recorded measurements:

0.251	0.248	0.241	0.251	0.249
0.248	0.249	0.243	0.240	0.245
0.244	0.250	0.251	0.249	0.253
0.246	0.254	0.256	0.258	0.251
0.249	0.253	0.257	0.259	0.250

The loss due to a defective exchanger is \$50 (cost of adjusting the spacing). What is the expected loss when 200 units per day are produced?

A new method for plate insertion is introduced that has a C_p index of 1.5. What is the quality loss after introduction of the new method? Relate the C_p index to the quality loss per unit.

2.5. The quality of a product is defined by two characteristics: Brinell hardness number (BHN) and circular diameter. The specifications of these characteristics are:

Hardness in BHN	250 ± 5
Diameter	1.0000 ± 0.002 in

The following BHN measurements were taken:

248	250	249	252	253
249	247	249	250	251
250	249	248	250	251
249	245	246	249	254

The following measurements of the diameter were also taken:

1.0010	1.0020	1.0015	1.0009	1.0019
0.9998	0.9999	1.0020	1.0011	0.9997
0.9980	1.0010	1.0009	0.9996	0.9990
1.0000	1.0013	1.0009	1.0009	1.0009

The loss caused by unacceptable BHN is \$20, and the loss caused by unacceptable diameter size is \$30. What is the total expected loss caused by deviations from target values?

2.6. A manufacturer of tungsten carbide rollers uses a microabrasive superfinish to obtain a surface finish of 2 μin (micro-inches). The specification for the diameter of the roller is 10 ± 0.0010 in. Measurements are taken on every roller after the finishing process to check for the quality of the process. The cost of a defective roller less than 9.9990 inches in diameter is \$100, and the cost of a defective roller greater than 10.0010 inches in diameter is \$30. The following measurements were taken in inches:

9.9991	9.9990	9.9995	9.9996
10.0010	10.0008	10.0009	9.9990
9.9995	10.0000	9.9998	10.0007
10.0010	9.9990	9.9999	10.0010
10.0010	9.9996	9.9991	10.0010

What is the quality loss per roller? Assuming that the lower limit of roller diameter is 9.9999 in, what is the quality loss per roller? Assume that there is a measurement error that has a uniform distribution with $\sigma = 0.001$. What is the true loss?

2.7. A producer of steel plug gauges requires the tolerances of the plug diameter to be $\pm^{0.0010}_{0.0000}$ mm. The loss due to a defective plug is $25 (a plug is considered defective if its diameter is larger than the target diameter by more than 0.0010 mm). A random sample of 20 units was taken and the following deviations (in millimeters) were recorded:

0.0010	0.0001	0.0002	0.0008	0.0010
0.0010	0.0000	0.0007	0.0008	0.0010
0.0000	0.0003	0.0004	0.0000	0.0010
0.0009	0.0000	0.0000	0.0001	0.0010

(*a*) What is the quality level of the manufacturing process?
(*b*) What is the effect of correcting the upper tolerance limit to 0.0009 on the quality level of the process?

2.8. A supplier of automotive electronics uses a laser interferometric transducer whose signals are processed by a microcomputer to ensure that a critical fuel injector's sphere-headed needle will perform through one billion cycles—or for 400,000 miles of driving. The upper tolerance limit on roundness between sphere and injector seat is 1 μm, while the lower limit is 0.5 μm. The target value of the roundness is 0.8 μm. The cost of a defective injector when the roundness exceeds 1 μm is $107, while the cost of a defective injector when the roundness is less than 0.5 μm is $160. The following tolerances (in microns) were recorded.

0.60	0.70	0.55	0.59	0.81
0.90	0.91	0.86	0.57	0.86
0.78	0.98	0.76	0.50	0.50
0.87	0.90	0.86	0.87	0.85
0.78	0.68	0.91	0.92	0.69

What is the quality level of the production process?

2.9. A producer of high-strength wire cables that are used in lifting spreaders and containers in a shipyard requires that the strength of the cable be more than 40,000 lb/in^2. The cost of a defective cable (200 ft in length) is $900. The annual production rate is 6000 cables. The following data (in pounds per square inch) were obtained from destructive tests performed on 15 cables:

41,000	42,000	50,000	46,000	70,000
42,096	41,250	51,000	60,000	49,000
46,000	41,039	40,085	70,000	65,000

What is the quality level of the production process? What is the total quality loss per year?

2.10. Burrs represent a common problem for all machining, punching, or casting processes. The method of removing burrs has a direct effect on the quality of the deburring operation. Manual deburring is a boring, tiring, and monotonous operation, and the quality of deburring varies from one workpiece to another.

In order to achieve uniformity among the workpieces, a manufacturer

installed an automatic system consisting of a robot equipped with fixtures, deburring tools, and quick-change device. The robot uses the deburring tool and follows the contour of the workpiece with high accuracy to obtain the desired quality. A measure of burring operation quality is the absence of burrs and sharp edges, that is, that material fragments are not visible and sharpness cannot be felt. This can be achieved when the method used to remove the burrs or sharpness produces a chamfer radius of 0.02 in (0.51 mm) maximum. Clearly this is an S-type tolerance. The cost of deburring the workpiece is $3.00. The unit cost of the workpiece before the deburring operation is $15.00, and the operating cost of the robot is $10 per hour. Assuming that there is a production rate of 100 workpieces per hour and that the following measurements of the chamfer radius of 20 workpieces are obtained:

0.015	0.017	0.020	0.021	0.010
0.013	0.022	0.015	0.009	0.015
0.019	0.020	0.012	0.003	0.001
0.020	0.010	0.007	0.013	0.018

what is the quality loss per unit? Assuming that the manufacturer specified that the radius of the chamfer must not be less than 0.002, what would the quality loss be? What should the tolerance be if the manufacturer wishes not to incur quality loss greater than $0.50 per workpiece?

2.11. Most early bearing and seal failures and excessive operating temperatures result from misalignment in coupled equipment. A manufacturer introduced a laser alignment system to align shafts of coupled motors. In this system, the laser is aimed directly at the target where both angular and parallel shaft/coupling displacement measurements are taken. Each measurement includes periodic sampling of the position of the beam from the target. The accuracy of the parallel displacement measurement is 0.001 in; and the accuracy for angular displacement is 0.0001 in/in.

Assume that the manufacturer intends to use this system to detect and measure misalignment between the headstock and tailstock of a lathe. The distance between the headstock and tailstock is 60 in. The quality of the workpieces produced on this lathe is affected by the amount of misalignment of the workpiece. The relationship between cost of quality and amount of misalignment is

$$c(y) = ay^2 + by$$

where c = quality cost per unit
y = amount of misalignment in inches
a, b are constants having values of 40 and 70, respectively

The following measurements of misalignments are taken at the beginning of a production period:

0.0020	0.0030	0.0010	0.0009
0.0001	0.0021	0.0012	0.0013
0.0022	0.0031	0.0009	0.0009
0.0025	0.0023	0.0017	0.0013
0.0019	0.0008	0.0040	0.0033

The optimal value of misalignment is zero, resulting in zero loss. Assume that there is a production rate of 800 units during the production period. What is the total quality loss per unit? What do you recommend to reduce this loss to a minimum?

2.12. One objective of the finishing operations on gears is to eliminate slight inaccuracies in the tooth profile, spacing, and concentricity. These inaccuracies are very small dimensionally (< 0.0005 in) and may increase wear of the gears and cause undesirable noises at high speeds. In order to eliminate the inaccuracies in gears that are not heat treated, such operations as shaving are used. Shaving is a machining process (cutting) that removes only a few thousandths of an inch of metal.

A manufacturer uses a shaving operation to eliminate the inaccuracies in the teeth of gears produced by a hobbing operation. The quality of the gear is measured by the maximum inaccuracy in any of its teeth. After the shaving operation, the gear is inspected for the inaccuracies. If any of the gear parameters (tooth profile, spacing, and concentricity) have an inaccuracy greater than 0.0003 in, the gear is reworked through the shaving process until the inaccuracies are eliminated. On the other hand, if the inaccuracies cannot be removed through reworking, the gear is discarded at a cost of $100. The cost of reworking a gear is $10. Assume that the manufacturer produces 300 gears per hour, the probability of the gear being reworked is 0.10, and the probability of the gear being discarded is 0.02. A sample of 30 gears is taken to check the quality after the shaving process. The following measurements represent the maximum inaccuracy of each gear in the sample:

0.0001	0.0002	0.0004	0.0002	0.0003
0.0002	0.0004	0.0006	0.0001	0.0000
0.0001	0.0005	0.0003	0.0002	0.0003
0.0002	0.0000	0.0004	0.0007	0.0003
0.0002	0.0005	0.0006	0.0001	0.0004
0.0003	0.0001	0.0002	0.0003	0.0006

What is the quality loss per gear? How do you propose to improve the quality of the gears?

REFERENCES

Athan, M., and P. Falb. *Optimal Control*. New York: McGraw-Hill, 1966.

Box, G. E. P., and G. M. Jenkins. *Time Series Analysis, Forecasting, and Control*. San Francisco: Holden-Day, 1976.

Feigenbaum, A. V. *Total Quality Control*. New York: McGraw-Hill, 1983.

Jessup, P. "The Value of Continuing Improvement." *IEEE Conference on Communications*, vol. 1 (1985), pp. 89–94.

Sullivan, L. P. "Reducing Variability: A New Approach to Quality." *Quality Progress*, vol. 17, no. 7 (1984), pp. 15–21.

Taguchi, G. *Introduction to Quality Evaluation and Quality Control*. Tokyo: Japanese Standards Association, 1978.

Taguchi, G. *On-line Quality Control during Production*. Tokyo: Japanese Standards Association, 1981.

Taguchi, G., and Y. Wu. *Introduction to Off-line Quality Control*. Tokyo: Central Japan Quality Control Association, 1979.

CHAPTER 3

TOLERANCE DESIGN AND TOLERANCING

Designers often believe that their job is to design a product emphasizing performance, appearance, and perhaps reliability, and that the production and manufacturing engineer's job is to produce whatever has been designed. Design engineers expect that their specifications will be accepted by the production engineers. They do not give serious consideration to production costs and feasibility. Often, there is a natural reluctance to change a proven design for the sake of a reduction in manufacturing cost (Boothroyd, 1982). Therefore, it is important to determine the optimal values of a product's parameters and their tolerances. The parameter design of products, an activity we refer to as off-line quality control, is beyond the scope of this text; however, the tolerance design of the parameters will be treated in detail in this chapter.

Tolerance design and parameter tolerancing are important tasks of design engineers. In fact, the methods stated in this chapter are important to design engineers as well as to production engineers. The objective of this chapter is to examine the effect of tolerance design on the quality evaluation of a product and to introduce methods for economical design of product tolerance and tolerancing.

3.1 FUNCTIONAL LIMITS AND THE SOCIETAL LOSS

The characteristics of a product are usually defined by nominal values and functional limits. Consider, for example, the door-fitting problem for an automobile. Assume that the nominal length of the door (using off-line quality control methods or methods for robust designs) and the nominal length for a window in the door are m_1 and m_2, respectively (measured from the center of the door) and that the clearance Δ_0 of each is 3 mm. In other words, if the door size were 3 mm larger than the nominal size (m_1), the door would fail to close. On the other hand, if the door size were 3 mm smaller than the nominal size, the door would not be tightly closed, causing rainwater leakage. This clearance of 3 mm is defined as the functional limit Δ_0 of door size. The functional limit of a product characteristic is determined from the engineering knowledge of the product or through experimentation.

Consider also the case of the ignition voltage of an automobile engine. Assume that the nominal voltage needed for ignition is 20 kV (determined in the parameter design phase). The functional limit of the ignition voltage is obtained by conducting experiments to find the threshold value of the voltage at which the ignition fails. Suppose it is found that the ignition fails when the voltage drops below 8 kV or when it exceeds 35 kV. Thus, the lower and upper limits of the ignition voltage are 8 kV and 35 kV, respectively. This is expressed in the following way:

$$20^{-12}_{+15} \text{ kV}$$

However, in many applications it is more appropriate to use the smaller limit to give both limits equal values from the nominal. Under this practice, the ignition voltage and its limits become

$$20 \pm 12 \text{ kV}$$

When the ignition voltage exceeds its functional limits, the spark plugs will fail to ignite, causing the engine to stall. In turn, the customer will incur repair costs and other inconveniences. These losses (tangible and intangible) to the customer are referred to as societal loss.

Let $L(y)$ be the quality loss of a product having a characteristic value y. It is defined as follows:

$$L(y) = \frac{1}{n} \sum_{i=1}^{n} \int_0^T L_i(t, y)\, dt \tag{3.1}$$

where

T = design life of the product
n = size of the market
$L_i(t, y)$ = actual financial loss to customer i at year t due to a product having an initial characteristic value y. Note that $L_i(t, y)$ is not a continuous function of time.

The quality loss $L(y)$ of Eq. 3.1 can be approximated by a continuous function for large values of n. As shown in Chap. 2, $L(y)$ can be expanded using the Taylor series and the nominal value m:

$$\begin{aligned} L(y) &= L(m + y - m) \\ &= L(m) + \frac{L'(m)}{1!}(y - m) + \frac{L''(m)}{2!}(y - m)^2 + \cdots \end{aligned} \tag{3.2}$$

Assume that $L(m) = 0$ and $L'(m) = 0$. Disregarding the higher orders $L(y)$ can be approximated by the third term in the above equation:

$$L(y) = k(y - m)^2 \tag{3.3}$$

The constant k is determined by estimating the loss A_0 when y deviates from m by Δ_0. Thus,

$$A_0 = k\Delta_0^2$$

or

$$k = \frac{A_0}{\Delta_0^2}$$

Substituting the above value of k into Eq. 3.3 results in the following loss function expression:

$$L(y) = \frac{A_0}{\Delta_0^2}(y - m)^2 \tag{3.4}$$

Assume that the loss A_0 caused by the failure of the ignition system is \$200. Assume also that the failure of the ignition system is caused by the deviation of the ignition voltage from its nominal value by 12 kV; then

$$L(y) = \frac{200}{(12)^2}(y - 20)^2$$

or

$$L(y) = 1.39(y - 20)^2 \tag{3.5}$$

When the ignition voltage is 18 kV, the quality loss becomes

$$\begin{aligned} L(18) &= 1.39(18 - 20)^2 \\ &= \$5.56 \end{aligned} \tag{3.6}$$

This is the predicted loss of each customer during the design life of the ignition system when the ignition voltage is 18 kV.

It is clear that the tolerances of the product characteristics significantly affect the total quality loss. The optimal tolerances are those that minimize the total quality loss.

There are three sources of variation of product characteristics from their nominal values. They are environmental factors, deterioration factors, and imperfections in manufacturing processes. The product designer must deal with the first two causes of variation. In fact, an important function of the product designer is to find countermeasures for all sources of variation by using the parameter design approach. After the parameter design step is completed, the product designer should perform tolerance design, that is, trade-off analysis between cost of tight tolerance and product quality. The production and manufacturing engineers focus on the variations caused by manufacturing process imperfections.

In the following sections, we present approaches for determining the optimal tolerances for the three types of cases:

1. The-Nominal-The-Best (N type)
2. The-Larger-The-Better (L type)
3. The-Smaller-The-Better (S type)

3.2 TOLERANCE DESIGN FOR THE-NOMINAL-THE-BEST (N TYPE)

This type of tolerance is required for many products, parts, and components when the nominal size of the characteristic is preferred; dimensions, clearance, and viscosity are typical examples. Under this type of tolerance, the product designer and the manufacturer should aim for the nominal measurement of product characteristics, and the variation should be reduced to a minimum. The following example illustrates the tolerance design for such product characteristics.

Example 3.1. Consider a product whose principal dimension is denoted by y. A deviation of 500 μm in the principal dimension from its nominal value causes product failure and a loss A_0 of \$300. The dimension is affected by the environmental temperature x and the wear of the product. Assume that the standard deviation of temperature is 5°F and the design life T is 10 years. Assume also that the dimension y at year t is given by

$$y = y_0 + b(x - x_0) + \beta t \tag{3.7}$$

where y_0 = initial dimension
b = coefficient of thermal expansion
β = wear rate per unit of time

The initial dimension y_0 is the nominal dimension (preferred dimension) at normal temperature x_0. The mean squared deviation ν^2 of the dimension y is

$$\nu^2 = E[y_0 + b(x - x_0) + \beta t - m]^2$$

$$= \nu_0^2 + b^2\nu_x^2 + \frac{T^2}{3} \times \beta^2 \tag{3.8}$$

Determine the quality loss at the end of the product design life.

Solution. The quality loss L is obtained as follows:

$$L = \frac{A_0}{\Delta_0^2}\nu^2$$

Substituting Eq. 3.8 into the above equation, we obtain

$$L = \frac{A_0}{\Delta_0^2}\left(\nu_0^2 + b^2\nu_x^2 + \frac{T^2}{3} \times \beta^2\right)$$

$$= \frac{300}{(500)^2}\left(\nu_0^2 + b^2 \times 25 + \frac{10^2}{3} \times \beta^2\right) \tag{3.9}$$

Consider four grades of materials M_1, M_2, M_3, and M_4 that can be used for manufacturing the product. The prices, heat expansion coefficients, wear rates, quality losses, and total costs $(P + L)$ are shown in Table 3.1. Since ν_0 is the same for all materials, it is not included in the quality loss calculations shown in the table.

The optimal grade of material is M_3, since it has the lowest total cost (\$13.88) of the material grade. Grade M_1, the least expensive grade caused by poor quality, typifies Japanese-made products of the 50s and 60s. On the other hand, when a designer chooses grade M_4 (the most expensive grade), the quality loss decreases significantly. However, the price is too expensive, resulting in a higher value of the total cost than the product made of grade M_3. Thus, choosing M_1 or M_4 will have a negative effect on the product market share.

Usually the price and the quality loss should be about equal, but a difference factor of two or three times is acceptable. A difference factor of more than several times is not acceptable and is a clear sign of poor tolerance design.

As shown in Table 3.1, the optimal grade for tolerance design in this case is M_3. Tolerances for the initial value of the principal dimension need to be determined. The following example shows the estimation of these tolerances.

Example 3.2. Determine the tolerance of the principal dimension y for the product given in Example 3.1.

Solution. The tolerance can be obtained by using the following equation:

$$\Delta = \frac{\Delta_0 \text{ (functional limit)}}{\phi \text{ (factor of safety)}} \tag{3.10}$$

$$\text{where } \phi = \sqrt{\frac{A_0}{A}} \tag{3.11}$$

A_0 = loss to the customer caused by the failure of the product

TABLE 3.1
Parameters of the tolerance design

Material grade	Price P (\$)	Expansion rate b µm/°F	Wear rate β µm/year	Mean squared deviation ν^2	Quality loss L	Total $P + L$
M_1	2.00	5	28	26758	32.11	34.11
M_2	4.50	4	20	13733	16.48	20.98
M_3	8.00	2	12	4900	5.88	13.88
M_4	18.00	1	5	858	1.03	19.03

A = manufacturer's loss when the product does not conform to the specification limits

The product in Example 3.1 has the following parameters:

$$\Delta_0 = 500\ \mu\text{m}$$
$$A_0 = \$300$$
$$A = \$8$$

The factor of safety ϕ is

$$\phi = \sqrt{\frac{A_0}{A}} = \sqrt{\frac{300}{8}}$$
$$\phi = 6.124$$

and

$$\Delta = \frac{\Delta_0}{\phi} = \frac{500}{6.124} = 82\ \mu\text{m} \tag{3.12}$$

3.3 TOLERANCE DESIGN FOR THE-LARGER-THE-BETTER CHARACTERISTICS (L TYPE)

The-Larger-The-Better is applicable to characteristics such as the strength of materials and fuel efficiency. In these cases, the larger the value of the characteristic, the better it is. Tolerance design for The-Larger-The-Better characteristics is explained by the example given below.

Example 3.3. Consider two types of cables, T_1 and T_2. The price and strength for either type are proportional to the cable's cross-sectional area. The prices are $P_1 = \$1750/\text{mm}^2$ and $P_2 = \$2250/\text{mm}^2$, and the strengths are $S_1 = 220\ \text{kgf/mm}^2$ and $S_2 = 265\ \text{kgf/mm}^2$ for types T_1 and T_2, respectively. The lower tolerance limit of the cable's breaking strength is 20000 kgf, and the loss caused by falling below the lower tolerance limit is $58 million. Perform tolerance design and determine the tolerance limits for the better cable.

Solution. Following the same steps as in Example 3.1, we first calculate the total cost for each cable (price + quality loss). Let x be the cross-sectional area of the cable, which is the parameter being sought.

Cable type T_1. The total cost C is obtained as the sum of the price and the quality loss.

$$C = P_1 x + \frac{A_0\Delta_0^2}{(S_1 x)^2}$$
$$= 1750x + \frac{58{,}000{,}000 \times (20{,}000)^2}{(220x)^2} \tag{3.13}$$

The total cost is minimized by taking the derivative of Eq. 3.13 with respect to x and equating it to zero.

$$\frac{dC}{dx} = P_1 - \frac{2A_0\Delta_0^2}{S_1^2 x^3} = 0$$

or

$$x = \left(\frac{2A_0\Delta_0^2}{P_1 S_1^2}\right)^{1/3}$$
$$= 818\ \text{mm}^2 \tag{3.14}$$

The price of this cable is

$$1750 \times 818 = \$1.43\ \text{million} \tag{3.15}$$

Cable type T_2. The cross-sectional area is

$$x = \left(\frac{2 \times 58{,}000{,}000 \times (20{,}000)^2}{2250 \times (265)^2}\right)^{1/3}$$
$$= 665\ \text{mm}^2 \tag{3.16}$$

The price of cable T_2 is

$$2250 \times 665 = \$1.50\ \text{million} \tag{3.17}$$

Cable type T_1 is selected, since the price of T_1 is less than T_2. The tolerance of this cable is obtained using Eq. 2.29, Eq. 3.11, and the concept of section 2.3.

$$\Delta = \phi\Delta_0$$
$$= \sqrt{\frac{A_0}{A}} \times \Delta_0$$
$$= \sqrt{\frac{58{,}000{,}000}{1{,}430{,}000}} \times 20\ \text{metric tons force}$$
$$\Delta = 127.4\ \text{metric tons force} \tag{3.18}$$

3.4 TOLERANCE DESIGN FOR THE-SMALLER-THE-BETTER CHARACTERISTICS (S TYPE)

A typical example of a Smaller-The-Better characteristic is the residual dynamic unbalance of a rotor. Wear, machine accuracy, deterioration, and noise level are also examples of this type. Tolerance design for The-Smaller-The-Better characteristics is illustrated by the following examples.

Example 3.4. A new machine, M_1, is being considered for the replacement of an existing machine, M_2. The processing speed of M_1 is twice that of M_2, reducing the manufacturing cost by $1.20 per unit of production. The tolerance on the roundness of the production unit produced by machine M_1 is 20 μm (or less). Defective products are scrapped at a loss of $6.00 per defective unit.

Compare the quality levels of the two machines, based on 20 units produced by each machine under the same conditions of control, with the roundness measured in μm. The measurements are given in Table 3.2.

Solution. Assuming that the tolerance limit of the current machine (M_2) is acceptable, the quality losses for M_1 and M_2 are calculated as follows:

Machine M_1:

$$\nu_1^2 = \frac{1}{20}(3^2 + 5^2 + 0^2 + \cdots + 10^2)$$

$$= 72.35$$

$$L_1 = \frac{A}{\Delta^2} \times \nu_1^2$$

$$= \frac{6.00}{20^2} \times 72.35$$

$$= \$1.09$$

Machine M_2:

$$\nu_2^2 = \frac{1}{20}(7^2 + 14^2 + 15^2 + \cdots + 14^2)$$

$$= 95.10$$

$$L_2 = \frac{6.00}{20^2} \times 95.10$$

$$= \$1.43$$

Comparing L_1 and L_2, the quality loss of M_2 exceeds that of M_1 by \$0.34. The manufacturing cost per unit produced on M_1 is \$1.20 less than the manufacturing cost of M_2. The net result is an improvement of \$1.54 per unit of production when M_1 replaces M_2.

If the annual production is 200,000 units, the annual gain becomes \$308,000. When M_1 replaces M_2, a new tolerance limit for the roundness of the products must be determined. The upper tolerance limit of products produced by the existing machine is related to the factor of safety ϕ_0.

TABLE 3.2
Measurements of roundness

Machine	Roundness (μm)									
M_1	3	5	0	14	12	5	10	7	4	8
	8	15	6	7	4	17	6	2	0	10
M_2	7	14	15	3	0	5	7	9	13	2
	16	2	6	13	7	4	14	13	2	14

$$\phi_0 = \sqrt{\frac{A_0}{6.00}}$$

The new machine, M_1, should have the following factor of safety:

$$\phi = \sqrt{\frac{A_0}{4.80}}$$

The new tolerance limit for the roundness is

$$\text{New tolerance limit} = \text{current tolerance limit} \times \frac{\phi_0}{\phi}$$

$$= 20 \times \frac{\sqrt{A_0/6.00}}{\sqrt{A_0/4.80}}$$

$$= 20 \times \sqrt{\frac{4.80}{6.00}}$$

$$\approx 18\ \mu\text{m}$$

Example 3.5. The nominal thickness of the plastic coating on an electrical wire is an important variable in determining the wear characteristics of the wire. The smaller the wear, the better the wire. The loss caused by the wear of the coating is \$20.00, when the coating thickness is ± 1 (in thousandths of an inch) from its nominal value. Ten measurements of the wire coating taken from selected points along a wire gave the following deviations from the nominal thickness:

1, 0, 0.5, 2, 1.5, 1, 2, 0.25, 0.5, 1.3

A new coating process will be introduced to reduce the loss by 50 percent. The following measurements of deviations are taken along a wire coated by the new process:

1.0, 1.0, 0.3, 0.5, 0.9, 1.0, 1.1, 0.8, 0.75, 0.9

Determine the tolerance of the coating thickness for the new process and the loss for each unit for the current and new processes.

Solution. The mean squared deviations of the current and new process, ν_1 and ν_2, are estimated as

$$\nu_1^2 = \frac{1}{10}(1^2 + 0^2 + 0.5^2 + 2^2 + 1.5^2 + 1^2 + 2^2 + 0.25^2 + 0.5^2 + 1.3^2)$$

$$= 1.45$$

$$\nu_2^2 = \frac{1}{10}(1^2 + 1^2 + 0.3^2 + 0.5^2 + 0.9^2 + 1^2 + 1.1^2 + 0.8^2 + 0.75^2 + 0.9^2)$$

$$= 0.74$$

The tolerance of the coating thickness for the new process is

$$\Delta_2 = \sqrt{\frac{A_2}{A_1}} \times \Delta_1$$

where Δ_1 = tolerance limit of the current process
A_1 = loss due to the wear of a wire produced by the current process
Δ_2 = tolerance limit of the new process
A_2 = loss due to the wear of a wire produced by the new process

Thus

$$\Delta_2 = \sqrt{\frac{10}{20}} \times \Delta_1$$

$$\Delta_2 = \pm 0.71$$

The losses per unit are

$$L_1 = \frac{A_1}{\Delta_1^2} v_1^2$$

$$= \frac{20}{1} \times 1.45$$

$$= \$29$$

$$L_2 = \frac{A_2}{\Delta_2^2} v_2^2$$

$$= \frac{10}{(0.71)^2} \times 0.74$$

$$= \$14.8$$

3.5 TOLERANCE ALLOCATION FOR MULTIPLE COMPONENTS

In the foregoing sections, methods of tolerance design for three types of product characteristics were presented. They placed emphasis on determining the tolerance for the product characteristics so that the total cost (price and quality loss) is minimized. This section is intended to determine the optimal tolerance allocation for all components that constitute the final product. In other words, when a product consists of k components, the characteristics of the product are affected by the tolerances of these components, and tolerance design for each component is needed so that the output characteristic of the product conforms to the functional limit.

Suppose that x is the characteristic of a component that affects an output y (the desired characteristic of the product) having functional limit Δ_0. Assume that y is linearly affected by x, and that β is the linear constant; that is,

$$y = m_0 + \beta(x - m) \tag{3.19}$$

where m_0 is the nominal value of y (the target value) and m is the nominal value of x. Let A_0 be the loss when the value of the output characteristic y does not

conform to the functional limits, $m_0 \pm \Delta_0$. The loss function is

$$L = \frac{A_0}{\Delta_0^2}(y - m_0)^2 \tag{3.20}$$

Substituting Eq. 3.19 into Eq. 3.20, the loss becomes

$$L = \frac{A_0}{\Delta_0^2}[\beta(x - m)]^2 \tag{3.21}$$

Replacing L by A, which is the price of the component with the characteristic x (or the loss when x deviates from its functional limits),

$$A = \frac{A_0}{\Delta_0^2}[\beta(x - m)]^2 \tag{3.22}$$

Solving Eq. 3.22 for x, we obtain

$$x = m \pm \sqrt{\frac{A}{A_0}} \times \frac{\Delta_0}{\beta} \tag{3.23}$$

which is the optimal tolerance specification for the component characteristic x.

Suppose a product has k components which have characteristics x_i nominal values m_i, and prices A_i $(i = 1, \ldots, k)$. In addition, the linear constant for component i, which affects the characteristic of the product, is β_i $(i = 1, \ldots, k)$. The tolerances $\Delta_1, \Delta_2, \ldots, \Delta_k$ for each component are given as follows:

$$\Delta_1 = \sqrt{\frac{A_1}{A_0}} \times \frac{\Delta_0}{\beta_1}$$

$$\Delta_2 = \sqrt{\frac{A_2}{A_0}} \times \frac{\Delta_0}{\beta_2}$$

$$\vdots$$

$$\Delta_k = \sqrt{\frac{A_k}{A_0}} \times \frac{\Delta_0}{\beta_k}$$

Note that $\beta_i \Delta_i$ is the contribution to the displacement of y caused by characteristic x_i when $x_i = m_i + \Delta_i$. Thus the square of the range of the output y caused by the variation of components' characteristics $x_1, x_2, \ldots, x_k$, is

$$\Delta^2 = (\beta_1\Delta_1)^2 + (\beta_2\Delta_2)^2 + \cdots + (\beta_k\Delta_k)^2$$

$$= \left(\sqrt{\frac{A_1}{A_0}} \times \Delta_0\right)^2 + \left(\sqrt{\frac{A_2}{A_0}} \times \Delta_0\right)^2 + \cdots + \left(\sqrt{\frac{A_k}{A_0}} \times \Delta_0\right)^2$$

$$= \frac{A_1 + A_2 + \cdots + A_k}{A_0}\Delta_0^2$$

thus (3.24)

$$\Delta = \sqrt{\frac{\sum_{i=1}^{k} A_i}{A_0}}\Delta_0$$

or

Consider the following three cases:

$$(a)\ \text{If} \quad \sum_{i=1}^{k} A_i << A_0, \quad \text{then} \quad \Delta << \Delta_0$$

This situation occurs when the final assembled product must be scrapped, because there is no way of repairing it, and the scrapping loss A_0 becomes several times the total price of all parts ($\sum_{i=1}^{k} A_i$). If, for example, A_0 is 5 times the sum of the total price for all components, then

$$\Delta = \frac{1}{\sqrt{5}}\Delta_0 = 0.447\Delta_0$$

This means that even when all components' characteristics have the same uniform distribution within the range of their respective tolerances (i.e., process capability index is 0.577), the output characteristic (characteristic of the assembled product) would have a distribution with standard deviation of $2 \times 0.447\Delta_0/\sqrt{12} = 0.258\Delta_0$, and a C_p of

$$\frac{2\Delta_0}{6 \times (0.447\Delta_0/\sqrt{3})} \approx 1.3$$

If $\sum_{i=1}^{k} A_i << A_0$ and since $\Delta < \Delta_0$, then the final assembly inspection should be discarded.

$$(b)\ \text{If} \quad \sum_{i=1}^{k} A_i >> A_0 \quad \text{then possibly} \quad \Delta >> \Delta_0$$

This situation often occurs when an assembled product can be adjusted after observing that the output y does not conform to the specification $m_0 \pm \Delta_0$. Naturally, it is expected that the adjustment cost A_0 is less than the total price of the product's components. The actual distribution of the output characteristic may satisfy its tolerance, $m_0 \pm \Delta_0$, because the control limits of components are usually smaller than $m_i \pm \Delta_i$. However, there is a high probability that the output y does not conform to its tolerance or adjustment limits even though all components conform to their tolerances, $m_i \pm \Delta_i$.

$$(c)\ \text{If} \quad \sum_{i=1}^{k} A_i \approx A_0 \quad \text{then} \quad \Delta \approx \Delta_0$$

Situations (a) and (b) are much more likely to occur than situation (c). In the last case, output tolerances are allocated among component characteristics that affect the output characteristic y. The tolerance for each component can be estimated using Eq. 3.23.

Example 3.6. A product consists of three components: C_1, C_2, and C_3. The nominal values of their characteristics are m_1, m_2, and m_3, and their prices are $p_1 = \$20$, $p_2 = \$60$, and $p_3 = \$130$, respectively. Each component characteristic has an approximately linear effect on a product's characteristics around nominal values m_1, m_2, and m_3, with linear constants 2, 3, and 1.5 for C_1, C_2, and C_3, respectively. The quality loss when the product's characteristic does not conform to the specification is \$500, and the functional tolerance of the product characteristic is 3 (i.e., $\Delta_0 = 3$). Assuming that losses caused by component failures are equal to component prices, what is the tolerance for each component?

Solution. Let y be the characteristic of the product. Then the relationships between the product's characteristics and the components' are (see Eq. 3.19):

For component C_1, $\quad y = m_0 + \beta_1(x_1 - m_1)$

For component C_2, $\quad y = m_0 + \beta_2(x_2 - m_2)$

For component C_3, $\quad y = m_0 + \beta_3(x_3 - m_3)$

The total quality loss for the product is

$$L = \frac{A_0}{\Delta_0^2}\left[\sum_{i=1}^{3} \beta_i(x_i - m_i)\right]^2$$

The above equation is solved to obtain

$$x_i = m_i \pm \sqrt{\frac{A_i}{A_0}} \times \frac{\Delta_0}{\beta_i} \qquad i = 1, 2, 3$$

or

$$\Delta_i = \sqrt{\frac{A_i}{A_0}} \times \frac{\Delta_0}{\beta_i} \qquad i = 1, 2, 3$$

Thus, the tolerance limits for the three components are

$$\Delta_1 = \sqrt{\frac{A_1}{A_0}} \times \frac{\Delta_0}{\beta_1} = \sqrt{\frac{20}{500}} \times \frac{3}{2} = 0.30$$

$$\Delta_2 = \sqrt{\frac{A_2}{A_0}} \times \frac{\Delta_0}{\beta_2} = \sqrt{\frac{60}{500}} \times \frac{3}{3} = 0.34$$

$$\Delta_3 = \sqrt{\frac{A_3}{A_0}} \times \frac{\Delta_0}{\beta_3} = \sqrt{\frac{130}{500}} \times \frac{3}{1.5} = 1.02$$

and the tolerance for the product characteristic, Δ, is obtained using Eq. 3.24 as follows:

$$\Delta^2 = \frac{\sum_{i=1}^{3} A_i}{A_0} \times \Delta_0^2 = \frac{20 + 60 + 130}{500} \times (3)^2$$

$$\Delta^2 = 3.78$$

or

$$\Delta = 1.94$$

3.6 NONLINEAR TOLERANCING

In the previous section, we discussed an approach for tolerance design when the characteristics of each component have an approximately linear effect on the product's characteristics. We now introduce a methodology for tolerance design when the component characteristic x has a nonlinear effect on characteristic y (functional limits $m_0 \pm \Delta_0$). In such a case, a graphical methodology as shown in Fig. 3-1 is recommended.

In order to determine the tolerance limits for x, we first obtain the functional limits $m - \Delta_1$ and $m + \Delta_2$, which are the projections of $m_0 - \Delta_0$ and $m_0 + \Delta_0$, respectively. Let A_1 and A_2 be the losses when the characteristic of the component does not meet the lower specification and the upper specification limits, respectively. Then

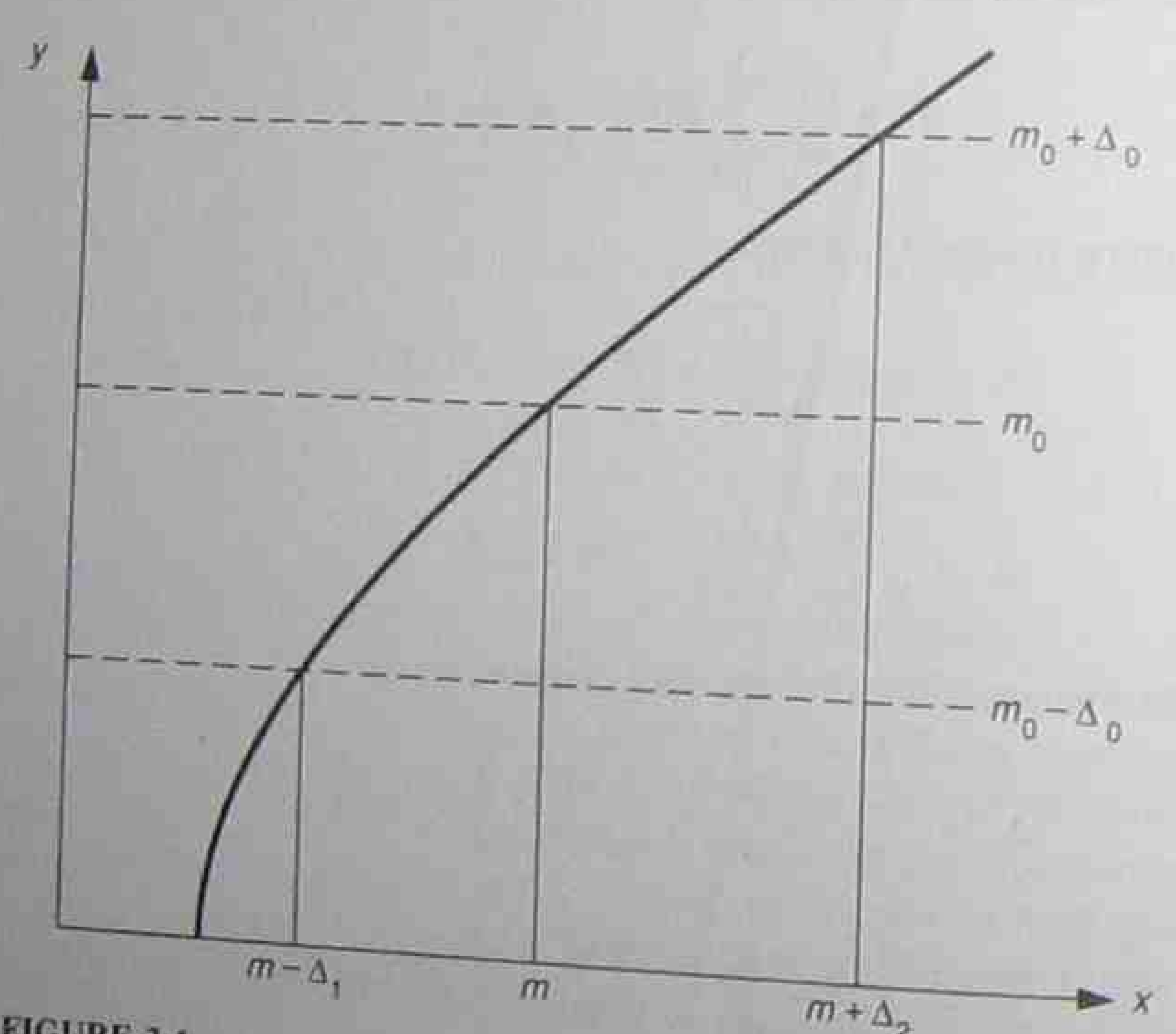


FIGURE 3-1
Nonlinear tolerancing.

$$\Delta_1 = \sqrt{\frac{A_1}{A_0}} \times \Delta_0$$

and

$$\Delta_2 = \sqrt{\frac{A_2}{A_0}} \times \Delta_0$$

To simplify the tolerancing, we choose the smaller tolerance limit as the tolerance limit for both sides of m. That is,

$$\Delta = \min(\Delta_1, \Delta_2)$$

is the tolerance limit for the characteristic x.

3.7 SUMMARY

This chapter introduced the concept of tolerance design and the methods of tolerancing. It discussed the effect of tolerance on the quality loss of a product.

Tolerance designs for The-Nominal-The-Best, The-Larger-The-Better and The-Smaller-The-Better characteristics were illustrated with examples. Finally, the chapter explained tolerance allocation for component characteristics having linear or nonlinear effects on a product's characteristics.

PROBLEMS

3.1 A manufacturer of tungsten carbide rollers requires that a critical dimension of the roller be 1 ± 0.1 in. The societal loss A_0 is \$100. The particular dimension is affected by environmental temperature x and rate of wear. The standard deviation of temperature x is 15°F. Design life T is 4 years, and $\nu = 0.03$ in.

There are four grades of material, S_1, S_2, S_3, and S_4, that may be used for the roller. These materials have different prices, heat expansion coefficients, and wear-out rates as shown below.

	Price p(\$)	Expansion rate (in/°F)	Wear-out rate (in/yr)
S_1	14.00	0.01	0.03
S_2	20.00	0.005	0.02
S_3	25.00	0.003	0.01
S_4	42.00	0.001	0.006

Choose a material from S_1, S_2, S_3, and S_4 that minimizes the overall cost. If A = \$0.40, calculate the manufacturing tolerance Δ for the best material. If a new heat treatment method is used to improve the wear-out rate of the S_1 grade material to 0.009 in/yr, what is the effect of the new method on the quality loss? Calculate the tolerance for S_1.

3.2 The existing tolerance on a certain clearance of a product is 15 μm or less. Products not conforming to certain specifications are scrapped at a loss of \$24.00. There are two types of machines, A_1 and A_2, that may be used to produce the product. The manufacturing cost using A_2 is \$2 less than that of A_1 for each piece. The following list contains data on 15 pieces of product produced by the two types of machines.

	Clearance (μm)				
A_1	3	7	4	13	8
	4	7	3	6	7
	5	9	4	4	3
A_2	8	9	11	9	7
	9	11	6	7	14
	14	11	11	9	7

Which machine should be chosen? What should be the tolerance on clearance?

3.3 Consider the power circuit of a television set that includes several resistors. The output voltage of the power circuit is 115 V with a functional limit of ± 25 V. The in-factory loss caused by a defective power circuit is 94¢. The resistance x of the power circuit affects its output voltage. An increase of the resistance by 1% results in an output-voltage decrease in the power circuit by 0.4 V. Consider three grades of resistors having the following mse (mean squared errors) from the nominal resistance. What is the best choice?

Grade	Price (¢)	mse
3rd	5	$(10\%)^2$
2nd	15	$(5\%)^2$
1st	40	$(2\%)^2$

3.4 Consider a manufacturer of automobile brake systems. A brake system is considered defective when the brake shoes wear by more than 3 mm, at which time the shoes are replaced at a cost of \$70 to the customer. The in-factory cost of replacing the brake shoes, if it is found that they do not meet the manufacturer's specifications, is \$20. The brake shoes currently used have a standard deviation of manufacture variation of 0.2 mm and a wear rate β of 0.5 mm/year, assuming preventive maintenance is done every 2 years. Estimate the quality loss.

3.5 A supplier of bumper bars (nickel plated) uses the electroplating process to improve appearance, protect base metal from corrosive attack, and to provide a wear-resistant surface. The thickness of the plated deposit depends upon the plating time and the amount of current passed. The specification of the thickness for the nickel plating ranges from 0.0002 to 0.0008 in, with a nominal value of 0.0005. If the plating is less than 0.0002 in thick, the supplier reworks the product at a cost of \$5. The following measurements of electric current are recorded below, in amperes (A):

16	15	12	13	14	13	14	14	16	14
14	16	15	14	13	14	13	15	15	13

The nominal value of electric current is 14 A. Find the quality level of the process, assuming the thickness is 0.0005 in at 14 amperes.

3.6 Assume that the supplier in Prob. 3.5 uses a more accurate electric power control system to ensure that the plating thickness is within the specified range. The cost of the new system is \$4000 each year. Consider that the annual production of bumpers is 50,000, and that the system improves the root mean squared deviation by a factor of 3.0. What is the total quality cost for the supplier?

3.7 Cemented carbide cutting tools used in a certain plant must be ground on their rake faces to resharpen them. This process can be completed in two minutes for each tool by an electrochemical grinding machine. The machine and equipment cost is \$6000. Depreciation for 10 years, interest, and taxes amount to \$1000 each year. The tolerance on the rake angle of the tool is $\pm 2°$. If the resharpened rake angle does not conform to the specified tolerance, it is discarded at a loss of \$150. Resharpening the tool can also be done in 4 minutes per tool using a diamond wheel on a conventional grinder, costing about \$1800 each year with an overhead cost of \$7.00 for each hour of operation. Twenty tools were resharpened, 10 using the electrochemical grinding and 10 using the conventional grinding. The following deviations of the rake angle from its nominal are obtained:

Electrochemical grinding: 1.5°, 1.2°, 1.3°, 1.0°, 1.0°, 0.9°, 0.8°, 0.9°, 0.7°, 0.9°

Conventional grinding: 1.5°, 1.3°, 1.4°, 1.6°, 1.4°, 1.3°, 1.0°, 1.0°, 0.9°, 0.9°

Which grinding method should be used so that the total cost is minimized? What is the annual loss if the number of tools to be resharpened annually is 3000?

3.8 A lamination consists of five layers, with the average dimensions shown below.

Layer	1	2	3	4	5
Dimension (0.001 in)	3	4	2	1	5

After the deposition of each layer, the mean squared deviations are found to be 0.0002, 0.0003, 0.0004, 0.0002, 0.0001, respectively. What are the optimal tolerances of each layer, if the overall tolerance is to be ± 0.0006 in?

3.9 A glass manufacturer uses ultrasonic machining where material is removed from the glass surface by microchipping or erosion with abrasive particles. The tip of the tool vibrates at low amplitude and high frequency. This, in turn, transmits a high velocity to the fine abrasive grains between the tool and the surface of the workpiece. With fine abrasives, tolerances of 0.0005 in or better can be achieved in the process. The manufacturer uses the ultrasonic machining to cut square glass plates with a principal dimension of 20 ± 0.0003 in. Glass plates having dimensions less than 19.9997 in are scrapped at a cost of \$60 per plate. On the other hand, glass plates with dimensions greater than 20.0003 in are recut with a 0.70-percent probability that the recut glass will meet the specifications of 20 ± 0.0003 in. The cost of recutting is \$10. A sample of 30 plates is taken, the principal dimension is measured, and the measurements are recorded as follows:

19.9990	19.9991	19.9998	19.9999	20.0000
19.9998	20.0001	20.0002	20.0001	19.9998
20.0006	19.9991	20.0005	20.0003	19.9999
19.9997	19.9998	20.0004	20.0002	20.0006
20.0000	19.9990	20.0003	19.9997	20.0000
19.9899	19.9998	19.9999	20.0000	20.0001

Assuming that the measurements follow a normal distribution, what is the quality loss per plate? What should the tolerance be so that the total quality loss per plate is minimized? What is the effect of recutting on the quality loss? Which of the following has greater effect on the quality loss:

(*a*) changing the tolerance of the principal dimension to 20 ± 0.0005 in; or

(*b*) reducing the cost of recutting to \$5 per plate?

3.10 An assembly having four components has a total tolerance on the assembly of 0.30 mm. A design engineer redesigned the assembly to have seven components, but for functional reasons the assembly must maintain the same tolerance of 0.30 mm. Assuming that the components have equal tolerances and that the loss of scrapping the assembly is \$25, what is the tolerance of each component? Suppose the cost of producing a component is directly related to its tolerance by the following equation:

$$C(\Delta) = a\Delta^2 + b\Delta^3$$

where Δ = tolerance of the component
a = a constant, 20
b = a constant, 100
$C(\Delta)$ = cost of producing the product with tolerance Δ

Which design for the assembly results in lower cost?

3.11 Single coatings of titanium carbide and titanium nitride, 5–10 μm thick, have improved the wear resistance factor of conventional cemented carbide tools. A manufacturer wishes to improve the wear resistance factor by layered coatings. However, the transition from substrate to coating, as well as between coating materials, usually involves abrupt changes in hardness and thermal expansion properties. During operation and during the cool-down period of manufacture, the difference in these properties produces stresses that often lead to cracking. Thick coatings of brittle refactory compounds also promote cracking when subjected to the heat caused by high speed and feed rates. The manufacturer devised a method that uses a microprocessor control on chemical vapor deposition furnaces to significantly improve the bonding processes. In addition, the microprocessor provides control of layer thickness to within a micron, resulting in uniform coatings.

It is extremely important that the coating thicknesses are within tolerance limits. The manufacturer intends to produce tools having eight layers—four layers of titanium nitride and titanium carbides separated by four layers of the bonding compound. The tolerance for thickness of the titanium layers is ± 0.001 in, and the tolerance for the bonding layers is ± 0.002 in. The tolerance in the tool should be ± 0.003 in. After using the microprocess for layer applications, 20 layers of the bonding compounds and 20 layers of the titanium were measured, and the tolerances of their thicknesses were as follows:

Titanium layers:

0.001	0.002	0.001	0.003	0.000
0.002	0.000	0.002	0.004	0.001
0.001	0.000	0.001	0.001	0.002
0.002	0.001	0.002	0.003	0.001

Bonding compound layers:

0.002	0.003	0.002	0.003	0.004
0.003	0.005	0.004	0.002	0.002
0.001	0.004	0.001	0.002	0.004
0.005	0.004	0.005	0.001	0.002

The loss caused by scrapping a defective tool (a tool that has tolerance greater than 0.003 in) is \$200. What is the quality loss per tool? What are the optimal tolerances for each type of layer that minimize the total quality loss per tool?

REFERENCES

Boothroyd, G., C. Poli, and L. E. Murch. *Automatic Assembly*. New York: Marcell Dekker, 1982.

Dallas, D. B. *Tool and Manufacturing Engineers Handbook*. New York: McGraw-Hill, 1976.

Farago, F. T. *Handbook of Dimensional Measurement*. New York: Industrial Press, 1982.

Kennedy, C. W., and D. E. Andrews. *Inspection and Gauging*. New York: Industrial Press, 1977.

Taguchi, G. "Off-line and On-line Quality Control Systems." *Proceedings of the International Conference on Quality Control*, vol. B4 (1978), pp. 1–5.

Taguchi, G., and Y. Wu. *Introduction to Off-line Quality Control*. Tokyo: Central Japan Quality Control Association, 1979.

CHAPTER 4

ON-LINE FEEDBACK QUALITY CONTROL: VARIABLE CHARACTERISTICS

Chapter 3 discussed the loss function technique for determining the tolerances of product characteristics. Keeping product characteristics close to the target values requires continuous monitoring and adjustments to the manufacturing processes during the production cycle.

There are many methods for controlling product quality during production cycles. Inspection of products during manufacturing, employment of diagnostic and adjustment processes, improvement of production processes, and the use of automatic control systems are some of the methods used. These methods constitute what is referred to as *on-line quality control*. An on-line quality control system that encompasses several of the above methods should be employed so that the desired target values of product characteristics can be controlled economically.

Unlike the traditional procedures for quality control, the methods explained in this chapter and those in Chaps. 5, 6, 7, and 8 determine the optimal control limits for parameters of the production process, and checking intervals that minimize the total loss of production. A principal objective of this chapter is to discuss the concept of on-line feedback quality control and its elements.

Before proceeding, a formal definition of on-line quality control would be useful. On-line quality control can be defined as *a set of quality control activities*

that are conducted during the production cycle of a product. One such activity is on-line feedback quality control, wherein measurements of product characteristics are obtained and analyzed, and the results fed back to upstream processes for adjustments. In this way, deviations of the characteristics of subsequent products are reduced.

4.1. FEEDBACK CONTROL WITH MEASUREMENT INTERVAL OF ONE UNIT OF PRODUCTION

In many of the companies belonging to the Toyota Group, the characteristic values of the products are automatically measured immediately after processing. When there is a deviation from the target value it is adjusted for the next piece of production. Moreover, the automatic measurements are often made by comparing every piece of production with an actual standard piece. This automatic control, with a measurement interval of one piece of production, is the unrecognized factor that made such drastic improvements in the quality level of Toyota's products. The reader might wonder: Is it really rational to make an adjustment in the production process even for the slightest deviations from the target values? That question and the basic concept of on-line feedback control systems are discussed in this section.

4.1.1 Mean Squared Drift

The mean squared drift per unit of production is obtained from the difference in measurements of two pieces of production produced in succession. It is denoted by σ_0^2. Assume that the measurements of quality characteristics of n pieces of production in succession are $y_1, y_2, \ldots, y_n$. The mean squared drift per unit of production is

$$\sigma_0^2 = \frac{1}{n-1}\left[(y_2 - y_1)^2 + (y_3 - y_2)^2 + \cdots + (y_n - y_{n-1})^2\right] \tag{4.1}$$

When measurements are taken at regular intervals on three pieces of production produced in succession, and the measurements of k sets of these pieces are denoted by

$$\begin{array}{ccc} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \vdots & \vdots & \vdots \\ y_{k1} & y_{k2} & y_{k3} \end{array}$$

where y_{ij} is the measurement of piece j in set i, then the mean squared drift per unit of production is

$$\sigma_0^2 = \frac{1}{2k}[(y_{12} - y_{11})^2 + (y_{13} - y_{12})^2 + (y_{22} - y_{21})^2 + (y_{23} - y_{22})^2 + \cdots + (y_{k2} - y_{k1})^2 + (y_{k3} - y_{k2})^2] \tag{4.2}$$

It should be noted that an estimation of σ_0^2, as given in Eqs. 4.1 and 4.2, may include measurement errors; that is, the change in the measurement values between successive pieces of production may contain a drift of measurement error. Suppose there is a three-minute interval between the production times of two successive pieces. The difference due to the drift of the measurement error during the three minutes will be included in the measurement values of the pieces.

Let the mean squared drift of the true characteristic value of two pieces of production in succession be denoted by σ_1^2 and the mean squared drift of measurement error be denoted by σ_m^2; then the following relationship holds:

$$\sigma_0^2 = \sigma_1^2 + \sigma_m^2 \tag{4.3}$$

The subject of large mean squared drift of measurement error will be discussed later. At the moment, it is assumed that the measurement error mean squared drift σ_m^2 is sufficiently small when compared with the mean squared drift from one piece to another, σ_1^2. That is, the following assumption is made:

$$\sigma_0^2 \approx \sigma_1^2 \tag{4.4}$$

The following two examples show how the mean squared drift and the mean squared drift of measurement errors are estimated:

Example 4.1. The dimensions of a certain product are measured immediately after its final manufacturing process. Measurements are taken twice daily on four workpieces processed in succession for a period of 10 days. Measurement data (deviation from target) are shown in Table 4.1. What is the mean squared drift per unit of production?

TABLE 4.1
Measurement data of dimension (μm)

Day	A.M.				P.M.			
1	0	−1	1	0	2	2	2	1
2	0	1	1	2	−1	0	1	1
3	−1	1	2	3	−1	−1	−2	−1
4	−1	−2	−3	−3	1	1	1	2
5	6	5	5	5	−3	−2	−2	−2
6	0	1	1	2	3	4	4	2
7	0	0	1	1	−3	−4	−5	−6
8	1	2	2	1	0	2	1	2
9	7	6	6	6	−5	−6	−7	−7
10	0	0	0	0	2	0	0	0

Solution. Using Eq. 4.1, the mean squared drift per unit of production, σ_0^2, is

$$\begin{aligned}\sigma_0^2 &= \frac{1}{20 \times 3}[(0+1)^2 + (-1-1)^2 + (1-0)^2 + \cdots + (2-0)^2 + (0-0)^2 + (0-0)^2] \\ &= \frac{1}{60} \times 50 \\ &= 0.833\ \mu\text{m}^2\end{aligned}$$

Example 4.2. In order to eliminate the impact of measurement error, the manufacturer in Example 4.1 randomly selects one unit of production and repeats the measurements on the same dimension of the product several times. Assume that the manufacturer performs four consecutive measurements of the dimension of one unit of production and that the measurements are made for six days, once in the morning and once in the afternoon. The results of these measurements are shown in Table 4.2. Find the true mean squared drift of the dimension.

Solution. Using Eq. 4.1 and the data of Table 4.2, the mean squared drift caused by measurement error, σ_m^2, is obtained as

$$\begin{aligned}\sigma_m^2 &= \frac{1}{12 \times 3}[(5-5)^2 + (5-4)^2 + (4-4)^2 + \cdots + (0-0)^2 + (0-0)^2 + (0-1)^2] \\ &= \frac{1}{36} \times 11 \\ &= 0.306\ \mu\text{m}^2\end{aligned}$$

The mean squared drift per unit of production as it relates to the product characteristic value, σ_1^2, is

$$\begin{aligned}\sigma_1^2 &= \sigma_0^2 - \sigma_m^2 \\ &= 0.833 - 0.306 \\ &= 0.527\ \mu\text{m}^2\end{aligned}$$

TABLE 4.2
Measurement data for Example 4.2

Day	Morning				Afternoon			
1	5	5	4	4	3	3	3	2
2	0	0	0	0	1	1	1	1
3	−2	−2	−1	−1	−5	−4	−4	−3
4	0	1	1	1	−1	−1	0	0
5	4	4	3	3	2	2	1	0
6	−3	−3	−3	−3	0	0	0	1

4.2 THE LOSS FUNCTION

The concept of the loss function, as discussed in Chap. 2, can be applied to estimate the losses due to the inclusion of both the mean squared drift of the product characteristic and the mean squared drift of measurement error. Let the annual cost of an automatic control system be Q, and the annual production be N units. The tolerance limit of the product characteristic value is $\pm\Delta$, and the loss due to a defective item is A. When the automatic control system operates properly without failure, its loss function is given by Eq. 4.5, provided that adjustments are made when there is a difference between the value of the measured characteristic and its target value.

$$L = \frac{Q}{N} + \frac{A}{\Delta^2}\sigma_0^2 \tag{4.5}$$

Example 4.3. The annual cost of the control system in Examples 4.1 and 4.2 is \$15,000, and the annual production is 800,000 units. The characteristic value to be controlled is the dimension of the product, which has a tolerance of $\pm$ 10 μm. The loss caused by a defective piece (rework cost) is \$6. What is the quality cost per unit of production?

Solution. The mean squared drift per unit of production, σ_0^2, which includes both the mean squared error among the manufactured units and the mean squared error caused by measurement, is 0.833.

Using Eq. 4.5, the quality cost per unit of production is obtained as

$$Q = \$15,000$$

$$N = 800,000$$

$$A = \$6$$

$$\Delta = \pm 10$$

$$L = \frac{15,000}{800,000} + \frac{6}{10^2} \times 0.833$$

$$= 0.01875 + 0.045$$

$$= \$0.064$$

4.3 FEEDBACK CONTROL WITH MEASUREMENT INTERVALS GREATER THAN ONE UNIT OF PRODUCTION

Automatic control systems for which every piece of production is measured immediately after processing, using a standard piece for comparison, were discussed in a previous section. In many situations, it is difficult (if not impossible) to have an automatic control system capable of such measurements and controlling the production system. In these cases, operators usually undertake the measurements

and the checking of the product. The cost of measurements, when made by operators, is usually higher than when automatic control systems are used. Moreover, the operator may not be able to measure every piece of production, resulting in a higher loss of quality per unit of production.

Decreasing the measurement interval to one piece of production is recommended when operators are performing the measurements, provided it is economically feasible. Obviously, if the operator is capable of measuring every piece, the loss function will be identical to that of an automatic control system, given by Eq. 4.5. This section is intended to determine, using the loss function, the optimal measurement interval when operators are performing the measurements so that the quality cost per unit is minimized. The following parameters are defined:

m = target value of the product characteristics

Δ = tolerance of the product characteristics

A = in-plant cost of reworking or scrapping a unit that falls outside of tolerance interval

B = cost per measurement of the product characteristics

C = cost per adjustment

n_0 = current measurement interval (units)

n = measurement interval

D_0 = current adjustment or control limit

D = adjustment or control limit

l = time lag of measurement

u_0 = current average number of products (units) between successive adjustments

$\overline{u}$ = predicted average number of products between successive adjustments

L = total cost of measurement and adjustment per unit: the sum of diagnosis cost, measurement cost, adjustment cost, and cost of time lag

σ_m^2 = measurement error

The diagnosis cost per product is given by

$$\text{Diagnosis cost per product} = \frac{B}{n} \tag{4.6}$$

The adjustment cost per product is

$$\text{Adjustment cost per product} = \frac{C}{\overline{u}} \tag{4.7}$$

The loss per product and the cost of time lag per product are obtained as follows. From Chap. 2, the loss caused by a product variation from nominal is

$$L_1 = k\nu^2 = \frac{A}{\Delta^2}\nu^2 \tag{4.8}$$

Assume that the control limits for production process adjustment are set at $\pm D$, and the product characteristics being measured follow a uniform distribution within this range ($\pm D$).

If the production process is found to be under control during diagnosis (the product characteristics are within the control limits), then the mean squared deviation from the target value is approximated by $D^2/3$ (the variance of the uniform distribution $= [(m + D) - (m - D)]^2/12$). The loss per unit is related only to the amount of deviation from the target value (no units are outside the control limits). Loss caused by deviation when the production process is under control is obtained by substituting $\nu^2 = D^2/3$ in Eq. 4.8.

$$L_1 = \frac{A}{\Delta^2} \times \frac{D^2}{3} \tag{4.9}$$

If the production process is found to be out of control during diagnosis, then the mean squared deviation is

$$\nu^2 = \left(\frac{n+1}{2} + l\right)\frac{D^2}{\bar{u}} \tag{4.10}$$

Equation 4.10 establishes that the variance of deviation is proportional to the number of defective products between successive diagnoses. Note that $(n + 1)/2$ is the average number of defective units between successive diagnoses and l is time lag (in units). (Figure 6-1 in Chap. 6 illustrates this in more detail.) By substituting Eq. 4.10 in Eq. 4.8, the following result is obtained:

$$L_2 = \frac{A}{\Delta^2}\left(\frac{n+1}{2} + l\right)\frac{D^2}{\bar{u}} \tag{4.11}$$

Measurement error is an independent source of variation, causing an increase of quality loss by

$$L_3 = \frac{A}{\Delta^2} \times \sigma_m^2$$

The total quality cost per product is obtained by adding Eqs. 4.6, 4.7, 4.9, and 4.11, and L_3.

$$L = \frac{B}{n} + \frac{C}{\bar{u}} + \frac{A}{\Delta^2}\left[\frac{D^2}{3} + \frac{D^2}{\bar{u}}\left(\frac{n+1}{2} + l\right) + \sigma_m^2\right] \tag{4.12}$$

The predicted average number of products between successive adjustments, $\bar{u}$, is

$$\bar{u} = \frac{D^2}{D_0^2} \times u_0 \tag{4.13}$$

This proportion is based on the Brownian motion principle, which is derived from the assumption that the average time for a randomly-moving particle to go a certain distance is proportional to the distance squared. When the current average adjustment interval u_0 and the current control limit D_0 are not available, the average adjustment interval can be estimated using the following equation (Taguchi, 1984)

$$u = \frac{(\text{distance})^2}{\text{variance of drift per unit product}} \tag{4.14}$$

The optimal diagnosis interval is obtained by substituting Eq. 4.13 into Eq. 4.12 and taking the derivative of Eq. 4.12 with respect to n and setting it to zero. This results in

$$n^* = \sqrt{\frac{2u_0 B}{A}} \times \frac{\Delta}{D_0} \tag{4.15}$$

The optimal control limit is derived by differentiating Eq. 4.12 with respect to D (assuming Eq. 4.13) and setting the derivative to zero:

$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4} \tag{4.16}$$

The predicted process capability index C_p, as defined in Eq. 2.1, of the production process is obtained by adding the deviation when the production process is under control ($D^2/3$) and the deviation given by Eq. 4.10:

$$C_p = \frac{2\Delta}{6\sqrt{\dfrac{D^{*2}}{3} + \left(\dfrac{n^*+1}{2} + l\right)\dfrac{D^{*2}}{\bar{u}}}} \tag{4.17}$$

Example 4.4. The manufacturer of integrated circuits for computers wishes to install an automatic measurement control system that has the same parameters as those given in Example 4.3. Presently, the measurements are taken by an operator, and the system has the following parameters:

Tolerance Δ:	10 μm
Loss due to a defective piece, A:	\$6.00
Measurement cost B:	\$1.50
Adjustment cost C:	\$12.00
Time lag l:	3 units
Current control limit D_0:	3 μm
Observed average adjustment interval u_0:	180 units

What are the quality losses per unit of production for both the automatic and the present measurement control systems?

Solution. Substitute the above information in Eqs. 4.15 and 4.16 to obtain the optimal measurement interval n^* and the optimal control limit D^*, respectively, for the present control system:

$$n^* = \sqrt{\frac{2\,u_0 B}{A}} \times \frac{\Delta}{D_0}$$

$$= \sqrt{\frac{2 \times 180 \times 1.50}{6}} \times \frac{10}{3}$$

$$\approx 30 \text{ units} \qquad (4.18)$$

$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4}$$

$$= \left(\frac{3 \times 12}{6} \times \frac{3^2}{180} \times 10^2\right)^{1/4}$$

$$\approx 2.3 \qquad (4.19)$$

The predicted average number of products between successive adjustments is

$$\bar{u} = u_0 \times \frac{D^{*2}}{D_0^2}$$

$$= 180 \times \frac{2.3^2}{3^2}$$

$$= 106 \qquad (4.20)$$

The quality cost per unit of production under the present measurement system with optimized measurement interval n^* and optimal control limit D^* is

$$L = \frac{B}{n^*} + \frac{C}{\bar{u}} + \frac{A}{\Delta^2}\left[\frac{D^{*2}}{3} + \left(\frac{n^*+1}{2} + l\right)\frac{D^{*2}}{\bar{u}}\right]$$

$$= \frac{1.50}{30} + \frac{12.00}{106} + \frac{6}{10^2}\left[\frac{2.3^2}{3} + \left(\frac{31}{2} + 3\right) \times \frac{2.3^2}{106}\right]$$

$$= 0.050 + 0.113 + 0.106 + 0.055$$

$$= \$0.324 \qquad (4.21)$$

In Eq. 4.21, the first term ($0.05) is the cost of measurement per unit, the second term ($0.113) is the cost of adjustment, the third term ($0.106) is the loss associated with deviation within the control limits, and the fourth term is the loss caused by drift from target between diagnoses. The quality loss caused by the measurement error has been ignored.

Comparison between quality cost of the automated measurement control system, as given in Example 4.3, and the cost of the present measurement system shows that the quality cost of the present system is 5.1 times the cost of the automated system. Consequently, we recommend that the present measurement system be replaced by the automatic measurement system, resulting in annual savings of

$$(0.324 - 0.064) \times 800{,}000 = \$208{,}000$$

4.3.1 Control Systems for Lot or Batch Types of Production

Example 4.5. Lot type production. An injection molding process produces 12 pieces at a time (12 pieces per "shot"). The average value of the width of each piece is checked once every 2 hours at a cost of $2. The checking process involves the immersion of the piece in ice water and then measurement of its width with error variance $(2.5\ \mu\text{m})^2$. The tolerance of the width is $m \pm 15\ \mu\text{m}$, where m is the target value of the width. The loss for each defective piece is $0.16, but if one piece in a shot of 12 measured is out of specification, all 12 pieces in the shot are discarded. The current control limit for the mean width is $m \pm 5\ \mu\text{m}$, and the average adjustment interval is 8 hours.

The adjustment cost is $15 and time lag l is 10 shots. Assuming there is an hourly production rate of 120 shots, 2000 working hours per year, and that one shot is the basic unit of production, find the parameters of the optimal control system and estimate the yearly savings.

Solution. The parameters of the injection molding process are:

$$\Delta = 15\ \mu\text{m}$$

$$A = \$0.16 \times 12 = \$1.92$$

$$B = \$2.00$$

$$C = \$15.00$$

$$n_0 = 2 \times 120 = 240 \text{ shots (current checking interval)}$$

$$D_0 = 5\ \mu\text{m}$$

$$u_0 = 8 \times 120 = 960 \text{ shots}$$

$$l = 10 \text{ shots}$$

$$\sigma_m^2 = (2.5\ \mu\text{m})^2$$

The quality cost per shot of the current control system, L_0, is

$$L_0 = \frac{B}{n_0} + \frac{C}{u_0} + \frac{A}{\Delta^2}\left[\frac{D_0^2}{3} + \left(\frac{n+1}{2} + l\right)\frac{D_0^2}{u_0} + \sigma_m^2\right]$$

$$= \frac{2.00}{240} + \frac{15.00}{960} + \frac{1.92}{15^2}\left[\frac{5^2}{3} + \left(\frac{241}{2} + 10\right) \times \frac{5^2}{960} + 2.5^2\right]$$

$$= 0.008 + 0.016 + 0.071 + 0.029 + 0.053$$

$$= \$0.179$$

The optimal checking interval n^* is

$$n^* = \sqrt{\frac{2\,u_0 B}{A}} \times \frac{\Delta}{D_0}$$

$$= \sqrt{\frac{2 \times 960 \times 2.00}{1.92}} \times \frac{15}{5}$$

$$= 134 \rightarrow 120 \text{ (once per hour)}$$

The optimal control limit is

$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4}$$

$$= \left(\frac{3 \times 15}{1.92} \times \frac{5^2}{960} \times 15^2\right)^{1/4}$$

$$= 3.4 \rightarrow 3.5$$

The average number of units between adjustments is

$$\overline{u} = u_0 \times \frac{D^{*2}}{D_0^2}$$

$$= 960 \times \frac{3.5^2}{5^2}$$

$$= 470$$

Quality cost per shot after the implementation of the optimal parameters of the control system, L, becomes

$$L = \frac{B}{n^*} + \frac{C}{\overline{u}} + \frac{A}{\Delta^2}\left[\frac{D^{*2}}{3} + \left(\frac{n^* + 1}{2} + l\right)\frac{D^{*2}}{\overline{u}} + \sigma_m^2\right]$$

$$= \frac{2}{120} + \frac{15}{470} + \frac{1.92}{15^2}\left[\frac{3.5^2}{3} + \left(\frac{121}{2} + 10\right)\frac{3.5^2}{470} + 2.5^2\right]$$

$$= 0.017 + 0.032 + 0.034 + 0.016 + 0.053$$

$$= \$0.152$$

The yearly savings are

$$(0.177 - 0.152) \times 120 \times 2000 = \$6,000$$

Example 4.6. Batch type production. In a typical batch type production system, 500 photo receptors are coated at a time. The critical characteristic of the coating is its thickness, which is specified to be 100 ± 30 μm. The loss of a defective unit is \$25. The quality of the process is checked by comparing the average thickness of the coating against the target value m at a cost of \$20. The difference from target m is adjusted when the amount of deviation is more than 2 percent of nominal value. The cost of adjustment is \$12 and the average adjustment interval is 2.5 batches with a time lag of zero. The annual production is 1200 batches. Assuming the estimated measurement error variance is $\sigma_m^2 = (1.2\ \mu\text{m})^2$, (1) find the optimal parameters of the control system and the annual savings (if any) when such a system is employed; and (2) find the annual savings if the control system obtained in (1) is improved by introducing a new device, so that the measurement cost is \$45, and the estimated measurement error variance is $(0.3\ \mu\text{m})^2$.

Solution. (1) The parameters of the present control system are

$$\Delta = 30\ \mu\text{m}$$

$$A = 500 \times 25 = \$12,500 \text{ per batch}$$

$$B = \$20$$

$$C = \$12$$

$$D_0 = 2\ \mu\text{m}$$

$$n = 1 \text{ batch}$$

$$u_0 = 2.5 \text{ batches}$$

$$l = 0$$

$$\sigma_m^2 = (1.2\ \mu\text{m})^2$$

The quality cost per batch for the present system is obtained by using Eq. 4.12:

$$L = \frac{B}{n} + \frac{C}{u_0} + \frac{A}{\Delta^2}\left[\frac{D_0^2}{3} + \left(\frac{n+1}{2} + l\right)\frac{D_0^2}{u_0} + \sigma_m^2\right]$$

$$= \frac{20}{1} + \frac{12}{2.5} + \frac{12500}{30^2}\left[\frac{2^2}{3} + \left(\frac{2}{2} + 0\right)\frac{2^2}{2.5} + 1.2^2\right]$$

$$= 20.00 + 4.80 + 18.50 + 22.20 + 20.00$$

$$= \$85.50$$

The optimal parameters of the control system are

$$n^* = \sqrt{\frac{2 u_0 B}{A}} \times \frac{\Delta}{D_0}$$

$$= \sqrt{\frac{2 \times 2.5 \times 20}{12500}} \times \frac{30}{2}$$

$$= 1.3 \rightarrow 1 \text{ batch}$$

$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4}$$

$$= \left(\frac{3 \times 12}{12500} \times \frac{2^2}{2.5} \times 30^2\right)^{1/4}$$

$$= 1.4 \rightarrow 1.5\ \mu\text{m}$$

and $\overline{u}$ is

$$\overline{u} = u_0 \times \frac{D^{*2}}{D_0^2}$$

$$= 2.5 \times \frac{1.5^2}{2^2}$$

$$= 1.4$$

The quality cost of the control system when operating at optimal parameters is

$$L = \frac{B}{n^*} + \frac{C}{\bar{u}} + \frac{A}{\Delta^2}\left[\frac{D^{*2}}{3} + \left(\frac{n^*+1}{2} + l\right)\frac{D^{*2}}{\bar{u}} + \sigma_m^2\right]$$

$$= \frac{20}{1} + \frac{12}{1.4} + \frac{12500}{30^2}\left[\frac{1.5^2}{3} + \left(\frac{2}{2} + 0\right)\frac{1.5^2}{1.4} + 1.2^2\right]$$

$$= 20.00 + 8.60 + 10.40 + 22.30 + 20.00$$

$$= \$81.30$$

The annual savings are only

$$(85.50 - 81.30) \times 1200 = \$5040$$

(2) The parameters of the new system are

$$A = \$12500$$

$$B = \$45$$

$$C = \$12$$

$$D^* = 1.5\ \mu\text{m}$$

$$n = 1 \text{ batch}$$

$$\bar{u} = 1.4 \text{ batches}$$

$$l = 0$$

$$\sigma_m^2 = (0.3\ \mu\text{m})^2$$

From Eq. 4.14, we obtain

$$n^* = \sqrt{\frac{2\bar{u}B}{A}} \times \frac{\Delta}{D_0}$$

$$= \sqrt{\frac{2 \times 1.4 \times 45}{12500}} \times \frac{30}{1.5}$$

$$= 2 \text{ batches}$$

$$D^* = 1.5\ \mu\text{m (unchanged)}$$

$$\bar{u} = 1.4 \text{ (unchanged)}$$

$$L = \frac{B}{n^*} + \frac{C}{\bar{u}} + \frac{A}{\Delta^2}\left[\frac{D^{*2}}{3} + \left(\frac{n^*+1}{2} + l\right)\frac{D^{*2}}{\bar{u}} + \sigma_m^2\right]$$

$$= \frac{45}{2} + \frac{12}{1.4} + \frac{12500}{30^2}\left[\frac{1.5^2}{3} + \left(\frac{3}{2} + 0\right) \times \frac{1.5^2}{1.4} + 0.3^2\right]$$

$$= 22.50 + 8.57 + 10.42 + 33.48 + 1.25$$

$$= \$76.22 < \$81.30$$

The new device lowers the quality cost of the control system obtained in (1) by \$5.08.

4.4 SUMMARY

This chapter introduced the concept of on-line feedback quality control systems and discussed the effect of mean squared drift of measurements on the parameters of the feedback quality control systems. Two types of feedback control systems were also discussed: (1) feedback control systems that require measurements on every piece of production; and (2) feedback control systems that require measurements after the production of a number of products. In all cases, total quality cost, including the loss function, is used as a criterion for design and evaluation.

PROBLEMS

4.1. A feedback control system is used in measuring the surface roughness of gauge blocks. The following successive measurements (in microns) are taken:

10	11	7	9	11	10	8	7	6	12
10	7	4	9	8	8	7	6	10	11
11	14	3	7	9	8	8	6	10	12
11	14	7	11	9	10	8	9	4	7

The specification for the surface roughness is that it not be more than 15 μm, and the loss caused by a defective gauge block is \$30. What is the quality loss per block?

4.2. A quality characteristic for the diameter of steel shafts is defined by the specification 0.750 ± 0.002 in, and defective shafts are scrapped at a cost of \$0.8. The production processes for producing such shafts involve sawing them to their required lengths, and then turning them to their specified diameters. The shafts are then inspected to ensure that their diameters meet the defined specifications.

The turning operation is diagnosed once every 50 shafts. The direct cost of a measurement is \$1.5, and the time lag of the adjustment procedure is represented by 5 shafts.

The adjustment cost of the turning operation is \$30, and the average number of shafts produced between successive adjustments (with adjustment limit of 0.75 ± 0.001 in) is 800. Find the optimal parameters of the control system, and compare the total cost per unit.

4.3. The paint department of an automobile manufacturer uses 20 robots for spraying paint on the metal frames of automobiles. The robot requires operator service to unclog the spray nozzle or to replace the nozzle when necessary. The quality characteristic of the painting process is thickness uniformity of the applied paint within the specification limits of ±30 μm. A robot takes 2 minutes to complete the painting process needed per frame. The spraying process is evaluated by measuring the paint thickness at intervals of 50 automobiles at a cost of \$4 per measurement.

A frame that has a thicker or thinner layer of paint than required is repainted (after preparation) at a cost of \$70, and the time lag of diagnosis is 2 frames. The average number of automobiles painted between successive adjustments is 1000, and the average adjustment time is 1 hr, at a cost of \$20 per adjustment. Find the optimal diagnosis interval and the optimal number of operators to be assigned for robot diagnosis and robot adjustment.

4.4. Consider an EMD (electro-machining discharge) method for razor-blade production. The quality characteristic of the process is measured by the length of a slot to be made in the blade with a tolerance of ±0.01 in. The cost of a defective unit is \$1, and the cost of measuring is \$0.50. The time lag is 20 units, and the direct cost of adjustment is \$10. The current control limit is ±0.005 in, and the current average adjustment interval is 2000 units. The process is diagnosed once every 500 units of production. Find the total loss per unit under the current operating conditions. What are the optimal n^* and D^* that minimize quality cost per unit?

4.5. Assume that there is a measurement error in Prob. 4.1. The measurement error has been checked once a week with calibration limit $D = 0.5\ \mu$m. The average calibration interval is 7 days. Assuming the rms (root mean square error) of the gauge block used as the standard is 0.1 μm, and using the following formula:

$$\sigma_m^2 = \frac{D^2}{3} + \left(\frac{n+1}{2}\right) \times \frac{D^2}{u} + \sigma_s^2$$

where σ_s^2 is the rms of the standard block, obtain the error variance of the measuring device.

4.6. An industrial engineer wishes to determine whether a fully automated feedback control system or a semiautomated system should be used in the control of a manufacturing process. The following are the parameters of the two systems:

Parameter	Fully automated system	Semiautomated system
B	\$2.50	\$1.00
C	\$12.00	\$10.00
σ_0^2	1.5 μm^2	2.5 μm^2
u_0	300 units	220 units
l	0 units	2 units

The tolerance of the dimension to be measured, Δ, is ±10 μm, and the loss due to a defective unit, A, is \$6. The annual operating costs of the fully automated and semiautomated systems are \$100,000 and \$75,000, respectively. The production rate is 200,000 units per year. Which system should be used?

4.7. Assume that the semiautomated feedback control system described in Prob. 4.6 can be improved by introducing a personal computer and additional sensors. The cost of such improvement is \$7000. This results in a mean squared drift per unit production of 1.85 and reduces the time lag l to 1 unit of production. What is the effect of the improvement on the decision making of the engineer?

4.8. In electrostatic powder spraying, dry powder is pneumatically fed from a supply reservoir to a spray gun, where a low-amperage, high-voltage charge is imparted to the powder particles. The part to be coated is electrically grounded so that the projected charged particles are firmly attracted to the part's surface and held there until melted and fused into a smooth finish in a baking oven.

An automobile manufacturer uses the electrostatic powder spraying operation to deposit a polishing layer of paint on car door handles. One of the major factors that affect the quality of the paint finish is the powder flow rate. Low flow rate will result in the application of a nonuniform layer of paint powder. On the other hand, excessive flow rate will result in a thicker layer than desired. Therefore, the manufacturer installed a microcomputer system to monitor the powder flow rate and to ensure that it is within a range of 25–28 lb/hr. The thickness of the paint layer on one door handle is measured at different locations, and the following measurements are recorded over the course of five days.

Day	Measurements (mm)				
1	0.5	0.3	0.2	0.1	0.4
2	0.3	0.3	0.2	0.2	0.1
3	0.1	0.2	0.1	0.2	0.3
4	0.2	0.4	0.5	0.4	0.5
5	0.3	0.4	0.2	0.6	0.1

The mean squared drift per unit of production is found to be 0.120. When the thickness of the paint layer falls outside the tolerance limits of 0.1 and 0.3, the door handle is scrapped at a cost of \$8.00. What is the quality loss per unit of production?

Assume that the manufacturer installed a microprocessor with high-accuracy sensors to detect any deviations in the powder flow rate. The cost of the system is \$7000, and its annual operating cost is \$3000. Measurements of paint layers after the installation of the new system are as follows:

Day	Measurements (mm)				
1	0.10	0.05	0.20	0.30	0.40
2	0.10	0.20	0.10	0.30	0.20
3	0.12	0.32	0.05	0.20	0.10
4	0.13	0.40	0.31	0.30	0.28

What is the quality loss per unit of production?

4.9. Optical fiber cables are replacing copper cables in the modern telecommunication systems. In order to protect the optical fiber as much as possible from external influences and to keep it functioning within mechanically permissible limits, it is protected by a loose buffer that is actually a small plastic tube. Each optical fiber is placed in one tube to ensure that it is protected against deformation and friction. To manufacture the loose buffers, two extruders are put together, so that the production of the single-fiber loose buffer's inner and outer tube is achieved in one continuous process. For this purpose, a control system is needed that functions very precisely and guarantees a constant extrusion rate of the buffer tube materials at about 250° C. This will also ensure maintenance of the wall thickness of the tubes, measuring only a few tenths of a millimeter. Assume that the control system has the following parameters:

Tolerance of the wall thickness	5 μm
Loss due to a defective tube	$20.00
Measurement cost per tube	$5.00
Cost of the system adjustment	$17.00
Current control limits of the process	4 μm
Time lag for system adjustment	7 tubes
Observed average adjustment interval	200 tubes

Find the quality loss per tube. What is the optimal adjustment interval of the control system? After the implementation of the optimal adjustment interval, the following measurements of the wall thickness are taken for 20 tubes (in mm):

0.008	0.009	0.007	0.005	0.003
0.006	0.007	0.009	0.008	0.006
0.004	0.006	0.004	0.009	0.006
0.005	0.007	0.008	0.008	0.004

What is the effect of the new adjustment interval on the quality loss per tube?

4.10. A workpiece 10 inches in diameter is to be faced down to a 4 inch diameter on the end. This can be performed by using a lathe equipped with an electronic device that controls the spindle speed and maintains the cutting speed at 200 ft/min (for the diameter being cut at any instant). Deviations in the cutting speed result in a final diameter different from the required 4 inches. The tolerance of the cutting speed is 10 ft/min. A workpiece that has a final diameter greater than 4 inches is reworked at a cost of $25. A workpiece with a diameter less than 4 inches is discarded at a cost of $70.00. The cost of adjusting the control system is $5, and the average adjustment interval is 25 workpieces. The following measurements are recorded for the cutting speeds and the final diameters of workpieces produced during the measurement period.

Cutting speeds (ft/m):

210	211	200	215	199	190	198	190
200	220	210	215	199	210	205	200
210	230	205	200	195	196	200	197
200	200	210	195	199	200	196	215

Final diameter (in):

4.001	4.002	4.001	4.005	4.001
4.000	3.996	3.999	4.000	3.998
3.997	3.999	4.000	4.003	4.006
3.990	4.005	4.007	4.001	3.998
3.993	4.000	4.002	4.006	4.000

The tolerance for the workpiece diameter is ±0.002 in. The time lag for the control system is one workpiece. What are the parameters of the optimal control system?

REFERENCES

Bodenstab, J. "Machine Vision for Electronics Manufacturing." *Robotics*, vol. 8, Sept. 1986, p. 21.

Devore, J. L. *Probability and Statistics for Engineering and the Sciences*. Pacific Grove, CA: Brooks/Cole Publishing Company, 1982.

Eliason, L., and J. West. "Vision-Based Robotics for Process Management." *Robotics*, vol. 8, Sept. 1986, p. 12.

Elsayed, E. A., and T. O. Boucher. *Analysis and Control of Production Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

Hunter, R. P. *Automated Process Control Systems: Concepts and Hardware*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Taguchi, G. "The Role of Metrological Control for Quality Control." *Proceedings of the International Symposium on Metrology for Quality Control in Production*, 1984, pp. 1–7.

CHAPTER 5

ON-LINE PROCESS PARAMETER CONTROL: VARIABLE CHARACTERISTICS

Chapter 4 introduced the use of on-line feedback systems to ensure the quality of manufactured products during production. The quality characteristics of the product are measured, and the results of the measurements are fed back to upstream manufacturing processes of the production line. Appropriate adjustment decisions are made based on the amount of deviations (as shown by the measurements) from the target characteristics. The concept of controlling the parameters of production processes in order to control the quality of the product characteristics during production is common knowledge among Japanese industries and has been gaining wider recognition among American industries. This recognition is reflected by the widespread use of SPC (statistical process control) in many industries.

It is important to determine the parameters of the production process to be controlled and the optimal control system for each parameter. With the exception of the traditional literature on automatic control engineering, there is, unfortunately, only limited material available on parameter control of the production process. In addition, the literature on automatic control engineering lacks the evaluation of product (or process) quality as it relates to the parameters of the automatic control systems.

The objective of this chapter is to address different methods of controlling process parameters and to discuss the relationships between process parameters and quality characteristics of the product.

5.1 PROCESS PARAMETER TOLERANCE

In this section, the effect of variations in process parameters on the quality characteristics of a product is discussed. Consider a case where the process parameter x affects the quality characteristic y of a product. Assume that the specification of the product characteristic y is given by

$$m_o \pm \Delta_0 \tag{5.1}$$

where m_0 is the target value of the product characteristic y, and Δ_0 is the tolerance of y.

The defective loss per unit of production found to be out-of-specification is A. When the process parameter x changes by one unit, it changes the product characteristic y by b units. Thus, the process produces defective units when the process parameter x deviates from its nominal value m by

$$m \pm \Delta \qquad \Delta = \frac{\Delta_0}{b} \tag{5.2}$$

We refer to Δ in Eq. 5.2 as the tolerance of the production process parameter. In other words, when process parameter x deviates by Δ from its nominal value m, the product characteristic reaches its specification limit, resulting in a loss of A per unit.

Obviously, there are other forms of relationships between the value of the product characteristic and the value of the process parameter. In the case where the effect on the characteristic value is nonlinear, the relationship is approximated by a quadratic polynomial, and the tolerance limits of the process parameter are obtained graphically as shown in Fig. 5-1.

Figure 5-1 shows the relationship between the process parameter x and the product characteristic y. In order to determine the tolerance limit of the process parameter, the target value of the product characteristic, m_0, and its tolerance limits $m_0 - \Delta_0$ and $m_0 + \Delta_0$ are placed on the vertical axis. Horizontal lines are then drawn through these values to the points where they intersect the x-y curve. Vertical lines are drawn at the points of intersection to meet the horizontal (x) axis at values $m - \Delta_1$, m, and $m + \Delta_2$, which represent lower tolerance limit, nominal value, and upper tolerance limit of the process parameter, respectively. Then,

$$m^{-\Delta_1}_{+\Delta_2} \tag{5.3}$$

is the tolerance for the process parameter x. However, the actual control limit for x is usually very small, and the following symmetrical tolerance is often recommended:

$$m \pm \Delta \qquad \text{where} \qquad \Delta = \min(\Delta_1, \Delta_2) \tag{5.4}$$

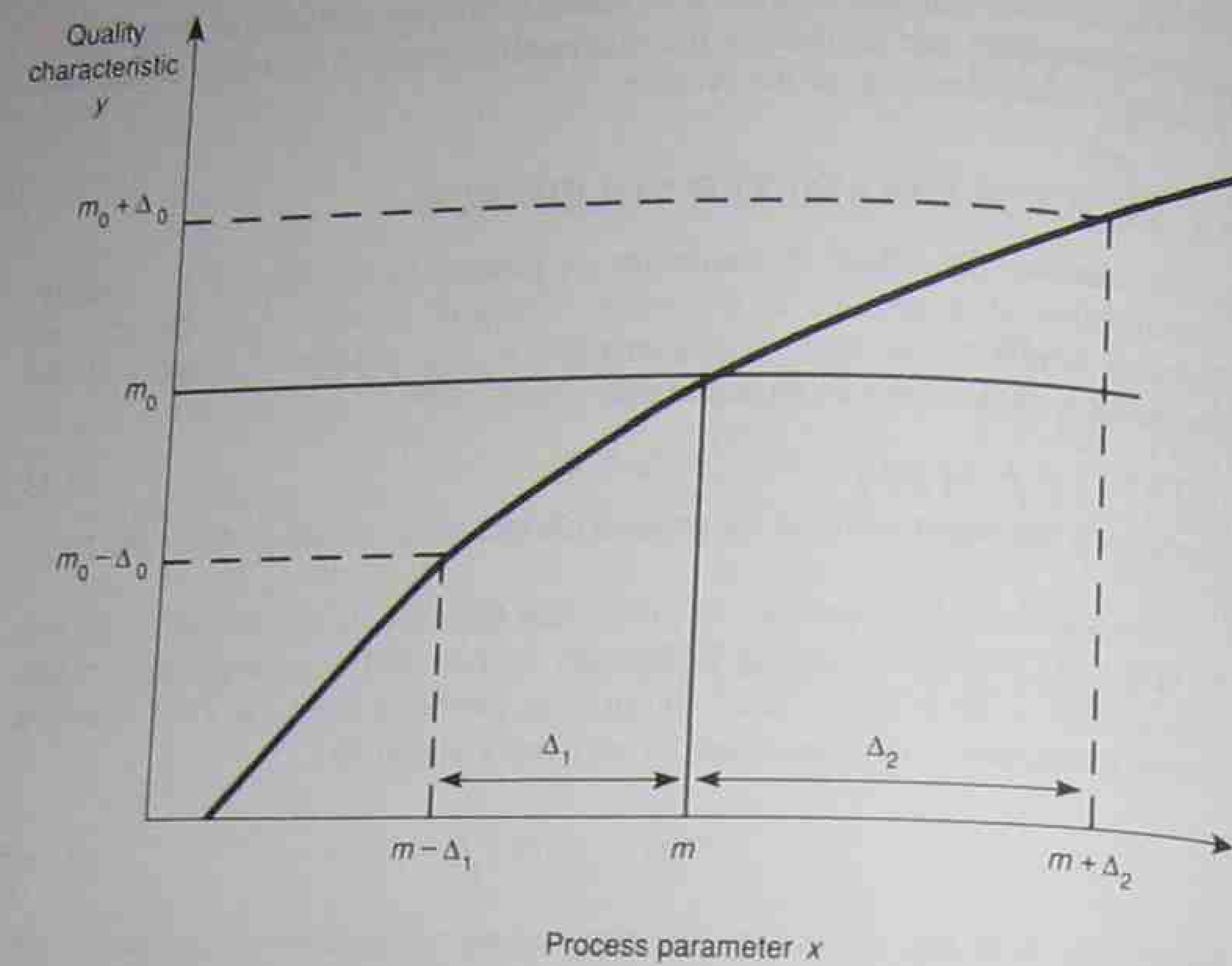


FIGURE 5-1
Estimation of the process parameter when the tolerance of the quality characteristic is given.

5.2 PROCESS PARAMETER FEEDBACK CONTROL SYSTEMS

In the previous section, the tolerance limit of a process parameter was estimated based on m_0 and Δ_0 of the product characteristic. In this section, the design of production process parameter control is presented in the same way the quality control system was discussed in Chap. 4. Parameters to be used for the present discussion are:

A = loss due to product characteristic being out-of-specification

B = measurement cost of production process parameter x

C = adjustment cost of production process parameter x

Δ = tolerance of production process parameter x

n_0 = current measurement interval of production process parameter x

D_0 = current control limit of production process parameter x

u_0 = current average adjustment interval of production process parameter x

The parameters of the optimal control system for production process parameter x are obtained by using the following steps:

1. Determine the tolerance (Δ) of the production process parameter, and the target value m, as discussed earlier in this chapter.
2. Obtain the optimal measurement interval n^* and the optimal control limit D^*:

$$n^* = \sqrt{\frac{2u_0 B}{A}} \times \frac{\Delta}{D_0} \tag{5.5}$$

$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4} \tag{5.6}$$

If n^* and D^* are found to be significantly different from the current n_0 and D_0, then obtain new values of n and D as follows:

$$n = (n_0 + n^*)/2$$

$$D = (D_0 + D^*)/2$$

These values are then used as the optimal parameters of the system.

3. Compare the loss of the current control system and that of the optimal control system using Eqs. 5.7 and 5.8, respectively:

$$L_0 = \frac{B}{n_0} + \frac{C}{u_0} + \frac{A}{\Delta^2}\left[\frac{D_0^2}{3} + \left(\frac{n_0+1}{2} + l\right)\frac{D_0^2}{u_0}\right] \tag{5.7}$$

$$L = \frac{B}{n^*} + \frac{C}{\bar{u}} + \frac{A}{\Delta^2}\left[\frac{D^{*2}}{3} + \left(\frac{n^*+1}{2} + l\right)\frac{D^{*2}}{\bar{u}}\right] \tag{5.8}$$

where $$\bar{u} = u_0 \times \frac{D^{*2}}{D_0^2} \tag{5.9}$$

These equations are the same as those given in Chap. 4, with the exceptions of Δ being the tolerance of the production process parameter and A being the loss due to the product characteristic being out-of-specification. The following examples illustrate these steps:

Example 5.1. Consider a product that is produced by a single-shot injection-molding process. The tolerance limit of the critical dimension of the product is

$$m_0 \pm 30\ \mu\text{m}$$

where m_0 is the target value of the dimension. The loss caused by out-of-specification is \$0.20, and the number of shots per hour is 250. The temperature of the molding pattern is checked every 2 hours and is controlled within a control limit of

$$m_t \pm 2°\text{C}$$

where m_t is the target value of the temperature.

The temperature measuring cost B is \$2.50. The temperature adjustment cost C is \$1.50, and the average adjustment interval is approximately once per day (i.e., 8 working hours). A change in the temperature by 1°C results in a change of 6 μm in the dimension of the product. Since this is a single-shot injection molding operation, the time lag is zero.

Obtain the optimal measurement interval n^* for the temperature and the optimal control limit D^*. Assuming 40 operating hours per week, what are the weekly savings?

Solution.

Step 1. Obtain temperature tolerance Δ as

$$\Delta = \frac{\text{tolerance of dimension}}{\text{effect of temperature change by } 1°\text{C}}$$

$$= \frac{30}{6} = 5.0$$

The parameters of the molding process are

$$\Delta = 5.0°\text{C}$$

$$A = \$0.20$$

$$B = \$2.50$$

$$l = 0$$

$$C = \$1.50$$

$$n_0 = 2 \text{ hr} \times 250 \text{ shots/hr} = 500 \text{ shots}$$

$$u_0 = 8 \times 250 \text{ shots} = 2000 \text{ shots}$$

$$D_0 = 2°\text{C}$$

Step 2. Obtain the optimal measurement interval n^* and the optimal control limit D^* using Eqs. 5.5 and 5.6, respectively.

$$n^* = \sqrt{\frac{2u_0B}{A}} \times \frac{\Delta}{D_0}$$

$$= \sqrt{\frac{2 \times 2000 \times 2.5}{0.2}} \times \frac{5}{2}$$

$$= 559 \approx 500 \text{ shots}$$

$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4}$$

$$= \left(\frac{3 \times 1.5}{0.2} \times \frac{2^2}{2000} \times 5.0^2\right)^{1/4}$$

$$= 1.03 \approx 1.0°\text{C}$$

Consequently, the optimal measurement interval is once every 2 hours, and the optimal control limit is $\pm 1.0°$C.

Step 3. The loss in the current process is

$$L_0 = \frac{B}{n_0} + \frac{C}{u_0} + \frac{A}{\Delta^2}\left[\frac{D_0^2}{3} + \left(\frac{n_0+1}{2} + l\right)\frac{D_0^2}{u_0}\right]$$

$$= \frac{2.5}{500} + \frac{1.5}{2000} + \frac{0.20}{5^2}\left(\frac{2^2}{3} + \frac{501}{2} \times \frac{2^2}{2000}\right)$$

$$= 0.005 + 0.000075 + 0.0146$$

$$= \$0.020$$

The loss when the process operates under optimal conditions is

$$L = \frac{B}{n^*} + \frac{C}{u} + \frac{A}{\Delta^2}\left[\frac{D^{*2}}{3} + \left(\frac{n^*+1}{2} + l\right)\frac{D^{*2}}{u}\right]$$

where $$u = u_0 \times \frac{D^{*2}}{D_0^2}$$

$$= 2000 \times \frac{1^2}{2^2}$$

$$= 500$$

and therefore $$L = \frac{2.5}{500} + \frac{1.5}{500} + \frac{0.20}{5^2}\left(\frac{1^2}{3} + \frac{501}{2} \times \frac{1^2}{500}\right)$$

$$= 0.005 + 0.003 + 0.006$$

$$= \$0.014$$

This results in an improvement of

$$L_0 - L = 0.020 - 0.014$$

$$= \$0.006 \text{ per unit of production}$$

The weekly savings are

$$0.006 \times 250 \times 40 = \$60$$

Example 5.2. In a welding process wherein 8000 spots are welded hourly by several robots, the electric power, which affects the quality of the welds, is controlled by checking its value once during an 8-hour workday. The adjustment limit is ± 8 percent of the nominal power, and the average adjustment interval is approximately one week, that is, 40 hours.

Given the following parameters, find the optimal values of the parameters for an on-line process control system, and estimate the annual gain.

Functional limit Δ	50% of nominal power
Cost of rewelding A	\$0.20
Checking cost B	\$2.40
Adjustment cost C	\$20.00
Current control limit D_0	8% of nominal power
Current checking interval n_0	$8000 \times 8 = 64000$
Observed average adjustment interval u_0	$8000 \times 40 = 320,000$
Time lag l	150 spots

Solution.

$$L_0 = \frac{B}{n_0} + \frac{C}{u_0} + \frac{A}{\Delta^2}\left[\frac{D_0^2}{3} + \left(\frac{n_0+1}{2} + l\right) \times \frac{D_0^2}{u_0}\right]$$

$$= \frac{2.40}{64,000} + \frac{20.00}{320,000} + \frac{0.2}{50^2}\left[\frac{8^2}{3} + \left(\frac{64,001}{2} + 150\right) \times \frac{8^2}{320,000}\right]$$

$$= 0.000038 + 0.000062 + 0.001707 + 0.000514$$

$$= 0.002231$$

for a yearly loss of

$$0.002231 \times 8000 \times 8 \times 250 = \$35,696 \text{ per year}$$

The optimal parameters are

$$n^* = \sqrt{\frac{2u_0 B}{A}} \times \frac{\Delta}{D_0}$$

$$= \sqrt{\frac{2 \times 320,000 \times 2.40}{0.20}} \times \frac{50}{8}$$

$$= 17,320 \approx 16,000 \text{ (once every 2 hours)}$$

$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4}$$

$$= \left(\frac{3 \times 20.00}{0.20} \times \frac{8^2}{320,000} \times 50^2\right)^{1/4}$$

$$= 3.5 \approx 4\%$$

$$\bar{u} = u_0 \times \frac{D^2}{u_0^2} = 320,000 \times \frac{4^2}{8^2}$$

$$= 80,000 \text{ (once every 10 hours)}$$

The loss when the welding process operates at optimal conditions is

$$L = \frac{2.40}{16,000} + \frac{20.00}{80,000} + \frac{0.2}{50^2}\left[\frac{4^2}{3} + \left(\frac{16,001}{2} + 150\right) \times \frac{4^2}{80,000}\right]$$

$$= 0.000150 + 0.000250 + 0.000427 + 0.000130$$

$$= 0.000957$$

or

$$0.000957 \times 8000 \times 8 \times 250 = \$15,312 \text{ per year}$$

The annual savings derived from the improvement in the process are

$$35,696 - 15,312 = \$20,384$$

Example 5.3. The viscosity of the emulsion in a continuous emulsion-coating process is carefully controlled in order to maintain the coating thickness at the desired level. The specification of the coating thickness is $m \pm 12\ \mu$m, and the loss incurred when the thickness exceeds the product specification is \$3 per squared meter. Current control limits are set at ± 5.0 poise (a measure of viscosity). A change in the viscosity of the emulsion by one poise affects the coating thickness by 0.6 μm. The average adjusting interval is 12,000 m^2, with a direct adjustment cost of \$10 and a time lag l of 30 m^2. The coating thickness is checked every 6000 m^2 (once during an 8-hour workday) at a direct cost of \$5. Determine the optimal diagnosis interval of the process, n^*, and the optimal control limit D^*.

Solution. The tolerance of the emulsion process, Δ, is

$$\Delta = \frac{12.0}{0.6} = 20 \text{ poise}$$

and the other parameters are

$$A = \$3$$

$$B = \$5$$

$$C = \$10$$

$$n_0 = 6000 \text{ m}^2$$

$$D_0 = 5.0 \text{ poise}$$

$$u_0 = 12,000 \text{ m}^2$$

$$l = 30 \text{ m}^2$$

The loss for the current checking system is

$$L_0 = \frac{B}{n_0} + \frac{C}{u_0} + \frac{A}{\Delta^2}\left[\frac{D_0^2}{3} + \left(\frac{n_0+1}{2} + l\right)\frac{D_0^2}{u_0}\right]$$

$$= \frac{5}{6000} + \frac{10}{12,000} + \frac{3}{20^2}\left[\frac{5^2}{3} + \left(\frac{6001}{2} + 30\right)\frac{5^2}{12,000}\right]$$

$$= 0.0008 + 0.0008 + 0.0625 + 0.0474$$

$$= \$0.1115 \text{ per m}^2$$

The optimal checking interval is

$$n^* = \sqrt{\frac{2u_0 B}{A}} \times \frac{\Delta}{D_0}$$

$$= \sqrt{\frac{2 \times 12,000 \times 5}{3}} \times \frac{20}{5}$$

$$= 800 \approx 750 \text{ (once per hour)}$$

The optimal control limit is

$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4}$$

$$= \left(\frac{3 \times 10}{3} \times \frac{5^2}{12,000} \times 20^2\right)^{1/4}$$

$$= 1.7 \approx 2.0$$

and the predicted interval between adjustments becomes

$$\overline{u} = u_0 \times \frac{D^{*2}}{D_0^2}$$

$$= 12,000 \times \frac{2^2}{5^2}$$

$$= 1920$$

The loss after the implementation of the optimal parameters is

$$L = \frac{B}{n^*} + \frac{C}{\overline{u}} + \frac{A}{\Delta^2}\left[\frac{D^{*2}}{3} + \left(\frac{n^*+1}{2} + l\right)\frac{D^{*2}}{\overline{u}}\right]$$

$$= \frac{5}{750} + \frac{10}{1920} + \frac{3}{20^2}\left[\frac{2^2}{3} + \left(\frac{751}{2} + 30\right)\frac{2^2}{1920}\right]$$

$$= 0.0067 + 0.0052 + 0.0100 + 0.0063$$

$$= \$0.0282$$

This results in an annual savings of

$$(0.1115 - 0.0282) \times 6000 \times 250 = \$124,950$$

5.3 MEASUREMENT (PREDICTION) ERROR AND PROCESS CONTROL PARAMETERS

The objective of this section is to discuss the effect of measurement errors on parameter control, as they were not fully covered in earlier sections.

The larger measurement errors are, the larger the variations in the quality of the products manufactured. For example, if there is a measurement error of $+3\ \mu$m, the adjustment of the process parameters is in error by the same degree. Suppose that the characteristic value of a product deviates from the target value by 7 μm, which includes 3 μm as a measurement error. The adjustment of the production process will experience a corresponding error if its calibration is made at 7 μm instead of the true deviation of 4 μm. Consequently, after adjustment the process parameter will be set at a value 3 μm from the target value. This will contribute to deviations of subsequent pieces of production from the target value before the process reaches the next measurement point. Thus, the amount of time that the process takes to go beyond the control limits is shortened.

Measurement error can be taken into account when the optimal values of the parameters of the quality control system are obtained by the method explained earlier in this chapter. We define the following:

u_0 = observed average adjustment interval of the current process

σ_m^2 = mean squared error of the measurement

D_0 = current control limit of the production process

Denoting the measurement cost of the current measurement method by B, adjustment cost of the process by C, and the time lag of adjustment by l, the loss function is

$$L = \frac{B}{n_0} + \frac{C}{u_0} + \frac{A}{\Delta^2}\left[\frac{D_0^2}{3} + \left(\frac{n_0+1}{2} + l\right)\frac{D_0^2}{u_0} + \sigma_m^2\right] \tag{5.10}$$

where Δ = tolerance of the process parameter
A = loss due to out-of-specification product
n_0 = current measurement interval

The optimal measurement interval n^* and the optimal control limit D^* are

$$n^* = \sqrt{\frac{2u_0B}{A}} \times \frac{\Delta}{D_0} \tag{5.11}$$

$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4} \tag{5.12}$$

The predicted average adjustment interval $\overline{u}$, can be approximated by

$$\overline{u} = u_0 \times \frac{D^{*2}}{D_0^2} \tag{5.13}$$

The following examples illustrate the above procedure:

Example 5.4. Assume that the dimensions of certain parts are measured once every two hours, and that the production process has the following parameters:

Control limit	$\pm 8\ \mu m$
Loss due to a defective product, A	\$3
Measurement cost B	\$2
Time lag l	10 parts
Variance of measurement error, σ_m^2	$3\ \mu m^2$
Tolerance of dimension, Δ	$\pm 25\ \mu m$
Hourly production	800 parts
Adjustment cost C	\$12
Average adjustment interval u_0	1.5 times per day

(*a*) Find the parameters of the optimal control system, and calculate its annual savings in comparison with the current system. (Assume 2000 production hours per year.)

(*b*) An engineer proposed a new measurement method that reduces the measurement error variance to one-fifth of its current value while it increases the time lag to 1.5 times the current value. The measurement cost is \$3. Calculate the loss per unit of production for the proposed method.

Solution.
(*a*) The parameters of the current process are

$$\Delta = 25\ \mu m$$
$$A = 300¢$$
$$B = 200¢$$
$$C = 1,200¢$$
$$n_0 = 1,600 \text{ parts}$$
$$D_0 = 8\ \mu m$$
$$u_0 = \frac{800 \times 8}{1.5} = 4300 \text{ parts}$$
$$l = 10 \text{ parts}$$
$$\sigma_m^2 = 3\ \mu m^2$$

The loss function of the current process is obtained by substituting the above parameters in Eq. 5.10:

$$L = \frac{B}{n_0} + \frac{C}{u_0} + \frac{A}{\Delta^2}\left[\frac{D_0^2}{3} + \left(\frac{n_0+1}{2} + l\right)\frac{D_0^2}{u_0} + \sigma_m^2\right]$$
$$= \frac{200}{1600} + \frac{1200}{4300} + \frac{300}{25^2}\left[\frac{8^2}{3} + \left(\frac{1601}{2} + 10\right)\frac{8^2}{4300} + 3\right]$$
$$= 0.12 + 0.28 + 10.24 + 5.72 + 0.07 + 1.44$$
$$= 17.87¢$$

The optimal measurement interval n^* and the optimal control limit D^* are

$$n^* = \sqrt{\frac{2u_0 B}{A}} \times \frac{\Delta}{D_0}$$
$$= \sqrt{\frac{2 \times 4300 \times 200}{300}} \times \frac{25}{8}$$
$$= 240$$
$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4}$$
$$= \left(\frac{3 \times 1200}{300} \times \frac{8^2}{4300} \times 25^2\right)^{1/4}$$
$$= 3.3\ \mu m$$

A predicted average adjustment interval $\bar{u}$ is calculated by using Eq. 5.13:

$$\bar{u} = u_0 \times \frac{D^{*2}}{D_0^2}$$
$$= 4300 \times \frac{3.3^2}{8^2}$$
$$= 740$$

The loss function is

$$L = \frac{200}{240} + \frac{1200}{740} + \frac{300}{25^2}\left[\frac{3.3^2}{3} + \left(\frac{241}{2} + 10\right) \times \frac{3.3^2}{740} + 3\right]$$
$$= 0.83 + 1.62 + 4.10$$
$$= 6.55¢$$

The annual savings are

$$\frac{1}{100} \times (17.78 - 6.55) \times 800 \times 2000 = \$179,680$$

(*b*) The loss function L for the new system is

$$L = \frac{B}{n^*} + \frac{C}{\bar{u}} + \frac{A}{\Delta^2}\left[\frac{D^{*2}}{3} + \left(\frac{n^*+1}{2} + l\right)\frac{D^{*2}}{\bar{u}} + \sigma_m^2\right]$$
$$= \frac{200}{240} + \frac{1200}{740} + \frac{300}{25^2}\left[\frac{3.3^2}{3} + \left(\frac{241}{2} + 15\right)\frac{3.3^2}{740} + 0.6\right]$$
$$= 0.83 + 1.62 + 2.99$$
$$= 5.44¢$$

The proposed control system is preferred to the current system, since a saving of 1.11¢ per piece is realized.

Thus far, we have used the loss function approach to determine the optimal parameters for both quality control systems and production processes. This approach can also be applied to determine the optimal calibration interval of measurement systems, as illustrated by the following example.

Example 5.5. An air gauge is used for the measurement of the principal dimension of a multiturn encoder. The specification of the dimension is $m \pm 10\ \mu m$, where m is the target value of the dimension, and the loss for a defective part is \$3.50. The air gauge is checked once each day, using a standard gauge having mean squared error $\sigma_s^2 = (0.5\ \mu m)^2$, to determine whether it needs calibration or not. The checking cost is \$1.80, and the calibration limits are $\pm 2.0\ \mu m$. The average calibration interval is 3 days, and the calibration cost is \$0.20. The time lag of calibration is 0 units and the daily production is 5000 units. Given 250 working days per year, find the optimal parameters of the calibration system and the annual savings (if any) when the optimal parameters are used in the system.

Solution. The parameters of the current checking and calibration system are

$$\Delta = 10\ \mu m$$
$$A = \$3.50$$
$$B = \$1.80$$
$$C = \$0.20$$
$$n = 5000 \text{ units}$$
$$D_0 = 2.0\ \mu m$$
$$u_0 = 15,000 \text{ units}$$
$$l = 0$$
$$\sigma_s^2 = (0.5\ \mu m)^2$$

The loss per unit of the current system is

$$L = \frac{B}{n} + \frac{C}{u_0} + \frac{A}{\Delta^2}\left[\frac{D_0^2}{3} + \left(\frac{n_0+1}{2} + l\right)\frac{D_0^2}{u_0} + \sigma_s^2\right]$$
$$= \frac{1.80}{5000} + \frac{0.20}{15,000} + \frac{3.50}{10^2}\left[\frac{2^2}{3} + \left(\frac{5001}{2} + 0\right) \times \frac{2^2}{15,000} + 0.5^2\right]$$
$$= 0.0004 + 0.0000 + 0.0467 + 0.0233 + 0.0088$$
$$= \$0.0792$$

The optimal checking interval of the air gauge is

$$n^* = \sqrt{\frac{2 u_0 B}{A}} \times \frac{\Delta}{D_0}$$

$$= \sqrt{\frac{2 \times 15000 \times 1.80}{3.50}} \times \frac{10}{2}$$
$$\approx 625 \text{ or once each hour}$$

The optimal control limit D^* is

$$D^* = \left(\frac{3C}{A} \times \frac{D_0^2}{u_0} \times \Delta^2\right)^{1/4}$$
$$= \left(\frac{3 \times 0.20}{3.50} \times \frac{2^2}{15000} \times 10^2\right)^{1/4}$$
$$= 0.3 \approx 0.5\ \mu m$$

and

$$\bar{u} = u_0 \times \frac{D^{*2}}{D_0^2}$$
$$= 15,000 \times \frac{0.5^2}{2^2}$$
$$= 938$$

The measurement error σ_m^2 is given by

$$\sigma_m^2 = \frac{D^{*2}}{3} + \left(\frac{n^*+1}{2} + 1\right)\frac{D^{*2}}{\bar{u}} + \sigma_s^2 = \frac{0.5^2}{3} + \left(\frac{626}{2} + 0\right) \times \frac{0.5^2}{938} + 0.5^2$$
$$= 0.417$$

The loss per unit of production for the optimal control system is

$$L = \frac{B}{n^*} + \frac{C}{\bar{u}} + \frac{A}{\Delta^2}\left[\frac{D^{*2}}{3} + \left(\frac{n^*+1}{2} + l\right)\frac{D^{*2}}{\bar{u}} + \sigma_s^2\right]$$
$$= \frac{1.80}{625} + \frac{0.20}{938} + \frac{3.5}{10^2}\left[\frac{0.5^2}{3} + \left(\frac{626}{2} + 0\right) \times \frac{0.5^2}{938} + 0.5^2\right]$$
$$= 0.0029 + 0.0002 + 0.0146$$
$$= \$0.0177$$

The yearly savings are

$$(0.0792 - 0.0177) \times 5000 \times 250 = \$76,875$$

5.4 SUMMARY

This chapter discussed the effect of variations in process parameters on the quality characteristics of the product, and reviewed the methods for determining the optimal values of those process parameters. These methods were also used to determine the optimal parameters of measurement systems in order to minimize the effects of measurement error and the calibration cost of measurement systems.

Two types of on-line feedback quality control were discussed, namely:

1. Repairing or discarding of defective products found by an inspection
2. Adjusting or recovering processes found out-of-control by checking either quality characteristics or process conditions

There is a third type of action for controlling product quality; it is used on processes in which predicted quality differs from the target or targets. This method, called feed-forward control, is introduced briefly in Appendix B.

PROBLEMS

5.1 The specification for the dimension of an injection molding product is $m \pm 50\ \mu m$. The parameters of the process are as follows:

Loss caused by out-of-specification	30¢
Current measurement method	Measurement is taken after every 10 pieces of production
Measurement cost *B*	50¢
Time lag *l*	300 pieces
Control limit	$\pm 20\ \mu m$
Adjustment cost *C*	$90
Measurement error variance	18 μm
Average adjustment interval	One adjustment every 2 days
Daily production	2400 pieces (8 hours daily production)

(*a*) Determine the optimal measurement interval n^*, the optimal control limit D^*, and the annual savings.

(*b*) A new measurement method is to be used immediately after processing is devised: measurement cost is $5 per measurement, measurement error standard deviation is 3.3 μm, and time lag is 1. Compute savings for each piece and annual savings, respectively. Annual production hours are 2000.

5.2 A hot-dip plating process requires putting a protective coating on a cast-iron casing by dipping the casing into molten zinc. This process requires that the thickness of the coating be $150 \pm 10\ \mu m$. The temperature of the molten zinc is measured once a day. A 1° change in the temperature of the molten zinc results in a 2-μm decrease in the thickness of the coating. The cost to measure the temperature is $3, and the cost of an adjustment is $30. The average daily production is 800 casings. Assume that the control limit on the temperature is currently set at $m \pm 5°C$, with no time lag. The cost of a defective unit is $10, and the variance of the temperature measurement error is 3 μm^2. What are the optimal control limit, the measurement interval, and the gain per unit of production against the current control system?

5.3 Seamless steel tubes are made by a hot extrusion process. The specification of the inner diameter of the tube is 20.000 ± 0.001 in. The control limits of the pressure

required for the extrusion are $100{,}000 \pm 5{,}000$ lb/in². The pressure is currently checked twice a day. Pressure in the extrusion process is adjusted once every 5 days at an average cost of $100. The cost of checking pressure is $2, the time lag is 5 tubes, and the loss caused by a defective tube is $200. The daily production is 300 tubes. What are the annual savings if a quality control system with optimal parameters is implemented?

5.4 A manufacturer uses cylindrical center type grinders for grinding the surface of a workpiece. The amount of grinding is specified to be ± 0.002 in from the target dimension of the workpiece. The control limits are set at ± 0.001 in, and the loss caused by a defective product is $20. Measurements of the surface of the workpiece are made at a cost of $10. If the process is found to be out-of-control, it is adjusted at a cost of $30 and a time lag of 3 workpieces. The measurement error is checked once a day with a standard mean squared error $\nu_i^2 = (0.0002\ \text{in})^2$ at a cost of $3.00. The calibration limit is currently 0.0005 in, and average calibration interval is 3.5 working days. Find the optimal control system, including the calibration of the measurement system, and estimate annual earnings.

5.5 Assume that the manufacturer in Prob. 5.4 uses an automatic measurement system with an error variance of $(0.0003\ \text{in})^2$ at an annual cost of $8000. The new system measures each item. What is the effect of the automatic measurement on the quality loss of the process?

5.6 Waterjet cutting involves the removal of material by the cutting action of high-velocity (sometimes two to three times the speed of sound), extremely high pressure water or water-based fluid with abrasive additives. The clean edge cuts produced through waterjet cutting lend themselves to improved product characteristics. Two critical factors affect the accuracy of the waterjet cutting process: (1) the velocity of the abrasive additives, and (2) the pressure of the waterjet.

A numerical control waterjet system is developed to accurately control the contour of the cut, the velocity of the abrasive additives, and the pressure of the waterjet. Cavities shaped as equilateral triangles are to be made in a sheet of 6-in-thick mild steel. The lengths of each side of the triangle equal 4.5200 in. The velocity of the abrasive additives is fixed at 150 ft/s, and the pressure of the waterjet is 55,000 lb/in². A change of the velocity by 10 percent will result in a 2-percent change in the length of the triangle sides. Also, a change of the waterjet pressure by 10 percent will result in a change of the final dimension of the triangle by 5 percent.

Observations of the current setup are given below:

Waterjet pressure in thousandths of a pound per square inch:

54	55	53	55	56	60	52	50
61	55	61	62	55	57	53	55
50	51	52	53	55	54	56	59

Velocity of the abrasive additives in feet per second:

120	130	132	135	125	140	145	150
150	152	150	141	148	139	142	155
156	150	149	146	145	150	143	150

Length of the triangle sides in inches:

4.5000	4.5100	4.4901	4.4530	4.5301
4.6120	4.5320	4.5321	4.5000	4.5120
4.5210	4.5301	4.5325	4.4981	4.5000
4.6000	4.5301	4.5212	4.5100	4.6102

The loss caused by a defective sheet is \$360, the average adjustment intervals for the water pressure and velocity are 300 and 250, and the time lags are 5 and 7, respectively. The cost of inspection is \$20. Determine the optimal parameters of the waterjet system.

5.7 A producer of screws and bolts uses a flat-die machine to roll the screws and bolts. The flat-die machines operate by rolling a blank across the face of a stationary die with a traversing stroke of a reciprocating die. A pneumatically operated starting finger positions the blank in the dies. The blank then rolls between the die faces and is penetrated progressively, so that the final size is reached before the blank rolls off the finish end of the dies. The quality of the final thread is affected by the number of revolutions of the blank between the dies and the pressure applied on the dies. The quality of the thread is measured by the depth of the crest seam produced as a result of the metal displacement. The tolerance of the crest depth is ±0.1 mm.

The production rate of the machine is 5000 bolts per hour. The machine's pressure is adjusted every 20,000 bolts at a cost of \$27.00. Bolts are then inspected. If found defective, they are discarded at a cost of \$0.50 per bolt. The checking cost is \$0.10 per bolt. The time lag of the machine is 300 bolts. Assume that the measurement errors for the depth of the crest follow a normal distribution with mean 0.02 mm and standard deviation of 0.01 mm. A 200-bolt sample is taken, and the mean depth of the crest is found to be 0.103 mm with a standard deviation of 0.05 mm. Find

(*a*) the quality loss per bolt, and

(*b*) the optimal adjustment interval of the machine's pressure.

5.8 The producer mentioned in Prob. 5.7 wishes to replace the flat-die system with a cylindrical-die machine capable of accomplishing the same operation. The advantages of the cylinder-die machine are versatility in producing a wide array of bolts with different lengths, and the ability to produce bolts with crest seam tolerances ±0.05 mm. In order to determine the optimal pressure for the process, the production engineer set the pressure at P_1 and measured the crest depths of 20 consecutive bolts. The pressure was changed to P_2 and 20 other measurements are taken. The measurements are given below:

Pressure	Depth of crest				
P_1	0.03	0.02	0.01	0.00	0.01
	0.02	0.01	0.03	0.04	0.01
	0.05	0.04	0.03	0.05	0.02
	0.06	0.05	0.02	0.01	0.03
P_2	0.01	0.02	0.03	0.02	0.01
	0.00	0.03	0.02	0.01	0.00
	0.04	0.03	0.01	0.02	0.01
	0.00	0.05	0.06	0.06	0.01

Using the same cost values and assuming the same measurement errors as given in Prob. 5.7, what is the quality loss per bolt for pressures P_1 and P_2?

Assuming that the manufacturer uses the pressure determined above, what is the corresponding optimal adjustment interval?

5.9 Commonly, in the boring process, a workpiece is mounted on a lathe and rotated. During this time, a boring bar, mounted rigidly on the carriage, is fed into the workpiece. When an initially round workpiece is clamped in a chuck with a given amount of force, the workpiece will be elastically deformed. As the workpiece is bored, the initial surface produced will be a true circle; that is, it will have zero roundness error. Upon releasing the chucking force, the workpiece, as a result of its elastic nature, will tend to return to its initial shape. Thus, permanent roundness error is unavoidable. The roundness error is perhaps the most important geometrical error because it affects dimensional accuracy and other important factors, such as fitting machine parts and wear in rotating elements.

A machining center is used for boring the inner diameter of tubes 2.5 inches in length. The tolerance of the inner diameter ±0.0001 in. The workpiece is held in position by an automatic chuck whose clamping force is adjustable. The inner diameters of the tubes are measured during manufacturing. The out-of-roundness is measured radially from the center of the circle of minimum radial separation.

It is found that the roundness error increases quadratically with the chucking force as given below:

$$r = aF^2 + bF + c$$

where r = roundness error
$a = 2 \times 10^{-8}$
$b = 3 \times 10^{-5}$
$c = 1 \times 10^{-4}$

F is the force in lbf. Assume that the nominal chucking (clamping) force is 150 lbf, and that the loss caused by a defective tube (roundness error outside the range from 0.0000 to +0.0001) is \$37.00. The cost of measuring the roundness is \$10 per tube, and the clamping force is adjusted at intervals of 200 tubes to ensure that the roundness error is within the specified range.

(*a*) Determine the tolerance of the clamping force.

(*b*) Develop a control system that adjusts the clamping force after the roundness measurement is taken.

(*c*) What are the parameters of the control system?

(*d*) Assume that the process is adjusted every 100 tubes at a cost of \$30.00. What should the optimal adjustment interval be?

REFERENCES

Divine, C. A., Jr., and A. R. Walsh. "Forging vs. Machining." *Machine Design*, May 13, 1965, pp. 192–196.

Doyle, L. E. *Manufacturing Processes and Materials for Engineers*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

Hunter, R. P. *Automated Process Control System: Concepts and Hardware*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Jamieson, J. L., and L. A. Holmes. "Electrogel Machining of Honeycomb Core." *The Tool and Manufacturing Engineer*, vol. 51, no. 5 (November 1963), pp. 105–106.

CHAPTER 6

ON-LINE QUALITY CONTROL: ATTRIBUTE CHARACTERISTICS

Chapters 4 and 5 discussed on-line quality control for variable characteristics. They also emphasized the relationship between product tolerance and the optimal process diagnosis and process adjustment intervals of the production process. To determine those relationships, one must make measurements of product characteristics and calculate deviations of those measurements from their target values. Some of the product characteristics that can be measured are dimensions, hardness, operating temperature in degrees, tensile strength in pounds per unit area, percentage of a particular impurity in a chemical compound, weights, and mean time to adjustment of items.

Specifications for most variable characteristics are given with upper and lower limits for the measured value of the characteristic. Other variable characteristics may only have an upper or a lower limit, such as strength (lower limit only), or impurity in a chemical compound (upper limit only). Because of time, cost, or availability of technology, measurements of some product characteristics

are not possible, or are not made even though they could be. These characteristics are referred to as attribute characteristics. Some of these characteristics may be judged by visual inspection. Typical examples of these may be whether a plastic cover of a stereo system is scratched, joints of integrated circuits have been soldered properly, lithographed labels have certain desired colors, or soft-drink bottles have labels.

As in the case of variable characteristics, an attribute characteristic may have upper and lower limits; it may have one of the limits or none. The presence of a label on a bottle has no limits; however, the diameter of a bored hole can have upper and lower limits. For example, consider a plug gauge used to inspect a 1.000/1.002-in diameter specification: the *go* gauge member inspects the lower limit of 1.000 in, and the *no-go* gauge member inspects the upper limit of 1.002 in.

This chapter examines methods for on-line quality control of attribute characteristics. These methods focus on the determination of checking intervals and preventive maintenance controls. The chapter also reviews feedback control systems with periodic checking, using one sample.

6.1 CHECKING INTERVAL FOR ATTRIBUTE CHARACTERISTICS

Attribute characteristics are verified in a manner similar to the variable control systems introduced in Chaps. 4 and 5. The production process is diagnosed at fixed intervals (usually expressed as a number of products). When a defective product is found, the process is adjusted so that the characteristics of the subsequent products (excluding those produced during time lag) are brought within the product specifications. This section presents a method for estimating the optimal checking interval for products with attribute characteristics.

Because of the lack of information for the attribute characteristic case compared to the variable characteristic case, it is difficult to consider the loss for nondefective products in the quality loss formula. However, if the loss caused by nondefective products is taken into account, it would be $A/3$ as predicted in Chap. 4, calculated as follows:

$$\frac{A}{\Delta^2} \times \frac{\Delta^2}{3} = \frac{A}{3}$$

The methods described in Chaps. 6 and 7 should be used only in cases where there is no way to measure the variable characteristics of the product, or no knowledge of the parameters of characteristics that should be checked, aside from whether the product works or not. We define:

A = loss caused by producing one unit of a defective product

B = cost of checking a product

C = recovery cost for abnormality in the production process. This cost consists of both the recovery cost and the cost of stopping the production process in order to recover it. C is expressed as

$$C = C_1 t + C_2$$

where C_1 is the cost of stopping the process for one unit of time, t is the average recovery time, and C_2 is the direct recovery cost including labor, material, and equipment.

$\bar{u}$ = average predicted time between recoveries (expressed in production units)

l = number of products produced from the time an out-of-control product is found during the checking process until the time the production process is stopped for recovery (time lag)

n = diagnosis interval; number of products produced between two successive diagnoses

L = total cost of process checking and recovery per unit; this is the sum of checking cost, cost of defective products, recovery cost, and loss caused by time lag.

DIAGNOSIS COST PER PRODUCT. Diagnosis cost per product is the cost of checking a product divided by the number of products produced between two successive diagnoses.

$$\text{Diagnosis cost per product} = \frac{B}{n} \qquad (6.1)$$

COST OF DEFECTIVE PRODUCTS. This cost is obtained by multiplying the average number of defective products by the cost of a defective product. The result is then divided by the total number of products produced between successive recovery actions.

The average number of defective units between successive diagnoses is approximately $(n + 1)/2$. This is explained as follows: After the i^{th} process diagnosis is performed, the process may follow any of the patterns shown in Fig. 6-1. (It is assumed that once a process begins to produce defective units, every unit produced will be defective, until the problem is corrected.) For example, Pattern 1 represents the case wherein nondefective products are produced until the $(i + 1)^{\text{th}}$ process diagnosis, where the first defective unit is produced. In Pattern n the process begins to produce defective products immediately after the i^{th} process diagnosis is performed. The total number of defective units for all n patterns is $\frac{1}{2}n(n + 1)$, and so the average number of defective products is $\frac{1}{2}n(n + 1)/n$ or $\frac{1}{2}(n + 1)$, as mentioned earlier.

$$\text{Cost per product because of defective products} = \frac{n + 1}{2} \times \frac{A}{\bar{u}} \qquad (6.2)$$

n products

Pattern	Diagnosis *i*						Diagnosis (*i* + 1)
1	o	o	o	....	o	o	x
2	o	o	o	....	o	x	x
⋮	⋮						
n − 1	o	o	x	....	x	x	x
n	o	x	x	....	x	x	x
Defective products		(1)	(2)		(*n* − 2)	(*n* − 1)	(*n*)

o Nondefective product

x Defective product

FIGURE 6-1
Distribution of defective products.

RECOVERY COST. This cost represents the direct cost of recovering the process to normal operations:

$$\text{Recovery cost per product} = \frac{C}{\bar{u}} \qquad (6.3)$$

COST OF TIME LAG. This cost is a delay cost caused by the number of defective units produced from the time a sample is taken for checking its characteristics until the production process is stopped for recovery:

$$\text{Cost of time lag per product} = \frac{lA}{\bar{u}} \qquad (6.4)$$

The total cost of diagnosis and recovery per unit of production, L, is obtained by adding Eqs. 6.1 through 6.4.

$$L = \frac{B}{n} + \frac{n + 1}{2} \times \frac{A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}} \qquad (6.5)$$

Example 6.1. An automatic spot-welding machine is diagnosed by sampling one of every 100 products. The diagnosis involves checking the welded parts of the product. If a product is found to be defective because parts are not welded, or the welding is incomplete, the welding machine is stopped to permit adjustments, and all defective parts found are scrapped at a cost of \$0.50 per defective product. Direct cost of a diagnosis is \$1.60, and it consumes 8 minutes, during which 30 units of production could have been produced. Therefore, the time lag is l = 30 products. Once the machine is found to be abnormal, it must be stopped for 20 minutes. The loss of stopping the welding machine is $12.00 per hour. An inspection of 5 products is required in order to detect and scrap all defective products from the 100 products produced between successive diagnoses. The inspection cost equals $8.00 ($1.60 × 5 = $8.00). The direct cost of recovery of the welding process is, C_2, $8.00. Therefore, the total recovery cost C is

$$12.00 \times \frac{20}{60} + 1.60 \times 5 + 8.00 = \$20.00$$

Assume that 16 adjustments to the welding machine have been performed during the last two months, and the total number of products produced during the same period was 84,000 units. What is the total cost (per product) of process diagnosis and recovery of the welding process?

Solution.

A = loss caused by production of a defective product = $0.50

B = diagnosis cost = $1.60

C = recovery cost = $20.00

$\bar{u}$ = average number of units produced between adjustments

$$= \frac{84,000}{16} = 5250$$

l = 30 units (products)

n = 100 units (products)

The total cost per unit of production is obtained by substituting the above values in Eq. 6.5:

$$L = \frac{B}{n} + \frac{n+1}{2} \times \frac{A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}}$$

$$= \frac{1.60}{100} + \frac{101}{2} \times \frac{0.50}{5250} + \frac{20.00}{5250} + \frac{30 \times 0.50}{5250} = \$0.0275 \text{ per unit}$$

If a new automatic welding machine exists that does not require adjustments, then the total cost per product would be zero. Though such a machine does not exist, let us assume that the interest and the capital recovery cost of such a machine is $0.10 per product. Then the cost per product of the new machine would be $0.0725 higher than that for the present machine.

One parameter that affects the total cost per product in Eq. 6.5 is the number of products produced between successive diagnoses. Example 6.2 illustrates the effect of the length of the diagnosis interval on the quality loss per unit of production.

Example 6.2. Consider the same automatic welding machine described in Example 6.1. Assume that the parameters of the production process are the same except that the number of products between successive diagnoses is (a) increased to 1500 units from its present value of 100 units; or (b) decreased to 50 units from its present value of 100 units. Determine the costs per product for (a) and (b).

Solution. (a) Consider the case when n is increased to 1500 units. The total cost per product is obtained by using Eq. 6.5:

$$L = \frac{1.60}{1500} + \frac{1501}{2} \times \frac{0.50}{5250} + \frac{20.00}{5250} + \frac{30 \times 0.50}{5250} = \$0.0792$$

The cost increase caused by increasing the time between diagnoses is

$$0.0792 - 0.0275 = \$0.0517 \text{ per unit}$$

(b) Consider now decreasing the time between successive diagnoses to 50 units. The cost per product is

$$L = \frac{1.60}{50} + \frac{51}{2} \times \frac{0.50}{5250} + \frac{20.00}{5250} + \frac{30 \times 0.50}{5250} = \$0.0411$$

Again, the cost per product is higher (by $0.0136) than that for the original diagnosis interval (100 units).

From this example, one can conclude that there is an optimal diagnosis interval n that minimizes the cost of diagnosis and adjustments per product. How the optimal interval is determined is explained next.

6.1.1 Optimal Interval between Successive Diagnoses

Equation 6.5 relates the total cost per product to the interval between successive diagnosis, time lag, time between recovery actions, and to the cost values A, B, and C. The optimal interval between diagnoses can be determined by setting the derivative of Eq. 6.5, with respect to n, to zero:

$$\frac{dL}{dn} = -\frac{B}{n^2} + \frac{A}{2\bar{u}} = 0$$

$$n^* = \sqrt{2\bar{u}\frac{B}{A}} \tag{6.6}$$

The total cost per product for the optimal n^* is obtained by substituting Eq. 6.6 into Eq. 6.5.

$$L = \sqrt{\frac{AB}{2\bar{u}}} + \frac{1}{2}\sqrt{\frac{2\bar{u}B}{A}}\,\frac{A}{\bar{u}} + \frac{A/2 + C + lA}{\bar{u}}$$

$$= \sqrt{\frac{2AB}{\bar{u}}} + \frac{A/2 + C + lA}{\bar{u}} \tag{6.7}$$

Example 6.3. Determine the optimal diagnosis interval for the automatic spot-welding machine of Example 6.1. What is the cost per product corresponding to the optimal diagnosis interval?

Solution. The optimal diagnosis interval is determined by substituting the values of $\bar{u}$, B, and A in Eq. 6.6:

$$n^* = \sqrt{2\bar{u}\frac{B}{A}} = \sqrt{\frac{2 \times 5250 \times 1.60}{0.50}} = 183 \text{ units}$$

and the cost per product is obtained by using Eq. 6.7:

$$L = \sqrt{\frac{2 \times 50 \times 1.60}{5250}} + \frac{0.50/2 + 20.00 + 30 \times 0.50}{5250} = \$0.0242$$

Since Eq. 6.6 ignores the effect of l, C, and $\bar{u}$, the value of n^* will not be accurately estimated unless the following conditions are satisfied. First, the time lag l should be much smaller than the average time between adjustments. Second, the loss caused by a defective product, A, should be much greater than $C/\bar{u}$, the adjustment cost per product. What happens if either of these conditions or both of them are not satisfied? Another expression is developed for n so that these conditions need not be satisfied. The alternative expression for the optimal diagnosis interval is obtained by setting the diagnosis cost per product approximately equal to the loss caused by producing defective products in Eq. 6.5. This equalization results in

$$n^* = \sqrt{\frac{2(\bar{u} + l)B}{A - C/\bar{u}}} \qquad (6.8)$$

Solutions to Eq. 6.8 exist when $A \gg C/\bar{u}$; otherwise one should use Eq. 6.6 instead. The derivation of Eq. 6.8 is as follows:

Let P_i ($i = 1, 2, \ldots$) be the probability that the first defective unit is the i^{th} unit of production. The interval between two successive diagnoses is n units of production, as depicted in Fig. 6-2. The probability of observing a defective unit for the first time after the k^{th} diagnosis is:

$$P_{n(k-1)+1} + P_{n(k-1)+2} + \cdots + P_{n(k-1)+n} \qquad (6.9)$$

The average number of defective units produced between the $(k-1)^{\text{th}}$ and k^{th} diagnoses is

$$\frac{nP_{n(k-1)+1} + (n-1)P_{n(k-1)+2} + \cdots + P_{n(k-1)+n}}{P_{n(k-1)+1} + P_{n(k-1)+2} + \cdots + P_{n(k-1)+n}} \qquad (6.10)$$

Assume that the probabilities of the first, second, third, . . . , n^{th} product being the first defective are equal. In other words,

$$P_{n(k-1)+1} = P_{n(k-1)+2} = \cdots = P_{n(k-1)+n} \qquad (6.11)$$

The average number of defective units in this case is $(n+1)/2$. This

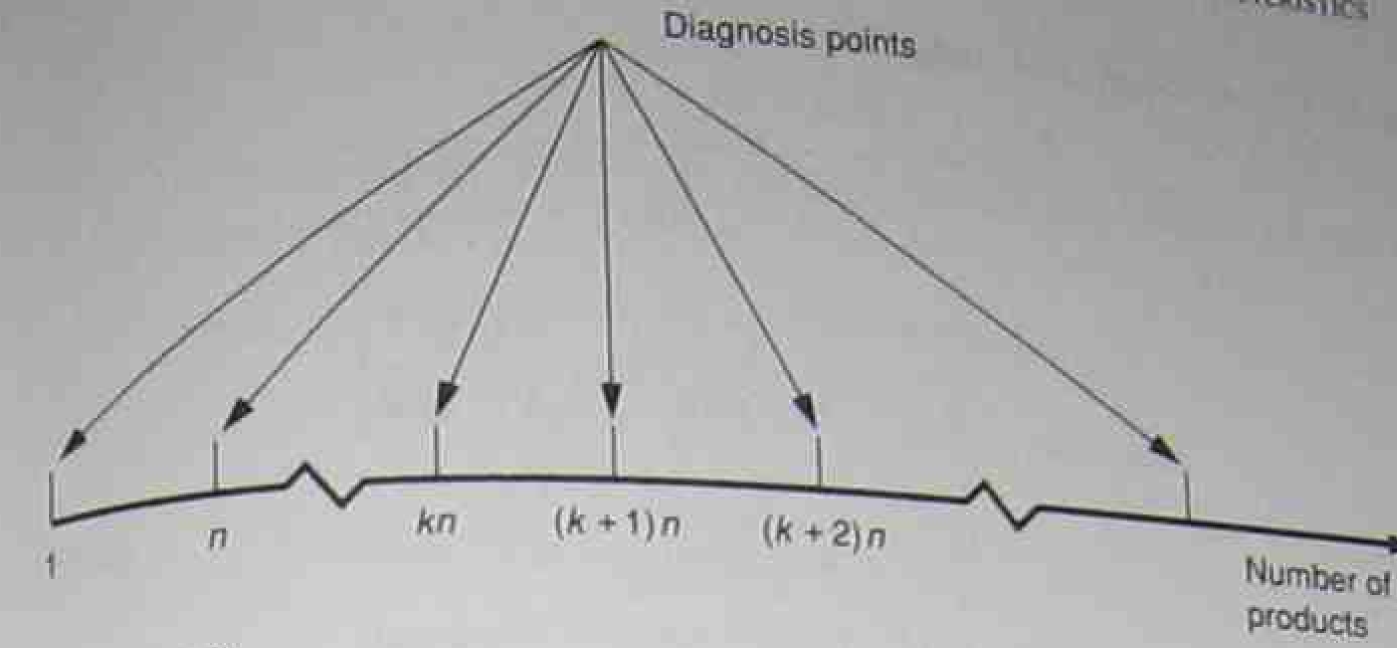


FIGURE 6-2
Diagnosis points.

explains the second term in Eq. 6.5. We now calculate the average number of units produced between two successive adjustments.

As mentioned above, we assume that the probabilities of the first, second, third, . . . , n^{th} product being the first defective are equal. When a defective unit is found at the k^{th} diagnosis, then the cause of defect (trouble) must have occurred between the $(k-1)^{\text{th}}$ and the k^{th} diagnoses, and the average number of defective units after the trouble has occurred is $n/2$. Thus, the actual number of products between two successive adjustments should be $\bar{u} + n/2$, instead of $\bar{u}$ as given in Eq. 6.5.

Replacing $\bar{u}$ by $\bar{u} + n/2$ in Eq. 6.5 results in

$$L = \frac{B}{n} + \frac{n+1}{2} \times \frac{A}{\bar{u} + n/2} + \frac{C}{\bar{u} + n/2} + \frac{lA}{\bar{u} + n/2} \qquad (6.12)$$

By using the following approximation in Eq. 6.12:

$$\frac{1}{\bar{u} + n/2} \approx \frac{1}{\bar{u}}\left(1 - \frac{n}{2\bar{u}}\right) \qquad (6.13)$$

and by taking the derivative of resulting equation with respect to n, the following is obtained:

$$\frac{dL}{dn} = -\frac{B}{n^2} + \frac{A}{2\bar{u}} - \frac{2n+1}{2} \times \frac{A}{u^2} - \frac{C}{2\bar{u}^2} - \frac{lA}{2\bar{u}^2} \qquad (6.14)$$

Since the third term, on the right-hand side of Eq. 6.14 is much smaller than the second term, the third term is neglected. The equation is then set to zero, giving

$$-\frac{B}{n^2} + \frac{A}{2\bar{u}} - \frac{C}{2\bar{u}^2} - \frac{lA}{2\bar{u}^2} = 0$$

or

$$n^* = \sqrt{\frac{2\bar{u}B}{A - C/\bar{u} - lA/\bar{u}}} \qquad (6.15)$$

Because $A \gg C/\bar{u}$ and $A \gg lA/\bar{u}$,

$$\frac{1}{A - C/\bar{u} - lA/\bar{u}} = \frac{1}{(A - C/\bar{u})\left(1 - \dfrac{lA/\bar{u}}{A - C/\bar{u}}\right)}$$

$$\approx \frac{1}{A - C/\bar{u}}\left(1 + \frac{lA/\bar{u}}{A - C/\bar{u}}\right)$$

$$\approx \frac{1}{A - C/\bar{u}}\left(1 + \frac{l}{\bar{u}}\right) \tag{6.16}$$

Substituting Eq. 6.16 into Eq. 6.15, the following is obtained:

$$n^* = \sqrt{\frac{2(\bar{u} + l)B}{A - C/\bar{u}}} \tag{6.17}$$

This completes the derivation of Eq. 6.8.

Example 6.4. Use Eq. 6.8 to obtain the optimal diagnosis interval for the automatic welding machine in Example 6.1. Determine the total cost per product.

Solution. By Eq. 6.8, n^* is

$$n^* = \sqrt{\frac{2(5250 + 30) \times 1.60}{0.50 - 20.00/5250}} = 185$$

and the cost per product is

$$L = \frac{1.60}{185} + \frac{186}{2} \times \frac{0.50}{5250} + \frac{20.00}{5250} + \frac{30 \times 0.50}{5250} = \$0.0242$$

By changing the diagnosis interval from 100 products to 185 products, the total cost per product is reduced from \$0.0275 to \$0.0242—a cost reduction of \$0.0033 per product. The expected cost reduction per month is 0.0033 multiplied by 42,000 products per month, or \$138.60.

A significant difference in results between the use of Eq. 6.6 and Eq. 6.8 to determine the optimal diagnosis interval is shown in Example 6.5.

Example 6.5. Berry (1974) reports the results of production monitoring of a large number of machine tools. A computerized monitoring system was installed to minimize loss for a total of 160 automatic screw machines in a single plant. The diagnosis procedure for the production process was similar to the procedure given in Example 6.1. The following are the parameters of the production process:

$$A = \$5$$

$$B = \$150$$

$$C = \$4000$$

$$l = 1000 \text{ units}$$

$$\bar{u} = 4000$$

Determine the optimal n by using Eqs. 6.6 and 6.8. What are the corresponding costs per product?

Solution. By using Eq. 6.6, the following is obtained:

$$n^* = \sqrt{2\bar{u}\frac{B}{A}} = \sqrt{\frac{2 \times 4000 \times 150}{5}} = 490 \text{ units}$$

$$L = \frac{B}{n} + \frac{n+1}{2} \times \frac{A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}}$$

$$= \frac{150}{490} + \frac{491}{2} \times \frac{5}{4000} + \frac{4000}{4000} + \frac{1000 \times 5}{4000} = \$2.86$$

Substitution of the parameters of the production process in Eq. 6.8 results in

$$n^* = \sqrt{\frac{2(\bar{u} + l)B}{A - C/\bar{u}}} = \sqrt{\frac{2 \times 5000 \times 150}{5 - 4000/4000}} = 612 \text{ units}$$

$$L = \frac{150}{612} + \frac{613}{2} \times \frac{5}{4000} + \frac{4000}{4000} + \frac{1000 \times 5}{4000} = \$2.88$$

Although the difference in cost per product appears to be insignificant, the difference between the two diagnosis-interval values is significant. The difference between the results of Eqs. 6.6 and 6.8 diminishes when $A \gg C/\bar{u}$ and $l \ll \bar{u}$. It is better to use Eq. 6.8 than Eq. 6.6, when either or both of these conditions are not satisfied.

6.1.2 Optimal Interval between Successive Diagnoses for Processes with Small Numbers of Defects

When a process operates under abnormal conditions and consistently produces a relatively large number of defective units, as has been assumed so far, Eq. 6.8 can be effectively used in determining the optimal diagnosis interval. However, when a process produces only a small fraction of defective units, Eq. 6.8 may lead to an overestimation of the optimal diagnosis interval. Therefore, in this section, new expressions are derived to determine the optimal diagnosis intervals under these conditions.

Assume that the fraction of defective products is p. Also, assume that the loss D incurred when a defective unit of production is not disposed of at the diagnosis point, but is sent on to the next production process, is much larger than A. The probability of detecting the deviation of the production process from the normal condition at the diagnosis point is p, and the probability of failure to detect such deviation is $1 - p$.

The average number of defective units produced from the moment the process begins to operate improperly until the first diagnosis point thereafter is $(n+1)p/2$. The probability of detecting the trouble at the second diagnosis point after failing to detect it at the first is $(1-p)p$. If the problem is detected at this point, the average number of defective units that have gone undetected is $(n+1)p/2$, while the number detected is np. Table 6.1 carries this analysis forward to further diagnosis points.

From Table 6.1, the average loss due to defective units when the production process deviates from normal conditions is

$$\left\{\frac{(n+1)p^2}{2} + np[(1-p)p + (1-p)^2p + \cdots]\right\}A$$

$$+\left\{\frac{(n+1)p}{2}[(1-p)p + (1-p)^2p + \cdots] + np[(1-p)^2p + 2(1-p)^3p + \cdots + (i-2)(1-p)^{i-1}p]\right\}D$$

$$= \left[\frac{n+1}{2}p^2 + np(1-p)\right]A + \left[\frac{n+1}{2}p(1-p) + n(1-p)^2\right]D \quad (6.18)$$

Substituting $p = 0$ in Eq. 6.18, the loss becomes nD. Assuming p is near zero, then, the loss for each abnormal condition is nD, which corresponds to the loss $\frac{1}{2}(n+1)A$ for $p = 1$ as was seen in the preceding section.

Substitution of the loss $2D$ for A in Eq. 6.5 and 6.8 results in the following respective equations:

TABLE 6.1
Number of undetected defective units at diagnosis points

Diagnosis point	Probability of detecting deviation in the process	Number of defective units found	Number of undetected defective units
1	p	$(n+1)p/2$	0
2	$(1-p)p$	np	$(n+1)p/2$
3	$(1-p)^2p$	np	$(n+1)p/2 + np$
4	$(1-p)^3p$	np	$(n+1)p/2 + 2np$
.	.	.	.
.	.	.	.
.	.	.	.
i	$(1-p)^{i-1}p$	np	$(n+1)p/2 + (i-2)np$
.	.	.	.
.	.	.	.
.	.	.	.

$$L = \frac{B}{n} + \frac{n+1}{2} \times \frac{2D}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}} \quad (6.19)$$

and

$$n = \sqrt{\frac{2(\bar{u}+l)B}{2D - C/\bar{u}}} \quad (6.20)$$

When $p \neq 1$ (the number of defective units during deviation of the production process is less than 100-percent), it is possible to detect all defective units by tracing back to the point when the production process begins to deviate from normal conditions. If this is done, there are no undetected defective units, and $D = A$. Therefore, Eq. 6.18 reduces to

$$\left[\left(\frac{n+1}{2}\right)p^2 + np(1-p) + \left(\frac{n+1}{2}\right)p(1-p) + n(1-p)^2\right]A \quad (6.21)$$

Using the approximation $n + 1 \approx n$ in Eq. 6.21 results in

$$\left[\frac{n}{2}p^2 + \frac{3}{2}np(1-p) + n(1-p)^2\right]A = n\left(1 - \frac{p}{2}\right)A \quad (6.22)$$

Therefore, the loss after tracing the defective units is nA (maximum) when $p = 0$ and $\frac{1}{2}nA$ (minimum) when $p = 1$. The corresponding losses per defective unit are $2A$ and A, respectively. When the loss per defective unit is A, the loss per unit and the optimal diagnosis period are given by Eqs. 6.5 and 6.8, respectively. The corresponding equations, when the loss per defective unit is $2A$, are

$$L = \frac{B}{n} + \frac{n+1}{2}\frac{2A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}} \quad (6.23)$$

and

$$n^* = \sqrt{\frac{2(\bar{u}+l)B}{2A - C/\bar{u}}} \quad (6.24)$$

Equation 6.24 results in an overdiagnosis of the production process. Practically, it is appropriate to use Eqs. 6.5 and 6.8 when p is about 0.5, since these equations are not significantly affected by the error estimation of A, as shown in the following section.

Example 6.6. In a production process, the surface-mounted devices are held in place on the circuit board by their solder connections. However, with surface mounting, there is no solder plug surrounding a pin in a hole of the PCB (printed circuit board) to give the connection added strength. Instead, the solder alone bonds the device to the circuit board. Thus, the electrical integrity of the PCB depends on the structural integrity of the solder connection. Therefore, the soldered connections of each board are inspected to detect missing devices, bridges outside the devices, and the absence of solder fillets. Inspection is made at different diagnosis points along the production line. The loss caused by producing a defective circuit board is \$30, and the cost of checking a board is \$10. The process is adjusted with an average interval of 500 circuit boards, at a cost of \$25. The time lag of the adjustment process, as expressed in units of production, is 5. When a defective board is detected, a tracing-back screening is performed. What is the optimal diagnosis interval and the corresponding loss per unit of production?

Solution. We first determine the diagnosis interval while disregarding the effect of the probability of detecting abnormality. The optimal diagnosis interval and corresponding loss per unit of production are obtained using Eq. 6.8 and 6.5, respectively:

$$n = \sqrt{\frac{2(\bar{u} + l)B}{A - C/\bar{u}}}$$

$$n = \sqrt{\frac{2(500 + 5)10}{30 - 25/500}} \approx 18$$

and
$$L = \frac{10}{18} + \frac{19}{2} \times \frac{30}{500} + \frac{25}{500} + \frac{5 \times 30}{500} \approx \$1.48$$

When the possibility of the worst-case occurrence at the diagnosis points is taken into account, then the optimal diagnosis interval and the loss per defective unit are obtained by using Eqs. 6.24 and 6.23, respectively.

$$n = \sqrt{\frac{2(\bar{u} + l)B}{2A - C/\bar{u}}}$$

$$= \sqrt{\frac{2(500 + 5)10}{2(30) - 25/500}} = 13$$

and
$$L = \frac{B}{n} + \frac{n+1}{2} \times \frac{2A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}}$$

$$= \frac{10}{13} + \frac{14}{2} \times \frac{60}{500} + \frac{25}{500} + \frac{5 \times 30}{500} = \$1.96$$

6.1.3 Sensitivity Analysis

Although errors in estimating parameters A, B, C, $\bar{u}$, and l may affect the optimal diagnosis interval, they may not significantly affect the total cost per product.

Example 6.7. Assume that an error was made in estimating the loss caused by a defective product in the case of Example 6.1, and that the actual loss is $1.00 instead of $0.50. What is the total cost of diagnosis and adjustment per product?

Solution. By using Eq. 6.8 and substituting $A = \$1.00$, the following is obtained:

$$n^* = \sqrt{\frac{2(5250 + 30) \times 1.60}{1.00 - 20.00/5250}} = 130 \text{ units}$$

and the true L is given by

$$L = \frac{1.60}{130} + \frac{131}{2} \times \frac{0.50}{5250} + \frac{20.00}{5250} + \frac{30 \times 0.50}{5250}$$

$$= \$0.0252 \text{ per product}$$

Although there is a 100-percent error in estimating A, the error in quality loss is only [(0.0252 - 0.0242)/0.0242] × 100 = 4.1 percent

Example 6.8. Assume that the direct cost of diagnosis B in Example 6.5 was erroneously underestimated by 100 percent and its actual value is $300. What is the true quality cost per product?

Solution. By Eq. 6.8,

$$n^* = \sqrt{\frac{2 \times 5000 \times 300}{5 - 4000/4000}} = 866 \text{ units}$$

and
$$L = \frac{300}{866} + \frac{867}{2} \times \frac{5}{4000} + \frac{4000}{4000} + \frac{1000 \times 5}{4000} = \$3.13$$

A 100-percent error in estimating the direct cost of diagnosis results in an 8-percent error in the total cost per product.

It is important to perform sensitivity analyses for different cost elements. In Example 6.7, a 100-percent error in estimating A results in a 4.1-percent error in quality cost per unit, whereas in Example 6.8, a 100-percent error in estimating B results in an 8-percent error in quality cost.

Results of sensitivity analyses can be useful in manpower allocation in the quality control area. For example, changing n from 490 to 612 in Example 6.5 increases the total quality cost per product from $2.86 to $2.88. Therefore, if there is a shortage in manpower at any time period, the diagnosis interval can be changed from 490 to 612 or higher during that period without significantly affecting the total cost of quality.

6.1.4 Number of Operators Required for Process Diagnosis and Process Recovery

In recent years, there have been several trends in automation—both in the automation of machines and in the automation of their controls and integrations. These trends have altered many of the functions that operators used to perform, such as continuous tending of machines, loading and unloading, and tool changing. Continuous functions such as diagnosis, adjustment, and inspection are difficult, costly, and perhaps even impossible to automate, thus requiring operator attention. The following example illustrates how the number of operators required for such functions is determined.

Example 6.9 A manufacturer of LP (long-playing) records uses 40 press machines. Each machine is capable of producing one record per minute. Defective records may be produced either because of drift of machine condition or because of the existence of foreign particles (such as dust) in the production environment. Therefore, the production process is diagnosed at intervals of 100 records. The loss caused by producing a defective record is $0.80; the cost of each diagnosis is $10; the number of records produced during the time lag is 30; the average number of records produced between successive recoveries is 8000; the average recovery time is 2 hours, and the direct recovery cost is $30. Find the optimal diagnosis interval and the optimal number of operators that should be assigned for process diagnosis and process adjustment.

Solution. The parameters of the production process are

$$A = \$0.80$$

$$B = \$10$$

$$C = \$30$$

$$l = 30 \text{ records}$$

$$\bar{u} = 8000 \text{ records}$$

$$n = 100$$

The number of records produced per week (using a 40-hour work week) by the manufacturer is determined as follows: 1 record/min × 60 min/hr × 40 hrs/week × 40 press machines = 96,000 records/week. Because a diagnosis is made every 100 records, the number of records inspected weekly is 960. The total time for the weekly diagnosis is 960 × 30 min = 28,800 min = 480 hr.

Therefore, the number of operators required for process diagnosis is

$$480 \text{ hrs} \div 40 \text{ hr/operator} = 12 \text{ operators}$$

In addition, the number of adjustments performed per week is 96,000 ÷ 8000 = 12, and the total hours required for adjustment per week is 12 × 2 hr = 24 hr which is equivalent to one operator's time per week. Thus, the total number of operators required for both diagnosis and recovery is 13.

The optimal diagnosis interval can be determined by using Eq. 6.8:

$$n^* = \sqrt{\frac{2(\bar{u} + l)B}{A - C/\bar{u}}}$$

$$= \sqrt{\frac{2(8000 + 30)10}{0.8 - 30/8000}} = 450 \text{ records}$$

The above result shows that the diagnosis interval should be increased from the current 100 records to 450 records, thereby reducing the number of diagnoses per week to 213. Consequently, the number of operators required for the weekly diagnosis will be decreased to 2.7 operators or a total of 3 operators for process diagnosis.

The present quality cost per record (diagnosis interval of 100 records), L_0, is

$$L_0 = \frac{B}{n} + \frac{n+1}{2} \times \frac{A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}}$$

$$= \frac{10}{100} + \frac{101}{2} \times \frac{0.8}{8000} + \frac{30}{8000} + \frac{30 \times 0.8}{8000} = \$0.112$$

and the quality cost per record when the diagnosis interval is increased to 450 is

$$L_1 = \frac{10}{450} + \frac{451}{2} \times \frac{0.8}{8000} + \frac{30}{8000} + \frac{30 \times 0.8}{8000} = \$0.052$$

Comparison between L_1 and L_0 shows a cost reduction of $0.060 per record, or a total of $5760 per week, when the diagnosis interval is increased to 450 units. The average defective rate of this manufacturing facility, P, is

$$P = \frac{\text{expected number of defectives between recovery actions}}{\text{number of products produced between recovery actions}}$$

$$= \frac{(n+1)/2 + l}{\bar{u}} = \frac{451/2 + 30}{8000} = 0.032$$

$$= 3.2 \text{ percent}$$

The cost of reducing the defective rate is not linearly proportional to the defective rate itself.

Table 6.2 shows some of the relationships between the length of the diagnosis interval, the quality cost per product, and the defective rate corresponding to the diagnosis interval for the cost data given in Example 6.9.

The defective rate, as illustrated in Table 6.2, increases with the length of the diagnosis interval, whereas the quality cost per unit is a convex function of the length of the diagnosis interval. These different relationships between the factors may result in an unacceptable defective rate at the optimal diagnosis interval, and managerial decisions will have to be made regarding the trade-off between quality cost per product and defective rate.

Example 6.10. An electrostatic spray finishing machine uses the particle-attracting method for applying primer and underseal to the chassis of medium-sized cars.

TABLE 6.2
Relationship between the diagnosis interval, quality cost per product, and defective rate

Diagnosis interval n	Quality cost per product L	Defective rate P
100	$0.112	1.00%
200	$0.068	1.63%
300	$0.055	2.25%
400	$0.051	2.89%
500	$0.049	3.50%
600	$0.049	4.13%

The coating of the chassis is continuously monitored by checking for uniformity and consistency. The electrostatic attraction of any coating material is greater on outer edges and hole edges, causing a heavier build-up of coating in these areas. This build-up and the uniformity of coating can be controlled by adjusting the applied charge. The diagnosis cost per unit is \$30. If a chassis coating is found defective (nonuniform buildup of coating), the chassis is repaired and recoated at a cost of \$250. If the electrostatic spraying process causes nonuniformity or build-up of coating material at edges or holes of the chassis, the process is adjusted to eliminate the causes of these problems. The number of chassis coated between the time of problem diagnosis and the recovery time of the spraying process is 2, and the average number of coated chassis between successive recoveries is 800. The recovery cost is \$155. What is the optimal diagnosis interval of the process? What is the quality cost per unit of production?

Solution.

$$A = \$250$$
$$B = \$30$$
$$C = \$155$$
$$\bar{u} = 800$$
$$l = 2$$

Substituting the above values in Eq. 6.8, the optimal diagnosis interval is calculated as:

$$n^* = \sqrt{\frac{2(800 + 2)(30)}{250 - 155/800}}$$

$$n^* \approx 14 \text{ chassis}$$

The quality cost per unit of production is obtained by substituting the above values in Eq. 6.5:

$$L = \frac{B}{n^*} + \frac{n^* + 1}{2} \times \frac{A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}}$$

$$= \frac{30}{14} + \frac{15}{2} \times \frac{250}{800} + \frac{155}{800} + \frac{2(250)}{800}$$

$$= \$5.30$$

6.2 FREQUENCY OF PROCESS DIAGNOSIS

The previous two sections have presented approaches for estimating the optimal diagnosis intervals and the optimal sampling inspection frequency, based on the loss function. This section presents a different approach for determining the optimal frequency of process sampling for attribute characteristics, as a function of the production rate of the process. This approach is applicable in situations

where cost and time constraints dictate the use of sampling inspection instead of 100-percent inspection and where administrative difficulties preclude the use of other traditional process control methods.

The production process is checked for acceptance (of products) or control (of the process) by periodically inspecting a number of products and determining the status of the product and/or process. Clearly, the frequency of diagnosis (inspection) is affected by the cost incurred by defective products and the cost of process control. As the frequency of diagnosis increases, the cost incurred by defective products decreases, whereas the cost of process control increases, as shown in Fig. 6-3.

Process checking intervals (diagnosis intervals) can be established according to the rate of production. When a production process is operating at a high production rate, it requires a greater frequency of diagnosis than when it is operating at medium or lower production rates. Obviously production processes with a high production rate will produce many more units in the time lag period compared to production processes with low production rates. The objective is to find the frequency of diagnosis as a function of production rate so that the total quality cost for each unit of production is minimized. The following examples illustrate this point.

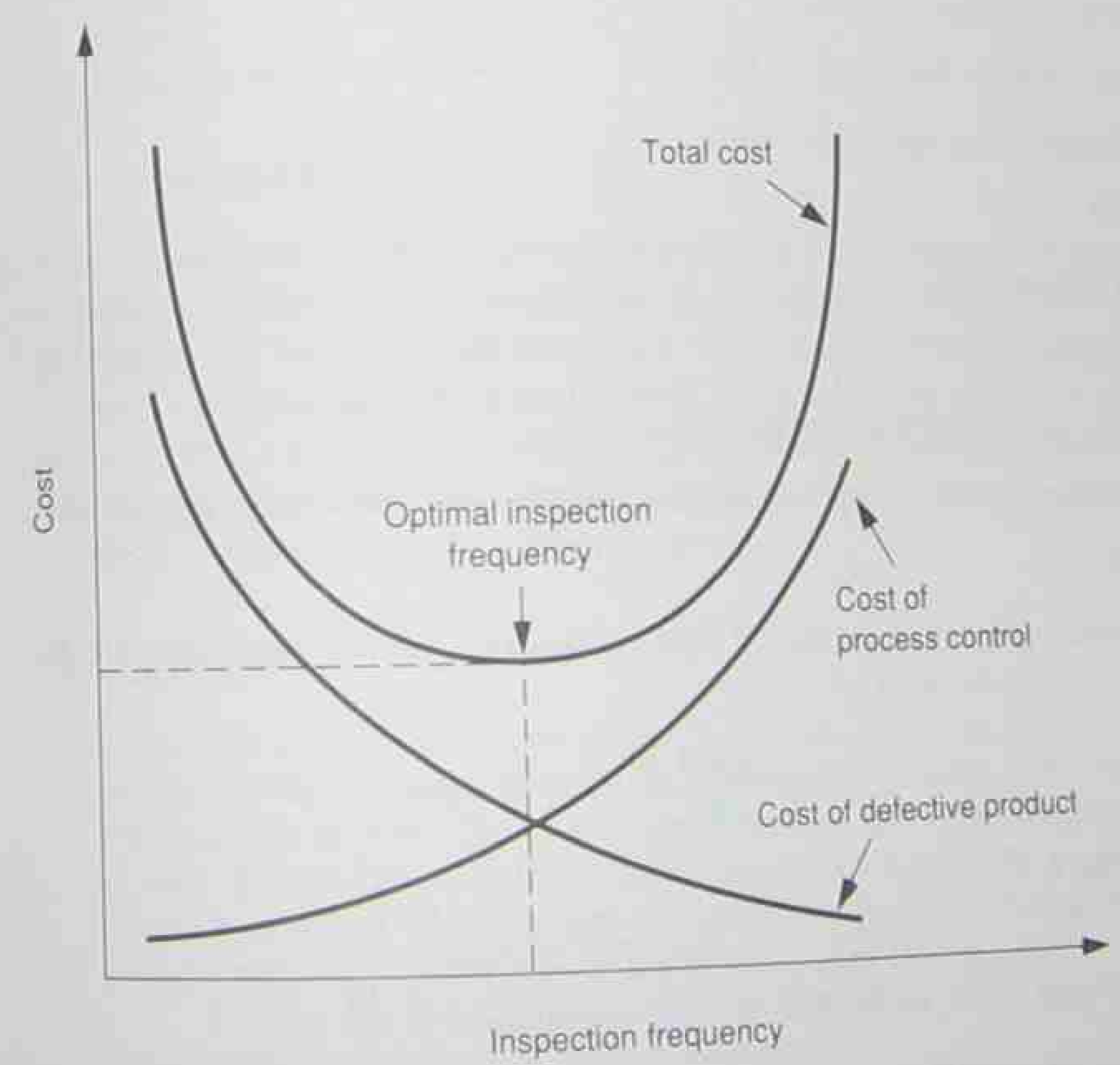


FIGURE 6-3
Optimal diagnosis frequency.

Example 6.11. The plinth of a turntable forms the unit base. It contains the turntable mechanism, electronics, and tone arm, and is molded in polystyrene by using an injection molding machine. It is inspected after molding for defects, such as flow lines (flow lines occur when the plastic "knits" together after it has flowed around either side of a core pin or similar obstruction during the molding process), sink marks (sink marks usually occur whenever there is a change in section thickness: ribs, locating pegs, mounting pillars), and shadow marks (shadow marks are caused by dramatic changes in direction of the plastic flow during injection into the mold; when it strikes a core pin or goes around a sharp corner to fill a recess or changes from one level to another [Shaw, 1985]).

If a plinth is found to be defective, it is scrapped to be reground and remolded. The loss caused by producing a defective plinth is \$50 when remolding is done and \$80 when remolding cannot be done. The former case occurs 60 percent of the time, and the latter occurs 40 percent of the time. The diagnosis cost is \$25 per diagnosis. The direct cost of process adjustment is \$100, and the time lag between the moment a sample is taken (and a defective product detected) and the moment the process is adjusted (so that no further defectives are produced) is 10 minutes. The average trouble occurrence interval of the injector is 400 products. Assuming that the process is in a controlled condition, determine the frequency of sampling inspection when the production rates are 2000, 1000, and 500 products per day.

Solution.

A = cost of a defective product

$= 50 \times 0.60 + 80 \times 0.40 = \$62$

B = diagnosis cost (inspection cost)

$= \$25$

C = adjustment cost

$= \$100$

l = time lag expressed as number of products

$\bar{u}$ = average number of products produced between two successive adjustments

$= 400$

Consider a production rate of 2000 units/day. Assume that there are 480 working minutes per day; then

$$l = \frac{2000}{480} \times 10 = 41 \text{ units}$$

By Eq. 6.8,

$$n^*_{2000} = \sqrt{\frac{2(400 + 41)(25)}{62 - 100/400}} = 19$$

where n^*_x is the optimal diagnosis period in terms of number of products when the production rate is x. In this case, the frequency of diagnosis is 105 per day.

Consider a production rate of 1000 units per day:

$$l = \frac{1000}{480} \times 10 = 21 \text{ units}$$

$$n^*_{1000} = \sqrt{\frac{2(400 + 21)(25)}{62 - 100/400}} = 18 \text{ units}$$

and the frequency of diagnosis is 56 per day.

Now consider a production rate of 500 units per day:

$$l = \frac{500}{480} \times 10 \approx 10 \text{ units}$$

and

$$n^*_{500} = \sqrt{\frac{2(400 + 10)(25)}{62 - 100/400}} = 18 \text{ units}$$

and the frequency of diagnosis is 28 per day.

The frequencies of process diagnosis at different production rates are

Production rate (per day)	Frequency (per day)
2000	105
1000	56
500	28

6.3 SUMMARY

This chapter discussed on-line quality control for attribute characteristics. It also examined methodologies similar to those used in the on-line quality control procedure for variable characteristics. They permit one to estimate the optimal diagnosis interval and the loss per unit for production processes when operating under normal conditions or when large numbers of defective items are produced. The chapter also discussed the frequency of production process diagnosis as a function of the production rate of a process.

PROBLEMS

6.1 An optical inspection system is used to identify defects in multilayer printed wiring boards (PWB) at a cost of B = \$3. The defects include open, short, and nonconductive wiring in the boards. The cost incurred because of a defective printed wiring board is \$90. If a defective board is found during diagnosis, the production process is adjusted at a cost of \$150, and 10 boards are produced during the time lag. The average number of boards produced between adjustments is 300.

What is the optimal diagnosis interval in terms of the number of boards produced? What is the quality cost per board?

6.2 A manufacturer of electronic assemblies uses a machine vision system to inspect resistor networks with 6, 8, or 10 pins. To determine whether or not a network is within specification, the system does the following:

1. It inspects the pins to determine that the correct number is present and that the pins are properly spaced.
2. It inspects the network's two ceramic standoffs for presence and proper shape.

The loss caused by producing a defective resistor is $25. The direct cost of process adjustment is $150, and the time lag from the diagnosis to the adjustment of the process (if the resistor is found to be defective) is 5 minutes. The average trouble occurrence of the production process is 60 products. Determine the frequency of sampling inspection when the production rates are 5000, 3000, and 1000 resistors per day.

6.3 Consider the manufacturer in Prob. 6.2. The average number of resistors produced between successive adjustments may be 150, 300, or 600. Determine the frequency of diagnosis for each of these process adjustment intervals when the production rates of the process are 10,000, 5000, and 2000 per day.

6.4 A manufacturer of high-volume integrated circuits is experiencing difficulty with the soldering process. Soldered connections may have problems with insufficient or excess solder, lead projection, or incorrect positioning of a device or lead.

Automated X-ray inspection systems are installed at different diagnosis points of the production line to check the solder quality of the product and to provide quantitative data for feedback control of the soldering process. Although defect detection is most often thorough, defective products occasionally go undetected and are sent to the next step in the production process. The loss caused by a defective unit is $20, the cost of the diagnosis is $15. The loss incurred when a defective unit is not disposed of at the diagnosis point is $35. The soldering process is adjusted at intervals of 1000 units of production at a cost of $80 per adjustment. The time lag of adjustment as expressed in terms of production units is 10. What are the optimal diagnosis interval and the loss per unit of production?

6.5 The production engineers of a cold reduction mill produce sheet steel by reducing the thickness of hot rolled steel coils to consumer specifications. The sheets are checked for surface defects. Because of the very high traveling speeds of the cold rolled strips (6000 ft/min), the engineers devised an optical system that checks for the surface defects. The average adjustment interval of the system is 4 hours, and the time lag is 3 minutes. The inspection cost is $20/min, the adjustment cost is $120, and the loss per unit of production (one foot of the sheet steel represents one unit of production) is $0.05. What are the optimal parameters of the on-line control system that minimize the quality loss per unit?

If a new device costing twice as much as the current device reduces the time lag by one-half, and the quality loss per unit is kept the same as that of the current system, what is the adjustment interval of the new system?

6.6 A manufacturer of high-temperature, high-pressure textile dyeing machines uses a high-speed microprocessor to monitor the dyeing process. Change in moisture content, temperature, and textile lint cause quality and maintenance problems. Sensors are connected to measure the change in temperature and moisture content of the dyeing process. Probes are also used to measure the amount of lint in the dyeing solution. The tolerance limits for these parameters are:

Temperature	± 3°C
Moisture content	± 5%
Amount of lint	$+0.1$ oz/cubic foot of solution

The quality of the dyeing process is checked by observing faded or darker-colored patches of the textile material.

The temperature and moisture content are adjusted at intervals of 10,000 and 15,000 feet of textile, respectively. The cost of adjusting the temperature is $5, and the cost of adjusting the moisture content is $20. Once the amount of lint is greater than 0.1 oz/ft^3, the process is stopped for maintenance at a cost of $250. The average time between maintenance (when converted to feet of textiles) is 200,000 feet.

The quality of the dyeing process is checked by using a high-speed vision system which takes continuous snapshots of the textile and compares the color of the material with a preprogrammed color. The cost of inspection is $0.10 per foot. What are the optimal control parameters for both temperature and moisture content? Assume that the loss per defective unit (one foot) of textile is $0.20.

6.7 A manufacturer of light accessories for automobiles checks the quality of taillight bulbs produced in a high-speed production line by testing them for light intensity. A defective bulb is automatically ejected from the line and discarded at a cost of $0.15. The ejection mechanism is monitored for accuracy. The mechanism may fail to eject the defective bulbs, or it may eject acceptable bulbs. The cost of inspecting the mechanism is $0.02, and the current adjustment interval is 20,000 bulbs. The probability that the mechanism ejects a good bulb is 0.01.

Assuming that there is a production rate of 10,000 bulbs per hour, what is the maximum acceptable adjustment cost for the mechanism if the quality loss per unit must not exceed $0.015 per bulb?

6.8 In a sheet metal forming operation, variations in process parameters are known to cause considerable variations in the unloaded shape of the formed part, even if the part is formed to the exact specifications before leaving the forming die. The phenomenon in which a formed part takes an unloaded shape different from the loaded shape is called "springback." This phenomenon may cause significant misalignment problems during the assembly of the affected part with other parts.

In order to reduce the "springback" effect, a producer developed an on-line quality control system for the sheet metal forming operation. Depending on variations in the sheet metal material, and on-line measurements of the friction coefficients at the punch (tool) and die interfaces, the control system adjusts the applied force at the punch interface and the speed of the rolling die. Assume that the tolerance of the applied force is ± 10 lbf, and the forming process produces sheets with thickness specifications of ± 0.002 in. The sheet is continuously monitored to ensure its quality. A change in the applied force by 10% will result in a 20% change in thickness of the formed sheet. The process is adjusted once every 1000 feet of sheet production at a cost of $25.00. The cost of measuring sheet thickness is $0.05 per foot. The time lag of the process expressed in sheet production is 26 feet of sheet metal, and the maximum quality loss allowed by the producer is $0.10 per foot. Assuming there is a defective loss of $2.00 per foot, determine the optimal tolerance of the sheet thickness.

REFERENCES

Berry, S.A. "Techniques in the Application of Computers to Industrial Monitoring." In *CAD/CAM and the Computer Revolution*, Society of Manufacturing Engineers, 1974, pp. 221–239.

Durrant, N.F. "Some Case Studies in Acceptance Sampling by Attributes." *Quality Assurance*, vol. 7, June 1981, pp. 21–22.

Kirkpatrick, E.G. *Quality Control for Managers and Engineers*. New York: John Wiley and Sons, 1970.
Shaw, M.B. "Process Study for an Injection-Moulded and Painted Hi-Fi Turntable Deck." *Quality Assurance*, vol. 11, December 1985, pp. 98–102.
Wadsworth, H.M., K.S. Stephens, and A.B. Godfrey. *Modern Methods for Quality Control and Improvement*. New York: John Wiley and Sons, 1986.

CHAPTER 7

ON-LINE QUALITY CONTROL: METHODS FOR PROCESS IMPROVEMENTS

The effects of production process parameters on product quality for variable and attribute characteristics were discussed in previous chapters. The total quality cost of a product is a function of many parameters, such as $\bar{u}, A, B, C$, and l. These can be classified according to the processes of the on-line quality control of the product. For example, A and $\bar{u}$ are parameters associated with the production process. B and l are parameters associated with the checking (diagnosis) process, and C is associated with the adjustment and recovery processes. These parameters have a direct effect on the estimation of the optimal checking interval and the optimal control limits of the production process. Table 7.1 lists the parameters and their associated processes.

The objective of this chapter is to present methods to improve the parameters of the production, diagnosis, and adjustment processes so that the total quality loss is minimized. We first introduce methods for improving the parameters of the production process.

7.1 PRODUCTION PROCESS IMPROVEMENT METHODS

As shown in Table 7.1, the parameters of the production process that affect the total quality cost per product are $\bar{u}$ and A. These parameters can be improved by

TABLE 7.1
Processes and parameters of the total quality cost

Process	Parameters
Production process	$A, \bar{u}$
Diagnosis process	B, l
Adjustment process	C

using any of the following methods individually or in combination:

1. Introduction of preventive maintenance measures such as tool changes (increases $\bar{u}$)
2. Use of tools with long lives (increases $\bar{u}$)
3. Improvement in the production process itself to achieve longer time between successive adjustments
4. Improvement of scrap and disposition methods for defective products (decreases A)

7.1.1 Preventive Tool Change

Preventive maintenance methods can be classified as either deterministic or probabilistic. Deterministic problems are those in which the timing and the outcome of the replacement (or maintenance) action are assumed to be known with certainty. For example, a cutting tool may not break down, but the loss caused by producing defective products increases with use. To reduce this loss, a replacement (or preventive maintenance) can be implemented.

Probabilistic problems are those where the timing and outcome of the replacement action depend on probability. In the simplest situation, the tool may be good or bad. The probability describing the state of the tool may be obtained by using a random variable whose distribution may be termed the tool's failure distribution.

Though failure distributions of tools play a major role in deciding on their optimal preventive maintenance schedules, this section will emphasize the preventive maintenance methods for deterministic problems. Throughout the chapter, there are references made to the cost caused by tool change, C', and the corresponding adjustment interval u'.

The effect of periodic tool changes as an approach to preventive maintenance is illustrated by the following example:

Example 7.1. A supplier of blocks for truck engines produces 3000 units per month. The production of the cylinder blocks requires a reaming operation. Defective products are produced when one of the reamers makes a defective hole. Process diagnosis is performed at constant intervals at a cost of \$4 per diagnosis. All defective products found are scrapped at a loss of \$60 per defective product.

The adjustment cost is \$150, and the average number of products produced between successive adjustments (failures of the process) is 2500 units. Assuming that the time lag is 1 cylinder block, what is the total quality control cost per product?

Solution. The parameters of the reaming process are

$$A = \$60$$
$$B = \$4$$
$$C = \$150$$
$$l = 1 \text{ unit}$$
$$\bar{u} = 2500 \text{ units}$$

By using Eqs. 6.8 and 6.5 the following results are obtained:

$$n^* = \sqrt{\frac{2(\bar{u} + l)B}{A - C/\bar{u}}}$$
$$= \sqrt{\frac{2(2500 + 1) \times 4}{60 - 150/2500}} \approx 18 \text{ units}$$
$$L = \frac{B}{n^*} + \frac{n^* + 1}{2} \times \frac{A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}}$$
$$= \frac{4}{18} + \frac{19}{2} \times \frac{60}{2500} + \frac{150}{2500} + \frac{1 \times 60}{2500}$$
$$= 0.222 + 0.228 + 0.060 + 0.024$$
$$= \$0.534$$

The total quality cost of the reaming process per month is \$0.534 × 3000 = \$1602.

As mentioned previously, the wearing of the reamers causes the production of defective products. Therefore, changing all reamers after the production of 1500 cylinder blocks instead of the current average failure interval of 2500 products will reduce the number of failures. In other words, the reamers are replaced by a new set, whether or not they are capable of producing nondefective products. This replacement procedure (a preventive maintenance measure) is called a periodic tool change. Of course, periodic changes of the reamers at smaller intervals of production u' will increase the cost per product. Let C' be the cost incurred by the change of the reamers. (This cost includes tool cost, labor cost, and the loss incurred by stopping the production process; usually $C' = C$.) It is estimated that if all reamers are replaced at every 1500th product, the probability of tool failure is 0.02.

The average interval between failures (after introducing the periodic tool change) is now expected to be

$$\bar{u} = \frac{1500}{0.02} = 75{,}000 \text{ units}$$

and the corresponding optimal diagnosis interval is

$$n^* = \sqrt{\frac{2 \times 75{,}000 \times 4}{60 - 150/75{,}000}} \approx 100 \text{ units}$$

Using Eq. 6.5 and adding the cost of tool change per unit, the total quality loss L is

$$L = \frac{C'}{u'} + \left(\frac{B}{n^*} + \frac{n^* + 1}{2} \times \frac{A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}} \right) \tag{7.1}$$

$$= \frac{150}{1500} + \left(\frac{4}{100} + \frac{101}{2} \times \frac{60}{75{,}000} + \frac{150}{75{,}000} + \frac{1 \times 60}{75{,}000} \right)$$

$$= 0.100 + 0.083 = \$0.183$$

Introduction of periodic tool changes results in a reduction of total quality cost per unit by \$0.36. The reduction in the annual cost is

$$0.36 \times 3000 \times 12 = \$12{,}636$$

In Example 7.1, the introduction of a preventive tool change at every 1500 units resulted in a significant cost reduction. It is usually more important to improve uniformity of tool life than to extend the average life length of the tool, and preventive maintenance should be considered (if applicable) in all production processes.

7.1.2 Tools with Longer Lives

Another method for reducing quality cost per product in a production process is to increase the time intervals between successive adjustments, $\bar{u}$, by employing cutting or manufacturing tools with longer lives. Example 7.2 illustrates this point.

Example 7.2. A manufacturer uses turret lathes to bore the housings of four-way solenoid valves. The cutting tool used is made of HSS (high-speed steel). Two types of defective products occur after the boring is complete; among all defectives, reworkable defectives occur 20 percent of the time, and nonreworkable defectives occur 80 percent of the time. The cost of reworking a defective housing is \$6 and the cost of scrapping a nonreworkable defective is \$50. Diagnosis cost is \$1.50 and adjustment cost is \$980. The monthly production is 8500 valves. Two adjustments are performed per month, the tool cost within the adjustment cost is \$250 and the time lag is 4 units.

A new tool with longer life has been introduced in the market with a cost of \$1200. Find the minimum ratio between the expected life of the new tool and the expected life of the current tool that will make it profitable to use the new tool.

Solution. The parameters of the boring process are

$$A = 6 \times 0.2 + 50 \times 0.8 = \$41.2$$

$$B = \$1.5$$

$$C = \$980$$

$$\bar{u} = \frac{8500}{2} = 4250 \text{ units}$$

$$l = 4 \text{ units}$$

The optimal diagnosis interval n^* is

$$n^* = \sqrt{\frac{2(4250 + 4) \times 1.5}{41.2 - 980/4250}} \approx 18 \text{ units}$$

and the present quality cost per unit is given by

$$L = \frac{1.5}{18} + \frac{19}{2} \times \frac{41.2}{4250} + \frac{980}{4250} + \frac{4 \times 41.2}{4250}$$

$$= \$0.44$$

Consider the new tool. Let the average number of units between successive adjustments be $\bar{u}$. Since l is much smaller than $\bar{u}$, and $C/\bar{u}$ is much smaller than A, n^* can be estimated by using the approximation of Eq. 6.6:

$$n^* = \sqrt{2\bar{u}\frac{B}{A}}$$

and substituting the equation above into Eq. 6.5, we obtain

$$L = \sqrt{\frac{2AB}{\bar{u}}} + \frac{\frac{1}{2}A + C + lA}{\bar{u}} \tag{7.2}$$

Equations 6.5 and 7.2 are equivalent.

Substituting A, B, l, and C for the new tool ($C = 980 + (1200 - 250) = \$1930$) in the above equations yields

$$L = \sqrt{\frac{2 \times 41.2 \times 1.5}{\bar{u}}} + \frac{\frac{1}{2} \times 41.2 + 1930 + 4 \times 41.2}{\bar{u}} \tag{7.3}$$

$$= \sqrt{\frac{123.6}{\bar{u}}} + \frac{2115.4}{\bar{u}}$$

If the cost L in Eq. 7.3 is less than \$0.44, then it is profitable to use the new tool. Therefore, $\bar{u}$ in Eq. 7.3 must be determined so that

$$\sqrt{\frac{123.6}{\bar{u}}} + \frac{2115.4}{\bar{u}} \leq 0.44$$

or

$$0.44\,\bar{u} - \bar{u}\sqrt{\frac{123.6}{\bar{u}}} - 2115.4 \geq 0$$

Substituting $x = \sqrt{\bar{u}}$ in the above equation gives

$$0.44x^2 - \sqrt{123.6}\,x - 2115.4 \geq 0 \tag{7.4}$$

The roots of Eq. 7.4 are

$$x = \frac{\sqrt{123.6} \pm \sqrt{123.6 + 4 \times 2115.4 \times 0.44}}{2 \times 0.44}$$

$$x = \frac{11.11 \pm 62.02}{0.88}$$

$$x = 83.10 \text{ or } -57.85$$

Since x must be positive, $x = 83.10$, and the average number of products between successive adjustments is

$$\bar{u} = x^2 = (83.10)^2$$

$$\bar{u} = 6906$$

The required ratio between the average life of the new tool and the present tool is

$$\frac{6906}{4250} = 1.624$$

That is, when the average life of the new tool is at least 1.624 times the average life of the current tool, it becomes profitable to replace the current tool with the new one.

If the new tool's average life is twice as long as the present one's, namely, $4250 \times 2 = 8500$ units, the optimal diagnosis interval becomes:

$$n^* = \sqrt{\frac{2 \times 8504 \times 1.5}{41.2 - 1930/8500}} \approx 25 \text{ units}$$

and the quality cost per product, L, is given by

$$L = \frac{1.5}{25} + \frac{26}{2} \times \frac{41.2}{8500} + \frac{1930}{8500} + \frac{4 \times 41.2}{8500}$$

$$= \$0.37$$

This is a cost reduction of \$0.07 per product from the present quality control cost. Thus, tool life plays an important role in the cost of quality per product, and it may be more economical to use tools with longer lives, even though their costs are higher. Of course, a more significant reduction in quality cost might be attained when periodic tool changes are performed for those tools with longer lives.

Example 7.3. Assume that the new tool introduced in Example 7.2 has an average life of 8500 units and that a periodic tool change is made after every 8000 units of production. With the new tool, the probability of producing a defective product before producing 8000 units is 0.10. Assuming $C = C'$, what would the quality cost be?

Solution. After combining the use of longer-life tools with periodic tool changes, the process production parameters are

$$u' = 8000 \text{ units}$$

$$C' = \$1930$$

$$A = \$14.20$$

$$B = \$1.50$$

$$C = \$1930$$

$$\bar{u} = \frac{8000}{0.10} = 80{,}000 \text{ units}$$

$$l = 4 \text{ units}$$

The optimal diagnosis interval is

$$n^* = \sqrt{\frac{2(80{,}004) \times 1.5}{41.2 - 1930/80{,}000}} = 76 \text{ units}$$

and L is obtained by substituting the above parameters in Eq. 7.1:

$$L = \frac{1930}{8000} + \left(\frac{1.5}{76} + \frac{77}{2} \times \frac{41.2}{80{,}000} + \frac{1930}{80{,}000} + \frac{4 \times 41.2}{80{,}000}\right)$$

$$= 0.24 + 0.066 \approx \$0.31$$

This loss is less than the loss (\$0.37) obtained in Example 7.2.

Because the cost of periodic tool changes in this equation is much greater than the combined costs of process diagnosis and process adjustment, it would be wise to find new methods for reducing excessive periodic tool changes. Such methods can be investigated experimentally by either reducing the cost of toolings or by using tools with different tool lives.

7.1.3 Automatic Process Diagnosis and Production Process Adjustment

This section introduces an alternative approach for improving the process parameter $\bar{u}$ by considering the use of automatic devices in process diagnosis, process adjustment, and inspection.

In recent years, the use of the digital computer in process monitoring, inspection, adjustment, and control applications has been expanded to include many production areas. Computer use has led to changes in production processes, and to the introduction of a variety of process diagnosis, adjustment, and control strategies.

Other control systems are also used. In systems such as feed-forward control systems, potential disturbances are measured before they disturb the process, and anticipatory corrective action is taken. Ideally, the corrective action compensates completely for the disturbance, thus preventing any deviation of the production process from the desired values.

In addition to process monitoring and adjustment systems for disturbances, much of the production equipment now available has sophisticated diagnosis and maintenance systems. When malfunctions or breakdowns occur, it is often difficult for a maintenance crew to immediately identify the source of trouble. Computers and sensors are often used to assist in diagnosing the problem and determining which machine component is malfunctioning. For example, current sensors can be used to detect current draw on motors in order to predict either motor failure or some other consequence. A current sensor, with output to a logic system, can help eliminate costly motor repairs by slowing a machine down or otherwise relieving stress on the motor (Death, 1986). This type of breakdown analysis may significantly reduce equipment downtime, thus increasing machine efficiency.

Diagnosis of the production process can also be performed by using automatic measuring devices attached to (or part of) the production equipment. Such devices permit automatic recording and calculating functions with the gauging of each single piece, thereby producing statistical evaluation virtually simultaneously with the gauging of individual parts. Both activities are executed without human action. There are many varieties of such devices and gauges; however, their detailed descriptions are beyond the scope of this book. The following example shows the effect of the introduction of automatic control systems on quality loss per unit of production.

Example 7.4. A manufacturer is considering the introduction of an automatic inspection device for helical gears. Each gear is inspected for concentricity, pitch, tooth profile, tooth thickness, and gear-tooth shape. The device can inspect all gears and detect 90-percent of all of the defective products. The purchase price of this device is $30,000 and its annual operating cost is 50 percent of its purchase price. The annual production of the process is 600,000 gears and the loss caused by producing a defective gear is $7. The present production process setup requires an operator to perform diagnosis at a cost of $8 per diagnosis. The adjustment cost of the process is $120 and the average number of gears produced between successive adjustments is 4800. The time lag caused by the current practice of the diagnosis method is 5 products, while the time lag of the automatic device is zero. Determine whether it is profitable to introduce the automatic inspection device.

Solution. The parameters of the present setup are:

$$A = \$7$$

$$B = \$8$$

$$C = \$120$$

$$l = 5 \text{ units of production}$$

$$\bar{u} = 4800 \text{ units}$$

The optimal diagnosis interval n^* is

$$n^* = \sqrt{\frac{2(\bar{u} + l)B}{A - C/\bar{u}}}$$

$$= \sqrt{\frac{2(4800 + 5) \times 8}{7 - 120/4800}}$$

$$= 105$$

Using Eq. 6.5, we obtain the quality loss per product:

$$L = \frac{8}{105} + \frac{106}{2} \times \frac{7}{4800} + \frac{120}{4800} + \frac{5 \times 7}{4800}$$

$$= \$0.185$$

The diagnosis cost of the automatic inspection device B' is

$$B' = \frac{\text{annual operating cost}}{\text{annual production}} = \frac{30{,}000 \times 0.5}{600{,}000} = \$0.025$$

The device can only detect 90 percent of all of the defective gears. In order to estimate $\bar{u}$ for adjusting the process after finding the defective products, it can be assumed that the operator makes the diagnosis to find the remaining 10 percent of the defective products that are ejected by the automatic inspection device. Thus, it is appropriate to give $\bar{u}$ a value of 10 times the present $\bar{u}$. Therefore, the diagnosis and adjustment cost associated with the remaining 10 percent of the defective gears is calculated as follows:

$$n^* = \sqrt{\frac{2(48{,}005) \times 8}{7 - 120/48{,}000}} = 331$$

$$L = \frac{8}{331} + \frac{332}{2} \times \frac{7}{48{,}000} + \frac{120}{48{,}000} + \frac{5 \times 7}{48{,}000}$$

$$= \$0.052 \tag{7.5}$$

The diagnosis and adjustment cost of the automatic device L' is obtained in the following way: The adjustment interval of the device, assuming that the automatic inspection device inspects one unit of product at a time, is

$$\bar{u}' = \frac{4800}{0.9} = 5333$$

and

$$L' = B' + \frac{1+1}{2} \times \frac{A}{\bar{u}'} + \frac{l'A}{\bar{u}'}$$

where

$$B' = \$0.025$$

and

$$l' = 0$$

Therefore,

$$L' = 0.025 + \frac{1+1}{2} \times \frac{7}{5333} + \frac{120}{5333} + 0$$

$$= \$0.0488 \tag{7.6}$$

The total cost of diagnosis and adjustment per product after the introduction of the automatic inspection device is the sum of Eqs. 7.5 and 7.6. Thus,

$$L = 0.052 + 0.0488 = \$0.1008 \approx \$0.1$$

The reduction of quality cost per product when the inspection device is introduced is $0.085. Economic justification of the device should include the cost of capital investment, life of the device, interest rate, and its controlling cost.

7.2 PROCESS DIAGNOSIS IMPROVEMENT METHODS

Improvement of diagnosis and inspection methods as well as their location in the production process may lead to reductions in total quality cost per product, as illustrated in the following examples.

Example 7.5. Example 6.3 showed that the optimal process diagnosis interval of the spot-welding machine was 183 units and the quality cost per product was

\$0.0242. The current inspections of the welded parts involve testing the parts by checking their appearance and by using a tensile test to determine the tensile strength of the welded joints. The inspection process consumes 8 minutes from the time a part is sampled until results are obtained and a decision is made with respect to the quality of the welding process. The tensile test is destructive. Thus, parts that undergo the tensile test are scrapped with no recoverable costs. The cost of the diagnosis is \$1.60, and the time lag is 30 parts. Consider replacing the current inspection method with a nondestructive ultrasonic method that uses high-frequency vibrations to detect all types of flaws in welds. The diagnosis cost is \$0.4, and the time lag l, is 2 parts. What is the effect of the ultrasonic inspection method on the quality cost per part?

Solution. The parameters of the ultrasonic inspection process are

$$A = \$0.5$$
$$B = \$0.4$$
$$C = \$20$$
$$\bar{u} = 5250 \text{ parts}$$
$$l = 2 \text{ parts}$$

The optimal diagnosis interval is

$$n^* = \sqrt{\frac{2(5252)(0.4)}{0.5 - 20/5250}} = 92 \text{ parts}$$

Using Eq. 6.5:

$$L = \frac{0.4}{92} + \frac{93}{2} \times \frac{0.5}{5250} + \frac{20}{5250} + \frac{2 \times 0.5}{5250}$$
$$= \$0.0128$$

With the ultrasonic inspection method, the optimal diagnosis and adjustment cost per part becomes \$0.0128. This is a realization of \$0.0242 − \$0.0128 = \$0.0114 per part, or a monthly cost reduction of \$479.

Generally, production process diagnosis is done immediately after processing; however, there are situations where it is preferable to diagnose a process after information from later stages of production (e.g., assembly, shipping, and customer service) has been gathered. Consider the following situation:

Example 7.6. A production process's diagnosis is currently done at the assembly station. Because any defective part is readily found at the assembly station, the diagnosis cost is assumed to be negligible. The loss caused by a defective part, A, is \$0.50. Adjustment cost C is \$50, and time lag is 800 parts. During the last three months, 300,000 parts were produced, and 20 adjustments were made to the production process.

The manufacturing department is considering moving the diagnosis process from the assembly area to an area where it will be done immediately after the production process. The associated costs of this decision are as follows: Diagnosis cost B = \$1.00; time lag l decreases to 5 parts; and the cost incurred by the production of a defective product, A, decreases to \$0.40 (reduction in the value-added loss caused by a defective product at subsequent production stations). The adjustment cost is the same as the present cost (\$50.00).

(a) Discuss the impact of changing the location of the diagnosis process on the quality cost per product.

(b) Assume that a periodic tool change is made at u' = 10,000 parts and that the probability of producing a defective product (part) is 0.02. Discuss the combined effect of moving the diagnosis process and introducing the tool change on the total quality cost (assume that $C = C'$).

Solution.

(a) In order to discuss the effect of moving the diagnosis process, the quality cost per product for the present system must be determined. The parameters of the present diagnosis process are:

$$B = \$0$$
$$A = \$0.5$$
$$C = \$50$$
$$n = 1 \text{ part}$$
$$\bar{u} = \frac{300{,}000}{20} = 15{,}000 \text{ parts}$$
$$l = 80 \text{ parts}$$

The loss for the present system is

$$L = \frac{B}{n} + \frac{n+1}{2} \times \frac{A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}}$$
$$= \frac{0}{1} + \frac{2}{2} \times \frac{0.5}{15{,}000} + \frac{50.00}{15{,}000} + \frac{800 \times 0.5}{15{,}000}$$
$$= \$0.0300 \qquad (7.7)$$

When the diagnosis process is performed immediately after the production process, the parameters become

$$B = \$1.00$$
$$A = \$0.4$$
$$C = \$50.00$$
$$\bar{u} = 15{,}000 \text{ parts}$$
$$l = 5 \text{ parts}$$

Now the optimal diagnosis interval is

$$n^* = \sqrt{\frac{2 \times 15{,}005 \times 1.00}{0.4 - 50.0/15{,}000}} = 275 \text{ parts}$$

The loss for the new system is

$$L = \frac{1}{275} + \frac{276}{2} \times \frac{0.4}{15{,}000} + \frac{50}{15{,}000} + \frac{5 \times 0.4}{15{,}000}$$

$$= \$0.0108 \qquad (7.8)$$

When Eqs. 7.7 and 7.8 are compared, it can be concluded that performing process diagnosis immediately after production results in a cost reduction of $0.0192 per part, or $1920 per month.

(b) Next, the effect of both changing the location of the diagnosis process and introducing periodic tool changes is investigated. Since $u' = 10{,}000$ and the probability of the processing becoming faulty is 0.02,

$$\bar{u} = \frac{10{,}000}{0.02} = 500{,}000 \text{ parts}$$

The loss after the introduction of periodic tool change and without a relocation of the diagnosis process is

$$L_1 = \frac{C'}{u'} + \left(\frac{B}{1} + \frac{1 + l}{2} \times \frac{A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}}\right)$$

$$= \frac{50.00}{10{,}000} + \left(0 + \frac{1 + 80}{2} \times \frac{0.5}{500{,}000} + \frac{50.00}{500{,}000} + \frac{80 \times 0.5}{500{,}000}\right)$$

$$= \$0.0052 \qquad (7.9)$$

The combined loss caused by the relocation of the diagnosis process to a position immediately after the production process and the introduction of the periodic tool change is calculated as follows:

$$n^* = \sqrt{\frac{2(500{,}000 + 5) \times 1.00}{0.4 - 50.00/500{,}000}} = 1582 \text{ parts}$$

$$L = \frac{C'}{u'} + \left(\frac{B}{n^*} + \frac{n^* + 1}{2} \times \frac{A}{\bar{u}} + \frac{C}{\bar{u}} + \frac{lA}{\bar{u}}\right)$$

$$= \frac{50.00}{10{,}000} + \left(\frac{1.00}{1582} + \frac{1583}{2} \times \frac{0.4}{500{,}000} + \frac{50.00}{500{,}000} + \frac{5 \times 0.4}{500{,}000}\right)$$

$$= \$0.0064 \qquad (7.10)$$

The combined effect of periodic tool changes and relocation of the diagnosis process increases the quality cost per part by $0.0012, or $120 per month, over the loss achieved by periodic tool change alone.

In general, when the average time between successive adjustments is short, diagnosis immediately after the production process is recommended. When the average interval between adjustments is long, it is less costly to diagnose the process at later processing stages where the diagnosis cost is usually nil.

7.3 PROCESS ADJUSTMENT AND PROCESS RECOVERY IMPROVEMENT METHODS

Methods for improving the production and diagnosis processes in order to minimize the quality control cost per product have been discussed in previous sections of this chapter. The recovery process (which involves bringing a production process to normal production after a stoppage) is the third kind of process that has a direct impact on quality control cost. This process, and methods for improving it, will be discussed in this section.

Although there are many improvement methods for the recovery process, this discussion is limited to two of the most commonly used methods. These methods are: (1) use of redundant (spare) machines; and (2) the implementation of an automatic diagnosis-recovery system.

Consider the cost of recovering a process, C. This cost is composed of two elements: the direct cost of recovery, and the cost incurred by the interruption of the production process while the adjustment is being made. This can be expressed as follows:

$$C = C_r t + C_d \qquad (7.11)$$

where C = cost of recovery
C_r = cost incurred by the interruption of the production process per unit time
t = average interruption time
C_d = direct recovery cost

Equation 7.11 indicates that the cost of recovery can be reduced by decreasing the cost of interrupting the system to perform the recovery, by reducing direct recovery cost, or by reducing both costs together. The average interruption time can be decreased by using redundant machines that can immediately continue production when a machine in the production line produces defective products. This decrease can also be accomplished by using automated facilities that can complete the recovery process in a shorter time. The use of redundant machines and the use of automatic recovery methods are discussed below.

7.3.1 Use of Redundant Machines

One of the major problems associated with automated flow production lines (production machines arranged in series) is reliability. Since the line often operates as a single mechanism, failure of one operating machine often results in stoppage of the entire line. Two methods that can improve the reliability of the production line are the provision of buffer storage between production stages and the use of redundant machines at some of the production stages.

This discussion concerns the use of redundant machines as a means of reducing production loss caused by the recovery process. Redundant machines are usually assigned to critical stages of the production system, as well as to bottleneck stages (a bottleneck stage has a cycle time as long as the cycle time of the entire production system). Redundant machines are also assigned to production stages that experience frequent machine failure. Redundant machines are grouped under three classifications:

1. *Hot standby:* a spare machine operating parallel to another similar machine. If one machine fails or needs adjustment and recovery procedures, the other machine continues production without interruption of the entire production system, reducing the time lag for recovery to zero.
2. *Cold standby:* a spare machine is available to start production as soon as the operating machine requires recovery procedures. The time lag for recovery may not necessarily be zero because start-up or setup time is needed to bring the spare machine on line.
3. *Warm standby:* This is similar to a hot standby with the exception that the portion of the production output of the spare machine at the production stage is always less than 50-percent of the total production output of that specific production stage. When the principal machine (machine producing more than 50-percent of production) requires repairs, the spare machine increases its production output to compensate for the portion of production lost while the principal machine is adjusted. In this case, the production loss caused by the time lag for recovery of the principal machine is zero.

The following example illustrates the use of redundant machines in reducing production loss (the quality cost per product) caused by the recovery process.

Example 7.7. An injection molding process is used for the production of print heads for high-speed printers. A mold that contains the pattern of the print heads is set up at the beginning of the production period. The mold must be checked for cleaning at an interval of 2500 units of production with an average cleaning time of 5 minutes and cleaning cost of \$6.00. The cost of interrupting the production process is \$4.00 per minute. In addition, the average interval of trouble occurrence is 20,000 items, and setting the machine up for resumed production after trouble occurs requires 8 hours at a direct cost of \$500. The parameters of the current process are

$$A = \$6$$

$$B = \$5$$

$$l = 1 \text{ product}$$

In order to decrease the quality cost per product, the manufacturing engineering department introduced a redundant machine in a cold standby state: the redundant machine would carry on production only when the mold required either cleaning or replacement. The cost of the redundant machine is \$200,000. The total annual cost is \$10,000 per year, and the financing interest is 15 percent. It takes 5 minutes to prepare the redundant machine to resume production.

Assume that the annual production is 3,000,000 print heads. What is the effect (on total quality cost per product) of introducing the spare machine?

Solution.

Current process. Two types of adjustments are required for the current process. Their parameters and costs are shown in Table 7.2.

Since type 2 adjustment requires checking the process, the average number of units between successive troubles, $\bar{u}$, is

$$\bar{u} = \bar{u}_2 \approx 20{,}000 \quad (7.12)$$

The optimal diagnosis interval n^* is

$$n^* = \sqrt{\frac{2(20{,}000 + 1) \times 5}{6 - 2420/20{,}000}} \approx 184 \text{ products} \quad (7.13)$$

The optimal diagnosis interval obtained by using Eq. 7.13 is based on the assumption that diagnosis is performed without interruption of the production process. If the process must be stopped and interrupted for diagnosis, the loss incurred by interrupting the process should be included as part of the diagnosis cost B. The diagnosis cost including cost of interrupting production ($l = 1$ product or 0.5 min) is

$$B = 5 + 1 \times 0.5 \times 4 = \$7$$

The corresponding optimal diagnosis interval is 218 products. The quality cost per product for $n^* = 218$ is given by Eq. 7.14:

$$L = \frac{7}{218} + \frac{219}{2} \times \frac{6}{20{,}000} + \frac{2420}{20{,}000} + \frac{1 \times 6}{20{,}000} = \$0.19 \quad (7.14)$$

TABLE 7.2
Parameters of the injection molding process

Type of adjustment required	Average interruption time	Direct adjustment cost	Number of units between adjustments	Total adjustment cost
1. Reset of mold	5 min	\$6	$\bar{u}_1 = 2500$	$C_1 = 4 \times 5 + 6 = \$26$
2. Repair or replacement of mold	8 hr	\$500	$\bar{u}_2 = 20{,}000$	$C_2 = 4 \times 480 + 500 = \$2420$

The adjustment cost is the component of Eq. 7.14 that can be most significantly reduced by the introduction of a redundant machine. The total quality cost for each product after the introduction of the spare machine is estimated below.

New process. The annual interest for financing the machine is $30,000. The annual cost of the machine is $10,000. The total annual cost of the redundant machine is $40,000. The increased cost per product incurred as a result of the introduction of the new machine is 40,000/3,000,000 = $0.013.

Thus, the introduction of the redundant machine results in a cost increase of $0.013 per product, as well as continuous production of products, because the redundant machine replaces the principal machine whenever the mold requires cleaning or replacement. When the principal machine is down, preparation time for the redundant machine to start production is 5 minutes (that is, the production process is interrupted for 5 minutes). Therefore,

$$C_2 = 4 \times 5 + 500 = \$520 \tag{7.15}$$

It should be noted that the redundant machine may not necessarily be used to replace the principal machine when type 1 adjustments are performed. (Table 7-2 shows that the time of type 1 adjustment equals 5 minutes, the same amount of time needed to prepare the redundant machine to start production.)

The optimal diagnosis interval is

$$n^* = \sqrt{\frac{2(20{,}000 + 1) \times 7}{6 - 520/20{,}000}} \approx 216 \text{ units}$$

The quality cost per product is obtained as

$$\begin{aligned} L &= \frac{7}{216} + \frac{217}{2} \times \frac{6}{20{,}000} + \frac{520}{20{,}000} + \frac{1 \times 6}{20{,}000} \\ &\quad + (\text{cost of spare machine per product}) \\ &= 0.091 + 0.013 \\ &= \$0.104 \end{aligned} \tag{7.16}$$

The introduction of the redundant machine reduces quality cost by $0.086 for each product, or $258,000 per year.

7.3.2 Automatic Diagnosis: Adjustment and Recovery Systems

The total cost of adjustment and recovery as given by Eq. 7.10 can also be reduced by the introduction of automated diagnosis-recovery systems. Increasingly, such systems are being widely used in many automated production systems because of their economic viability and flexibility. Presently, there is a large variety of automated diagnosis and adjustment systems available that are capable of sorting defective parts, and controlling and adjusting the production process.

Typical applications of automatic inspection and process control systems are similar to those described below. After each production process, an automatic inspection and diagnosis device inspects dimensional and electrical characteristics of the product, and then takes the following action:

1. Control of process: If the value of inspected characteristics deviates from the nominal by the control limit, the control factor level of objective characteristics is changed automatically to ensure that the next unit will be near to nominal.
2. Automatic stoppage: If the actions described above are followed and if the next product is found defective, that is, out of specification limit, the automatic inspection device stops the production process and calls for operator attention.

Recent developments in computer technology, artificial intelligence, and expert systems have resulted in the introduction of advanced diagnosis and adjustment systems capable of diagnosing the causes of defective products and isolating the sources of the problems. In addition, such systems are capable of instituting necessary adjustment and recovery procedures so that the production process is minimally interrupted. The following example illustrates the economic justification for the introduction of automatic diagnosis-adjustment and recovery systems in a production line.

Example 7.8. Assume that it is technically feasible to use an automatic diagnosis-adjustment system for a production process. The cost of such a system is Q, and its annual operating cost is $0.5Q$. The parameters of the current diagnosis-adjustment system are

$$A = \$50$$
$$B = \$160$$
$$C = \$2000$$
$$\bar{u} = 5250 \text{ products}$$
$$l = 30 \text{ products}$$
$$n = 184 \text{ products}$$
$$\text{Annual production} = 504{,}000 \text{ products}$$

What must the cost of the diagnosis-adjustment system be, to make it profitable to introduce it into the production process?

Solution. The cost of quality per unit before the introduction of the automatic diagnosis-adjustment system is

$$\begin{aligned} L &= \frac{160}{184} + \frac{185}{2} \times \frac{50}{5250} + \frac{2000}{5250} + \frac{30 \times 50}{5250} \\ &= \$2.42 \end{aligned} \tag{7.17}$$

The annual cost of the automatic-adjustment system per unit is given by:

$$\frac{0.5Q}{504{,}000} \tag{7.18}$$

Assume that this automatic system continuously diagnoses the unit, and that once a defective unit is found, it adjusts the production process so that the characteristics of the ensuing units are within their specified limits (the time lag is zero). The cost of quality incurred by the introduction of the automatic adjustment system becomes

$$L = \frac{0.5Q}{504{,}000} + \frac{1+1}{2} \times \frac{50}{5250}$$

$$= \frac{0.5Q}{504{,}000} + 0.01 \qquad (7.19)$$

It is economically feasible to introduce the automatic diagnosis-adjustment system when L in Eq. 7.19 is less than or equal to \$2.42. That is,

$$\frac{0.5Q}{504{,}000} + 0.01 \le 2.42$$

or

$$Q \le \$2{,}429{,}280 \qquad (7.20)$$

It has been assumed so far that the automatic diagnosis-adjustment system has no time lag and that the automatic system itself has no failures or adjustments. Consider a case where the automatic system finds a defective unit, stops the production process, and makes an adjustment with a time lag of 50 units. In effect, defective units would not be produced; rather, the production process would be interrupted for a production period equal to 50 units.

Consider an automatic diagnosis-adjustment system that costs \$2,000,000 and has a time lag of 1 unit. The cost of quality per unit is obtained by substituting the process parameters in Eq. 7.1.

$$L = \frac{0.5 \times 2{,}000{,}000}{504{,}000} + \frac{1+1}{2} \times \frac{50}{5250} + \frac{1 \times 50}{5250}$$

$$= \$2.00$$

A saving of \$0.42 per unit, or \$211,680 per year, is realized.

7.4 SUMMARY

This chapter examined methods for improving product quality during production. Methods for improvement of production processes included periodic tool changes, automatic diagnosis and recovery of the production process, process diagnosis improvement, and the scrap and disposition of defective products. Employing any of these methods separately, or in combination, causes significant improvements in the quality of products. Obviously, the cost of employing any of the methods presented must be considered economically in terms of the increased improvement in product quality. Numerous examples were cited to evaluate the economic consequences of introducing the methods mentioned above.

PROBLEMS

7.1 A production line produces 1000 printed circuit boards (PCBs) per day. Inspection of the boards is conducted after the last production process. If a board is found

defective, it is scrapped at a cost of \$50 per unit. The process is diagnosed once every 200 units of production at a cost of \$100 (this cost includes diagnosis, review, and evaluation of the process). In addition, the diagnosis procedure requires 10 minutes.

If diagnosis determines that the production systems or machines require adjustment, it is immediately performed at a direct cost of \$100 per adjustment. The average number of PCBs produced between successive adjustments is 6000 units.

(*a*) What is the current quality cost per unit?

(*b*) What will the quality cost per unit be if the diagnosis interval is optimized?

7.2 An injection molding machine is diagnosed at every 100-shot interval (12 products are produced in every shot). The number of units produced per day is 800 shots (250 working days per year). Diagnosis cost B is \$2, and the time lag l is 2 shots. The loss per defective unit, A, is \$6, and the average number of shots produced between successive adjustments is 4000.

(*a*) Find the optimal diagnosis interval. What are the annual savings because of the adoption of the optimal interval?

(*b*) Assume that it takes 5 minutes for each diagnosis and 30 minutes for adjustment. What are the staff-hours required for the following quality control systems? (i) Existing system; and (ii) quality control system when using only the optimal diagnosis interval.

7.3 Suppose the parameters of the diagnosis-adjustment system for a production process are $A = \$8$, $B = \$1.5$, $C = 120$, $\bar{u} = 6000$ units of production, $l = 2$ units. The distribution of the number of units produced before failure, for the last 50 process failures, is shown below:

Number of units produced	0–1000	1001–2000	2001–3000	3001–4000	4001–5000	5001–6000
Number of occurrences	1	2	3	5	6	33

Currently, the diagnosis is performed immediately after the injection process. It can also be performed at the assembly process, and the parameters for this case are $A_0 = \$10$, $B_0 = \$0$, $C_0 = C = \$120$, $\bar{u}_0 = \bar{u} = 6000$ units, and $l_0 = 1200$ units. A periodic tool change can be implemented at a cost of $C' = 100$, with any of the replacement intervals $u' = 1000, 2000, 3000, 4000$, or 5000.

Design quality control systems with the following combinations:

(*a*) present diagnosis system without tool changes

(*b*) new diagnosis system with periodic tool changes, using $u' = 1000$, $u' = 2000$, $u' = 3000$, $u' = 4000$, and $u' = 5000$

(*c*) new diagnosis system without tool changes

(*d*) new diagnosis system with tool changes as in (*b*)

Compare the quality control systems in (*a*), (*b*), (*c*), and (*d*).

7.4 A manufacturer of LVDTs (linear variable differential transducers), used for the inspection of linear dimensions of products, inspects the transducers at the final stage of production. If a transducer is found defective, it is discarded at a cost of \$25. The parameters of the production process are

$$B = \$3$$

$$C = \$500$$

$$\bar{u} = 800 \text{ units}$$

$$l = 20 \text{ units}$$

(a) What is the optimal diagnosis interval of the process? What is the quality cost per unit?

(b) In order to minimize the cost of defective units, the quality control department recommended that a check should be made at the third production process. Implementation of this recommendation would reduce the cost of a defective unit to \$10. In addition, the present method of inspection should be upgraded at an increased cost of \$4 per unit. Determine the cost of quality per unit if these recommendations were implemented. (Annual production is 10,000 units.)

7.5 A manufacturer of light bulbs uses an accelerated life test to determine the average life of the bulbs. The light bulbs are tested at the final production stage with inspection costs of \$10 per bulb. (Because of the destructive nature of the test, only samples of the bulbs are subjected to the accelerated life test.) If the average life of bulbs in a sample is less than 750 hours, the production process is interrupted for adjustments for a duration of 90 minutes. The direct cost of adjustment is \$50, and the cost of interrupting production is \$20 per minute.

The checking process can be improved by using an electron microscope capable of detecting 80 percent of the defects in a bulb filament, at a cost of \$0.70 per unit. The following costs are associated with the checking process:

Inspection process	A	l	$\bar{u}$
Accelerated life test	\$1/sample	200	2500
Electron microscope	\$1/unit	10	—

Assume that annual production is 2,000,000 units. Is it economical to introduce the new inspection and diagnosis process?

7.6 The parameters of a stamping process are as follows:

$$A = \$12$$

$$B = \$2$$

$$C = C_r t + C_d = \$2 \times 180(\text{minutes}) + \$60$$

$$\bar{u} = 5000 \text{ units}$$

$$l = 10 \text{ units}$$

(a) Find the optimal diagnosis interval, and obtain the quality cost per year, assuming a yearly production of 300,000 units.

(b) What are the savings (if any) in quality cost per year if a redundant stamping machine with operating cost of \$150,000 per year is acquired, and the interruption time of the production process is reduced to 2 minutes instead of 180 minutes?

7.7 A producer of fiber optic filaments to be used for optical transmission of communications signals, television programs, and telephone calls developed an inspection system for the filaments. The fiber optic filament consists of a single filament of glass drawn to 1 mm in diameter and 3000 ft in length. Maintaining a consistent diameter over 3000 feet is difficult with the filament constantly in motion, making physical-contact measuring methods impossible to use.

The producer uses a 2-mW laser, scans the laser beam across the moving fiber optic filament, and focuses the reflected laser light into a photo receiver. Any change in the filament diameter can be detected by comparing the signal of the photo receiver with data stored in a comparator. Any increase in diameter will cause a draw motor control impulse to increase the speed of the draw; if there is a decrease in diameter, the processed signal provides a draw motor signal to reduce the speed of the draw. The diameter of the filament is continuously measured. If it is found to be out of the 1 ± 0.05 mm specification, a control system adjusts the speed of the draw motor. However, there is a 1-minute delay, causing a loss of \$8.00. The control system is adjusted every 4 hours at a cost of \$73.00. The cost of inspection is \$5.00 per 1000 feet. Assuming that the production rate is 10,000 feet per hour, calculate the quality loss per unit of production (1 foot).

In order to increase the production rate of the system, the manufacturer intends to introduce a new inspection system. The cost of the new inspection system is \$80,000, and the average time between adjustments is 6.5 hours. However, the inspection cost is \$3.00 higher than that of the current system. As a result of the introduction of the new system, the production rate would be 20,000 feet per hour. What are the optimal parameters of the new system? Compare the quality losses for both current and new systems.

7.8 An assembly plant for light and heavy truck axles requires accurate torque on 20,000 bolts per day. Every torque on an axle requires a precise torque value. Application of excess torque results in teeth stripping of the bolt thread, and insufficient torque results in bolt loosening and axle disassembly.

The operator uses a tool to apply the required torque. The tool is periodically calibrated in order to ensure the quality of the torquing operation. The required torque specification for every bolt is 1200 ± 20 lb · ft. The bolt-tightening operation is continuously monitored. When the torque limits are exceeded (by insufficient torque or excessive torque), the axle is diverted to another assembly station for adjustments at a cost of \$10 per axle. The torque monitoring system has an accuracy of ± 10 lb · ft. This may cause the diversion of axles having torque within specification limits to the adjustment station. Also, it may not divert axles that have torque values outside the specification limits. However, the monitoring system is always calibrated to ensure that the probability of not diverting axles with proper torque is 0.999.

The two processes under study are the torque application process and the monitoring process. If an axle is shipped without having proper torque, it causes a \$300 loss. Every 500 bolts, the tool is calibrated at a \$5 cost. The cost of calibrating the monitoring system every 2000 bolts is \$100. An axle has approximately 50 bolts.

Determine the tolerance limits for the torque to ensure that no axle is shipped without having the proper torque applied to its bolts. Suggest methods to improve the current system. Compare the performance of the improved systems with the current one.

7.9 Microfinishing, the final operation before assembly on crankshafts, camshafts, and plain bearing journals, achieves surface finishes of 2 μin. The microfinishing process uses a noncompressible, film-backed abrasive that is held in a precision-machined toolholder. The toolholder consists of two halves that conform to the shape of the workpiece, which is held between centers, rotated, and oscillated. The parameters of the process are the following: contact pressure between the toolholder and the workpiece, rpm (revolutions per minute) of the workpiece, degree of oscillation, and heat generated by friction between the toolholder and the workpiece.

It is extremely important to keep tight control on the part's dimensions. This is achieved by measuring, before finishing, the part's dimensions and temperature. The system's in-process controller uses this information to determine the amount of material to be removed and to bring the part within size tolerance. Temperature is measured because during processing, friction generates heat, affecting the size of the part. The data obtained during gauging are fed back to the controller, which constantly compares it with the required size. Once the correct size is reached, the tool is automatically released.

A manufacturer uses the microfinishing process to achieve diameters of plain journal bearings with tolerances of ±0.004 in. Two critical factors have direct impact on the quality of the process (quality is measured by the percentage of bearings having diameters within the specified tolerance). These factors are temperature between the bearings and the toolholder, and the holding pressure of the workpiece. The temperature is kept at 30°C, with the pressure at 100 lb/in^2. A 1°C change in the temperature results in a 0.0001-in change in the bearing diameter. A 5-lb/in^2 change in the pressure will result in a 0.0002-in change in the diameter. The temperature is controlled by adjusting the amount of the coolant; the pressure is controlled by adjusting the distance between the two halves of the workpiece.

Delays exist from the moment gauging takes place until the required parameter is changed. Because of the demand requirements, each bearing spends 10 minutes in the microfinishing process. Afterwards, its diameter is measured. If it is found that the diameter is smaller than the lower specification limit, the bearing is discarded at a loss of $175. However, if the diameter is greater than the upper specification limit, the bearing is reworked at a cost of $25. The controller is checked after the production of 200 bearings at a cost of $110. The costs for temperature and pressure adjustments are $10 and $15, respectively. Four sets of bearing tolerance measurements are taken after the microfinishing process and are as follows:

Set 1:
Temperature = 30.5°C
Pressure = 103 psi
Bearing tolerances:

+0.0001	−0.0003	+0.0004	−0.0003
−0.0002	+0.0004	+0.0002	+0.0001
−0.0003	−0.0005	+0.0003	+0.0006
+0.0001	+0.0001	−0.0001	−0.0004
−0.0003	+0.0002	+0.0004	−0.0005

Set 2:
Temperature = 30.5°C
Pressure = 105 psi
Bearing tolerances:

+0.0002	−0.0004	+0.0003	−0.0004
+0.0001	+0.0005	+0.0005	+0.0002
+0.0004	−0.0003	−0.0003	−0.0005
+0.0005	−0.0002	+0.0006	−0.0004
+0.0004	−0.0001	+0.0003	+0.0003

Set 3:
Temperature = 31°C
Pressure = 104 psi
Bearing Tolerances:

+0.0003	−0.0001	−0.0006	+0.0002
−0.0005	+0.0003	−0.0004	−0.0001
−0.0004	+0.0005	+0.0005	−0.0003
+0.0006	−0.0004	+0.0003	+0.0001
−0.0003	+0.0002	−0.0004	−0.0005

Set 4:
Temperature = 29.5°C
Pressure = 104 psi
Bearing tolerances:

+0.0002	−0.0003	+0.0003	−0.0003
+0.0002	+0.0002	+0.0001	−0.0002
+0.0001	+0.0001	−0.0001	+0.0004
−0.0003	−0.0004	−0.0002	+0.0001
+0.0001	+0.0001	−0.0004	−0.0005

Determine the controller's optimal adjustment intervals for both temperature and pressure. If the two adjustment intervals are different, can you replace them by one interval? If so, what are the quality losses per bearing?

7.10 A military ammunition plant produces hand grenades. The quality control system for the grenades involves inspecting grenade caps and testing parts for dimensional accuracy and cracks. The dimensional accuracy of the parts is checked by using linear variable differential transducers (LVDTs). As soon as the grenade body arrives at the inspection station, the LVDTs move from their reference positions to touch the part. The amount of travel is electronically compared to preprogrammed limits in the microprocessor. All LVDTs return to their reference positions after each measurement is taken. Sensors detect an LVDT's failure to return to its reference position, indicating that the tolerances of the dimension being measured have exceeded their limits.

After the check for dimensional accuracy, the cap is inspected for proper installation. If a cap is missing or improperly installed, a limit switch is tripped, which in turn directs the grenade to a reject conveyor. The final inspection of the grenade involves the detection of cracks in the grenade's shell. This is done in the following way: After the grenade is positioned it is spun 360° in proximity to a pair of eddy-current crack transducers. If a crack exists, indicating a defective grenade, the transducers detect the eddy currents. Cracked grenades are automatically transferred to an eject conveyor.

Thus, the inspection station for the grenade conducts three types of inspections: dimension accuracy, which is a variable case; proper installation of the cap, which is an attribute case; and crack existence, which is also an attribute case. The parameters of these processes are as follows:

Dimensions accuracy:

$$l = 3 \text{ parts}$$

$$A = \$18.00$$

$$B = \$1.00$$

$$\overline{u} = 100 \text{ parts}$$

$$\text{Tolerance of the dimensions} = \pm 0.002 \text{ in}$$

$$C = \$25.00$$

Cap installation:

$$l = 2 \text{ parts}$$

$$A = \$25.00$$

$$B = \$0.50$$

$$\overline{u} = 200 \text{ parts}$$

$$C = \$40.00$$

Crack detection:

$$l = 5 \text{ parts}$$

$$A = \$29.00$$

$$B = \$3.00$$

$$C = \$20.00$$

$$\overline{u} = 80 \text{ parts}$$

What are the optimal parameters of the inspection system for each of the above characteristics?

Given the following information for a new system capable of all the functions of the three current systems, what are the optimal parameters of the new system?

$$l = 1$$

$$A = \$26.00$$

$$B = \$4.00$$

$$C = \$80.00$$

Compare the new system with the existing ones.

REFERENCES

Berry, S. A. "Techniques in the Application of Computers to Industrial Monitoring." In *CAD/CAM and the Computer Revolution*, Society of Manufacturing Engineers, 1974, pp. 221–239.

Besterfield, D. H. *Quality Control*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

Death, M. "Sensors: Keys to Automation." *Manufacturing Engineering*, June 1986, pp. 53–57.

Farago, F. T. *Handbook of Dimensional Measurement*. New York: Industrial Press, 1982.

Feigenbaum, A. V. *Total Quality Control*. New York: McGraw-Hill, 1983.

Kirkpatrick, E. G. *Quality Control for Managers and Engineers*. New York: John Wiley and Sons, 1970.

Peters, M. H., and W. W. Williams. "Location of Quality Inspection Stations, an Experimental Assessment of Five Normative Heuristics." *Decision Sciences*, vol. 15, 1984, pp. 389–408.

Pryor, T. R., and W. North, eds. *Applying Automated Inspection*. Dearborn, MI: Society of Manufacturing Engineering, 1985.

Stout, K. *Quality Control in Automation*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

Taguchi, G. *On-line Quality Control during Production*. Tokyo: Japanese Standards Association, 1981.

CHAPTER 8

INTRODUCTION TO PREVENTIVE MAINTENANCE

In Chaps. 4 through 7, we discussed on-line real-time quality control systems during production. We also showed that there are many factors that affect the quality of products during production. One of these factors is the production machine itself. The production machine's accuracy and breakdown rate may directly affect product quality. From the viewpoint of the machine producer, a poorly functioning machine has a quality problem.

Similarly, in a communication system where the quality level of information, measured by the signal-to-noise ratio, is poor, an on-line quality control system may be required. There is no great conceptual difference in the quality control systems applied to communication and production systems.

In this chapter, the effect of preventive maintenance on the product characteristics is discussed, and optimal preventive maintenance schedules that minimize the quality loss per unit of production are introduced. Relationships between tolerances of the product characteristics and optimal preventive maintenance schedules are also introduced.

Preventive maintenance involves repairs, replacement, and maintenance of equipment and products before their failures in order to avoid unexpected failures during their use. The objective of preventive maintenance is to minimize the downtime of equipment. However, excessive preventive maintenance results in unnecessary repair and maintenance costs. Therefore, an optimal preventive

maintenance schedule exists that minimizes the total cost of repair and downtime of equipment.

Preventive maintenance as it affects on-line quality control systems may involve two areas of applications. The first is the quality control of characteristics of the products. The second is the reduction of the expected failures of the machine during the production operation. A machine may fail in that it is not able to produce products that meet the quality requirements. A machine failure may also be a sudden breakdown of the machine during operation. Failures of either type can be reduced by employing preventive maintenance schedules. This chapter focuses on the development of preventive maintenance schedules for the first type of machine failures. Preventive maintenance schedules for the second type are discussed in detail in the reliability, replacements, and maintenance literature (e.g., Jardine, 1973).

The purpose of the following discussion is to develop a relationship between functional limit values and the average time consumed before a production process produces products having deviations equal to the upper or lower functional limits.

Assume that the loss due to a product failure (when deviations of the product characteristics are greater than the functional limits) is proportional to the amount of downtime of the process.

Let t = average amount of downtime

P = loss due to product failure per unit of time

c = repair cost of the process

u = average time between product failures (mean time between failures, MTBF)

The total loss per failure, C^*, is

$$C^* = Pt + c \tag{8.1}$$

The total loss of product failure per unit time, assuming there is no preventive maintenance, is

$$L = \frac{C^*}{u^*} \tag{8.2}$$

where u^* is the long-term average time between product failures.

As presented in Chap. 3, there are two cases of product deviations from the target value. First, the deviations (drifts) can be on both sides of the target value. Second, the deviations can only be on one side of the target value, such as the wear of mechanical parts.

In the first case, we can use a random walk model as an approximation to determine the average time to reach the upper or lower limit of a product characteristic, at which point failure will occur. When using the random walk model, the average time u^* is assumed to be proportional to the square of distance between the target value and the upper or lower limits (the functional limit, Δ^*, is equal for both upper and lower limits of the target value). Thus, u^* is

$$u^* = \lambda(\Delta^*)^2 \tag{8.3}$$

where λ is the constant of proportionality, determined by observing the average time u_0 to reach a predetermined deviation Δ_0. Thus, the value of u^* can be obtained using Eq. 8.3 as follows:

$$u^* = u_0 \frac{(\Delta^*)^2}{\Delta_0^2} \tag{8.4}$$

In the second case, where the average time to reach the functional limit Δ^* is proportional to the amount of deviation from the target value, u^* is

$$u^* = u_0 \frac{\Delta^*}{\Delta_0} \tag{8.5}$$

In the following sections, we discuss the use of preventive maintenance schedules in reducing the loss caused by the deviations in product characteristics from the target values. Maintenance schedules will be determined for products when deviations occur on one or both sides of the target values.

8.1 PREVENTIVE MAINTENANCE SCHEDULES: DEVIATIONS ON BOTH SIDES OF TARGET VALUES

In this section, we consider the occurrence of deviations on both sides of the target values. The preventive maintenance schedules discussed below apply when a product experiences deviations caused by extended use of the operations. The schedules also apply when the parameters of a production machine deviate from their target values, thus causing deviations in quality characteristics of the products being produced. The preventive maintenance schedules that will be discussed below are based on the rule: *Perform preventive maintenance when the amount of deviation in the product characteristic being monitored reaches* Δ. We define the following parameters:

Δ^* = functional limit of a product or of a parameter of the production machine

C^* = loss per failure due to deviations greater than Δ^*

u^* = average time between product (or machine) failures when no preventive maintenance is performed

n = checking interval (to check the amount of deviation)

B = checking cost

Δ = preventive maintenance limit (the amount of deviation or drift at which preventive maintenance should be performed)

C = preventive maintenance cost

u = average preventive maintenance interval

l = time lag of the production process, if any

n_0 = current (or initial) checking interval

Δ_0 = current (or initial) maintenance limit

u_0 = current average preventive maintenance interval

Following the derivation of Eq. 4.12, we derive the loss function of the current maintenance schedule as follows:

$$L_0 = \frac{B}{n_0} + \frac{C}{u_0} + \frac{C^*}{u^*} \times \frac{1}{(\Delta^*)^2}\left[\frac{\Delta^2}{3} + \left(\frac{n}{2} + l\right)\frac{\Delta^2}{u}\right] \tag{8.6}$$

The first two terms of the above equation are straightforward. The third term represents the loss caused by the use of the current preventive maintenance schedule and is derived in the following way:

Assume that the time required for the parameter of the production process to reach Δ is proportional to the squared distance from the target value, and that the time required to reach the maintenance limit is

$$u_0 \frac{\Delta^2}{\Delta_0^2}$$

We also assume that the parameter of the production machine being measured follows a uniform distribution within the range $\pm\Delta$. No preventive maintenance is done as long as the value of the parameter is within this range. The mean squared deviation of the parameter of the production machine within $\pm\Delta$ is

$$\frac{[(m + \Delta) - (m - \Delta)]^2}{12} = \frac{\Delta^2}{3} \tag{8.7}$$

Since the parameter of the production machine is checked at intervals of n units of time, then the average time the parameter is outside the control limit Δ is $n/2$. When the time lag l is considered (not negligible), then the average time the parameter is outside the maintenance limit Δ is $(n/2 + l)$. So, the mean squared deviation becomes

$$\frac{\Delta^2}{3} + \left(\frac{n}{2} + l\right)\frac{\Delta^2}{u} \tag{8.8}$$

When Δ in Eq. 8.8 becomes Δ^*, failures will occur at intervals of u^*, causing a loss C^*. Thus, the expected loss as a result of preventive maintenance (using a checking interval n and a preventive maintenance limit Δ) is

$$\frac{C^*}{u^*} \times \frac{1}{(\Delta^*)^2}\left[\frac{\Delta^2}{3} + \left(\frac{n}{2} + l\right)\frac{\Delta^2}{u}\right] \tag{8.9}$$

The following example illustrates how the checking interval, the loss function, and the optimal preventive maintenance interval are determined for a microwave transmission system.

Example 8.1. A microwave transmission system is checked once every three months. The system consists of five smaller systems, and a sixth one, to be used when one of the five primary systems fails. Average repair frequency of the system is once a year, with an average repair time of 3 hours. During repair, another system may fail, causing a transmission failure and a loss of $P' = 5600$ (channels per system) × \$1.5. The transmission failure rate is one failure per minute. So the expected loss P when failure lasts for 1 hour is

$$P = 5600 \times \$1.5 \times 0.25(\text{working ratio}) \times 60(\text{min}) \times \frac{5}{365 \times 24}$$
$$= \$71.9 \qquad (8.10)$$

The cost of checking a system is \$60. The current control (maintenance) limit is 39.5 dBm (decibel margin) against functional failure limit 12.6 dBm. Maintenance cost is \$220, and the average maintenance interval observed is once a year. Determine the optimal checking interval n and optimal limit D, assuming that the parameters of system are

$$\Delta^2 = 12.6 \text{ dBm} = 10^{1.26} = 18.2$$
$$P = \$71.9$$
$$B = \$60$$
$$C = \$220$$
$$D_0^2 = 39.5 \text{ dBm} = 10^{3.95} = 8913$$
$$n_0 = 90 \text{ days} \times 24 \text{ hr/day} = 2160 \text{ hr}$$
$$u_0 = 365 \text{ days} \times 24 \text{ hr/day} = 8760 \text{ hr}$$

Solution. Since the principle "the larger the margin, the better the system" applies here, the the current loss function is

$$L = \frac{B}{n_0} + \frac{C}{u_0} + P\Delta^2\left(\frac{1}{3D_0^2} + \frac{n_0 + 1}{2} \times \frac{1}{u_0 D_0^2}\right)$$
$$= \frac{60}{2160} + \frac{220}{8760} + 71.9 \times 18.2\left(\frac{1}{3 \times 8913} + \frac{2161}{2} \times \frac{1}{8760 \times 8913}\right)$$
$$= 0.028 + 0.025 + 0.049 + 0.018$$
$$= \$0.120 \qquad (8.11)$$

This represents quality loss per hour. When there are 1500 stations, each having 15 channels, total loss per year becomes

$$\$0.120 \times 24 \times 365 \times 15 \times 1500 = \$23.7 \text{ million}$$

The optimal checking interval n and optimal maintenance limit D are

$$n = \sqrt{\frac{2u_0 B}{P} \times \frac{D_0^2}{\Delta^2}}$$
$$= \sqrt{\frac{2 \times 8760 \times 60 \times 8913}{71.9 \times 18.2}}$$
$$= 2678 \rightarrow \text{once in four months}$$

$$D_2 = \left(\frac{P}{3C} \times u_0 \times D_0^2 \times \Delta^2\right)^{1/2} \qquad (8.12)$$
$$= \left(\frac{71.9}{3 \times 220} \times 8760 \times 8193 \times 18.2\right)^{1/2}$$
$$= 11{,}926 \rightarrow 40.8 \text{ dBm} \qquad (8.13)$$

The loss function for optimal maintenance system, L, is then

$$L = \frac{B}{n} + \frac{C}{u} + P\Delta^2\left(\frac{1}{3D^2} + \frac{n+1}{2} \times \frac{1}{uD^2}\right) \qquad (8.14)$$

and using

$$u = u_0 \times \frac{D_0^2}{D^2}$$
$$= 8760 \times \frac{8913}{11926}$$
$$= 6547 \text{ hr} \qquad (8.15)$$

the following is obtained:

$$L = \frac{60}{120 \times 24} + \frac{220}{6547} + 71.9 \times 18.2\left(\frac{1}{3 \times 12{,}022} + \frac{2881}{2} \times \frac{1}{6547 \times 12{,}022}\right)$$
$$= 0.021 + 0.034 + 0.036 + 0.024$$
$$= 0.115 \qquad (8.16)$$

The improvement \$0.005 is not much, nor is the current maintenance system near optimal. However, the yearly gain is expected to be

$$0.005 \times 365 \times 24 \times 15 \times 1500 = \$985{,}500$$

Example 8.2. Assume that the temperature of the cooling water in an atomic power station causes a major failure when it deviates from the nominal value by as much as 150°C, and that the corresponding loss is \$30 million. The current control limit Δ_0 is 20°C, and the average maintenance interval u_0 is once every 15 days (360 hours). The temperature is checked once every 8 hours at a cost of \$6. The average preventive maintenance cost for adjusting the control system is \$400. The effect of time lag is almost nil. Calculate the total loss of the current system.

Solution. The predicted average time between failures (assuming there is no preventive maintenance) is obtained by using Eq. 8.4.

$$u^* = u_0 \times \frac{(\Delta^*)^2}{\Delta_0^2}$$

$$= 360 \times \frac{150^2}{20^2}$$

$$= 20,250 \text{ hr}$$

The parameters for Eq. 8.6 are

$$B = \$6$$
$$n_0 = 8 \text{ hr}$$
$$C = \$400$$
$$u_0 = 360 \text{ hr}$$
$$C^* = \$30,000,000$$
$$u^* = 20,250 \text{ hr}$$
$$\Delta^* = 150°\text{C}$$
$$l = 0$$

Substituting the above values in Eq. 8.6, we obtain the following total loss per hour.

$$L_0 = \frac{B}{n_0} + \frac{C}{u_0} + \frac{C^*}{u^*} \times \frac{1}{(\Delta^*)^2}\left[\frac{\Delta_0^2}{3} + \left(\frac{n_0}{2} + l\right)\frac{\Delta_0^2}{u_0}\right]$$

$$= \frac{6}{8} + \frac{400}{360} + \frac{30,000,000}{20,250} \times \frac{1}{(150)^2}\left[\frac{20^2}{3} + \left(\frac{8}{2} + 0\right)\frac{20^2}{360}\right]$$

$$= 0.75 + 1.11 + 8.78 + 0.29$$

$$= \$10.93 \tag{8.17}$$

It is apparent in the above loss function that the first term is greater than the fourth term, indicating that more checking is being performed than what optimally should be. Also, the second term is much smaller than the third, implying that narrower maintenance limits should be used. The optimal checking interval and the optimal maintenance limit are obtained as follows:

Following Eq. 8.6, the loss function expression is given below.

$$L = \frac{B}{n} + \frac{\Delta_0^2 C}{u_0 \Delta^2} + \frac{C^*}{u^*} \times \frac{1}{(\Delta^*)^2}\left[\frac{\Delta^2}{3} + \left(\frac{n}{2} + l\right)\frac{\Delta_0^2}{u_0}\right] \tag{8.18}$$

The optimal n and Δ can be obtained by taking the partial derivatives of Eq. 8.18 with respect to n and Δ, and equating the results to zero. Thus,

$$\frac{\partial L}{\partial n} = -\frac{B}{n^2} + \frac{C^*}{u^*} \times \frac{1}{(\Delta^*)^2} \times \frac{(\Delta^*)^2}{2u^*} = 0$$

whence

$$n = u^* \sqrt{\frac{2B}{C^*}} \tag{8.19}$$

and

$$\frac{\partial L}{\partial \Delta} = -\frac{2(\Delta^*)^2 C}{u^* \Delta^3} + \frac{C^*}{u^*} \times \frac{1}{(\Delta^*)^2} \times \frac{2}{3}\Delta = 0$$

giving

$$\Delta = \Delta^* \left(\frac{3C}{C^*}\right)^{1/4} \tag{8.20}$$

Example 8.3. What are the optimal checking interval and preventive maintenance limit of the control system described in Example 8.2?

Solution. Using the parameters given in Example 8.2, and Eqs. 8.19 and 8.20, we obtain the optimal checking interval

$$n = u^* \sqrt{\frac{2B}{C^*}}$$

$$= 20,250 \sqrt{\frac{2 \times 6}{30,000,000}}$$

$$\approx 12 \text{ hr (twice per day)}$$

The optimal preventive maintenance interval is

$$\Delta = \Delta^* \left(\frac{3C}{C^*}\right)^{1/4}$$

$$= 150\left(\frac{3 \times 400}{30,000,000}\right)^{1/4} \approx 12°\text{C}$$

The total loss after the implementation of the optimal checking interval and the optimal preventive maintenance interval is obtained as follows:

$$u = u_0 \times \frac{\Delta^2}{\Delta_0^2}$$

$$= 360 \times \frac{12^2}{20^2} = 129.6$$

and

$$L = \frac{6}{12} + \frac{400}{129.6} + \frac{30,000,000}{20,250} \times \frac{1}{150^2}\left[\frac{12^2}{3} + \left(\frac{12}{2} + 0\right)\frac{12^2}{129.6}\right]$$

$$= 0.5 + 3.09 + 3.16 + 0.44 \tag{8.21}$$

$$L = \$7.19$$

Annual savings when the optimal checking and preventive maintenance intervals are implemented (assuming there are 6700 hours per year) are

$$(10.93 - 7.19) \times 6700 \approx \$25{,}000$$

8.2 PREVENTIVE MAINTENANCE SCHEDULES FOR FUNCTIONAL CHARACTERISTICS

In many instances, the tolerance for a functional characteristic of a product is expressed as a percentage of the functional target value. In this situation, the preventive maintenance limit can also be expressed as a percentage of the functional target value, and the optimal preventive maintenance schedules can be determined using the same equations given in the previous section.

Example 8.4. The characteristic y of a unit used in airplanes is immediately checked before every takeoff at a cost B of \$2.50. Should the value of y deviate by 60 percent of its nominal value m, the airplane will have a 0.01 probability of a failure that would cost \$180 million. The current maintenance limit is 10-percent of m and the repair (preventive maintenance) cost is \$800, including the loss caused by a delayed takeoff. The average maintenance interval is once every 120 takeoffs, and the average flight time is 6 hours. Time lag between testing and repairing the defective unit is negligible. Determine the optimal checking interval, optimal preventive maintenance limit, and the total loss per unit time.

Solution. The parameters of this system are

Functional limit Δ^*	$\pm 60\%\ m$
Loss due to failure C^*	\$1.8 million
Current control limit Δ_0	10%
Current checking interval n_0	6 hr
Current average maintenance interval u_0	$6 \times 120 = 720$ hr
Checking cost B	\$2.50
Maintenance cost C	\$800
Time lag l	0

The loss of the current preventive maintenance is obtained using Eq. 8.6:

$$L = \frac{B}{n_0} + \frac{C}{u_0} + \frac{C^*}{u^*} \times \frac{1}{(\Delta^*)^2}\left[\frac{\Delta_0^2}{3} + \left(\frac{n_0}{2} + l\right)\frac{\Delta_0^2}{u_0}\right] \tag{8.22}$$

where u^* is the expected mean time to airplane failure when no maintenance is used; u^* can be determined as follows:

(1) When y deviates on both sides of the target value,

$$u^* = u_0 \frac{(\Delta^*)^2}{\Delta_0^2} \tag{8.23}$$

(2) When y deviates only on one side of m,

$$u^* = u_0 \frac{\Delta^*}{\Delta_0} \tag{8.24}$$

In this example, we assume that deviations occur on both sides of m. Thus,

$$u^* = 720 \times \left(\frac{60}{10}\right)^2$$

$$= 25{,}920 \text{ hr}$$

Substitution in Eq. 8.22 results in:

$$L = \frac{2.50}{6} + \frac{800}{720} + \frac{1{,}800{,}000}{25{,}920} \times \frac{1}{60^2}\left[\frac{10^2}{3} + \left(\frac{6}{2} + 0\right) \times \frac{10^2}{720}\right]$$

$$= 0.417 + 1.111 + 0.643 + 0.008$$

$$= \$2.179$$

The optimal checking interval n and the optimal control limit Δ are

$$n = u^* \sqrt{\frac{2B}{C^*}}$$

$$= 25{,}920 \sqrt{\frac{2 \times 2.5}{1{,}800{,}000}}$$

$$= 43.2 \text{ (once every 7 flights)}$$

and

$$\Delta = \Delta^* \left(\frac{3C}{C^*}\right)^{1/4}$$

$$= 60\left(\frac{3 \times 800}{1{,}800{,}000}\right)^{1/4}$$

$$\approx 12\ (\%)$$

The total loss corresponding to the optimal maintenance schedule is

$$L = \frac{2.50}{42} + \frac{800}{5184} + \frac{1{,}800{,}000}{25{,}920} \times \frac{1}{60^2}\left[\frac{12^2}{3} + \left(\frac{42}{3} + 0\right) \times \frac{12^2}{5184}\right]$$

$$= 0.059 + 0.154 + 0.926 + 0.008$$

$$= \$1.147$$

Assuming the total number of airplanes is 200, and each airplane makes 600 flights yearly, then the annual savings resulting from the implementation of the optimal preventive schedule are

$$(2.179 - 1.147) \times 600 \times 6 \times 200 = \$743{,}040$$

When the functional limits of a product characteristic are equal and are expressed as a percentage of the functional target value, the procedure outlined in Example 8.4 can be used to obtain the preventive maintenance schedules for the production machine that produces the unit. On the other hand, when a product's functional limits are not at equal distances from the target value of the product characteristic (e.g., the functional limits are $m + \Delta_1$ and $m - \Delta_2$), two preventive maintenance checking intervals are determined independently for Δ_1 and Δ_2. The preventive maintenance schedule for the production machine is then based on the minimum of the preventive maintenance checking intervals for Δ_1 and Δ_2.

Example 8.5. The internal gear of the planetary reduction gears for a marine medium-speed diesel engine is 3 m in diameter. The gear is produced by a hopping machine with extreme accuracy. Because of noise and wear-out factors, the tolerances of the diameter are +0.2 mm and −0.3 mm. The checking interval is 20 hours, and the cost of checking the gear is $300. If the diameter of the gear deviates from the nominal diameter by +0.2 or −0.3 mm, the gear may cause failure of the system where it will be placed, resulting in a loss of $200,000 when Δ_1 is exceeded or $300,000 when Δ_2 is exceeded. The current maintenance limit for the hopping machine is ±0.05 mm, and the repair and adjustment cost is $250. The average maintenance interval is 400 hours. Neglecting the time lag of the checking process, determine the optimal preventive maintenance control limit for the hopping machine and the optimal preventive maintenance schedule.

Solution. The parameters of the system are

Functional limit Δ_1^*	+0.2 mm
Functional limit Δ_2^*	−0.3 mm
Loss due to failure when Δ_1^* is exceeded, C_1^*	$200,000
Loss due to failure when Δ_2^* is exceeded, C_2^*	$300,000
Current control limits Δ_0	±0.05 mm
Current checking interval n_0	20 hr
Checking cost B	$300
Current maintenance interval u_0	400 hr
Maintenance cost C	$250
Time lag l	0

In order to determine the optimal checking interval and the optimal preventive maintenance limit, we first determine the interval and the limit for Δ_1^* and Δ_2^* independently, as follows:

Preventive maintenance for Δ_1^*:

$$L_1 = \frac{B}{n_0} + \frac{C}{u_0} + \frac{C_1^*}{u_1^*}\frac{1}{(\Delta_1^*)^2}\left[\frac{\Delta_0^2}{3} + \left(\frac{n_0}{2} + l\right)\frac{\Delta_0^2}{u_0}\right]$$

Since the characteristic deviates on one side by Δ_1,

$$u_1^* = u_0\frac{\Delta_1^*}{\Delta_0}$$

$$u_1^* = 400 \times \frac{0.2}{0.05} = 1800 \text{ hr}$$

$$L_1 = \frac{300}{20} + \frac{250}{400} + \frac{200{,}000}{1600} \times \frac{1}{(0.2)^2}\left[\frac{(0.05)^2}{3} + \left(\frac{20}{2} + 0\right) \times \frac{(0.05)^2}{400}\right]$$
$$= 15.62 + 2.78$$
$$= \$18.40$$

The optimal checking interval n_1 and the optimal control limit Δ_1 are

$$n_1 = u_1^*\sqrt{\frac{2B}{C_1^*}}$$
$$= 1600\sqrt{\frac{2 \times 300}{200{,}000}}$$
$$= 87.6 \text{ hr}$$

$$\Delta_1 = \Delta_1^*\left(\frac{3C}{C_1^*}\right)^{1/4}$$
$$= 0.2\left(\frac{3 \times 250}{200{,}000}\right)^{1/4}$$
$$\Delta_1 = 0.0494 \text{ mm}$$

Preventive maintenance for Δ_2^*:

$$u_2^* = u_0\frac{\Delta_2^*}{\Delta_0}$$

$$u_2^* = 400 \times \frac{0.3}{0.05} = 2400 \text{ hr}$$

The loss for the current control system is

$$L_2 = \frac{300}{20} + \frac{250}{400} + \frac{300{,}000}{2400} \times \frac{1}{(0.3)^2}\left[\frac{(0.05)^2}{3} + \left(\frac{20}{2} + 0\right) \times \frac{(0.05)^2}{400}\right]$$
$$= 15.62 + 1.24$$
$$= \$16.86$$

The optimal checking interval n_2 and the optimal control limit Δ_2 are

$$n_2 = u_2^*\sqrt{\frac{2B}{C_2^*}}$$
$$= 2400\sqrt{\frac{2 \times 300}{300{,}000}}$$

$$= 107.3 \text{ hr}$$

and

$$\Delta_2 = \Delta_2^*\left(\frac{3C}{C_2^*}\right)^{1/4}$$

$$= 0.3\left(\frac{3 \times 250}{300{,}000}\right)^{1/4}$$

$$= 0.067$$

The optimal control limit for the hopping machine is ± 0.0494 mm, and the optimal preventive maintenance interval is $\min\{87.6, 107.3\} = 87.6$ hours.

8.3 PREVENTIVE MAINTENANCE SCHEDULES FOR LARGE SCALE SYSTEMS

In this section, we shall develop preventive maintenance schedules for large-scale systems, such as telephone communication networks, power transmission lines, and manufacturing systems. Consider, for example, the case where a preventive maintenance schedule is to be developed for electric power lines used for pantographs (devices for transferring current from an overhead wire to a vehicle, usually consisting of two parallel, hinged, double-diamond frames). This electric power wire is subject to wear. If the wear is even all over the wire, then only periodic replacements are needed for the wire. However, due to many external factors, such as temperature change from one location to another, and from time to time, the wear rate is nonuniform over the wire length. Therefore, periodic checking of the wire is needed in conjunction with a preventive maintenance schedule, which is developed as follows.

Suppose that the optimal periodic maintenance interval is found to be 3 years, and that the total loss L' is given by

$$L' = \text{preventive replacement cost} + \text{loss due to wire breakdown} \tag{8.25}$$

Let us also assume that the replacement cost of the total length of 350 miles of electric wire used for a bullet train is \$22 million, and the average loss per wire breakage is \$2.5 million. The average number of wire failures during a three-year period of operation is

$$\frac{22}{2.5} = 8.8 \tag{8.26}$$

(assuming that maintenance cost and quality loss are balanced at nominal wearing rate). If the actual number of wire failures is one or two during the three-year life period of the wire, it becomes apparent that the wire is "over-designed," and another wire should be used instead.

Assume that the starting thickness of the wire is m and the allowed amount of wear during three years is 30 mm. The average wear rate b is

$$b = \frac{\text{estimated wear during three years}}{\text{total operating hours during three years}}$$

$$b = \frac{30}{3 \times 365 \times 18 \text{ (hours per day)}} = 0.001522 \text{ mm/hr} \tag{8.27}$$

Let Δ^* be the functional amount of wear; that is, when the value of the thickness y becomes $m - \Delta^*$, wire breakage occurs. The expected wire breakage per mile per hour, N, is

$$N = \frac{8.8}{350 \times 3 \times 365 \times 18} = 0.000001275 \text{ failures/hr} \cdot \text{mi}$$

Thus, the average time between failures (assuming constant failure rate), u^*, is

$$u^* = \frac{1}{N} \approx 784{,}000 \text{ hr}$$

Preventive maintenance is achieved by first checking the wear of the wire at different locations. The following 20 observations of wear (mm) are taken:

0.03	0.20	0.07	0.23	0.06
0.13	0.22	0.01	0.37	0.15
0.20	0.16	0.02	0.14	0.03
0.02	0.07	0.45	0.23	0.24

The mean rate of wear is 0.152 mm, and its standard deviation is 0.120 mm. Because the value of the standard deviation is comparable to the mean, the use of periodic replacements will only result in high quality loss (many failures will occur between replacements).

If the wire is checked once every 18 hours at a cost of \$0.5 per mile, preventive maintenance is not done, and periodic maintenance is done at three-year intervals, the loss function for the entire length of the wire (after deleting maintenance terms from Eq. 8.6 and substituting $\Delta^2/u = (\Delta^*)^2/u^*$) becomes

$$L = \frac{B}{n} + \frac{C^*}{u^*}\left(\frac{n}{2} + l\right)\frac{350}{u^*} \tag{8.28}$$

The minimization of Eq. 8.28 with respect to n yields

$$n = u^*\sqrt{\frac{2B}{350 \times C^*}} \tag{8.29}$$

or

$$n = 784{,}000\sqrt{\frac{2 \times 0.5}{350 \times 2{,}500{,}000}}$$

$$= 26$$

Thus, the optimal checking interval is approximately once a day. The loss function of the current checking and maintenance system, L_0, is

$$L_0 = \frac{B}{n_0} + \frac{C^*}{u^*}\left(\frac{n_0}{2} + l\right)\frac{350}{u^*}$$

$$= \frac{0.5}{18} + \frac{2,500,000}{784,000}\left(\frac{18}{2} + 9\right)\frac{350}{784,000}$$

$$= \$0.0278 + 0.0256$$

$$= \$0.0278 + \$0.0256$$

$$= \$0.0533 \text{ per mile}$$

The annual savings resulting from the implementation of the optimal checking interval are nil.

8.4 SUMMARY

This chapter introduced the concept of employing preventive maintenance schedules to improve the quality of products. It also introduced optimal preventive maintenance schedules that minimize the quality loss per unit of production. Relationships were developed between the tolerances of product characteristics and optimal preventive maintenance schedules. Finally, optimal checking schedules for large-scale systems were offered to help minimize total quality loss.

PROBLEMS

8.1 One of the manufacturing processes of printed circuit boards (PCBs) is solder masking, performed by an automatic mask-dispensing machine. This keeps solder out of the areas where hardware is attached. The thickness of the solder mask is critical. If it is too thick, it can crack or seep through the board material into holes that should be open. Thin or uneven application may allow solder to fill the open areas. The nominal value of the solder thickness is 2 mm, and the loss caused by deviation from the nominal value is \$120. The current control limit Δ_0 is 1.5 mm, and the average maintenance interval of the mask-dispensing machine is once every 5 days (120 hours). The solder thickness is checked once every 5 hours at a cost of \$10, and the average preventive maintenance cost to adjust the machine is \$50. Time lag is not considered. What are the optimal checking interval, the preventive maintenance limit, and the loss per unit time? If you were to design a different masking system, what type of parameters would you consider?

8.2 A coordinate-measuring machine (CMM) is used to measure the diameter of gold-coated shafts. The coat thickness is $m \pm 0.02$ in. Assume that the coating process produces coated shafts with diameters within the specification limits; however, the coordinate measuring machine may reject products because of measurement errors. The CMM is checked every 20 hours for the amount of measurement error by measuring a standard gauge block. If the amount of errors is equal to or greater than

0.0001 in, preventive maintenance is performed to eliminate the sources of error. The cost of checking the CMM is \$20, and the preventive maintenance cost is \$120. The current preventive maintenance limit is 0.00005 in, and the average maintenance interval is 200 hours. Determine the optimal preventive maintenance schedule of the CMM. Define new methods to improve this measuring system.

8.3 A manufacturing facility for electroplating processes uses an independent electric power source, since regulated and filtered power is an important element that contributes to the uniformity of the electroplating process. The electric cables (50 miles long) that provide power for the process are checked for wear once every 90 days. The maximum allowable wear of the cable diameter is 10 mm in a four-year period. The replacement cost of the total cable system is \$5 million. Cable failure will result in a loss of \$1.9 million. There are 18 operating hours per day. The following 20 observations are taken of the wear (per 100 hours of operations):

0.02	0.03	0.14	0.17	0.10
0.06	0.14	0.07	0.19	0.04
0.09	0.12	0.06	0.17	0.14
0.13	0.14	0.17	0.06	0.15

The cables are checked once every 2 days at a cost of \$0.40 per mile. Determine the optimal checking intervals (in hours) and the loss per hour for both current and optimal maintenance policies.

8.4 Solve Prob. 8.2, assuming that the current preventive maintenance limit is 0.0001 in and the average maintenance interval is 250 hours.

8.5 Solve Prob. 8.3, assuming the cost of checking is \$0.50 per mile and a cable failure will result in a loss of \$3.5 million.

8.6 An optical transmission system is used for transmitting electrical signals via an optical fiber. Its components are an electro-optic transducer as the light transmitter at the beginning of the route, the actual fiber optic transmission route, and the opto-electric transducer as the light receiver at the end of the route. The quality of transmission is affected by the uniformity of the diameter of the optical fiber cable. The nominal value of the diameter is 1 mm, and the loss caused by deviation from the nominal value is \$100 per 1000 feet of cable length. The current control limit, Δ_0, is 0.98 mm, and the average maintenance interval of the drawing machine (it draws the cable to the required specifications) is 10 days (24 hours per day). The diameter of the cable is checked once every 4 hours at a cost of \$4, and the average time lag of the system is 2 minutes (production rate is 10,000 feet per hour). The cost of machine adjustment is \$70.00. What is the preventive maintenance cost for the drawing machine that results in minimal quality losses? What are the optimal parameters of the system such as checking interval, preventive maintenance limit, and the loss per unit time?

8.7 Solve Prob.8.6, assuming a maximum allowable quality loss per unit of \$0.10 per foot,

REFERENCES

Berry, S. A. "Techniques in the Application of Computers to Industrial Monitoring." *In CAD/CAM and the Computer Revolution*, Society of Manufacturing Engineers, 1974, pp. 221–239.

Jardine, A. K. S. *Maintenance, Replacement, and Reliability*. New York: John Wiley and Sons, 1973.

"Robotic PCB Masking Gives Fast ROI, Doubles Production." *Manufacturing Engineering*, vol. 98, April 1987, p. 57.

Stout, K. *Quality Control in Automation*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

Taguchi, G. *On-line Quality Control during Production*. Tokyo: Japanese Standards Association, 1981.

APPENDIX A

AREAS UNDER THE NORMAL CURVE

Proportion of the total area under the curve that is under the portion from $-\infty$ to $(X_i - \overline{X}')/\sigma'$. ($X_i$ represents any desired value of the variable X.)

$\frac{X_i - \bar{X}'}{\sigma'}$	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
−3.5	0.00017	0.00017	0.00018	0.00019	0.00019	0.00020	0.00021	0.00022	0.00022	0.00023
−3.4	0.00024	0.00025	0.00026	0.00027	0.00028	0.00029	0.00030	0.00031	0.00033	0.00034
−3.3	0.00035	0.00036	0.00038	0.00039	0.00040	0.00042	0.00043	0.00045	0.00047	0.00048
−3.2	0.00050	0.00052	0.00054	0.00056	0.00058	0.00060	0.00062	0.00064	0.00066	0.00069
−3.1	0.00071	0.00074	0.00076	0.00079	0.00082	0.00085	0.00087	0.00090	0.00094	0.00097
−3.0	0.00100	0.00104	0.00107	0.00111	0.00114	0.00118	0.00122	0.00126	0.00131	0.00135
−2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0017	0.0018	0.0019
−2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
−2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
−2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
−2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
−2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
−2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
−2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
−2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
−2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
−1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
−1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
−1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446
−1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
−1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
−1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
−1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
−1.2	0.0985	0.1003	0.1020	0.1038	0.1057	0.1075	0.1093	0.1112	0.1131	0.1151
−1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357
−1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
−0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
−0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
−0.7	0.2148	0.2177	0.2207	0.2236	0.2266	0.2297	0.2327	0.2358	0.2389	0.2420
−0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
−0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.0305	0.3085
−0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
−0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821
−0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
−0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
−0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000

*Reprinted from "Statistical Quality Control," E. L. Grant and R. S. Leavenworth, 1980, by permission from McGraw-Hill Book Company.

$\frac{X_i - \bar{X}'}{\sigma'}$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
+0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
+0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
+0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
+0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
+0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
+0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
+0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
+0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8079	0.8106	0.8133
+0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
+1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
+1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
+1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
+1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
+1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
+1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9384	0.9406	0.9418	0.9429	0.9441
+1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
+1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
+1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
+1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
+2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
+2.1	0.9821	0.9826	0.9830	0.0834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
+2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
+2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
+2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
+2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
+2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
+2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
+2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
+2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
+3.0	0.99865	0.99689	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
+3.1	0.99903	0.99906	0.99910	0.99913	0.99915	0.99918	0.99921	0.99924	0.99926	0.99929
+3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
+3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
+3.4	0.99966	0.99967	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
+3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983

APPENDIX B

FEED-FORWARD CONTROL

Let $\mathbf{y} = (y_1, y_2, \ldots, y_h)$ be an objective characteristic vector having target $\mathbf{y}_0 = (y_{10}, y_{20}, \ldots y_{h0})$, $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ be given conditions which affect output $\mathbf{y}$, and $\mathbf{z} = (z_1, z_2, \ldots, z_l)$ be a vector of process parameters.

For example, in the steel industry, quality characteristics of iron ore and other materials are observed. The quality characteristics of iron will be affected by the quality of raw materials, including ore. When material conditions are at nominal, standard process conditions are used to produce iron to meet target $\mathbf{y}_0$. When the material conditions are not at nominal, the quality of iron will deviate from nominal. In order to minimize the deviation, process parameters $\mathbf{z}$ are changed by solving the following equation:

$$f(\mathbf{z}, \mathbf{x}) = \mathbf{y}_0 \tag{B.1}$$

where
- $\mathbf{x}$ = material conditions
- $\mathbf{z}$ = process parameters to adjust in order to satisfy the above equation
- $f(\mathbf{z}, \mathbf{x})$ = prediction function of $\mathbf{y}$

Using Eq. B.1, y_i's can be predicted. Denote the difference between predictions and nominal values by

$$y_{ip} - y_{i0} \qquad i = 1, \ldots, h \tag{B.2}$$

where y_{ip} is the predicted value obtained for process condition z, and material condition $\mathbf{x}$, and y_{i0} is the target value of the ith quality characteristic. The feed-forward quality control focuses on the search for the right adjustment

of $\mathbf{z}$ from $\mathbf{z}_0$, the nominal values of process parameters, in order to minimize the loss. That is,

$$\underset{\mathbf{z}}{\text{Min}} \sum_{i=1}^{h} \frac{A_i}{\Delta_i^2}\left[\beta(y_{ip} - y_{i0})\right]^2 \tag{B.3}$$

where y_{ip} is the predicted y_i using $f(\mathbf{z}, \mathbf{x})$, and β, is the shrinkage or damping factor, commonly used in control engineering.

SHRINKAGE FACTOR. Suppose y is the true (unknown) value of a characteristic having target y_0, and suppose y_p is the prediction with an error variance σ_p^2. How much should this characteristic be adjusted? If the adjustment is as much as

$$-(y_p - y_0) \tag{B.4}$$

then error variance σ^2 after adjustment is

$$\sigma^2 = E[\underbrace{y - (y_p - y_0)}_{\text{True value after adjustment}} - y_0]^2 = \sigma_p^2 \tag{B.5}$$

If adjustment is made with a shrinkage factor β, then

$$\sigma^2 = E[\underbrace{y - \beta(y_p - y_0)}_{\text{True value after adjustment}} - y_0]^2 = (1-\beta)^2(y - y_0)^2 + \beta^2\sigma_p^2 \tag{B.6}$$

Here it is assumed that y_p has the error variance σ_p^2 around the true value of y.

Minimizing mean square error σ^2 in Eq. (B.6) with respect to β:

$$\beta = \frac{(y - y_0)^2}{(y - y_0)^2 + \sigma_p^2} = 1 - \frac{\sigma_p^2}{\sigma_p + (y - y_0)^2} = 1 - \frac{1}{F_0} \tag{B.7}$$

where

$$F_0 = \frac{\sigma_p^2 + (y - y_0)^2}{\sigma_p^2} \tag{B.8}$$

This means the adjustment of y should be as much as

$$-\beta(y_p - y_0) \tag{B.9}$$

where $\beta = 1 - 1/F_0$. Note that F_0 is always ≥ 1, and hence $0 \leq \beta \leq 1$.

However, since the true value of y is not known, F_0 has to be estimated. Suppose y_p is an unbiased estimation of y having error variance,

$$E(y_p - y_0)^2 = (y - y_0)^2 + \sigma_p^2 \tag{B.10}$$

Then $(y_p - y_0)^2$ may be used as an estimator of $\sigma_p^2 + (y - y_0)^2$, and

$$F_0 \approx \frac{(y_p - y_0)^2}{\sigma_p^2} \tag{B.11}$$

In the case of a vector, we have to calculate

$$F_{0i} = \frac{(y_{ip} - y_{i0})^2}{\sigma_{pi}^2} \quad (i = 1, 2, \ldots, h) \tag{B.12}$$

and we take

$$\beta_i = \begin{cases} 0 & \text{when } F_{0i} \leq 1 \\ 1 - \dfrac{1}{F_{0i}} & \text{when } F_{0i} > 1 \end{cases} \tag{B.13}$$

Note that it is possible for the estimated F_{0i} to be less than 1; but since the true value must be ≥ 1, we take $\beta_i = 0$ in such a case.

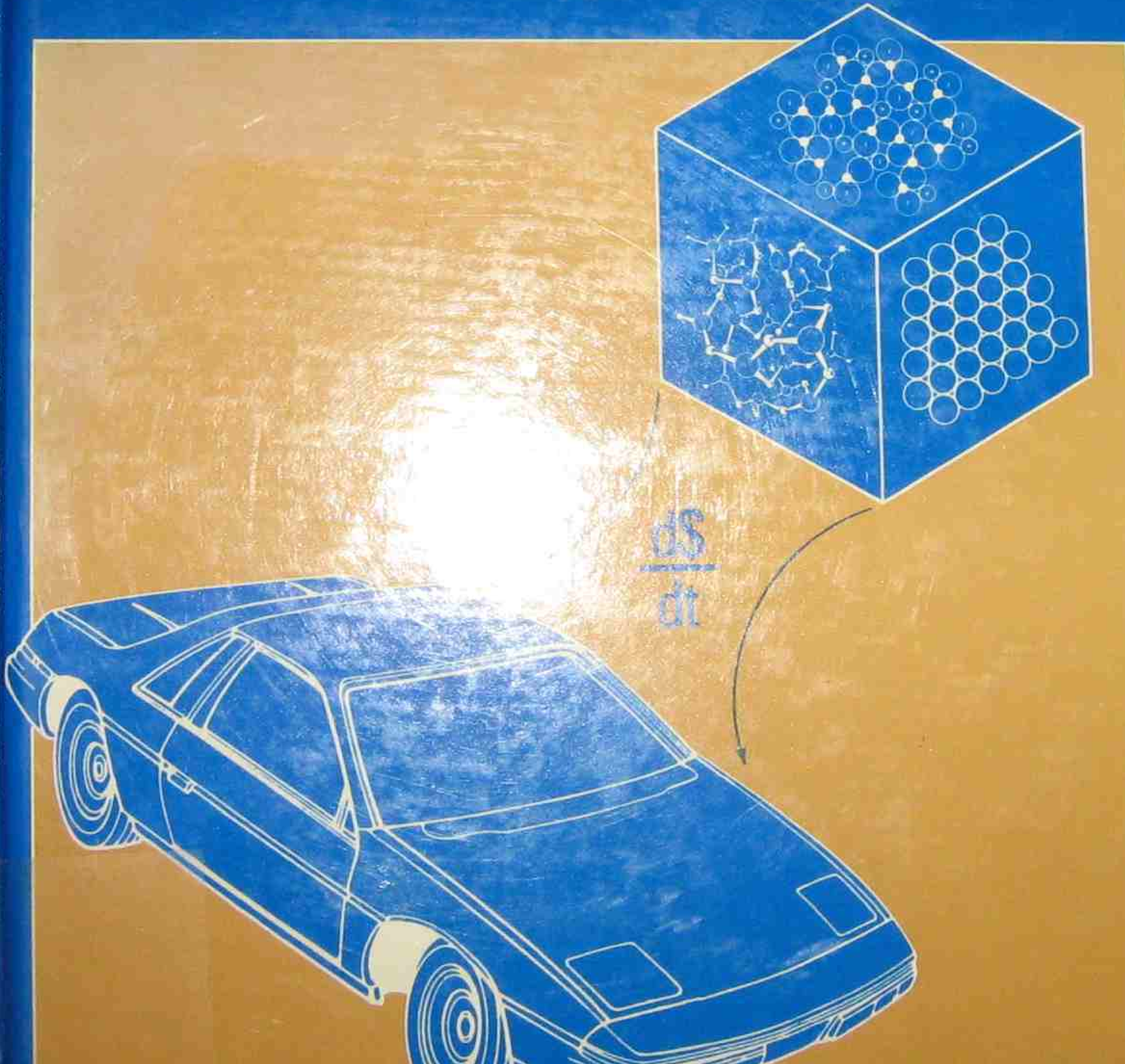
INDEX

ISBN 0-07-062830-0
90000
9 780070 628304

MANUFACTURING ENGINEERING

Economics and Processes

Kenneth C Ludema   Robert M. Caddell   Anthony G. Atkins



manufacturing engineering: economics and processes

KENNETH C LUDEMA

Professor of Mechanical Engineering and Applied Mechanics
The University of Michigan

ROBERT M. CADDELL

Professor of Mechanical Engineering and Applied Mechanics
The University of Michigan

ANTHONY G. ATKINS

Professor of Mechanical Engineering
University of Reading, United Kingdom

Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632

Library of Congress Cataloging-in-Publication Data

Ludema, K. C.
Manufacturing engineering.

Includes bibliographies and index.
1. Production engineering. I. Caddell, Robert M.
II. Atkins, Anthony G., . III. Title.
TS176.L84 1987 670.42 86-21282
ISBN 0-13-555582-5

Table 3-1
Used with permission of The American Society of Mechanical Engineers from "Surface Texture" (USASI-B46.1-1986)

Figure 4-3
From K. H. Moltrecht and R. M. Caddell, "How to Determine Production Tolerances–Part 1," *The Tool Engineer* (Oct. 1957), pp. 81–85. Reprinted courtesy of The Society of Manufacturing Engineers.

Continued, page 405

A division of Simon & Schuster
Englewood Cliffs, New Jersey 07632

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Cover design: George Cornell
Manufacturing buyer: Rhett Conklin

ISBN 0-13-555582-5 025

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Prentice-Hall of Southeast Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

contents

preface

Because of the diversity of disciplines involved in what is called manufacturing, it is not possible to give a simple definition of a ''manufacturing engineer.'' Yet, in our opinion, a manufacturing engineer is more concerned with the design of manufacturing systems than with their operation, per se. Depending upon the academic program in which courses related to manufacturing are offered, certain topics are presented in some depth, while others are ignored completely. Although subjects such as business practices, classical economics, and the like all address content that falls under the broad umbrella of manufacturing, our concern in this book addresses, primarily, subjects that are of an engineering nature. Historically, such coverage has usually involved the materials and processes used to manufacture products; in most instances a detailed description of hardware and a qualitative approach to problem solving were considered. Although useful, this approach does not address the fundamental aspects of the individual processing methods, nor does it pose the kind of problem solving that typifies most other undergraduate engineering subjects. Probably because of these reasons, manufacturing courses are often held in low esteem by those whose specialties are narrower in scope but more analytical in nature. This is truly ironic, since to analyze any of the major processes in depth demands the use of a number of those specialties. Such intensive study of any individual process must be left to those specialized texts that provide an in-depth coverage of a given process such as casting, machining, and the like. Our intent is to address the major principles of processing methods plus certain other topics of concern that engineers face in manufacturing industries, and to include *numerical* problem solving wherever possible. Any essential descriptive coverage is kept to a necessary minimum. We note that this field is currently laced with acronyms and catchy phrases, but acronyms come and go, while fundamental concepts remain.

This book is divided into three major sections, each of which carries a brief foreword. Part A includes some philosophical aspects of manufacturing concerns and stresses the importance of economic considerations. The latter does not require any formal background in business procedures or classical economics per se. A concise coverage of surface finish and tolerances, which find importance in all processes, is also included here. Part B starts with a review of mechanical properties and engineering materials and then uses this background as it applies to the four major and traditional processing methods. It concludes with a chapter on what has come to be called nontraditional or special processing techniques. Part C includes a number of types of case studies in which the principal goal is to study the economics of producing parts by different manufacturing methods. It concludes with some ideas related to integrated design and manufacturing.

Since courses of this type are usually offered in the latter part of typical engineering curricula, we assume that users of this text will have earlier covered the topics of statics, strength of materials, mathematics, and a traditional course in materials science. If at all possible, we suggest that self-paced computer instruction packages be available to students for review of topics concerning the above. TV tapes and, possibly, plant trips can be used to demonstrate many of the hardware aspects of processes and systems, thereby negating the need for extensive descriptive coverage during classroom lectures.

With regard to units, a few comments are pertinent. Although most students are familiar with the SI System, much of industry in the United States has not followed this usage. Even the traditional metric system has met with much resistance. It is likely that a number of graduates who enter a manufacturing industry will encounter situations where the units involved may be SI, metric, or English. For that reason we have interspersed different units throughout this text with the hope that the reader will feel comfortable regardless of which system is posed. Following this preface is a table of useful conversion factors that relate to the subject matter in this text.

To give equal and uniform coverage to all chapters in a typical one-semester course is unlikely. Depending upon which engineering department offers such a course, instructors may decide to extend the material in certain chapters while reducing or even skipping the content of others. Naturally, we think it best to address all topics, but time constraints may make this too difficult. In our experience mechanical engineers seem to think that the business and human factors of manufacturing are not their concern, whereas industrial engineers tend to regard processes and equipment as things that can be handled via catalogs and handbooks. Our hope is that this text will show each group that each attitude fulfills only a part of the complete picture.

1. Industrial engineering students usually have a strong background in statistics and economics but have a lesser exposure to mechanics and materials. Chapters 1 through 5 and selected parts of Chapters 13 and 14 may find greatest interest, and can be extended as an instructor sees fit. If less emphasis is to be placed upon processing methods the most logical groupings are Chapters 6, 8, and 9, *or* 7, 10, and 11.
2. Metallurgical or material engineers, who have a deeper background in materials, could consider a brief review of Chapters 1 to 5 and then proceed with detailed

coverage of Chapters 6 to 12; in fact, an extension of Chapters 7, 10 and 11 might be considered. Selected topics from Chapters 13 and 14 may then be covered.
3. Mechanical engineers, who have a deeper background in mechanics, should have at least a reasonable exposure to Chapters 1 and 2, and a definite coverage of Chapters 3, 4, and 5. Then Chapters 6 through 12 should be addressed in full, followed by selected portions from Chapters 13 and 14.

A number of individuals have provided information and suggestions to us as this book was being compiled; some was quite recent, while some was accrued over years of contact. Professors W. G. Ovens and W. R. DeVries reviewed the original manuscript and made a number of suggestions and constructive criticisms, almost all of which we have incorporated into the final version. We are grateful for their help. Our close association with Professor W. F. Hosford shows clearly in Chapter 8. The late Professor L. V. Colwell was our friend and colleague, and a teacher to two of us. A bit of what he taught us makes up much of Chapters 4 and 9, and we are grateful to have known him for so many years. A group of people supplied much information that has been included in Chapters 13 and 14, and we thank them collectively for their help. They include Professors J. M. Alexander, J. A. G. Kals, W. A. Knight, K. Lange, and G. W. Rowe, and Drs. A. Beevers, B. Lengyel, and C. Ruiz.

Kenneth C Ludema
Robert M. Caddell
Anthony G. Atkins

SOME BASIC UNITS AND THEIR ABBREVIATIONS FOR THE SI SYSTEM

Unit	Standard	Abbreviation
length	meter	m
mass	kilogram	kg
time	second	s
*force	newton	N = kg m/s^2
*stress	newton/meter^2	N/m^2
*stress	pascal = 1 N/m^2	Pa
*energy	joule	J = Nm

* These are derived from basic units.

MULTIPLICATION FACTORS USED IN THE SI SYSTEM

Factor	Prefix	Symbol	Factor	Prefix	Symbol
10^{18}	exa	E	10^{-3}	milli	m
10^{15}	peta	P	10^{-6}	micro	μ
10^{12}	tera	T	10^{-9}	nano	n
10^{9}	giga	G	10^{-12}	pico	p
10^{6}	mega	M	10^{-15}	femto	f
10^{3}	kilo	k	10^{-18}	atto	a

USEFUL CONVERSION FACTORS—ENGLISH TO SI UNITS

To convert from	to	Multiply by
inch (in.)	meter (m)	2.54×10^{-2}
feet (ft)	meter	3.048×10^{-1}
inch^2	meter^2	6.452×10^{-4}
feet^2	meter^2	9.29×10^{-2}
inch^3	meter^3	1.639×10^{-5}
feet^3	meter^3	2.832×10^{-2}
pound-force (lbf)	newton (N)	4.448
pounds/inch^2 (psi)	pascal (Pa = N/m^2)	6.895×10^{3}
kilopounds/inch^2 (ksi)	pascal	6.895×10^{6}
horsepower (hp)	foot-pounds/minute	33×10^{3}
horsepower	watts (W)	7.457×10^{2}
foot-pound (ft-lbf)	joules (J)	1.356

SOME OTHER USEFUL RELATIONS

1 micron = 10^{-6} meter
1 angstrom = 10^{-10} meter
10 angstroms = 1 nm

MANUFACTURING ENGINEERING:
Economics and Processes

FOREWORD TO PART A

The manufacture of consumer products is a dynamic profession. Winning or losing may involve "shaving pennies" from the cost of a product, and this requires much ingenuity. A slight redesign of the part, the choice of a better material, the use of a processing method that produces less scrap, and a more efficient arrangement of machinery are some of the considerations that can lead to lower costs of production. Aspects of these topics are addressed in Chapter 1.

Decisions in designing both products and manufacturing systems are really interdependent, and they have one common, vital relationship. This has to do with economic advantage. Chapter 2 covers some of the important factors that engineers should consider, and presents the importance of the time value of money. Engineers need not know business or accounting procedures to gain an appreciation of this topic.

Chapters 3 through 5 cover topics that are vital in manufacturing but often catch engineers by surprise. In Chapter 3 ideas on product quality attributes are discussed; although of great importance, the defining and measuring of "quality" are often considered to be the responsibility of "someone else." Certain aspects of tolerances are presented in Chapter 4. Again this is sometimes viewed as a problem that is handled by others, but both quality and assembly must be considered in the design of a manufacturing system. In Chapter 5, the use of automation and computers is introduced. This section is intended to illustrate, in a somewhat philosophical way, topics of importance to engineers who are involved with the production of consumer products—topics that are often viewed as being the sole responsibility of others. We disagree with such an attitude, since all designed products must be processed from selected materials, and the manufacturing processing of said materials is the gate through which all designs must pass before products reach the consumer. The production of quality items at a competitive price is of prime importance to manufacturing engineers.

1 engineering in the manufacturing industry

1.1 INTRODUCTION

Engineering is the art of distilling relevant scientific knowledge for the purpose of bringing useful items, materials, or services into being. The engineer seeks to quantify all of the variables that are germane to a project, makes informed estimates where hard data are not available, and sets down a logical progression of thought toward solving problems, in terms and language that others can readily understand and use as a basis for action. Obviously, engineers are important people.

Engineers in manufacturing industries have an added responsibility, and that is to include economics in their calculations. Many engineers are spared of economic concerns, such as those in government, in academia, and in the junior positions in design and analysis. Certainly everyone must be concerned about available funds for projects (or pay), but relatively few engineers need to give as much thought to company profits as do engineers who design manufacturing systems.

Manufacturing is, after all, done to make a profit. In fact, in planning the manufacture of products, there should be a prospect of higher profit than could be realized from investing the capital needed for the manufacturing facilities into stocks, for example. Only upper management has the authority to decide what level of profits are worth an investment, but they depend on others to provide the information for decision making. Manufacturing engineers are the major source of information on manufacturing costs.

A manufacturing enterprise can be very complex, involving many people of various skills. It may be surprising to know that many companies employ no professional engineers at all. The expertise of skilled tradesmen and hard-working managers is often sufficient to keep a company profitable, particularly if the products have a long *life cycle*.

However, some products require continual updating, which may involve mathematical skills (airplane design), materials expertise (jet engines and oil drill bits), chemical skills (plastics and oil refining), or computer skills (adaptive control and information systems) in their design or manufacturing. These are the skills in which engineers are proficient to varying degrees.

In the same way, manufacturing systems must be designed, and this often requires engineers. The design of a manufacturing system would seem to be merely the selection and arrangement of commercially available machines, conveyors, robots, and computers. Certainly a part of the work of manufacturing engineers is just that. However, success in business is not ensured by using ideas and items available to everyone. Innovation provides the vital margin, and engineers are the authors of most technical innovation.

Traditionally engineers concentrated on the hardware of products and processes; some have been directly involved in the manufacturing of products. A new and vital area for engineers is the design or specifying of *manufacturing information systems*. Manufacturing of products is done by the coordination of the efforts of a great number of people and machines. However, the system is prone to error, some of which are human in origin and others not. Each error is a threat to profitability. The minimizing of errors has usually been done by organizing people so that everything that is necessary to know is known by the proper people. It may be seen that there is a great amount of information concerning the operation, beginning with the design of the product and ending with the rate of sales of products, and it is widely distributed. Computers provide the capability of automating the transfer of such information, and engineers design these automated systems.

The sections following describe the manufacturing industry and the engineering challenges in it. There follows a description of the major new approach to the design of manufacturing systems, namely, to include the transfer of design information along with the design of the hardware of the system. Finally, the role of the computer in designing the systems and in the operation thereof will be discussed.

1.2 THE MANUFACTURING INDUSTRY

Manufacturing constitutes between 30 percent and 70 percent of the gross national product (GNP) in the United States, depending on what is defined as manufacturing. We can readily identify *manufactured* products. Both a nail and an automobile are manufactured. But the act of manufacturing requires defining. For our purposes it is the conversion of raw materials into *hardware*, both completed products and components thereof. This definition excludes the conversion of crude oil into gasoline, because gasoline is not hardware. It also excludes building houses, because in the building process there is little or no conversion of raw materials. The building of automobiles is not only a manufacturing operation, but a significant assembly activity as well. In fact, most large automotive companies are a large mixture of manufacturing and assembly operations. They may cast engines in iron, purchase sheet metal which they form into shapes for body panels, and purchase tires which they mount on wheels with no conversion at all.

Frequently manufacturing is an action defined as producing *discrete* parts, which is different from continuous processing as in making gasoline. But wire making is also a continuous process, in that the product is often miles long, hardly a discrete part in the usual sense of the term. Of course, we could argue that the making of gasoline and wire are continuous processes only while a batch of raw materials lasts, whereupon the process variables should be changed because two batches of raw material are seldom exactly alike.

Most manufacturing organizations begin with an idea for a product to sell for a profit. Someone finds the money to begin, and begins to make and market a product. Some manufacturers remain small, but others continually offer new products and become very large and diverse. Both types succeed and both types fail. Success depends on such mundane matters as having the right products and satisfying consumers. At times it would appear that the large corporation with its professional polish and elegantly appointed offices might be insulated from concern for consumers, but they cannot be.

New products must be suggested and reviewed regularly and continuously for a company to survive. The sequence of events and activities by which products come into being and enter the marketplace varies considerably according to the product and individual company practice. The concept or idea for a new or updated item (or product, or machine) may originate from any of many places, such as from market analysis groups, salespersons, individual entrepreneurs outside of industry, or from an engineer within a plant. Promising ideas are assessed by small groups of experienced people, often in a new product development department or equivalent. Most ideas are rejected for a variety of reasons. A few ideas, the ''sure winners'', are sent on to design groups and perhaps produced. Some of those sell well, and some do not.

The handling of the new design beyond the initial exploration stage varies by industry and company, but we may differentiate between two major classes of products, namely *capital* products and *consumer* products. Capital products include large airplanes, railroad cars, machine tools, cash registers, and other items and machines used in conducting commerce or business. The major concern in the making of capital items is that the item should perform to some well-defined specification. The buyer usually wants some guarantee that the item will not only perform a given task but will do so, over a specified period of time, below some specified cost; this includes the costs to buy, operate, maintain, and dispose of at the end of use.

Consumer products are treated very differently. Examples are lawn mowers, floor covering, automobiles, toasters, shoes, and small personal computers. With these items the major concern is first cost. Performance and life-cycle cost, though they should not be misrepresented, are not as clearly communicated from the manufacturer to the end user as in the case of capital items.

Capital products and consumer products are rarely made in the same factory. If one company makes both types of products, they are usually made in different divisions at least. The reason is that each type of product requires very different organization. The basis for the success of capital products is consistent quality and responsive service. This requires an emphasis of design and development. Capital products do not change model or style very much. They are usually made from a relatively small range of materials, by a staff of skilled craftsmen, and often with very specialized machinery.

Consumer products require good design and development as well, but competitive advantage is gained by efficient material use, material handling, and manufacturing processes. Consumer products are usually easier to make than are capital products, and thus more manufacturers will compete in the consumer market. The manufacturing system must then be both physically and economically flexible, and the people designing the systems must understand the entire system.

An outline of the steps required to design a manufacturing system for consumer products is shown in Figure 1–1 in terms of the evolution from the idea for a product to the preparation of a factory to make the product. The steps will vary according to industry and product, but the urgency to plan carefully is the same for all. An ill-conceived product line is a merciless sinkhole for money. Every step in the planning process must assure the feasibility of making a large profit (or minimizing losses in some cases). Figure 1–1 shows that the decision sequence involves the joint effort of people of many disciplines. The time dimension cannot be shown in the figure, because each product and each organization places different emphases on different steps. Indeed, the time dimension often depends on relationships between people and depends on the effectiveness in transferring information among them. The personal aspect is often minimized by engineers, but it is important. In a successful industry one finds a great deal of harmony and personal understanding among and between individuals in the sequence shown. Acrimony between groups and a confused organization structure tend to lengthen and garble the lines of communication, thus slowing the process and allowing errors to survive in the system. Losing sports teams often have the same problem.

Beside the purely human aspect of information transfer, the technical content of design information must also be efficiently transferred. Since a design document (for example, blueprint and specifications) contains information provided by people of very diverse technical skills and backgrounds, all participants in the design must understand each other adequately. This often requires a major shift in thinking from college, where engineering students learn how to specialize, and ignore broader areas of study.

It should be noted that Figure 1–1 shows close communication between engineers and business specialists. The reason is that many decisions in manufacturing are based on economic and business considerations. But many business decisions depend on technical input as well. Engineers usually avoid courses in economics and business, and business students rarely take courses in engineering. Thus, at the start of their careers there is little common means of communication between them. Yet they must communicate. Long-term prosperity in the manufacturing industry requires good communication, and when an industry is prosperous, its valuable employees also prosper.

Figure 1–1 can be viewed as an information flow network. Engineers do not usually think of their handiwork as an element in a flow of information. The engineer calculates and designs for the purpose of solving problems, and submits his conclusions in a report to his supervisor. Traditionally, designers acted as if their responsibility were ended when the design was first committed to paper, that is, the blueprint. The next group to deal with the design was usually the manufacturing group, whose responsibility was to make whatever was designed so that a third group could market it. There was little interaction between designers and manufacturers. The separation was *organic*, or planned organiza-

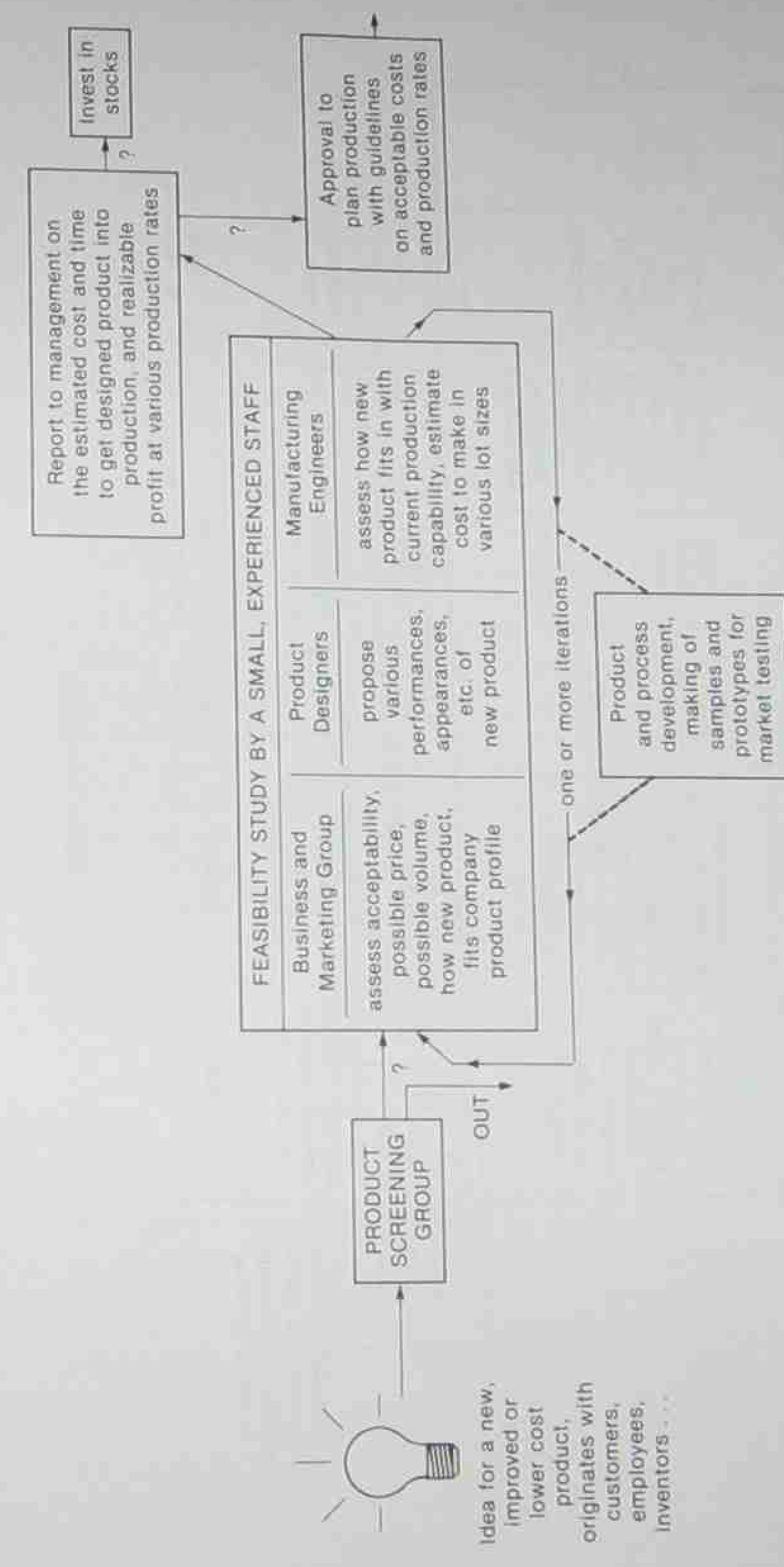


Figure 1–1 Chart of the steps in bringing product concepts into reality via engineering, business and marketing specialists, and labor.

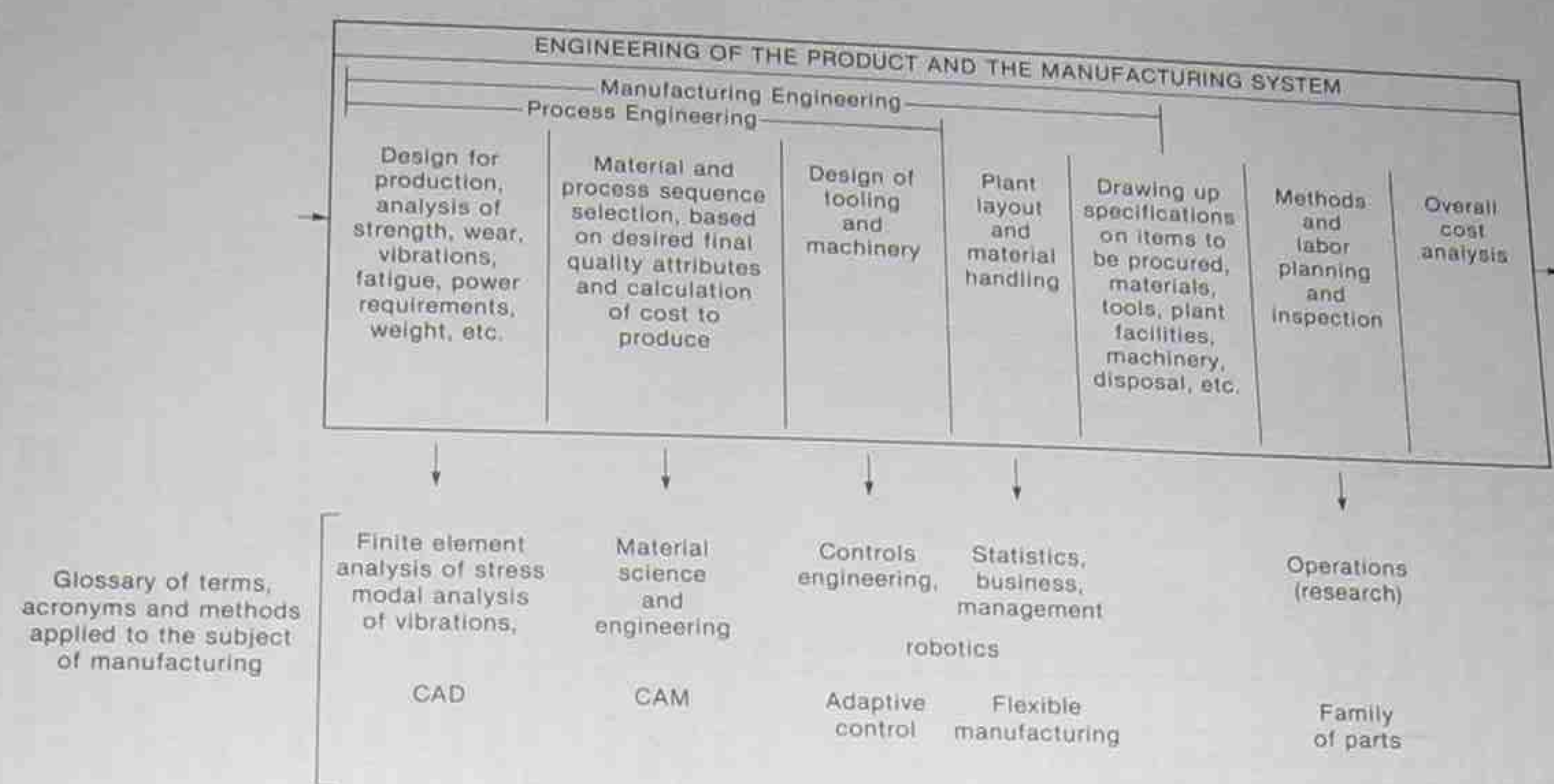


Figure 1–1 (con't.)

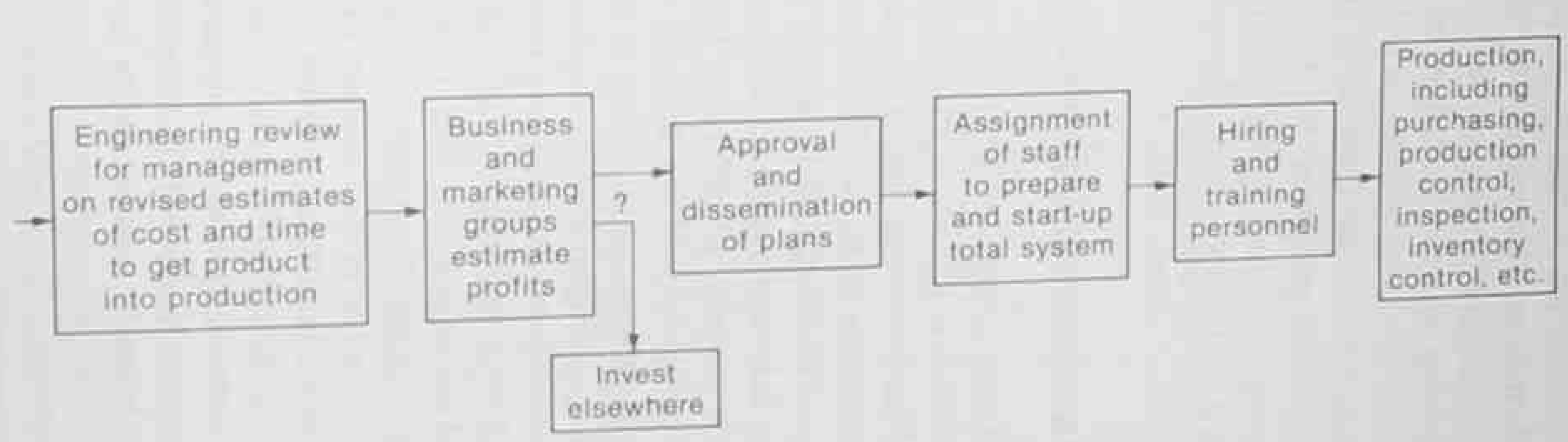


Figure 1–1 (con't.)

tionally by placing each group under a different vice president of the company. The result was a system that could not respond well to changes in the markets. Such companies fade from view, particularly during economic recessions. The modern view is that organizational walls should not be built, or conversely, all specialists who can contribute to the design of a product and manufacturing system are given free and orderly access to each other and to all relevant information. This means that each specialist must learn at least the basic terms and concepts of all other specialists, and each specialist must communicate his own conclusions and thinking in terms that others are likely to understand. Thus it is very important to regard ones work as an element in a flow of information. The flow of design information is as important to the survival of the company as is the flow of materials through the manufacturing processes. For that purpose, information, like materials, must be more completely *characterized* and its flow automated. This topic will be discussed at greater length in a later chapter. First, the technical role of engineers will be discussed.

1.3 ENGINEERING IN THE MANUFACTURING INDUSTRY

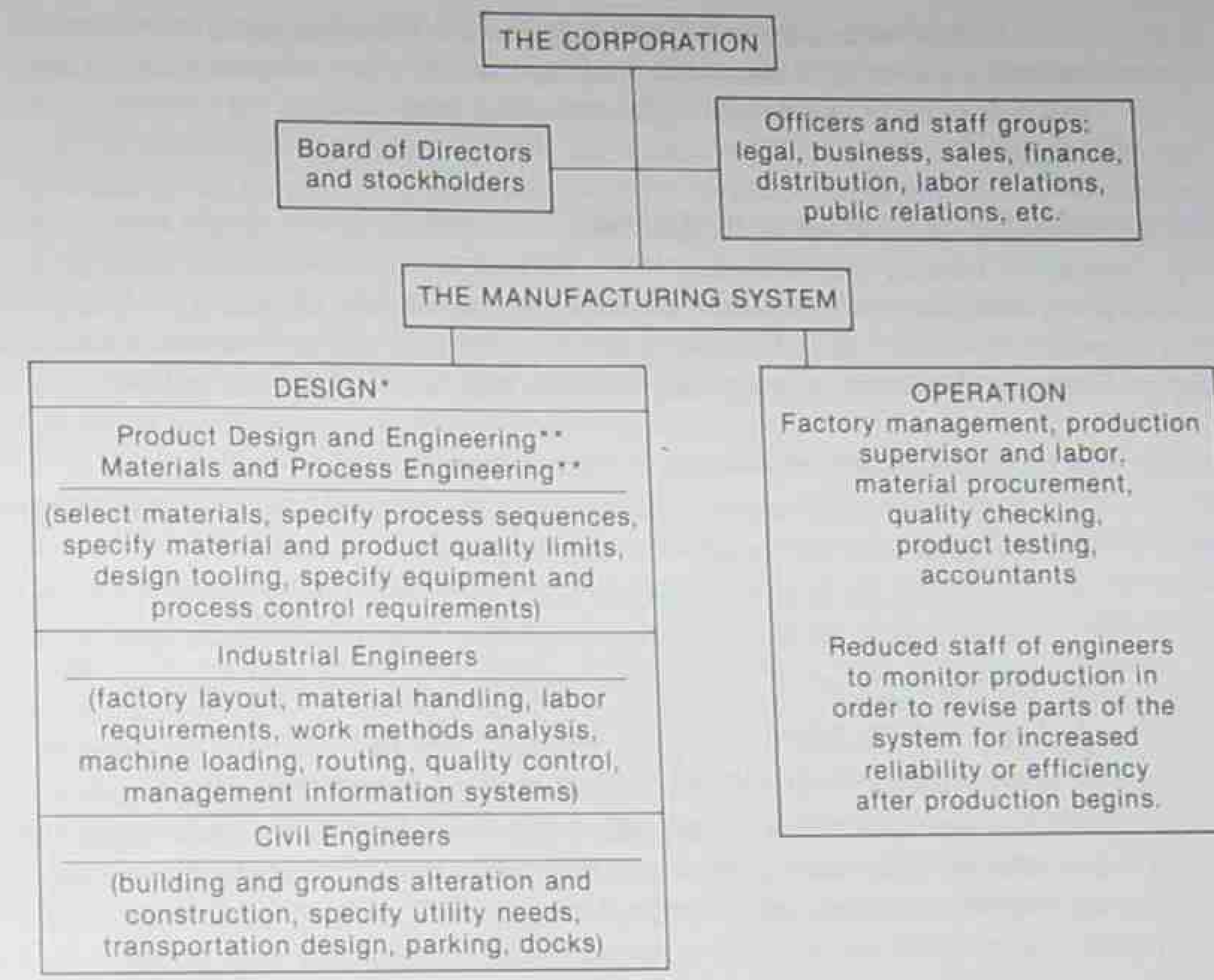
The manufacturing industry is very large, but a relatively small fraction of people in these industries are engineers. The reason is that many products do not need engineers to design or to make them. Clever and hard-working people can and do accomplish many things without engineers, and will continue to do so. Doubtless many products are made by companies that employ engineers but do not really need or use the capability of engineers. Doubtless too, there are others that need engineers but have not yet recognized the need.

The need for engineers is probably related to the complexity of the manufacturing operation as much as to the complexity of the product. In fact, we may differentiate between the product and the manufacturing operation as to the engineering need. Complicated products, such as an x–y plotter, require engineers for the design but skilled machinists to make them. The design of machinery to make shoes requires engineering, but the design of the shoe itself does not. The curious practice in many industries is that the design group is often called the Engineering Department. Most other engineers are in the Manufacturing Department or Plant Construction Department.

To place the work of engineers in the context of other people it is useful to distinguish between *operating* a manufacturing facility on the one hand, and *designing the product and designing the manufacturing system* on the other. Figure 1–2 illustrates the difference in expertise needed within the manufacturing system. The diagram shows the management staff to one side, but only to keep the illustration simple. It shows that the major engineering challenges in manufacturing are in the following areas:

1. Product design and engineering
2. Material and process engineering
3. Industrial engineering

The prominence of each of these groups depends strongly on the products produced. In fact, whether engineering will be prominent at all in a particular company may also



*Any or all of these functions could be contracted to outside companies, particularly Civil Engineering.
**The technical background of these groups depends on the product. Some products may require electrical or mechanical or computer skills above others.

Figure 1–2 Diagram of specialties required in manufacturing.

depend on business cycles. It is common to find that a company with a good product in a stable market decreases its engineering activity after some time. This has the benefit to a company of decreasing both the payroll and other expenses associated with product development. This may be followed by other steps to increase profit, such as to use raw materials of lower cost and to shortcut some of the steps in production. This tactic gains considerable credit for upper management, but only for a short time. Eventually product quality will decline, or not keep pace with the competition, and sales will decline or warranty costs increase very quickly. Management will call upon engineers for new ideas, but the few remaining engineers will probably be overwhelmed with emergency problems. Very likely the most competent person(s) to solve a particular problem will have been urged into early retirement, or some modern tools for problem solving will have not been purchased.

If the company survives such folly, a later folly that sometimes occurs is to give excessive authority to engineers, to design and produce a high-quality product—which will cost more to make than can be recovered from sales.

1.4 COMPUTERS IN MANUFACTURING

Programmable computers and electronic processors are becoming widely used in the manufacturing industry. Computers stand in a long line of developments that have changed industry. Historians of technology cite previous developments to include:

1. The use of animal power to replace human power.
2. The use of water wheels and steam engines to allow larger-scale and higher-speed operation than is possible with animal power.
3. The use of machinery to generate electricity, with which forces can be multiplied, energy can be widely distributed, and high temperature can be generated.
4. The discovery of oil and the development of the chemical industry.
5. Labor unions, which raised the standard of living of workers.
6. Automation, which allowed mass production of consumer goods.
7. Computers, which, when applied to process and inventory control, allow manufacturers to make short production runs economically in order to make several options available in many consumer products.

So far, the complete potential of computers has not yet been realized. Thus there is a great amount of exploration and research under way in this field, which will doubtless spawn several new engineering activities relevant to the manufacturing industry.

The presence of computers as *aids* in design and manufacturing has popularized the acronyms CAD (computer-aided design), CAM (computer-aided manufacturing) and CAE (computer-aided engineering). The simplest meaning of the wide use of acronyms is that the computer has become more useful than previous aids to engineers. CAD usually refers to automated drafting and a set of software packages for force and stress analysis, dynamic modeling, modal (vibration) analysis, and other aids to designers. CAD does not usually include that vital and creative step in the design activity which converts a concept for a product into the representation of feasible parts and shapes on paper or other storage media.

The most visible aid rendered by computers to the manufacturing enterprise is in controlling the motions of various parts of process machinery. Since the electronic means for effecting mechanical control is by way of (a number of) pulses sent to stepping motors (or equivalent), control by microprocessors has become known as *numerical control* or NC. The NC machines are particularly effective for medium and small batch sizes of products where labor costs in *job shop* scale of manufacturing are too high and high capital cost of *transfer machine* for *mass production* is unwarranted. Older NC machines are programmed to make tools or parts move along some predetermined simple path,

without measuring the quality of the work being done. Modern NC machines are made to optimize their motions along multiple complex paths, and a few are equipped with sensors for controlling the quality of the product being produced.

A technology somewhat separate but related to NC is robot technology. The construction of the robot is an extension of machine tool technology to a class of machinery that makes higher-speed movements in more degrees of freedom. Robots find particular application at present in material handling (≈30 percent), spray painting (3 percent), spot welding (30 percent), overlay welding (11 percent), and other large-range but low-precision motions. Research in this field will probably yield robots that *learn* new movements from manual patterning and robots that are controlled from computers which will be programmed to recognize geometric images via television.

Production management in real time has gone beyond human capability and requires computers. In manufacturing of automobiles, for example, there may be 30 options available to customers, including color, type of radio, engine size, etc. There is little point in assembling the car unless all of the parts are available at the proper time (except minor items such as wheel covers or bumper guards). After production start-up the salespeople project, in general terms, how many four-door bodies, large engines, FM radios and reclining seats will be needed, and these are ordered into production. The production of components in-house and the supply of components from outside are usually done in batches so that buffer storage regions are planned to hold a limited number of parts. One attempts to store mostly the low-cost, small-volume, hard-to-add-on-later, discontinuous, type of parts. When all parts are in hand, the signal is sent out to start the major selected component down the line. This may be the floor of the auto body, for example. All parts are timed to meet the major component at appropriate points along the assembly line, so that the correct parts are added to the proper car.

This complicated operation is coordinated by computers. When properly done, the car involving the largest investment in parts is completed first so that the cost may be recovered early. In other words, a properly controlled system is optimized relative to some quantity, often highest profit. There are many factors to consider, such as the changing costs of one material compared with another, or the cost of paint versus the cost of upholstering, etc., but all of these details are recorded somewhere and should be used in the decision making process. Primarily, this activity engages the industrial engineer, but with considerable initial input from manufacturing engineers.

It is apparent from this illustration that the economic aspects of production are of great concern to the manufacturing engineer, but there is yet one necessary interaction. That is the management of design information in the factory. This is the next area for aid by the computer.

1.5 TRANSFER OF DESIGN INFORMATION

Information on the design of products is difficult to control. The problem is not secrecy primarily, but rather it is a matter of insuring that everyone working on a new product or process is using exactly the same information. Very often, after an initial design is completed and circulated in a plant, a great number of changes must be made. Perhaps a different material or bearing should be used somewhere in the design because material costs have changed, or because the company happens to have a large supply of leftover material from a failed product line, or perhaps because an old machine must be replaced in order to fabricate parts with the material specified in the new design. All of this information should be available to the designer, but frequently it is not. Success or failure of a manufacturing plant often depends upon good management of such information. Formerly this information was managed by the organization of people into appropriate groups, with hierarchies of coordinators, and liaison personnel. Recently the technical content and volume of such information has greatly increased. This is due mostly to increased competition but partly also to high energy costs, to sudden variations of materials prices and availability, and to the emergence of many new processes. One emerging reality is that human organizations are not dynamic enough to channel or transfer the appropriate information to appropriate individuals with adequate lead time. Computers have been very helpful in this area. One example may be cited in which design information is transferred with minimum error through a plant. That example is the engine plant of General Electric at Evendale, Ohio. The jet engine is a product that is steadily upgraded part by part; thus there is a steady flow of new design details to coordinate with the entire system. Formerly a new design for a part was committed to paper for communication throughout a plant. The new design was reviewed by material buyers, process engineers, plant engineers, and by the designers of adjacent parts. Frequently someone suggests a way to revise a new design for increased versatility or convenience, so changes are made. Unfortunately, not all who have seen the early versions of a new design are immediately aware of the recent changes. The consequence is that materials may be purchased or tooling may be designed on the basis of unfinished designs. Good organization minimizes this risk but does not eliminate it. General Electric designers store their designs electronically. All who have need to know about the new design are notified, and they must respond with a note of the action they will take. When a design is revised, those that have requested and seen the earlier design are interactively notified of changes. All such transactions are recorded so that the responsibility may be fixed for any inaction or action.

In the GE plant, when a design is reasonably fixed, tooling is made, equipment is prepared, and raw material is purchased. Production is begun by checking the quality of incoming material. If the material has the correct alloy content and hardness, it receives a laser-etched number. The part is transferred to a machine where its number is read by automatic reader. If it is at the correct, previously prepared station, approval to proceed is given on a TV screen on the machine. On the same screen there is a *print* of the part, with a list of all of the tools and gages to be used at that station. If the processing machine is automated, it will make the proper motions to produce the shape shown on the screen. After the part is finished, it is moved to an automated inspection station, where its shape is compared with the original design; if all is well, the part is cleared for scheduling into subassemblies.

The GE technology is not directly applicable to all products or manufacturers, but the concept of design information control has considerable potential. An important beginning point in establishing such a system in a factory is to link all of the computers at

design stations, processing stations, and inspection stations together. Through such links information can be gathered that would simply overburden human information channels. Consider two examples of information control that are immediately possible in computerized plants. The first is to keep a running record of the accuracy of parts from each machine, and this serves two purposes: to monitor machine condition and to continually assess the cost of holding manufacturing tolerances. The second example is to instrument machines to detect material properties or process behavior that would produce out-of-specification parts. When such conditions are found the process can be stopped before more cost is added in later processing steps.

The full potential of the computer information systems now awaits several major developments. To date the computer has been used to automate the simpler tasks of humans, in a way that is suited to human senses and limitations. New systems must be designed so that computer based systems complement humans. In process control, for example, there are three such needs, namely,

1. To determine what other forms of data are obtainable from a process, in addition to what is now obtained.
2. To develop new methods of direct sensing of product quality, rather than simply to adapt existing methods.
3. To develop fundamental models of processes.

Designers are said to make between 80 and 90 percent of the first choices of materials and processes for a product. Within this figure, there is considerable latitude on the definition of a designer, but optimal choices on processes are not made until very late in the progression toward the start up of manufacturing. The exact value of having early and accurate information for making the final decisions on processes and materials is not known. Managers feel very vulnerable in this area. Most would welcome progress in gathering accurate information, but few are able to express the form of aid required. The core of the problem is that competing processes are not well enough modeled or described so that a short analysis can be made on their relative costs and capabilities. It would be very useful if some of such information were available on a computer, but linked with a data base under the jurisdiction of specialists in various areas for information beyond the understanding of the designer. In other words, the designer should have access to a computerized consultant (either human or an expert system) in materials and processes in such a way that he is compelled to ask well-focused and vital questions of the proper specialists. This aim serves the dual purpose of formalizing material and process data, which has not been done to a significant extent, and it will relieve designers of excessive dependence on familiar materials and processes.

The latter point is not widely recognized by designers as a need. Perhaps the selection of materials and processes is not considered a part of the design process by very many, but this topic has its obverse side. Process engineers and production engineers too often discover that a design is finalized or frozen before consideration for the best materials and processes is included.

1.6 ECONOMIC AND MANAGEMENT CONCERNS IN MANUFACTURING

The engineering of a manufacturing system is a very broad field and it includes some consideration of every aspect of a manufacturing enterprise, including the financial and marketing activities. The engineering covers every detail of products, from the conceptual design to the shipment of a product to the customers. It involves intermediate redesign of a product for economic manufacture, material monitoring and handling, manufacturing process control, and some aspects of plant engineering and waste management.

Of major importance in manufacturing is consideration of the effective use of capital assets such as buildings and machinery. If a product will expend the useful life of a machine over the expected production run of the product, an optimum has probably been achieved in the economic use of capital. More often, a simple optimum is not possible. For example, a machine will usually outlast a production run of one product, be altered and adjusted for a different product, etc. If a very expensive and versatile machine is purchased such that it can be adapted to many products, the engineer faces the possibility that the machine will never be properly amortized if a product line changes drastically. Specifically, a company may purchase ten large presses expecting to produce stamped metal parts for ten years. However, after five years it may become necessary to produce the parts in plastics, which will require retiring the presses early and purchasing plastic molding machines. Or it may occur that after the purchase of ten presses of a particular capacity, it may be found that for competitive reasons, larger and more accurate parts are required, rendering the purchased machines obsolete earlier than expected.

The same applies to the use of buildings. At one time it may seem best to store materials at the ground floor and assemble products at the second floor level, but the product line may change in such a way that it would be most economic to intersperse storage and production on the same floor. A third example, one in material handling, involves trucks and railroads. A plant may be constructed with little storage area for raw materials or finished products, expecting materials to be transported by highway truck. However, freight rates may change over a short period of time in such a way that rail transportation would be recommended. This may require construction of a rail bed and more storage space than is currently available.

Engineers must make decisions that involve materials, plant facilities, and in-house capability of people as well. Take as an example the manufacture of coffee makers. Several of its parts could be made of plastics; others must be made of metal or glass. Some important points to consider in making this product are the shipping weight, the method of packaging in order to protect different types of materials, differing warranty considerations, and different production methods. Parts may be made in-house or purchased outside. The decision on which parts to produce in-house depends upon what machinery is either available or may be purchased with the expectation of future products requiring the same machines. For the water heater tank it may be cheaper to use the services of a

supplier who has both the experience and the machinery for deep drawing of metal. If your company is not skilled in deep drawing, it may require six months to get up to production. On the other hand, when you rely on a supplier, you must be assured that the supplier's labor problems will not halt production of your parts. By this example it is seen that manufacturing engineering includes consideration of labor relations and perhaps an analysis of the economic climate of the city or of the entire nation.

The economic analysis and decisions of engineers are very little different from that of the individual consumer, except in the scale of concern. In deciding whether to replace a troublesome but operable TV set the consumer is mindful of the possibility that new and revolutionary developments in TV may be available in one or two years; they may be worth waiting for. Likewise, before purchasing an automobile it is useful to consider whether the maker of a particular brand will maintain convenient dealerships in the future, or if there may be a problem in parts supply.

In manufacturing it is paramount that all decisions be tempered by considerations of economics. Everyone must participate in lowering costs, at every stage of manufacturing. For example, when the Ford Motor Company of Europe was planning the plant in Valencia, Spain for building the Fiesta, they considered it necessary to reduce the cost of "redesigning components, sorting out the plant and tooling, small improvements in the product, changes in material . . . ," from the usual "18% of the launch cost . . . (to no) more than 7 or 8%." To this end the management stated that "(We) must have . . . total discipline, . . . can't have any nonsense . . . (about lowering) the cost of changes, in putting things right after production has begun . . ." (John McDougall, chairman, Ford-Europe).* On the latter point management usually expects about 1000 changes of one kind or another during the first year of production of automobiles.

Most designs are not immediately convertible to profitable products. Even where a product is useful, reliable, and attractive, a low-cost substitute may capture the market. There are many reasons why a product may be too expensive. Some of the reasons are out of the range of technical solution. Some of these include an excessive overhead rate because of large staff groups or other high expenses to maintain a company in existence. More direct causes of high prices may be that employees are paid too much or are underutilized. Or the design may be too difficult to make economically. The specified materials may be too expensive to purchase and/or to process, process control may be inherently poor, or the production yield may be low if the part is made as designed. Or, again, the product may deteriorate in service more quickly than expected.

Redesigning a product for efficient and economic manufacture requires a direct and delicate interaction between two major groups in any company, the conceptual designers and the manufacturing personnel. Each group performs a vital function in the company, and until a product is on the market each can argue for the importance of their own position with some impunity. To lower the cost often requires a slight change in appearance, shape, function, smoothness of operation, or some other feature of the product.

* E. Seidler, *Let's Call It Fiesta* (Lausanne, Switzerland: Haessner Publishing Co., 1976).

Designers resist such compromises and urge manufacturing groups to *dig* the economies out of their own domain.

In deciding between the concerns of designers and manufacturers, management often makes decisions that involve the lower risk among several alternatives. For example, if a proposed product is elegant but overpriced, it may be projected that production costs can be reduced after production begins. If the company is in good financial condition, it may be able to temporarily *carry* the difference between the first manufacturing cost and the return from the price at which the product must be sold. If production were held up until the manufacturing group is satisfied that the product can be made to meet market price, it may be of such low quality that it quickly establishes a poor reputation on the market. It requires a well-informed and resolute management to arrive at the proper compromise in product design.

To a great extent, the strategy used by a manufacturing enterprise may depend upon one's position in an industry. For example, in the early 1920s, General Motors was recovering from a total financial failure. Alfred P. Sloan, then president of the company, adopted a strategy for recovery using the philosophy, "In order to gain market shares against a competitor, it is not necessary to have greater than competitive quality."* This philosophy is now widespread and was restated, critically, in 1981 in the words, "if a product works well for an extended period it may be overdesigned in some of its aspects. If such is the case, consider downgrading the materials or processing where you can."†

The compromise between designers and manufacturers is often carried out in a spirit of caution, faithfully reflecting organizational divisions. It has been found that the most profitable products are likely to come from a group consisting of a mature conceptual designer, a mature manufacturer, a person with experience as a shop foreman, and one or two younger people with analytical skills. Product designs developed by such groups are often ready for production with little alteration. Such products go into production months and perhaps years earlier than those from divided activities simply because there are fewer errors to purge from the system and because the latest information on most aspects of product production was used in the design.

PROBLEMS

1–1. What engineering skills are required to bring the following products from the idea stage to the market place?

- **(a)** Wooden lead pencils.
- **(b)** Pads of paper.
- **(c)** Cans of beverage.
- **(d)** Electrical power.
- **(e)** Bicycles.

* Alfred P. Sloan, *My Years with General Motors* (Anchor Books, Doubleday, 1963).

† H. E. Chandler, *Metal Progress* (Feb. 1981), p. 9.

1–2. How much more would *you* pay for the following improvements in consumer products? Justify your answer in terms of both subjective and economic terms.

- **(a)** A rustproof automobile.
- **(b)** A quiet vacuum cleaner.
- **(c)** A TV set that is only 30 mm thick.
- **(d)** A personal computer that operates at twice the speed as present models costing $1000.

2 economic principles applied to manufacturing

2.1 INTRODUCTION

The simplest expression for the economics of manufacturing is (on a unit basis):

$$\text{profits} = \text{selling price} - \text{marketing cost} - \text{cost to manufacture} - \text{overhead cost}$$

Many details are left out of the equation, but in practice, there are many costs assessed against the selling price. Only the *cost to manufacture* will be of major concern in this book. The other variables in the equation are the concern of management, and for their purposes some very intricate accounting schemes are used. Management is continuously adjusting accounting and information systems, in order to be completely informed on the status of every activity that contributes to the profit and loss of the company.

The cost to manufacture is determined most confidently after a product has been in production for some time. But a thriving business must continually develop new products, using new materials and new processes. One of the main responsibilities of manufacturing engineers is to *predict* the cost to make such products, and this is done when a manufacturing system for the products is being designed. The goal is not always to find the absolute minimum cost at the moment, but rather the lowest range of cost for several alternatives which depend on ever-changing markets. Some of these alternatives involve considerations of the size of production lots, production rates, choices of materials and processes, and various amounts of plant automation. Predicting product cost requires a detailed study of both the hardware and the software (i.e., the information) of a proposed system and also requires contact with many specialists in business and engineering. It is

from the others that manufacturing engineers obtain the interest rates, depreciation rates, overhead rates, market projections, and new developments in material properties that affect the cost to manufacture.

2.2 THE DIRECT COST OF MANUFACTURING

The factors in the direct cost to manufacture products include the following:

a. The cost to buy, inspect, store, move, and inventory raw material or purchased components.
b. The cost of production machinery and material-handling equipment.
c. The cost to prepare, install, and maintain the entire facility.
d. The cost to operate and service equipment.
e. The cost to assemble, package, store, and inventory finished products.
f. The cost of utilities, transportation, disposal, and protection.

The sum of all these costs divided by the number of units the system will make should give the cost to make each unit of product. But such an overall approach does not provide information with which costs can be *shaved* from products. A more thorough approach is to determine the *value added* to the product by each operation. If the cost of a particular operation exceeds the value added by that process, then the operation should be studied carefully. Some engineers, usually the design-materials-process engineers, will study the materials and processes themselves. Industrial engineers will probably determine whether the labor content or material-handling aspects of the operation can be revised, and civil engineers will consider the cost of space and facilities in the equation.

The cost of a manufacturing operation is not a fixed or intrinsic amount. That is, the cost to bend a sheet of metal into an electronic chassis, 100 mm × 300 mm, is not simply $3.50, for example. If process costs are given in handbooks, or as examples in textbooks, they will usually be given in ranges of cost, because the actual cost depends on many local conditions. The cost of equipment, personnel, and facilities, plus the time value of investment, are added to products with the passage of time, whether the facilities operate or not. To avoid the buildup of cost it may be necessary to cease manufacturing and dispose of the facilities. To reverse the buildup of cost it is necessary to sell products. Management regards the liquidating of the business as a real possibility at all times. They have little loyalty to a product line or to the plant or equipment. Rather, management is required, by the owners, to provide a return on the money invested in the company, by legal means, of course, and with consideration of the social consequences of their decisions. Thus career opportunities for engineers and all employees are inextricably tied to the success of the products produced. This is more often discovered in times of economic recession than during full employment and prosperity.

2.3 THE TIME VALUE OF MONEY

The specific time value of money is the interest rate. Borrowing money entails paying back the principal plus interest; investing or lending money entails receiving it back again after some time with interest added. Discussions of interest are complicated by the several methods of calculating interest, and by the terminology used by business specialists.

Interest may be calculated either by the simple method or the compound method. With simple interest one assesses a fixed amount of interest over a period, for example, 5 percent per year. Thus a present loan of 100 dollars is paid back in one year plus the 5 dollars interest; or it can be paid back in two years plus 10 dollars interest, and so on. Compound interest is a series progression in interest accumulation; it may be compounded quarterly, monthly, or any other time period. If, for example, 100 dollars is borrowed at 5 percent interest compounded quarterly, the interest builds up as follows: 5/4 percent or \$1.25 interest is owed after three months, making the amount to be repaid at that time \$101.25; over the next quarter 5/4 percent is charged on the \$101.25, making the amount owed after six months \$102.52; over the next quarter 5/4 percent is charged on the \$102.52, making an amount owed after nine months of \$103.80, and finally 5/4 percent is charged on the new amount, so that \$105.09 is owed at the end of the year. This is equivalent to a simple interest of ≈5.1 percent. If the interest is compounded monthly, the final amount is \$105.12, or an interest equivalent to ≈5.12 percent.

Since most interest-bearing transactions cover several interest periods, it is convenient to use equations instead of the method used above. There are two common problems in calculating the relationship between a present and a future sum as altered or influenced by interest. One is for calculating the single payments to make at the end of one interest period, and the other is for calculating payments each interest period so that repayment of a loan occurs over a given number of interest periods. The converse of borrowing money is to invest it in order to receive a return. An investor may wish to know how much to invest now to receive a desired amount at the end of one interest period, etc. The equations for calculating these amounts are:

To calculate what sum of money S will be available after a periodic rate of interest i accumulates for several interest periods n on a single present investment, P use:

$$S = P(1 + i)^n \tag{2-1}$$

(referred to as the single-payment compound amount)

The reciprocal calculates the amount P that must be invested now in order to receive a desired sum S at the end of n periods:

$$P = S\{1/(1 + i)^n\} \tag{2-2}$$

(referred to as the single-payment present worth amount)

To calculate what sum S will be available at the *end* of n years when an amount R is invested at the beginning of each interest period, use:

$$S = R\,\frac{\{(1 + i)^n - 1\}}{i} \tag{2-3}$$

(referred to as the uniform annual series compound amount)

The reciprocal calculates the amount R that must be set aside at the beginning of every year to accumulate a desired sum S by the end of n years:

$$R = \frac{Si}{\{(1 + i)^n - 1\}} \tag{2-4}$$

(referred to as the sinking fund deposit amount)

To calculate the uniform end-of-the-year payment R that can be realized for n years from an investment of P, use:

$$R = P\,\frac{\{i(1 + i)^n\}}{\{(1 + i)^n - 1\}} \tag{2-5}$$

(referred to as the capital recovery amount)

The reciprocal gives the amount P that should be invested now so that n uniform payments of R can be received each period to dissipate the investment:

$$P = R\,\frac{\{(1 + i)^n - 1\}}{\{i(1 + i)^n\}} \tag{2-6}$$

(referred to as the uniform annual series present worth amount)

Equation (2–5) for capital recovery amount is particularly useful for calculating how much return a machine should generate in order to "pay for itself." Usually the value and the production capability of the machine are known. The process engineer should estimate, where possible, how much value is added to a part being processed by the machine in question; if that value is greater than the cost of the operation, then the operation will contribute to the profitability of the company.

Occasionally a question arises as to whether investment in a machine is a good business decision. In such a case the expected investment in the machine may be inserted into the equation as the value P. The likely return from the machine in production is R, and some projection can be made on the number of interest periods n that the machine will be used. Either Eq. (2–5) or (2–6) can then be used to calculate a value of i, which is the return on the investment. If this value of i is greater than the interest rate offered in the money or stock markets, it would be wiser to invest in and use the machine rather than invest in stocks.

2.4 THE INFLUENCE OF PRODUCTION LOT SIZE ON COST

The cost of manufacturing includes the cost of the facilities prorated or spread over a production lot. For example, if a factory and equipment cost \$$10^7$ and 10^7 parts will be made in these facilities, leaving worn-out facilities that have sufficient value to cover only the salvage cost, the amortized cost of facilities is $1 per part. Such simple economics are rare outside of the sidewalk lemonade industry. In most industry the $$10^7$ has time value. That is, the $$10^7$ could earn $$10^6$ per year, if the interest rate is 10 percent, without the bother of setting up a manufacturing facility. If one chooses to set up a manufacturing plant, some of the $$10^7$ must be allocated for plant construction up to three years before production begins. In that three years, whatever money is expended is not available to yield a return; perhaps it was borrowed and interest must be paid for its use. This time value of invested money must be added to the price of the product. For example, assume a linear investment rate in plant construction and equipment purchasing over three years, at an annual interest rate of 12 percent. The accumulated cost of the money at the time production begins may be calculated in several ways. If the money is paid at a linear rate at the time of monthly billing by contractors and suppliers, the time span will be 36 months with a monthly interest rate of 1 percent per month. The added cost of money then may be seen by the use of Eq. (2–3), where $n = 36$, $i = 0.01$ and $R = 10^7/36$. $S = \$1.197 \times 10^7$.

Further, the invested $$1.197 \times 10^7$ is unavailable at the beginning of the year for profitable investment elsewhere, so the production run must absorb the interest again. If we assume a linear monthly recovery of the invested money, by Eq. (2–5) we can calculate the monthly payback required from sales. For this calculation, $n = 12$, and $i = 0.01$. Then $R = \$0.106 \times 10^7$. This amount must be recovered from $10^7/12 \approx 825{,}000$ parts per month, which is $1.289/part. Now assume that the cost of material per part is 55¢ and the cost of labor is expected to be $2.00. The cost to manufacture, then, is $3.839/part. It is simple to project what the cost/part would be if it should occur that sales are greater or less than projected.

The selling price will be much higher than the manufacturing cost, depending on the method of storing, transporting, and distributing the product. The markup from manufacturing cost to selling price is in the vicinity of a factor of 3 for one brand of home-appliances and a factor of 20 for some automotive replacement parts. One cost to be covered by the revenue from selling a product is the cost of doing business, and the cost of planning new products and production facilities. Ideally, all of the costs connected with bringing product A to market should be covered by the price of product A. In industry, the time of engineers, lawyers, business specialists, and upper management is so diffusely spread over all of the products that precise accounting is not possible. Furthermore, there are municipal taxes, insurance, lawsuits, and board meetings to cover. Often it is most reasonable that all costs other than materials, labor, and facilities directly chargeable to a product are added up and allocated to the products sold. This cost is called "indirect cost," or "burden," or "overhead." (Teachers are in this category.)

The indirect cost may be assigned in a simple way, as a figure to be added to the direct cost of manufacturing a part. With complicated items or in the event of inefficient management, the indirect cost may be applied as a multiple factor, as much as 3 or 5. The $3.839 direct cost of producing a product will have its price increased by the overhead cost. Assuming the overhead rate to be twice the direct cost of manufacture (that is, the overhead multiple factor is 3), the total cost associated with manufacturing the item is $11.517.

A profit margin must now be added. Recall that there is little point in setting up or continuing to run a manufacturing operation unless the financial return exceeds the potential return from investment in stocks or bonds. Recall also that the cost to manufacture includes the current cost of money. The profit margin is the *extra* return; however, profit margins cannot be set at will. The produce price must be set so that the product will sell in the marketplace at approximately the expected rate of production.

Price setting is a delicate art. Let us assume that an appropriate wholesale price in the case of our example would be $12.645, which is a profit of $1.128. This amount would appear to be an 8.9 percent return. Actually, the return can be applied only to the investment, which constituted $1.289 within the total manufacturing cost. The profit, $1.128, then becomes 87.5 percent of the investment share per part, and is worthy of some attention. In reality, a profit of 15 to 20 percent is to be expected.

But the selling price cannot be set by the seller. The marketplace sets the final price. Sometime the profit margin becomes unattractively small. At such times ways are explored to reduce the total cost of the product, direct cost first (because direct cost is leveraged by indirect costs) and then the indirect cost (requires sacrifices by managers). The limit on reducing direct cost is the point at which the buyer detects reduced quality or reliability. The limit on reducing indirect cost (engineering and management time) is the point at which smooth functioning of the plant is in danger or when future product plans receive inadequate attention. These are difficult trends to detect and very much depend on the projections one makes for future markets, future material, and labor cost and even future tax and inflation trends.

In fact, it sometimes happens that a price will be set below total cost to manufacture plus overhead. This may occur where a manufacturer wishes to *penetrate* a market. This may require three years to do. One make of small cars was estimated to lose between $700 and $800 per car in 1981, all in anticipation of a favorable sales position in the future. It seems foolish to deliberately manufacture cars at a loss, but if there are no products to present to the market, there will be no penetration of a market. The auto maker may not have expected to lose as much as $700 per car, but his annual loss may have been greater with an idle plant than with an operating plant. For the same reason, old established manufacturers experiencing severe competition may sell goods at a loss in order to maintain market share, or in order to test the strength of the competition (that is, to try to drive them out of the market).

2.5 THE INFLUENCE OF PRODUCTION RATE AND PRODUCTION METHODS

Marketing people have the responsibility to estimate the expected sales rate of products, but the estimates cannot be assured. Thus the manufacturing system designer must provide alternatives in processes and manufacturing methods to provide a range of profitable production rates. The most flexible economic system is one which has a minimum of capital investment and a minimum of indirect cost. This means a heavy, or almost exclusive, use of either direct labor or of outside suppliers. The main problem with outside suppliers is that some of the control of production schedules is lost. A secondary problem is that proprietary information is not guarded as carefully by outside suppliers as by the owner of the information.

The problem with nearly exclusive use of direct labor is that little labor-saving or automated machinery would be used, which can readily be shown to result in high manufacturing cost. An example of the costs for various degrees or levels of automation is given below, for making parts with four types of lathes. The machines, their costs, and the times to make the same part on each machine are given in Table 2–1.

The accounting can become very complicated if all possible factors are included, so the following assumptions are made in order to get to the point:

1. One normal 8-hour shift consists of 1500 production hours per year.
2. Each machine will be worn out after making a million parts. Thus the engine lathe could produce 9000 parts per year on one shift. It should therefore last 111 years, or 37 years on three shifts. The fully automatic lathe should last about 22 years on one shift, or about seven years on three shifts.
3. The annual interest rate is 10 percent.
4. Direct labor cost is $15/hour and applies only while a part is in production. (Overhead applied to direct labor equals direct labor cost and is included in Table 3–2 as labor cost.)
5. Material and cutting tool cost is 50¢ per part.
6. Setup cost, i.e., installation, initial tooling, and start-up cost of the machines, is 5 percent of the cost of the machine.
7. Since different machine lives introduce a factor that is very difficult to amortize, for a first approximation the actual rate of wearing out of machinery can be ignored in calculating the cost of money invested in machinery and an arbitrary linear 10-year depreciation life can be applied.

With the above figures we can calculate the cost/piece to make the parts in various annual lot sizes. With Eq. (2–5) we calculate an annual amount R that must be returned from the sale of the annual production. The investment in the machine is P; also, $i = 0.1$ and $n =$

Table 2–1

Machine type	Cost to purchase*	Labor time per piece	Production rate/hour
a. Engine lathe (manual)	$ 15,000	10 min	6 pieces
b. Semi-automatic lathe	20,000	5 min	12 pieces
c. Automatic lathe	60,000	1 min	30 pieces
		(the operator serves two machines)	
d. Automatic lathe with robot to load and unload	120,000	0	30 pieces

* 1985 prices.

Table 2–2

Machine type	Interest (R)	Setup cost	Labor cost per piece (inc. ov'hd)	Material and tool cost/pc	Cost per part at various annual prod. lot sizes		
					100	10^4	10^{6}*
a.	$ 2,441.18	$ 750	$5.00	50¢	$ 37.41	$5.82*	$5.50
b.	3,254.91	1000	2.50	50¢	45.55	3.43	3.00
c.	9,764.72	3000	.50	50¢	128.65	2.28	1.01
d.	19,529.45	6000	0	50¢	255.79	3.05	.53

*Beyond the capacity of the machines.

10. The interest cost and the setup cost must both be recovered; both are shown in Table 2–2.

Note that with large annual production lots and low labor cost, the cost of material becomes a large fraction of the cost to make the part. Material cost will be discussed further.

We can also calculate the cost per piece for full annual production of each machine for one shift and for three shifts by using the figures in Table 2–2 (see Table 2–3). With alternative "a" the cost per piece is nearly the same over a wide range of production rates; any advantage could be offset by a second and third shift "premium" for labor. With alternative "d" there is a clear advantage in using the machinery many hours per day.

The cost picture often becomes confused by the rate of depreciation applied to equipment. In engineering terms, the rate of depreciation is the rate at which the machine life is dissipated, but in accounting terms it is the rate at which the Internal Revenue Service allows one to declare a reduction in value of the machine. The IRS rate is usually faster than the actual rate for general machinery. Fast write-off does serve three purposes, but it adds one problem. The first advantage is that a fast depreciation reduces property values for local taxes. Second, depreciation amounts are subtracted from profits when income is calculated; thus high rates of depreciation reduce income taxes. Third, management is less reluctant (different from *more likely*) to dispose of a well-depreciated machine than others in order to purchase an upgraded variety. The main disadvantage of an artificially high depreciation rate is that the final cost to manufacture is set arbitrarily

Table 2–3

Machine type	1 shift		3 shifts		Difference in cost
	production	cost/piece	production	cost/piece	
a.	9000	$5.85	27,000	$5.62	− 4%
b.	18,000	3.24	54,000	3.08	− 5%
c.	45,000	1.28	135,000	1.09	− 7%
d.	45,000	1.07	135,000	.69	−36%

high (depreciation is included in overhead rates). In the above examples, recall that the engine lathe would have lasted much longer (say ten times longer) than the assumed depreciation life even when operated three shifts per day. The cost per piece for a full production rate would have been less than $5.62 if a lower overhead rate were used.

With reference to the machinery in the above example, the best choice for very large annual production lots is clearly the automated system. However, relatively few items are made in optimum annual lot sizes. First, the optimum annual lot size may exceed the needs of the market, for example, size 16 shoes. Second, the American consumer, at least, wants a very wide range of options to choose from. Thus for many products a production lot may be an uneconomic 10,000 per year or even 1000 per year. There are then two alternatives available for keeping costs low.

The first alternative is to buy components from an outside supplier. That supplier may be making similar components for others, and together all of the components can be made at low cost. This is a common practice. Consider the many manufacturers of lawn mowers; the great majority buy engines from only two sources. The same applies to refrigeration compressors, electric motors, integrated circuits, light bulbs, computer printers, tires, and components in most other consumer products.

The second alternative is to contract work from the outside and use your idle machines. This means that your company will be making products that do not carry your brand name, but recall that the purpose of the company is to make a profit, not to hold to high principles of exclusiveness from supplier status.

2.6 INFLUENCE OF MATERIAL AND PROCESS SEQUENCE ON COST

The choice of a material for a product will usually influence the needs in manufacturing facilities. Take, as an example, a portable tape recorder, which may be marketed with one of three types of covering:

1. With a sewn leather case, which, in this example, would be purchased from outside. This decision would be based on the reliability of outside suppliers and on a policy that your company could not become competitive in the area of leathercraft. Assume that the suppliers offer two prices, depending on the number purchased.
2. With a removable case of molded polypropylene, molded in-house and available in two colors. Plastic molding is a very competitive business, but you expect to take advantage of the availability of a fully amortized molding machine and low-cost dies.
3. With a bonded coating of polyurethane, available in several colors. The urethane would be applied by a new process which has been patented by your company. It makes a more durable and attractive product than the other options, and no competitor is likely to offer a similar product in the near future. The cost for the urethane option involves an investment in machinery and therefore is expected to be higher

than for the polypropylene option if only tape recorders are coated, but lower if other products beside the tape recorders will be coated.

Assume that the cost per covering for each of the options will be a function of annual production, as shown in Fig. 2–1. Note that two prices are given for the urethane option.

Note that purchasing items from a supplier insulates one from the effect of sales volume within certain ranges, and also that the price for the urethane option depends on the sales of other products, an intangible quantity for the manufacturing engineer to work with in predicting production costs. Information of the above type must now be considered in the light of the probable price that the market will bring for each option.

Curves of the type in Fig. 2–1 assume that the employees and machinery are devoted to a single product for the year. Frequently it is good strategy to make products at a high rate for a short time, stop production for a time, and wait until the inventory decreases. These decisions are usually not made by the engineer. It is sufficient to provide management with the type of information shown in Fig. 2–1.

An important type of information that the design-materials-processing engineers must provide for management is the influence of changes in material and energy prices on processing costs. It happens frequently that some prices change drastically and quickly due to political upheaval in foreign lands. For example, the price of foreign crude oil increased by a factor of 3 in one day in 1973. This increase in price was *passed on* to the final user progressively as the suppliers sorted out what fraction of their crude oil had to be purchased from expensive sources versus the fraction that would be supplied from their own wells, the prices of which were controlled by the government. Crude oil prices directly affect the cost of plastics such as polyurethane, polyethylene, and polypropylene. Many companies that made plastic toys, milk cartons, sheet, and fiber could no longer compete with similar products made of wood, cotton, or coated paper and were forced out of business within a few months. The same increase in oil prices raised the cost of industrial fuel, and suddenly made energy-intensive processes expensive. Examples of such processes include heat treatment of steels, reduction of bauxite, and electroplating processes.

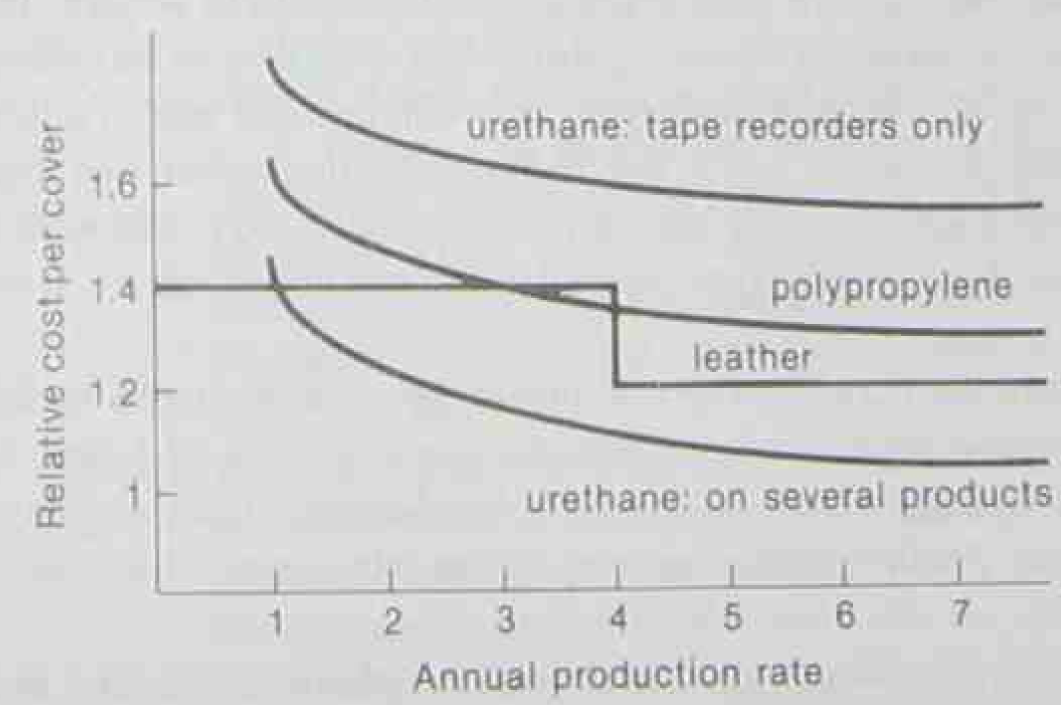


Figure 2–1 The relative costs of three options in covering for portable tape recorders.

The market in metals varied considerably as well in the 1970s. The price of cobalt increased from $4/lb to $14/lb because of civil war in Zaire. This suddenly increased the cost of the cemented carbide cutting tools, which made it *economic to slow the rate of production* of some products (see section 13–9). Chromium was another element that became expensive when unrest occurred in (then) Rhodesia, which increased the cost of stainless steel and some wear-resisting irons, both of which contain large fractions of chromium. At a time of such price changes, engineers must find alternative materials and processes, but always with the prospect of closing down an existing process for which the machinery and tooling are not amortized as planned. When stainless steel became expensive, there were several alternatives available for some applications. One was to use plain carbon steel and coat it or paint it; this required new operations. For other applications, glass may be substituted, but this requires not only new processes but also a redesign of some products.

One example in energy conservation is choosing between lasers and oxyacetylene flame for surface hardening. Lasers are expensive, and only about six percent efficient in terms of delivery of energy from the power line to the target surface, whereas the gas flame is a low-cost installation and 30 percent efficient. But upon review of the surfaces being processed, it was found that a few passes of the laser beam in selected areas was sufficient for the end use, and the general hardening with the flame was not necessary. Lasers therefore became preferred for many applications on the basis of energy costs.

Process cost estimating is thus seen to be a rather complicated activity, but it is also challenging. An engineer involved in cost calculations becomes aware of a broad range of concerns in the company, in the manufacturing industry at large, and in world affairs.

2.7 COST ACCOUNTING

Cost accounting is a means of arriving at the cost, and therefore value, of performing manufacturing operation. It is not simply the tabulation of the costs as an end in itself. Rather, cost accounting involves several closely connected activities which, together, are properly referred to as a *management information system*. Engineers have a prime responsibility for identifying the data that are both necessary and possible to acquire from the manufacturing system, and the means of acquiring those data. The data to be obtained are those necessary for checking on a regular basis that the actual costs of manufacturing follow the base-line prediction developed when the manufacturing system was being planned. In some cases it is sufficient to check once a week or once a day, but with the advent of computers it is possible to check some operations every minute or even continuously.

If there are variances from the norm, either real profits decrease (or even vanish), or profits increase. Variances arise for the several reasons given in Secs. 2.5 and 2.6. Negative variances may be due to unexpected incidences of breakdowns, work stoppages, and adverse weather conditions. Positive variances may occur when new processes are *learned* faster than expected.

Management information systems vary considerably in complexity according to the

size of a company, the variations of the products, and the personal preferences of upper management. Smaller enterprises often succeed with a very informal system, and some larger ones have failed though they have very elaborate systems. The goal of all is the same, however, and that is to determine as precisely as possible how profitable each and every employee and machine is. Some can be carried at a loss for a time if it fits in with a long-range plan. It would be expected that profitable elements would *cover* for the losing elements. But management must know precisely how much one element covers for another, and what the trend of such costs is.

The basic ingredients of a comprehensive cost accounting scheme are:

1. Identifying the physical areas of factories called *cost centers* for which costs can be reasonably well isolated and in which responsibility for cost incurrence can be assigned to responsible individuals.
2. Comparison of the actual costs in a center with the forecast results for a planned period.
3. The measured variances and the reasons for them.
4. The action needed to control and reduce costs to manufacture products.
5. Profit or loss, and cash availability, forecast at regular intervals.
6. Regular comparison of sales prices with unit or product costs in order to determine the competitive position of the company.
7. Planning for replacement of existing facilities, or provision of new facilities.
8. Planning for allocation of raw material and material sources.

Engineers are usually expected to work in harmony with the cost accountants, but not to be deeply involved with the methods of cost accounting. This section was provided for perspective on the continuum of expertise required to successfully operate a manufacturing enterprise.

PROBLEMS

2–1. Assume that the market value of automobiles driven 15,000 miles per year declines each year to 72 percent of its value at the beginning of that year, and that the operating costs remain substantially constant for the life of the car, but the annual maintenance costs to keep it in top condition increase linearly to equal the market value of the car at the end of the fifth year. Further assume that you can invest money or borrow money at the same rate, say 10 percent for simplicity. What is the most economic strategy for buying a car for personal use (at 15,000 miles per year)?

2–2. Financial institutions often advertise that they offer higher interest rates than others do. A common method is to offer monthly or even weekly compounding of interest, but in fine print an interest rate is given to three decimal places to make sure that there is no misunderstanding. Provide a table of interest rates that yield the same annual return for:

(a) Simple interest (assume 5 percent).

(b) Compounded quarterly.
(c) Compounded monthly.
(d) Compounded weekly.
(e) Compounded daily.

2–3. Assume that the material cost of hamburgers is 24 cents each, and all other costs per hamburger at a sales rate of 600 per hour is 36 cents. Normally the hamburger sells for 69 cents each, but to boost lagging sales the management advertises hamburgers at 39 cents each.
(a) At what rate must hamburgers be sold to avoid losing money?
(b) At what rate must hamburgers be sold to avoid a decline in hourly profits?

3 product quality attributes

3.1 INTRODUCTION

The quality of a product need not be high, only adequate, in order to sell well. Consumers usually abhor this attitude in principle, but they endorse it in their buying practices.

Manufacturers, like cooks, know approximately how well their product is made but do not always know how to assess the tastes and desires of the consumer. Some do so by watching the inventories in the warehouses rise or fall. Others measure *their* fraction of the market for similar products, and strive to hold or enlarge it. Still others monitor quality by determining customer loyalty, that is, the fraction of repeat sales to customers, if they can be identified. The auto industry, for example, uses numbers of repeat sales in the range of 60 to 70 percent as a guide. If fewer than 60 percent of previous customers buy the same brand, or another in the same family, of the new model cars, then special attention must be paid to new products. There may be a problem of styling, price, or quality. If customer loyalty exceeds 70 percent, perhaps too much is being paid for manufacturing the product. In other words, the product may have excessive quality.

Product quality is a broad subject, encompassing such attributes as the appearance of a product, its reliability, the noise it makes, how it *feels*, whether or not it builds up a charge of static electricity, or whether or not it fails gracefully. To get a perspective on some of the measures of product quality, it is instructive to sample the various magazines available on newsstands that cater to the fisherman, the sports enthusiast, the auto racer, the electronics and computer club, the flyer, the muscle builder, the debutante, and many other consumers. Most of such magazines contain, as a regular feature, an analysis of new products, assessed and criticized in intricate detail. Clearly, there is some bias in some evaluations, but there is also considerable truth. Ultimately there is limited benefit in attempting to separate bias from truth, because consumers will buy what they want to buy. Thus if the consumer thinks that a noisy vacuum cleaner is better than a quiet one, then by all means supply noisy ones. If customers *need* FM tape player sets of 200 watts power and with assured frequency response far beyond the capability of the human ear to receive, supply it. If car buffs insist that rack-and-pinion steering aids cornering ability, by all means build on that attitude with products and advertising. If consumers insists on buying a 10-megabyte computer for letter writing (word processing), then satisfy them quickly before they raise their *needs* to a 15-megabyte computer.

3.2 THE CONTROL OF QUALITY IN MANUFACTURING

Quality is usually designed into a product; it is strongly influenced by the choice of manufacturing process, and it is *controlled* by properly organizing the people within a manufacturing facility. If there is a deficiency in any of the three areas, costs can be high. One company found in 1983 that 25 percent of their manufacturing assets were tied up in reacting to quality problems, and about 20 percent for the inventory. It required company-wide action for a year to control the problem.*

The design-materials-process engineers write specification sheets for product parts in such terms as "part dimension," "flatness," "hardness," "surface roughness," "reflectivity," "density," and the like. These specifications are taken from some knowledge of the intended use or function of the part. The engineers also write guidelines on how to measure the *objective* quality attributes, and specify the tools and instruments to use. Some attributes must be evaluated *subjectively*, such as uniformity, texture, color, and the like. These are sometimes defined by the marketing group, and their measurement may involve psychologists and others who are not engineers.

During the design of a manufacturing system, industrial engineers usually determine the location in the system at which quality can and should be checked. They determine how many people will be required, what human factors are involved, the training the inspectors should receive, and the methods that the inspectors will use to report data on quality to management. When production begins, the factory managers effect the plan and supervise the inspectors.

Quality control managers, or people of equivalent title, analyze the data from inspectors. They frequently use statistical methods to determine a number of very useful quantities, such as:

1. Whether the system is drifting toward more or fewer rejects.
2. The cause and source of the most rejects, for example, worn machinery, poor raw materials, operator carelessness, and so on.

* Young, *Wall Street Journal* (1981).

3. Whether the tolerances on the dimensions specified by the design engineers can be economically achieved.
4. Whether inspection is too rigorous or too lax.
5. Whether the current inspection stations are optimally located in the system.

In parallel with a continual activity to check the quality of manufactured parts, there is usually a careful study of the reasons for products to be returned for refund or repair. If product durability is a problem, or if consumers frequently complain that a product does not function as expected, a new study is done by designers. Perhaps some part will be redesigned or perhaps a manufacturing process must be revised to improve the product.

3.3 QUALITY STANDARDS AND COST

Some products can no doubt be made at a profit whatever the quality. But for most products there must be an appropriate balance between quality and cost to manufacture. Thus it is difficult to set down fixed standards of quality for all parts for all time.

The balance of cost and quality varies with the type of market and the economic times. For example, the price of a Rolls-Royce sedan is ten or more times that of a Chevrolet sedan, whereas the quality ratio is far less than 10:1. But the Rolls-Royce continues to sell enough to remain in production. The reason is that the two brands are not in the same market; that is the same people are not *competing* for both products.

Difficult though it may be to obtain, some absolute measure of quality must be used by inspectors. Where an instrument is used to measure a quality attribute (e.g., with a scale to measure weight), an inspector has only to adjust the settings and assure the proper operation of the instrument. However, much inspection is subjective in nature, such as rubbing the hand over a surface to detect *bumpiness*, or other quality. With experience, human inspectors become very adept at finding even miniscule defects on very large or intricate parts. Examples are pinpoint-size defects on large sheets of glass, or disorder in the array of wires and terminals on computer boards.

In the subjective area, as inspectors become better trained to spot defects, they can and do begin to reject parts that few consumers would reject or complain about. Consumers vary considerably in their opinion of the quality of products, depending on many personal and societal factors. It is evident that, to prevent a gross mismatch of perceptions of product quality, there must ultimately be a *meeting of the minds* of the inspectors and consumers in some way. The ideal is to establish direct communication between them, which does occur between manufacturers selling finished or semi-finished components to one another. But communication with the final arbiter, the consumer, is not practicable. Indirect communication does occur, but through the very imperfect information network consisting of market analysts, factory management, and quality control supervisors. If this network is efficient, then good and flexible quality standards will be developed and used. Otherwise, profits will eventually and surely decline, either because parts are made too expensively, or because consumers will detect undesirable quality and not buy.

3.4 THE DESIGN-MATERIALS-PROCESS ENGINEER AND QUALITY CONTROL

Consumer attitudes are frequently dismissed by engineers, who prefer, and were trained, to be methodical and logical, at least in their professional lives. But the reality of the consumer product industry is that a number of subjective factors often govern the quality of products.

In defense of engineers it must be stated that the primary responsibility for quality control has not been theirs in many industries. Inspection is most thoroughly done near the end of the manufacturing sequences, just before shipping. It has therefore become the jurisdiction of the manufacturing managers, and engineers are thereby somewhat removed from such concerns. But again, design-materials-process engineers have not been well equipped to impinge upon the quality problem. In fact, they do not adequately understand quality themselves. Consider the usual method of specifying a quality attribute, for example, surface finish, on a new engineering drawing. This specification is frequently copied from the drawings of previous successful parts, not because it is known to be appropriate, but because this practice usually works well, probably for two reasons. The first is that surface finish is almost always overspecified in order to be safe. Second, a familiar specification will *signal* the processing people to use established processes. The importance of this practice cannot be overemphasized. Most everyone in industry knows of numerous examples of good parts being produced by an old machine or by a prototype machine under the watchful eye of process engineers, but when a new machine is installed numerous problems appear. One type of problem is that the new machine may make parts to specifications, but the parts do not function satisfactorily. The converse is found as frequently: The parts fail inspection, but testing shows that they function well. The genius of technicians and engineers eventually makes the process function satisfactorily, but not without some price in mental turmoil, cost overrun, and delay in start-up of the system.

Many problems could be averted if the processes were well understood, and if the product attributes resulting from the process were measured in terms that have some connection with the function of the part. Take the example of grinding of a cylindrical shaft. The finished product must meet specifications of diameter, hardness, straightness, and surface finish. Most grinding machines produce out-of-round and slightly bowed parts. After grinding, only the shaft diameter and surface roughness are measured, and the part is either rejected or accepted. Out-of-roundness is not measured explicitly, and yet this attribute is important in some applications. Surface roughness in terms of average asperity height is measured, but skew and kurtosis (discussed in Sec. 3.7) are usually not, though they are thought to be important attributes of lubricated surfaces. The list of such examples is endless, and most designers plead for relief from such concerns. These details are thought to be the domain of materials people or the manufacturing department. However, here is the problem in quality control: It is a complicated field, and requires good communication between specialists.

It will be increasingly necessary for engineers to formalize their understanding of product quality attributes because of the trends toward complete automation. There are two aspects of automation of consequence to quality. One is the automation of inspection, to escape some of the limitations of human inspectors. The other is the in-process inspection of parts for real-time control of automated production machinery. The first is obvious, but the second requires some explanation.

In-process inspection is some combination of taking measurements from a part while a manufacturing operation is taking place on that part, or taking measurements from a part that is between operations. One purpose of such inspection is to insure that the failure of one operation does not damage a part or the processing machinery. But there are two other important purposes. One is to find defects in parts as early as possible, to avoid investing further processing cost in the part. The other is to determine very quickly when a process needs adjustment to maintain quality, before several unacceptable parts are made. This is the way soup is made, and this is the procedure of processing parts in a job shop. Human operators of machinery realize when a process variable has changed during the process, by such senses as smell, sight and hearing. These senses have meaning to experienced machine operators. However, in high-production-rate factories, more of the processes are semiautomatic. The machines are attended by less skilled people and often by people with minimal experience with any particular machine. Thus the detailed conditions of the materials in-process, and important nuances in the processes are not fully perceived until formal inspection occurs by designated inspectors. Inspection stations are often widely distributed, and the result is that, depending on the inspection procedure, several bad parts could be made before the defect is caught and the process altered. Or worse, damage to the machine or its fixtures often results from unattended broken tools, or parts that have slipped from a fixture. Process sensing is often prescribed for such problems, but in this case the sensing is done to avert serious consequences to the machinery and is not aimed primarily at the problem of product quality.

3.5 RANK ORDERING OF PRODUCT QUALITY ATTRIBUTES

In the area of quality control and subjective inspection, the design-materials-process engineer will have increasing responsibility to provide much more formalized information than has been done in the past. Two major areas are:

1. The instrumentation and automation of inspection. This consists of two parts, namely,
   a. To find ways to use available sensors of light, temperature, pressure, and the like to detect product quality attributes now perceived by human senses.
   b. To identify areas in which new sensing techniques should be developed.
2. To use the output from the sensors to control processes in real time.

One of the first steps in formalizing the subject of product quality attributes is to translate common expressions for quality into technical terms. Product quality is often expressed in quaint ways. For example, when an auto enthusiast claims that car A *corners* *better* than car B does, this apparently has reference to some quality in a car by which a sharp curve in a road can be satisfactorily negotiated. The *ordinary* car may do so successfully, but not satisfactorily. The difference apparently is a hitherto unquantified quality of a car providing appropriate *feedback* whereby the driver purports to perceive when the limit of traction or friction between the tires and road is being approached.

The formal approach to quality control consists in developing a rank ordering of *product quality attributes*. That is, the designer should first determine which are the primary attributes of a part, or, in other words, those attributes that are vital for functioning or consumer acceptance of the product. Then, since primary attributes are often difficult or expensive to measure, one or more secondary attributes for each primary attribute should be given, with perhaps tertiary attributes as well. With each listed attribute there should be one or two methods given by which it may be measured, and there should be a comment on how practical it would be to measure that attribute. Also, with each listed secondary and tertiary attribute there should be some comment on the extent to which it connects or fails to connect with the higher-order attribute. Finally there should be an analysis of each process to determine the urgency of measuring primary attributes. Recall that there are two concerns in the engineering of process control. The one is the control of product quality, and the other is prevention of damage to the machine and fixtures. It would be useful to list all of the ways that a process can go wrong and to devise protection according to the probability and seriousness of such damage.

Practical examples of rank ordering of quality attributes are now given. To begin, the primary attributes of a product such as silk cloth may be appearance, or *feel*. But it may not be possible to automate the measurement of these attributes. In such cases relevant secondary attributes must be found that correlate satisfactorily with primary attributes. For example, part of the desirable appearance of a silk surface may be its shininess. This then could be a useful secondary attribute, and could be measured by reflectivity. However, reflectivity deals only with the intensity of reflected light as compared with the intensity of incident light. The intensity of light is measured with a photocell, but the photocell and associated electronics does not *see* the surface in the same way the eye does. The difference may be due to surface chemistry or composition which alters the polarity of light in a way not detected by photometers but in some instances is perceived by the human eye as a general surface attribute. Thus the secondary attribute, reflectivity, may not be a complete measure of the primary attribute, appearance, but it may detect a majority of the variations in the primary attribute. In the case of appearance it would seem that physicists and psychologists could already have defined all of the necessary variables so that an equation could be available to relate the relevant attributes, but that is not yet in hand.

Another example using reflectivity is found in the coffee industry. The ultimate test of coffee is the drinking thereof, apparently. Whereas advertisements refer to the *taste test*, that quality known as odor may also be important. But these are not the attributes used in the first quality check of coffee beans. Rather, *every* bean from the roaster is slid down a chute, past an incandescent light bulb and a photocell. If the bean is too light or too dark, it is rejected. It is difficult in this case to determine whether the measured attribute has rational meaning relative to the primary attribute. The measured attribute

may not be a secondary attribute as much as it is a *substitute attribute*, defining some quality pleasing to wholesale buyers, but not sought in the *taste test*.

Another example is in metal cutting. One problem of great importance is the progression of tool wear. It is sometimes found that after severe *crater wear* the tool tip breaks off and the part becomes damaged. If a tool could be taken out of service before it fails, it would reduce scrap rates and allow economic regrinding of the tool. However, there are many changes in the process that occur long before the tool is near the point of fracture. In particular, the surface finish changes considerably, and the residual stress state or the depth of work-hardening changes. In some instances a dull tool exerts a larger force against the part than does a sharp one, which bends the part away from the tool during the cut. This would affect the dimension of the finished part. The primary attributes listed so far for this problem are surface finish, residual stress state and dimension, but none of these are measured in process. Work on this problem has focused on the measurement of forces on the tool during cutting, and the analysis of noises generated during wear, to see if trends can be detected that correlate with tool wear. Some results are available from the noise analysis, but they are obscured by noises from many sources other than those that can be directly connected with cutting.

The primary attributes connected with tool wear are not measured because of the difficulties in doing so, particularly in process. Cutting force and noise are easy to measure, but these are not obviously connected with any of the attributes of a product, or of a process. In this case the ready availability of some sensing technologies obscures the reason for process sensing in the first place. Something of value will inevitably be found from research on force measurement and noise analysis, but a more direct approach to in-process sensing and design of real-time control begins with an analysis of relevant quality attributes.

A clearer example of rank ordering of attributes may be found in the assembly of electronic components. Soldering is a common method of connecting a wire to a connector tab, and this must be done with 99.999 percent reliability, because there are hundreds of such soldered joints in one board or subassembly. A failed solder joint ultimately is manifested as a high electrical resistance, the solder may have surrounded the wire but did not wet and bond metallurgically. In a poor joint the wire and the solder are separated by less than a micron of oxide and soldering flux, and the electrical resistance of such a joint usually varies with humidity and temperature and with the way the soldering was attempted. Therefore, measuring the electrical resistance of the joint may not reveal its true state. Consequently, it is common to measure a secondary attribute of the joint, namely, some evidence of wetting. The evidence of a liquid solder wetting a vertical wire is that the solder will *climb* the wire about 3 mm or so; lack of wetting is manifested by the solder's being depressed around the wire. If complete wetting is the only condition that produces the climb, and if there are no known cases of complete wetting without solder-to-wire bonding, then the secondary attribute is sufficient for quality measure. The climb of the solder can be determined by visual inspection, or it can be automated by analyzing the image from a TV camera with a shape analysis program in a computer.

The important point in the above examples is that the measurement of a secondary *better* than car B does, this apparently has reference to some quality in a car by which a sharp curve in a road can be satisfactorily negotiated. The *ordinary* car may do so successfully, but not satisfactorily. The difference apparently is a hitherto unquantified quality of a car providing appropriate *feedback* whereby the driver purports to perceive when the limit of traction or friction between the tires and road is being approached.

The formal approach to quality control consists in developing a rank ordering of *product quality attributes*. That is, the designer should first determine which are the primary attributes of a part, or, in other words, those attributes that are vital for functioning or consumer acceptance of the product. Then, since primary attributes are often difficult or expensive to measure, one or more secondary attributes for each primary attribute should be given, with perhaps tertiary attributes as well. With each listed attribute there should be one or two methods given by which it may be measured, and there should be a comment on how practical it would be to measure that attribute. Also, with each listed secondary and tertiary attribute there should be some comment on the extent to which it connects or fails to connect with the higher-order attribute. Finally there should be an analysis of each process to determine the urgency of measuring primary attributes. Recall that there are two concerns in the engineering of process control. The one is the control of product quality, and the other is prevention of damage to the machine and fixtures. It would be useful to list all of the ways that a process can go wrong and to devise protection according to the probability and seriousness of such damage.

Practical examples of rank ordering of quality attributes are now given. To begin, the primary attributes of a product such as silk cloth may be appearance, or *feel*. But it may not be possible to automate the measurement of these attributes. In such cases relevant secondary attributes must be found that correlate satisfactorily with primary attributes. For example, part of the desirable appearance of a silk surface may be its shininess. This then could be a useful secondary attribute, and could be measured by reflectivity. However, reflectivity deals only with the intensity of reflected light as compared with the intensity of incident light. The intensity of light is measured with a photocell, but the photocell and associated electronics does not *see* the surface in the same way the eye does. The difference may be due to surface chemistry or composition which alters the polarity of light in a way not detected by photometers but in some instances is perceived by the human eye as a general surface attribute. Thus the secondary attribute, reflectivity, may not be a complete measure of the primary attribute, appearance, but it may detect a majority of the variations in the primary attribute. In the case of appearance it would seem that physicists and psychologists could already have defined all of the necessary variables so that an equation could be available to relate the relevant attributes, but that is not yet in hand.

Another example using reflectivity is found in the coffee industry. The ultimate test of coffee is the drinking thereof, apparently. Whereas advertisements refer to the *taste test*, that quality known as odor may also be important. But these are not the attributes used in the first quality check of coffee beans. Rather, *every* bean from the roaster is slid down a chute, past an incandescent light bulb and a photocell. If the bean is too light or too dark, it is rejected. It is difficult in this case to determine whether the measured attribute has rational meaning relative to the primary attribute. The measured attribute

may not be a secondary attribute as much as it is a *substitute attribute*, defining some quality pleasing to wholesale buyers, but not sought in the *taste test*.

Another example is in metal cutting. One problem of great importance is the progression of tool wear. It is sometimes found that after severe *crater wear* the tool tip breaks off and the part becomes damaged. If a tool could be taken out of service before it fails, it would reduce scrap rates and allow economic regrinding of the tool. However, there are many changes in the process that occur long before the tool is near the point of fracture. In particular, the surface finish changes considerably, and the residual stress state or the depth of work-hardening changes. In some instances a dull tool exerts a larger force against the part than does a sharp one, which bends the part away from the tool during the cut. This would affect the dimension of the finished part. The primary attributes listed so far for this problem are surface finish, residual stress state and dimension, but none of these are measured in process. Work on this problem has focused on the measurement of forces on the tool during cutting, and the analysis of noises generated during wear, to see if trends can be detected that correlate with tool wear. Some results are available from the noise analysis, but they are obscured by noises from many sources other than those that can be directly connected with cutting.

The primary attributes connected with tool wear are not measured because of the difficulties in doing so, particularly in process. Cutting force and noise are easy to measure, but these are not obviously connected with any of the attributes of a product, or of a process. In this case the ready availability of some sensing technologies obscures the reason for process sensing in the first place. Something of value will inevitably be found from research on force measurement and noise analysis, but a more direct approach to in-process sensing and design of real-time control begins with an analysis of relevant quality attributes.

A clearer example of rank ordering of attributes may be found in the assembly of electronic components. Soldering is a common method of connecting a wire to a connector tab, and this must be done with 99.999 percent reliability, because there are hundreds of such soldered joints in one board or subassembly. A failed solder joint ultimately is manifested as a high electrical resistance, the solder may have surrounded the wire but did not wet and bond metallurgically. In a poor joint the wire and the solder are separated by less than a micron of oxide and soldering flux, and the electrical resistance of such a joint usually varies with humidity and temperature and with the way the soldering was attempted. Therefore, measuring the electrical resistance of the joint may not reveal its true state. Consequently, it is common to measure a secondary attribute of the joint, namely, some evidence of wetting. The evidence of a liquid solder wetting a vertical wire is that the solder will *climb* the wire about 3 mm or so; lack of wetting is manifested by the solder's being depressed around the wire. If complete wetting is the only condition that produces the climb, and if there are no known cases of complete wetting without solder-to-wire bonding, then the secondary attribute is sufficient for quality measure. The climb of the solder can be determined by visual inspection, or it can be automated by analyzing the image from a TV camera with a shape analysis program in a computer.

The important point in the above examples is that the measurement of a secondary attribute in the place of a primary attribute involves some risk. Unfortunately, the attraction of presently available instruments obscures the study of *needed* instruments with the result that most processes are not controlled as directly as they could be.

3.6 METHODS OF MEASURING PRODUCT QUALITY ATTRIBUTES

Most of the primary quality attributes, such as size, roundness, straightness, cylindricity, taper, reflectivity, and electrical resistance, can be measured with simple instruments which require little technical background. Some attributes, such as residual stress and submicroscopic cracks, are more difficult to measure and require specialized equipment. Some of the secondary attributes which are measured for process control involve the measurement of forces, vibration modes, and machine speed, but again require fairly simple equipment. Most of these measuring devices are available with the proper electronics and data display stations. However, surface texture is a much more complex attribute, both to measure and to understand. Its measurement will now be described.

3.7 SURFACE ROUGHNESS

Line drawings of objects often leave the impression that surfaces are perfectly smooth, but this is never achieved in practice. Except for cleaved materials, few surfaces are atomically smooth. Most surfaces are composed of arrays of small bumps, known as *asperities* or *protuberances*, which were formed in the manufacturing process. Furthermore, many processes *damage* the substrate beneath a surface to a small depth, which may affect the performance of the part. For example, severe grinding can reduce the fatigue life of a steel part by a factor of 10 or more.

Some processes such as polishing produce very smooth surfaces, but the asperities are still over 30 atoms high. Such surfaces are formed on hard memory disks and optical surfaces, but they are far too expensive for most other products. Economics, then, will limit the relative roughness or smoothness of manufactured parts, and economics will demand orderly ways to specify and verify surface smoothness. Table 3–1 shows the relative costs of surface smoothnesses as achieved by various common metal cutting and finishing processes. (See page 46 for Table 3–1.)

One surprising fact is that designers rarely know how smooth a surface should be, or how to make a surface of a particular quality. Useful surfaces have been manufactured for many years, but success usually continues only as long as every detail in making the surface is done as it has always been done. A major problem often occurs when it is necessary to change suppliers of tooling or materials during a production run. Though the product is made exactly the same *way* as in the past, and though the product passes inspection, some will suddenly fail in the hands of the consumer. An opposite type of problem can also occur, when a process is changed so that it produces a measurable change in surface quality without affecting the failure rate of the product. This produces

the suspicion that close control of product quality may not have been necessary all along, and that too much money was spent in making the product.

Surfaces are too complicated to describe with a single word or number. To illustrate, take the example of the highways on which we travel. We can describe at least five scales of geometry of roads, most of which are regular and apparently periodic, some of which are irregular. The largest scale relates to following the grand features of the earth surface, such as large hills and mountains. On a smaller scale we see roadways that follow the smaller contours of rolling countryside. A third and smaller scale of geometry has a wavelength of the order of the length of a car, which at 60 mph can produce uncomfortable oscillations of about two or three cycles per second. A fourth smaller scale, of the order of 50 to 100 mm, has little affect on ride comfort, but produces a loud rumble from vehicle tires. The fifth and smallest scale, of the order of order of 1 to 10 mm, is important for wet skid resistance of tires on the road. We see that, in describing road surfaces, the size scale of interest depends on the function or use that one has in mind for the road.

The topography of manufactured surfaces may be visualized as being similar to the contour of the earth surface. Surfaces contain irregular features called *flaws*. The regular features, minus the flaws, are referred to as the *surface texture*. For some purposes it is useful to define two scales of texture, a fine scale known as the *surface roughness* and a coarser scale known as the *waviness*.

Some aspects of *surface roughness* can be assessed either visually or by running a fingernail over a surface. Such subjective methods have limits but are still widely used. A less subjective method uses comparators which are four or five samples of surfaces made to different roughnesses. A visual comparison is made between a production part and the comparator or standard surfaces. The inspector declares that the production surface looks more like standard "c" than "d" for example, and thus may reject or accept the part.

The most common objective method of measuring surface texture uses a stylus tracer. It consists of a diamond or other hard projection much like a needle or stylus of a record player. The stylus moves along a solid surface, rising over the peaks and descending into valleys as it moves. The rise and fall of the stylus, that is, its vertical motion, is detected by an electronic system and amplified for various purposes. The stylus is moved horizontally along the surface to be traced by one of two different systems. The first and simplest is the *sled* arrangement, which involves two spheres of about 6 mm radius that ride on (slide over) the same surface as that being measured, to the sides of the tracer. The sled also rises and falls while sliding over the asperities, following the waviness mostly. The radius of the stylus tip is much smaller than that of the spheres. The actual effect of the vertical motion of the sled on measured roughness is rather small, but the sled is nevertheless not recommended for careful work. Its major appeal, however, is that a tracer with a sled can be set on any surface by hand and data can be obtained in less than a minute.

The second and superior (at least, more expensive) mechanical system guides the stylus on a remote sliding surface of high precision. To use the latter system it is necessary to bring a specimen surface to the stylus and align the specimen to be fairly parallel to the surface that guides the stylus. A sketch of a precision stylus tracer unit is shown in Fig. 3–1.

The vertical motion of the stylus is detected by a sensor connected to an amplifier.

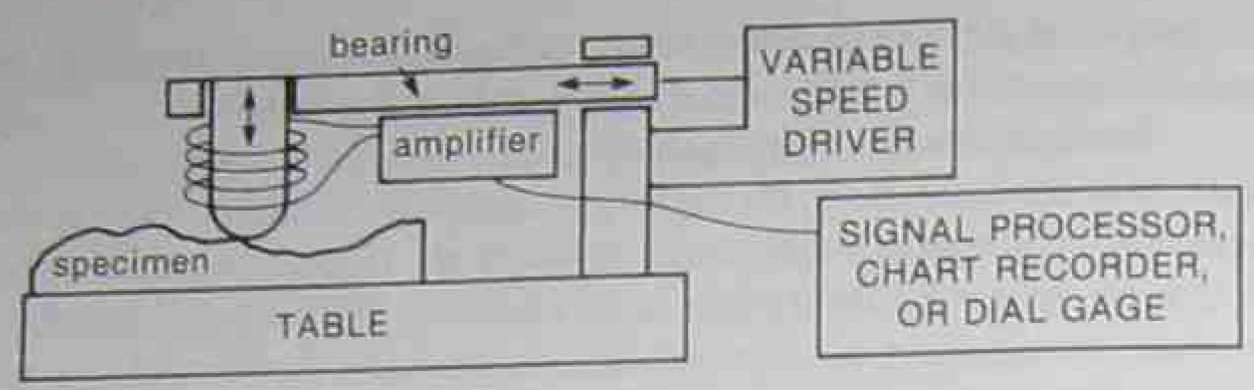


Figure 3–1 Sketch of a stylus tracer system for measuring surface roughness.

There are four end uses for the output from the amplifier. The simplest and most common older method is to connect the amplifier output to an averaging circuit and a dc voltage meter. Vertical oscillations of the stylus deflect the needle of the meter. The dial on the meter is calibrated in terms of an average asperity height for direct and convenient readout in the production environment.

A second objective is to record an amplified profile on a chart paper for visual characterization. If profiles are obtained along closely spaced parallel lines of tracer travel on the surface, one can construct a three-dimensional surface map. To map accurately it is necessary to move the stylus at the same speed during each traverse, to trace each pass in the same direction, and to know the location of the start and stop of each trace very accurately.

The usual output from a tracer system to a chart recording is shown in Fig. 3–2a. Such a recording is deceptive, in that the asperities appear as jagged peaks. This is the result of the common practice of amplifying the vertical motion from the stylus to the chart about 100 to 1000 times more than that for the horizontal displacement of the tracer carrier. When the two dimensions are properly represented, as shown in Fig. 3–2(b), the profile seems very smooth, but it is realistic. The slopes of most profiles are in the range of 3 deg to 10 deg. Actually, most surfaces have the approximate profile of the surface of a lake swept by a gentle breeze. On such a surface at least three scales of roughness are visible.

A third objective is to process the signals from the amplifier to produce numbers that characterize the roughness of a surface. The analog signals can be processed directly, or they can be digitized before processing, either with or without compensation for the nonlinearity of the electronics. This nonlinearity results in signal amplification and phase shift that varies with the frequency of the processed signal. Each manufacturer of surface tracer equipment has its own way of compensating for these effects.

The signal to be processed may be represented as a seemingly random wavy line,

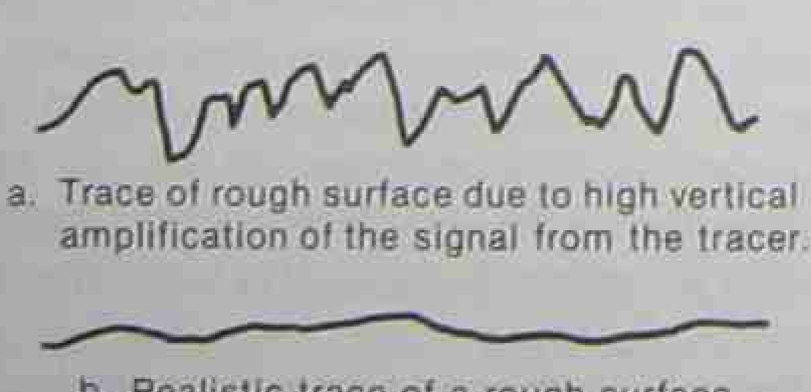


Figure 3–2 Strip chart recording from a stylus tracer system.

as shown in Fig. 3–2. The instrument "selects" a sample length, nominally about 0.030 in. or 0.75 mm.* This length should include a minimum of 30 crossings of the *mean*. The electronics are then designed to pass all wavelengths less than the sample length and *cuts off* the longer wavelengths to avoid extraneous data. Next the electronics selects a *mean line* such that equal areas are enclosed by the wavy line above and below the mean line. All vertical measurements are taken as absolute values from the mean line.

Signal processing is done in a number of ways, and more than 20 results or *surface profile parameters* have been proposed to describe a surface. These include parameters for asperity height, spacing, and slope, expressed as extreme values, average of selected values, and average of all values within a sample length. There are also some *statistical parameters* which provide a distribution of values and a measure of the randomness of such distribution. The most common parameters will now be described.

The ordinary parameters in use characterize asperity height, and they are listed with the most useful parameter first. The parameters are defined in terms of quantities shown in Fig. 3–3: y is the absolute value of a vertical dimension from the mean line, y_i indicates that each of N values of y is taken in turn, and the values p and v are the values of y of selected high peaks and deep valleys. The parameters are

$$R_q = \sqrt{\frac{1}{N}\sum (y_i)^2}$$ and is the *root mean square (RMS) average roughness height*

$$R_a = \frac{1}{N}\sum |y_i|,$$ and is the *arithmetic average roughness height*

$$R_z = \frac{(5\text{ highest peaks} + 5\text{ deepest valleys in one cutoff length})}{5}$$

and is the *ten point roughness height*

* Amstutz and Hu, "Surface Texture: The Parameters," a publication of Warner & Swasey, Dayton, Ohio, # MI-TP-003-0785.

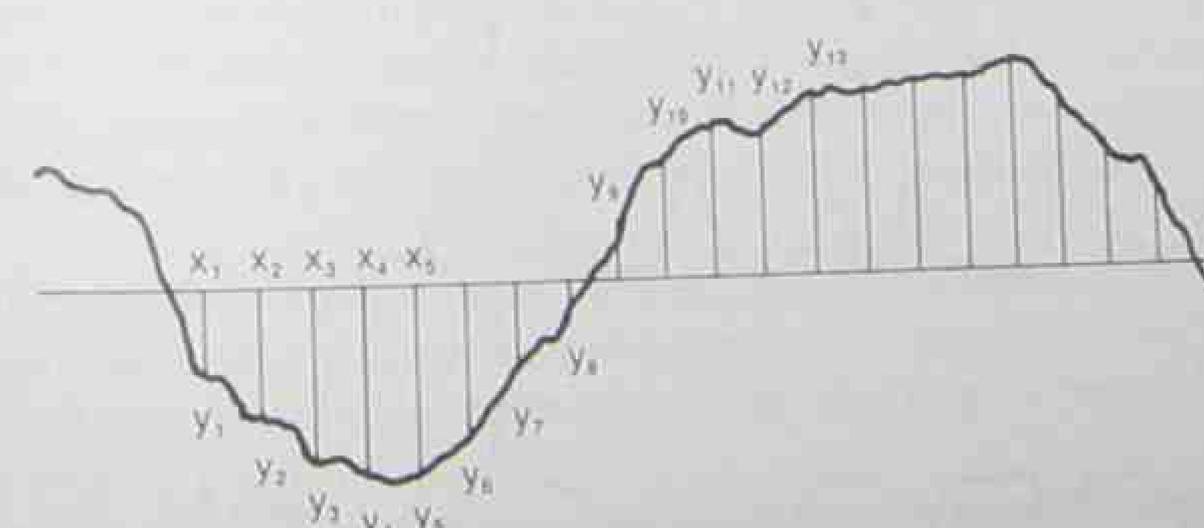


Figure 3–3 Topographic waveform and coordinates for analysis.

$R_t = p_{MAX} + V_{MAX}$ over 5 cutoff lengths, and is the *maximum height, over five cutoffs*

$R_p = p_{MAX}$, or $R_v = V_{MAX}$, and are the *maximum peak height or maximum valley depth in one cutoff*

The statistical parameters tabulate the y values at regular sample spacing shown as x values in Fig. 3–3(b). The amplitude density function (ADF) is often plotted to the right of the surface profile, as shown in Fig. 3–4. Two useful parameters relate to the shapes of the ADF curves. Figure 3–4(a) shows a Gaussian ADF: Fig. 3–4(b) shows an ADF *skewed* upward, and Fig. 3–4(c) shows an ADF that is skewed downward. The corresponding waveform illustrates the reason for the skew.

Mathematically *skew* is defined as:

$$S_s = \frac{1}{(R_q)^3} \times \frac{1}{N}\sum_{i=1}^{N} (y_i)^3$$

The ADF in Fig. 3–4(a) has a $S_s = 0$; in Fig. 3–4(b) $S_s < 0$, and in Fig. 3–4(c) $S_s > 0$. An $S_s < 0$ is said to be desirable as a bearing surface, whereas an $S_s > 0$ is preferred for electrical contacts.

A further characterization defines the peakedness or sharpness of the Gaussian curve in Fig. 3–4(a), and that is *kurtosis*. Mathematically it is

$$S_k = \frac{1}{(R_q)^4} \times \frac{1}{N}\sum_{i=1}^{N} (y_i)^4$$

A flattened curve has $S_K < 3$, a sharp one has $S_K > 3$, and a Gaussian curve has $S_K = 3$. A waveform with $S_s > 0$ automatically has $S_K < 3$ and vice versa.

The only wavelength parameter in use is the *correlation length*. It is a measure of repetitiveness of waveforms. A perfectly random waveform displays no repetition, whereas

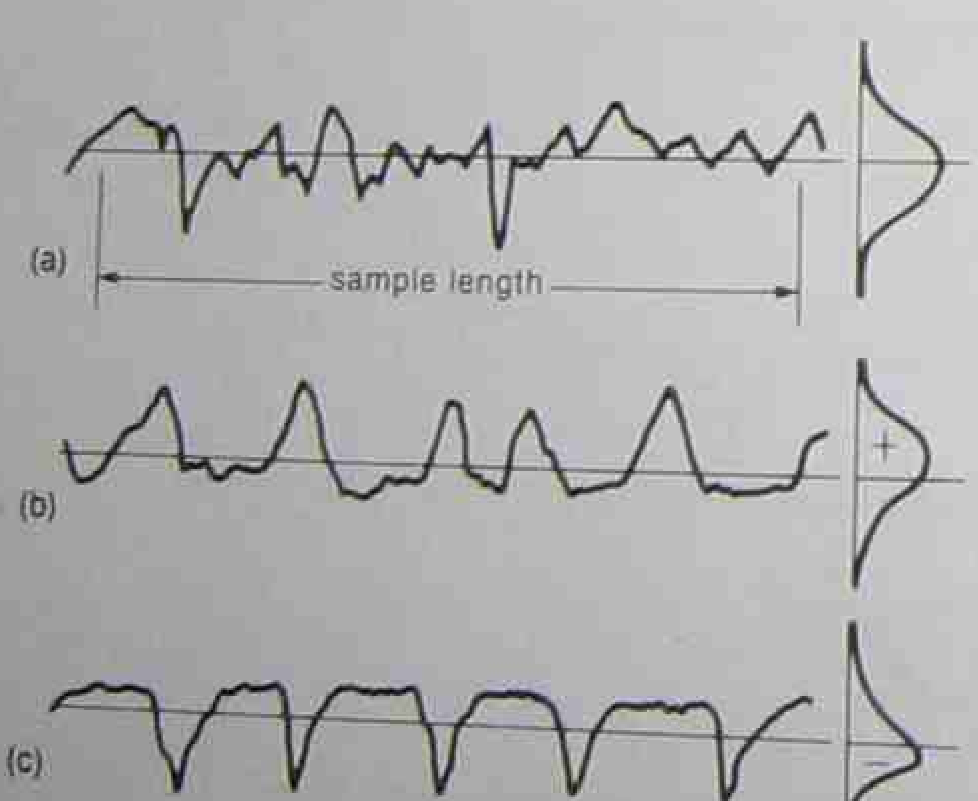


Figure 3–4 Three types of topographics and then amplitude density function, a. Gaussian, b. + skewed, and c. − skewed.

other surfaces may display a semblance of repetition every millimeter (for example) on average.

There is no advice available on which parameters one should use. The reason is that there is almost no connection between any of the parameters and the likelihood of a surface to function properly. Each corporation has different practices, probably based on the personal preference of a group leader. However, recall that surface roughness measurement is done mostly to control processes. Each process is done in a unique way in each plant, and this is the best reason to accept individual ways of characterizing surfaces. But virtually all of these methods can be improved considerably. For example, in principle, a single parameter cannot describe more than one surface feature or cannot adequately describe the shape of an ADF. At least two unique and independent parameters should be used, and perhaps as many as five.

In all cases, however, certain practical matters should be noted. In using the stylus tracer a single trace should not be taken because surface roughness is very directional, and it varies by as much as 20 percent across a surface. A minimum of five traces should be done, at least 2 mm apart and in various directions.

When working with soft materials the tracer can leave considerable damage behind. The stylus has a load applied to it, usually about 0.5 gram or 5 mN, which produces a contact stress between the stylus and sample surface sufficient to produce plastic flow of most materials. With the following equation one can calculate the contact radius a between a sphere and a flat plate by

$$a = \left\{0.75Wr\left(\frac{1-\partial_1^2}{E_1} + \frac{1-\partial_2^2}{E_2}\right)\right\}^{1/3}$$

where W = the load applied to the sphere.
∂ = the poisson ratio of the materials 1 and 2.
r = the radius of the sphere.
E = Young's modulus of the materials 1 and 2.

Most styli are diamond, and if we take the E of diamond to be four times that of the average metals, assume $r = 10\ \mu m$, and take poisson ratio of all materials to be 0.3, then $a \approx 0.00695/E$, where E is that of the surface to be traced. Now, the mean contact pressure between a sphere and a flat plate is $p_m = W/\pi a^2$, which finally is $p_m \approx 6.6E^{0.67}$. For steel the average contact stress is about 0.02 times E, which is about twice the tensile strength of the hardest steel. Clearly the stylus will scratch a groove in virtually all metals. These scratches are hardly visible on very hard surfaces but constitute defects on softer surfaces.

Another practical problem results from the mass of the stylus and mounting lever. At high rates of horizontal traversal the stylus jumps over some peaks and gives erroneous readings. Most commercial stylus tracer systems provide a range of traversal speeds of the order from about 0.1 to 1.5 mm/s.

There are several other commercial devices for measuring surface roughness. These methods include contact resistance, capacitive methods, and light scattering. Each is useful in some specialized area, but in every case the results are compared with the results of the stylus tracer. The stylus method is the standard of the world, and it was invented at the University of Michigan.

PROBLEMS

3-1. Express the surface roughness of your home county.

3-2. Explain the particular surface attributes that provide the basis for the resistance, capacitive, and light-scattering methods of surface characterization.

3-3. Give the primary, secondary, and tertiary product quality attributes for;

(a) A steel shaft being cut in a lathe.

(b) A silicon wafer being etched.

(c) A computer hard disk surface being coated with media.

(d) A plate of glass being ground.

(e) A watermelon that you are considering for purchase.

(f) A pad of paper that you are considering for purchase.

TABLE 3-1 AVERAGE RANGE OF SURFACE ROUGHNESS, R_A, ACHIEVED BY VARIOUS PRODUCTION PROCESSES, AND THE RELATIVE COST TO ACHIEVE THE CENTRAL VALUE OF ROUGHNESS GIVEN. THE FULL RANGE OF ROUGHNESS SEEN IN PROCESSED PARTS MAY BE THREE TIMES AS WIDE AS THOSE GIVEN. (TABLE 14-3 ON PAGE 386 GIVES VALUES FOR JOB SHOP OPERATIONS.)

Process	Average Range of R_a(*um*)	Relative Cost
(Reference)	50	1
Sand casting, hot rolling, flame cutting	25 to 12	≈ 1.5 to 2
Forging, sawing, planing and shaping	12 to 3	≈ 1.8 to 2.5
Permanent mold casting, extruding investment cast, die casting, cold rolling, drawing, drilling, electrical discharge milling, chemical milling, broaching milling, reaming, boring, turning	5 to 2	≈ 2 to 3
Barrell finishing, roller burnishing, electrolytic grinding, honing grinding	7 to 2	≈ 3 to 5
Polishing, lapping	3 to 0.5	≈ 4.5 to 7.5

4 tolerances

4.1 INTRODUCTION

Regardless of the method by which parts are processed, dimensions of concern *cannot* be produced to exact sizes. This is so regardless of the number of parts to be made. The use of *tolerances* is both a recognition and acceptance of that fact. Consider the following questions. If exact sizes could be produced, would that be necessary? From a *functional* viewpoint the answer is a qualified no, since most mechanisms, assemblies, and the like will perform satisfactorily if the desired *basic* sizes of individual components are bounded by permissible variations. Tolerances define such variations. It is admitted that in most cases, the smaller the tolerance, the better will be the quality of the product; however, a balance between quality and cost of the product must be considered. As shown in Fig. 4-1, costs escalate rapidly as tolerances are decreased. This leads to the next question. Is it economical to strive for *exact* sizes? Except in rare cases, the answer is definitely no.

One such rare case is the production of gage blocks, which industry uses as reference measuring standards. Essentially, each gage block consists of two flat and parallel surfaces, where the distance between these surfaces is produced to a certain basic size that constitutes the reference dimension; tolerances are maintained to a few *millionths* of an inch. No simpler geometry can be envisioned yet to maintain such exactness; the manufacturing methods necessitated plus the special measuring devices needed to verify this exactness lead to a cost of thousands of dollars for a set of such books. Because of their intended use, such close tolerances are justified, but this is an *exception* rather than a general rule. For most products, such exactness is not only unnecessary but also uneconomical in terms of putting competitive products in the marketplace. A sound rule to follow is that tolerances should always be as large as possible, consistent with func-

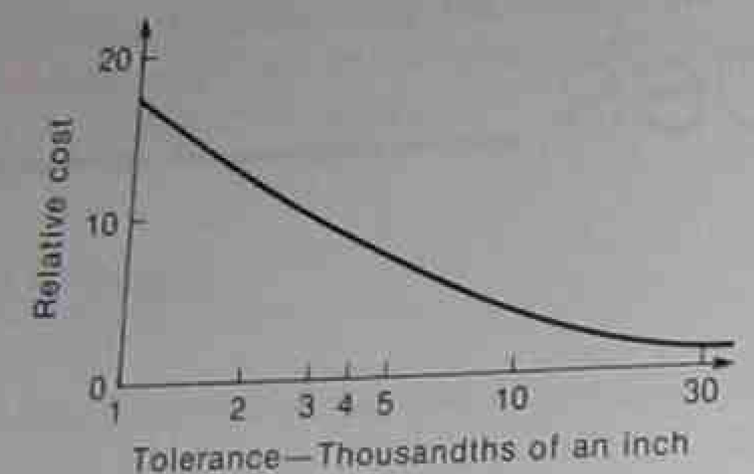


Figure 4–1 Schematic of the trend of relative cost versus specified tolerance.

tional demands. An old story, attributed to Sir Isaac Newton, is appropriate. When asked how he had his many experimental devices produced, his purported reply was that he made a freehand sketch of all dimensions to the closest inch and then turned it over to his assistant. Using a pencil and straight edge, the assistant refined the sketch and added tolerances of one-quarter inch to all dimensions. This was given to a junior draftsman who made an elegant drawing to full scale. To avoid future criticism, the draftsman assigned his own tolerances. He considered the smallest number he could think of and divided it by two. The veracity of this story is open to question, but its message is important. Designers, wanting assurance of proper function, believe that tight tolerances will achieve this aim, but this can lead to such high costs that the end product is not economically competitive.

There are three discrete attitudes involving tolerance specifications. The first concerns the *designer,* whose goal is to insure proper function. Next are those who must see that the item is *manufactured;* their goal is to produce the item as economically as possible. Finally come those responsible for *assembling* individual components into a unit; their main concern is to complete the assembly without problems. It is most sensible and economically efficient if these different viewpoints are considered *concurrently* during the initial design stage; otherwise, problems may often arise and redesigning for efficient manufacturing and assembly will be required.

4.2 THE DESIGN ASPECT

In certain situations, such as press fits or the allowable clearance between a shaft and journal bearing, proper tolerances can be calculated, but these are exceptions rather than a general rule. For the overwhelming majority of cases, there seems to be no *scientific* approach used to determine appropriate tolerances. The authors of this book have for many years sought an answer to the question of *how* tolerances are determined, and the conclusion drawn is that *past experience* provides the basis for such a decision.

Considerations of size, weight, strength, and the like are often studied initially. As an example, suppose that a gearbox must fit into a certain volumetric space and is to transmit certain loads or torques over a range of speeds. Maximum levels of noise may be of concern, in addition to strength and allowable deflections. Using equations related to

gear design, stress analysis, beam theory, and so on, an initial design that satisfies functional demands can be accomplished in a *reasonably* direct manner. Then tolerance specifications must be considered, and several questions arise here.

1. To satisfy noise levels, what surface finish and, in essence, what tolerances are required for the gears?
2. Where are the best locations for any bearings to satisfy deflection specifications? By what extent can these locations vary?
3. If shafts are supposed to be parallel, to what extent can their centerline distances vary and still provide proper meshing of gears?

The answers to such questions are not simple, so it is not surprising that a designer faces a degree of uncertainty in arriving at tolerance assignments. In numerous cases, some of the components to be used in the gearbox may be purchased from other companies who specialize in the manufacture of such items; in the above situation, bearings are an obvious example. Proper sized bearings might be selected from a vendor's catalog where the bearing tolerances are specified. A similar comment could apply to the gears. Thus, the designer must often consider the basic dimensions and manufactured tolerances of such purchased items in the overall design.

One final observation pertains to those situations where a finished product could be made by practically *hand-fitting* components as assembly proceeds. For example, a craftsman could make his own bolt with his own version of a thread, and then make an individual nut to fit that bolt, but think of the problems that would arise if replacement parts are needed. To avoid such a problem, national and international standards have been developed for thread sizes as well as many other engineering components. Even then, control of tolerances is demanded if any bolt of a certain thread size is to mate directly with any nut of the same size. This leads to the concept of *interchangeable assembly,* which is the prerequisite for mass production. Consider as an example the cost of a Rolls-Royce, made in rather limited quantities, compared with a Chevrolet which is mass-produced. The Rolls involves more costly parts in the engine, transmission, body, and trim, and much tighter control of mating parts is maintained (almost like hand fitting). Because of these higher quality standards, it is no wonder that the Rolls costs an order of magnitude more than the Chevrolet.

4.3 THE MANUFACTURING ASPECT

Once a product is designed, those who specialize in manufacturing give consideration to selecting processing methods that might be used to make the product. Economic considerations are *always* involved at this point. Most often the initial design includes a specification of the materials to be used as well as the basic dimensions and tolerances. Not all materials can be processed with equal ease, and very close tolerances require high costs, as shown in Fig. 4–1. Because of these facts, the manufacturing personnel will often raise questions with design people in order to be convinced that the original specifications are really essential. It is possible that the designer was unaware of the potential problems (and excessive manufacturing costs is one of the crucial concerns) as seen by the manufacturing engineer. Often a modification of the original specifications is made that does not impair functional demands yet leads to lower production costs. Such give and take, involving both viewpoints, leads to an acceptable compromise.

4.4 THE ASSEMBLY ASPECT

After components are manufactured they are usually combined with other items to form an assembly.* Perhaps the major concern of those involved at this stage is that randomly selected parts should assemble on the first attempt. That is what *interchangeable assembly* is all about, and a careful analysis of the dimensions and tolerances of individual items is required if assembly concerns are to be satisfied.

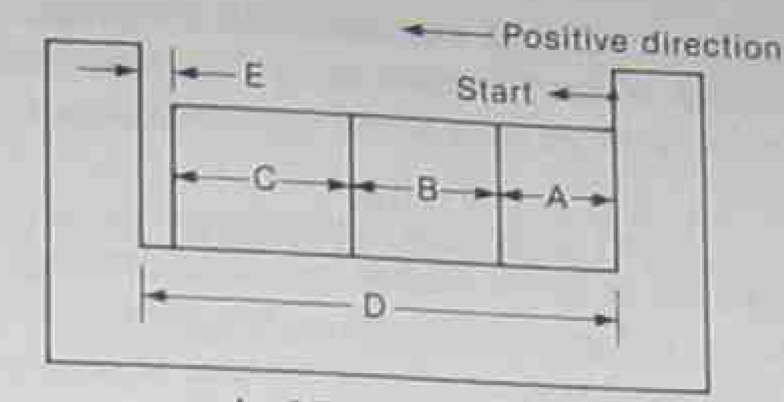
Three distinct approaches can be considered. The first is termed *complete* or *100 percent* interchangeability, where assembly will result regardless of the combination of components chosen at *random* for each assembly. *Statistical* interchangeability is the second approach. There it is *probable* that assembly will result, but 100 percent assurance is not guaranteed. Finally, with *selective assembly*, the critical dimensions of components are first measured; then the components are segregated into groups or batches of particular size variations. From these batches, pairs of components are selected so that desired tolerance limits are satisfied. The initial preinspection or measurement plus the need to categorize components into pertinent batches make this approach more costly than statistical interchangeability. Selective assembly is used only where close mating tolerances are deemed essential. With appropriate examples, each approach is now discussed. In *all* of the examples cited, the basic sizes and tolerances carry units of inches. This negates the need for constantly repeating units in this chapter and has no effect on the concepts. If SI units are preferred, all basic sizes, tolerances, and answers can be multiplied by 25.4 to convert to millimeters.

4.5 100 PERCENT INTERCHANGEABILITY

Consider Fig. 4–2, where three blocks, A, B, and C are to be assembled in a channel of dimension D. Except for the tolerances to be assigned to D, all other basic sizes and tolerances are shown. It is emphasized here that for this entire chapter we shall always use *equal, bilateral* tolerances;† the reason will become evident as various examples are

* Throughout this chapter the words *component*, *item*, or *part* indicate individual items and are used interchangeably. Simiarly, the words *assembly* or *unit* denote a combination of single items.

† Note, for example, that $1.000^{+0.004}_{-0.000}$ shows a unilateral (one-sided) tolerance, whereas $1.000^{+0.003}_{-0.001}$ indicates unequal, bilateral tolerances. In contrast, $1.000^{+0.002}_{-0.002}$ gives equal, bilateral tolerances. For each case, the *basic or nominal* size is 1.000.



A = 0.750 ± 0.003
B = 1.000 ± 0.005
C = 1.125 ± 0.004
D = 2.894 ± X
E = 0.005 (minimum)

Figure 4–2 Schematic assembly of three blocks in a channel; all basic dimensions and tolerances are in inches.

presented. For a detailed coverage of other ways to assign tolerances as well as many aspects of dimensioning, the reader can consult the text by Spotts.*

Suppose it is essential that the minimum gap size E must *never* be less than 0.005 and this is the only restriction involved. It is then necessary to determine the tolerances that must be assigned to D. Because of the simplicity of this problem it is possible to state the answer directly; it is 0.002. However, with more complicated situations, a systematic approach provides greater efficiency, so that approach is introduced here. We employ what is called the *path equation*. Usually, two decisions are required. The first is to determine what is the most extreme combination of limiting conditions that could be encountered. Then one must decide how the various surfaces involved must contact each other once the limiting conditions are determined. For the assembly in Fig. 4–2, the extreme conditions occur when A, B, and C have their *maximum* possible sizes and are assembled in a channel where D is the *smallest* possible size. Tolerance X can then be found. The path equation can be started at *any* contact point between the various items or at a free surface corresponding to any of the dimensions involved. Positive values in the path equation are chosen arbitrarily in either of two directions. As shown in Fig. 4–2, the equation has been chosen to start at the interface between A and the right side of the channel, and moving left is considered positive. Using the limiting conditions, we find that the path equation becomes

$$(0.750 + 0.003) + (1.00 + 0.005) + (1.125 + 0.004) + (0.005) - (2.894 - X) = 0 \quad (4\text{-}1)$$

so $X = 0.002$. This means that $D = 2.894 \pm 0.002$ *since bilateral tolerances* were used. Although not a specified requirement, the maximum value of E occurs when an assembly consists of the smallest sizes for A, B, and C and the largest size for D. Start the path equation at the left-hand side of the channel and use movement to the right as positive; then

$$2.896 - 0.747 - 0.995 - 1.121 - E = 0 \quad (4\text{-}2)$$

* M. F. Spotts, *Dimensioning and Tolerancing for Quantity Production* (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1983).

thus $E = 0.033$. If any assembly consisted of a combination of extreme dimensions, E would vary from a minimum of five to a maximum of thirty-three thousandths of an inch. The *average* value for E, from the average sizes of the four components, is 0.019; thus

$$E = 0.019 \pm 0.014$$

Now consider a case where it is not a simple matter of arriving at an obvious answer. The components are shown in Fig. 4–3(a); both are flat plates, one containing two pins and the other two holes. Assembly involves fitting the pins in the top plate into the holes of the lower one. The problem is to assign tolerance X to the centerline dimensions so that assembly will always occur. The extreme combination of dimensions results when the pin dimensions are maximum, the holes minimum, one centerline dimension is maximum and the other minimum; Fig. 4–3(b) illustrates this extreme case. The path equation is started arbitrarily at the centerline of the left pin, and moving left is considered negative. Then

$$-(0.500 + 0.001)/2 + (0.505 - 0.002)/2 + (2 - X) + (0.380 - 0.002)/2 - (0.375 + 0.001)/2 - (2 + X) = 0 \quad (4\text{-}3)$$

Note here that the full tolerance X enters into the equation, whereas only one-half of the tolerances on pins and holes pertains. This should be kept in mind when this problem is reanalyzed in the next section.

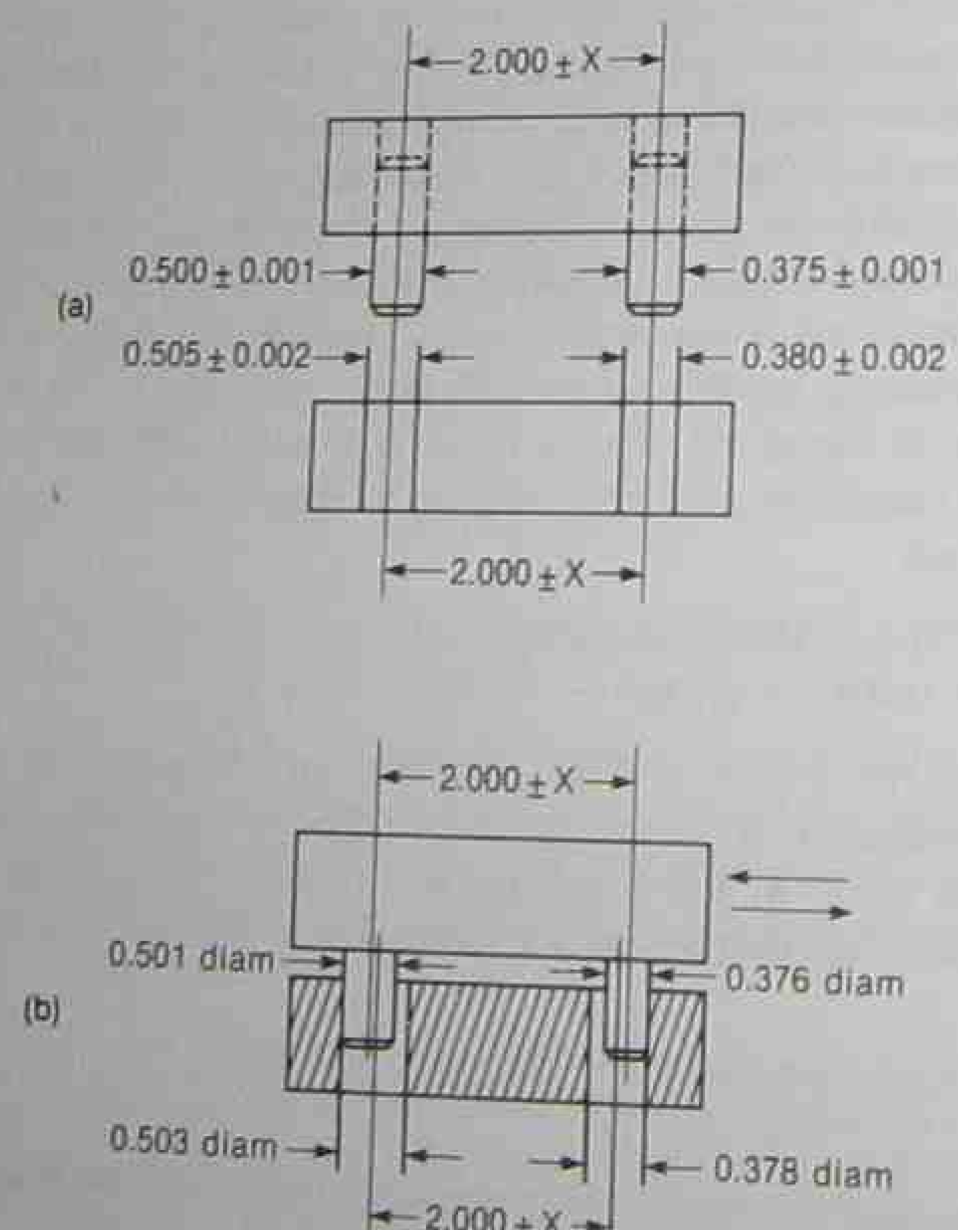


Figure 4–3 (a) Sketch of two components prior to assembly and (b) an assembly using one combination of extreme conditions.

Equation (4–3) reduces to $X = 0.001$, so the centerline dimensions are specified as 2.000 ± 0.001. In problems such as this, the path equation is most useful.

4.6 STATISTICAL INTERCHANGEABILITY

With regard to the two examples just discussed, it is appropriate to ask how probable it would be that one is likely to encounter a group of dimensions having extreme values occurring in any one assembly. The answer is not very likely, even though such a possibility *does* exist. In fact, as the number of items in a given assembly increases, the probability of such an occurrence can be shown to approach zero. For these reasons, the concepts of probability theory and the use of statistical methods for assigning tolerances should be considered since this leads to less stringent tolerances and lowered manufacturing costs. Rather than assuring positive assembly as discussed in Sec. 4–5, we are taking a chance that some assemblies will not be possible, but the *odds* are greatly in our favor that this is not a likely result.

In this book, our major intent is for the reader to gain an appreciation of why this approach is useful. We assume *no* prior background in statistical theory, nor do we expect that complete expertise will follow the minimal coverage to be presented. In addition, we point out that statistical concepts beyond those used here are available and used. But they are far beyond what is necessary for this basic introduction.

Although the type of problems of concern here will not require detailed calculations of certain basic statistical parameters, such parameters must be introduced if a physical understanding is to be meaningful. First is the term *average*. This is defined as the summation of the total number of individual measurements, regardless of what is being measured, divided by the total number of individual measurements. It is an *indication* of *central tendency*, or of the measurement that occurs most often. In equation form it is

$$\bar{X} = \sum X/N \quad (4\text{-}4)$$

where $\bar{X}$ is the average, $\sum X$ is the sum of all individual measurements, and N is the number of measurements.

The second parameter is called the *standard deviation*. This is related to the spread or dispersion of individual measurements both greater and less than the average. It carries the symbol σ and is defined as

$$\sigma = \sqrt{\sum (X - \bar{X})^2 / N} \quad (4\text{-}5)^*$$

In words, the difference between each individual measurement and the average is squared; these are all summed, divided by the number of individual measurements, and

* In later chapters, σ is used to denote stress, but since each chapter is self-contained, confusion should not result.

then taken to the one-half power (square root). We note that the term σ^2 is called the *variance*.

Suppose a part is assigned a certain basic size and *bilateral* tolerance. In producing many such parts, those responsible for manufacturing will *aim* at the basic size so that parts both larger and smaller than basic will still fall within the acceptable tolerance range. It is likely, therefore, that more items will have a final size closer to basic than would those near the extremes of the tolerance range. If a plot were made of the size versus the number of times it occurred (that is, a *frequency distribution*), the results might look like Fig. 4–4, where the points are then connected and described by some continuous curve as shown. The most widely used curve is called a *Gaussian* or *normal* curve and is the only one we shall use in this regard. The distribution of occurrences described by such a curve is called a *normal distribution*. Distributions other than normal can occur, but to become involved with them is beyond our aims. It is also added that a normal curve is described by a particular equation that provides a plot from minus to plus infinity. The equation itself is of little interest here; rather, it is the *area* beneath the curve that is of importance. The total area under the curve and bounded by the abscissa is considered to have unit area, while the area bounded by any two *discrete values* of the measured quantity is a fraction of one (that is, the total area) and indicates the probability or likelihood that an individual value will fall between the two discrete values chosen. Consequently, the shape of the curve (see Fig. 4–4) shows that there is a greater probability that values closer to the average will result as compared with values further from the average. Since the standard deviation is a function of the individual dimensions (i.e., X), we can superimpose units of σ as shown in Fig. 4–4. With a normal distribution, the area under the curve between limits of $\pm\sigma$ from the average encompasses about 68 percent of the *total* area from minus to plus infinity while limits of $\pm 3\sigma$ include over 99 percent of the total area. To be more precise, the probability is 99.73 percent, which means that if 10,000 measurements are made only 27 out of that entire number would be expected to fall outside the 3σ limits. So, although this could happen, the probability is extremely small. For our purposes, all bilateral tolerances will be related directly to $\pm 3\sigma$ limits; this is the reason that bilateral tolerances were chosen earlier.

Before we use the above ideas for specific problems, there are certain assumptions to be added.

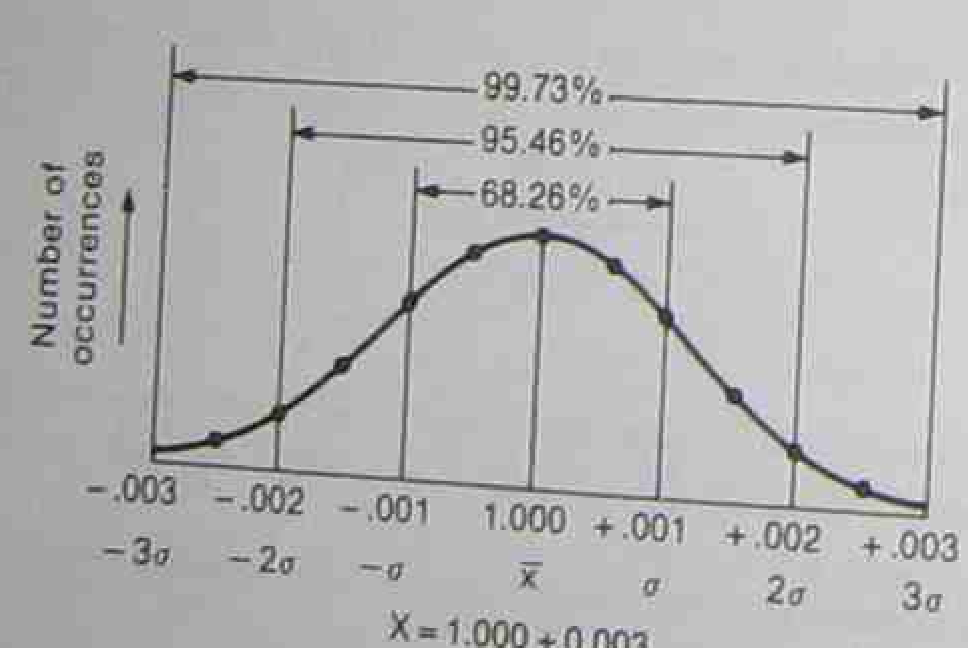


Figure 4–4 Illustration of a normal distribution showing the spread of tolerances and the corresponding multiples of the standard deviation σ.

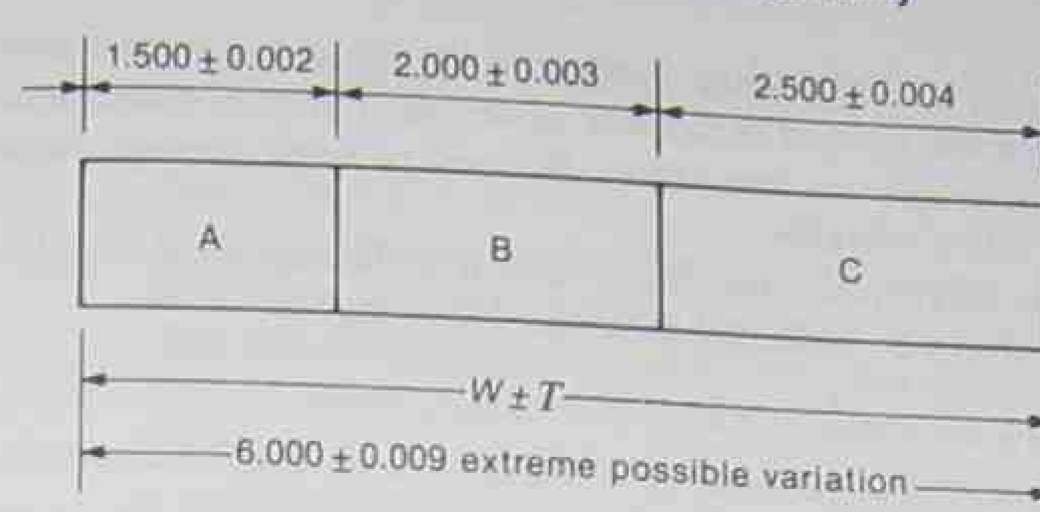


Figure 4–5 Stack of three blocks showing the extreme dimensions of the overall stack.

1. For *each dimension* involved, the variation of sizes displays a normal distribution.
2. The occurrence of the dimension of any individual component, due to the manner of manufacture, is *completely independent* of the individual dimension of any other component.
3. The selection of individual components to be assembled together is completely *random*.
4. The basic size of each component is equal to the average of all such parts, and the *total* tolerance spread is equal to 6σ for each individual part and its resulting distribution. Thus if a bilateral tolerance is given as ± 0.003, then 6σ equals 0.006.

As a first example, consider Fig. 4–5, where parts A, B, and C, whose basic sizes and tolerances are shown, are to be stacked together in an assembly whose basic size is shown as W with a tolerance of $\pm T$. The extreme variations that could *ever* occur fall between 6.009 and 5.991 and for that possibility, $W = 6.000$ and $T = \pm 0.009$. How likely is it that these extreme combinations might occur? This is where the idea of probability enters, and Fig. 4–6 will assist here.

Since each component displays a normal distribution and assembly is random in nature, there is no reason to expect that if a component of A happens to have a dimension less than average, it will be assembled with parts B, and C, having dimensions also less

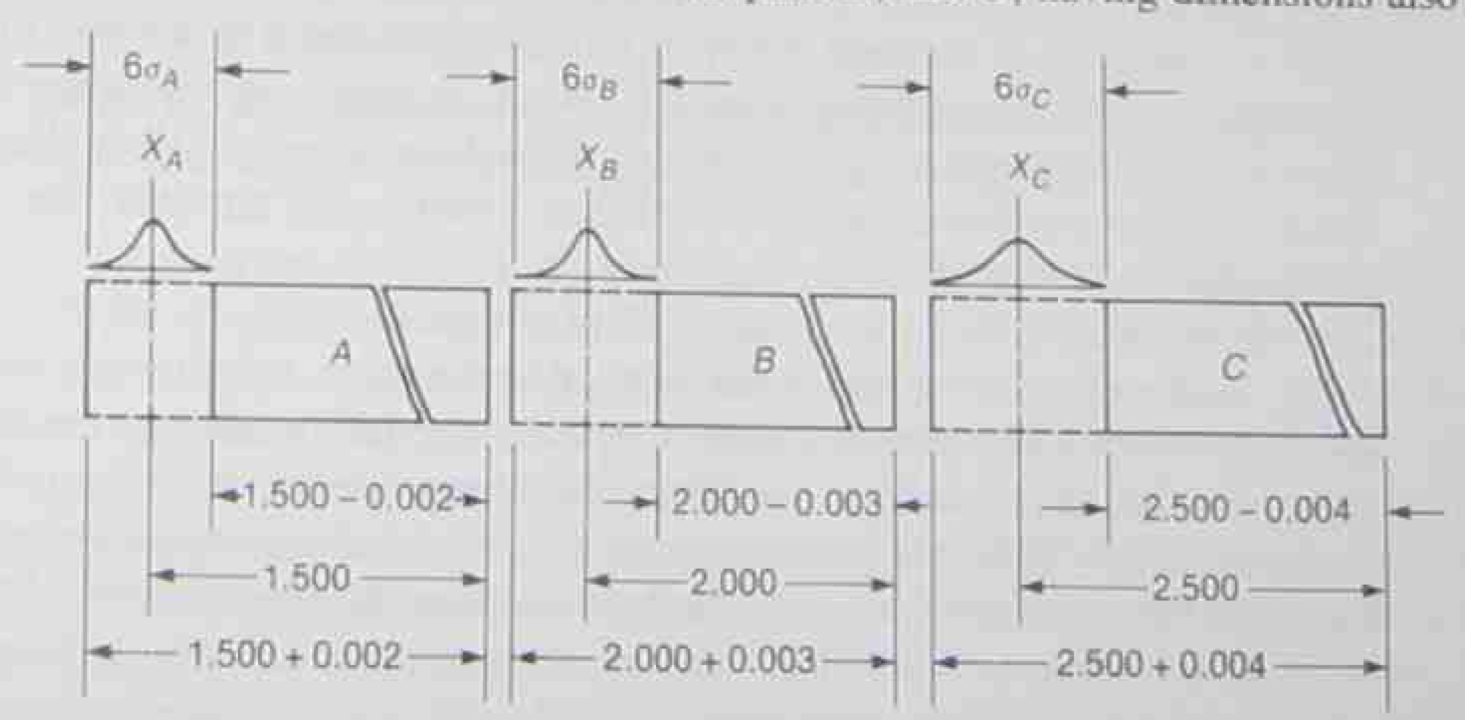


Figure 4–6 The three individual blocks from Fig. 4–5 showing the 6σ spread for each block.

than average. A mix of A, B, and C, where some dimensions are above and some below average, is more likely. Because the greatest probability of the occurrence of dimensions for each individual part is close to their average values, the most probable average value of W is the sum of those individual averages.

Here, A, B, and C are the independent variables, while W is the dependent variable and the probable variation of W can be found by using the following,*

$$\sigma_u^2 = \sigma_A^2 + \sigma_B^2 + \sigma_C^2 \tag{4-6}$$

where σ_u is the standard deviation of the assembly or unit and the other three relate to the components. Recalling that the total tolerance spread equates to 6σ, we revise Eq. (4–6) to

$$(T_u/6)^2 = (T_A/6)^2 + (T_B/6)^2 + (T_C/6)^2 \tag{4-7}$$

or simply

$$T_u^2 = T_A^2 + T_B^2 + T_C^2 \tag{4-8}$$

where T is the tolerance of each component and the overall unit, respectively.

This form will find major use and implies that the distribution of assembly measurements (i.e., W) will also display a normal curve. Returning to Figs. 4–5 or 4–6, and using Eq. (4–8), we have

$$T_u^2 = (0.002)^2 + (0.003)^2 + (0.004)^2 \tag{4-9}$$

so $T_u = 0.0054$. Thus from a statistical viewpoint W would vary as 6.000 ± 0.0054 rather than 6.000 ± 0.009 in., as found earlier.

Now let us reassess the situation connected with Fig. 4–2. There, the average size of the three parts when they are stacked together is 2.875, and since the average channel size is 2.894, the average value for E is the difference, which is 0.019. With the minimum value for E of 0.005, the 3σ spread is ± 0.014, which is the dependent variable; therefore,

$$(0.014)^2 = (0.003)^2 + (0.004)^2 + (0.005)^2 + X^2 \tag{4-10}$$

so $X = 0.012$ and D becomes 2.894 ± 0.012. Recall that X was found to be 0.002 for 100 percent interchangeability from Eq. (4–1), so assigning the larger tolerance has obvious economical advantages. Note that *if* the three largest possible blocks were combined with the smallest channel, interference would result; however, the probability of this happening is remote.

Next consider Figs. 4–3(a) and 4–3(b), where from Eq. (4–3), X was found to be 0.001. There are six independent variables, the tolerances, but note that only *one-half* of the values assigned to pins and holes applies in Eq. (4–3), whereas the full value of X pertains. When all dimensions are basic, the average clearance between either pin-hole combination is 0.005; that is the value of the possible horizontal translation. Since the minimum clearance must be zero, the 3σ value here is taken as 0.005. Then, with Eq. (4–8) there results

* This is given without proof. For a derivation, see M. F. Spotts, *Dimensioning and Tolerancing*.

$$(0.005)^2 = \left(\frac{0.001}{2}\right)^2 + \left(\frac{0.002}{2}\right)^2 + \left(\frac{0.001}{2}\right)^2 + \left(\frac{0.002}{2}\right)^2 + X^2 + X^2 \tag{4-11}$$

from which $X = 0.0033$ as compared with 0.0010 from Eq. (4–3). These introductory examples illustrate the benefits derived, in terms of larger tolerances, if one is willing to take the small risks involved. It is again stressed that if normal distributions do not exist, then σ is not equal to $T/6$ and different, but similar, factors must be used. Yet, even when the individual components do not display normal distributions, if enough components are involved in an assembly, the assembly dimension itself will show a distribution that approaches a normal one. Finally, we note that industrial concerns often use a multiplying factor as a "factor of safety" on tolerances to reconcile theory and practice, so complete reliance is not placed upon the results predicted by statistical procedures.

4.7 SELECTIVE ASSEMBLY

Occasionally, functional requirements are so demanding in regard to tolerance specifications that the method of selective assembly is used. As an introduction, assume that a shaft is to mate with a hole, that the minimum and maximum clearances have been specified by the designer, and that this restricted range of clearances *must* be maintained. To avoid possibly excessive manufacturing costs, the tolerances assigned to each component are of such a magnitude that random assembly would lead to many instances where the allowable clearance would not be attained. The following procedure is then necessary.

1. Establish the minimum and maximum mating sizes.
2. Compare the positions of each component relative to the basic size of each part.
3. Determine the interval size to be used for categorizing each part as to selective measurement. This is found by taking one-half of the difference between the maximum and minimum allowable variations.
4. Inspect *every* part and place each part in a particular group denoting maximum size variation of that group.
5. Select a pair of components from their respective groups such that the allowable variations in any individual assembly will not be exceeded.

Suppose the *clearance* between any shaft and hole must be within 0.001 to 0.003 in. In Fig. 4–7, the dimensions of the hole and shaft are given as 2.002 ± 0.002 and 1.998 ± 0.002, respectively, and the extreme clearances must be between 0.001 and 0.003 for satisfactory function. The interval size, as noted in step 3 above, is 0.001; this is shown on Fig. 4–7. Note that with random assembly many pairs of individual components, when assembled, would not only violate the stated requirements but could lead to interference (that is, no clearance). However, after the hole and shaft diameters are *inspected*, each is assigned to a particular group; these are indicated as A through D, as shown in Fig. 4–7. Then the assembly of *any* pair of components from similarly labelled groups will positively satisfy the acceptable clearance range. For example, two parts from group A would

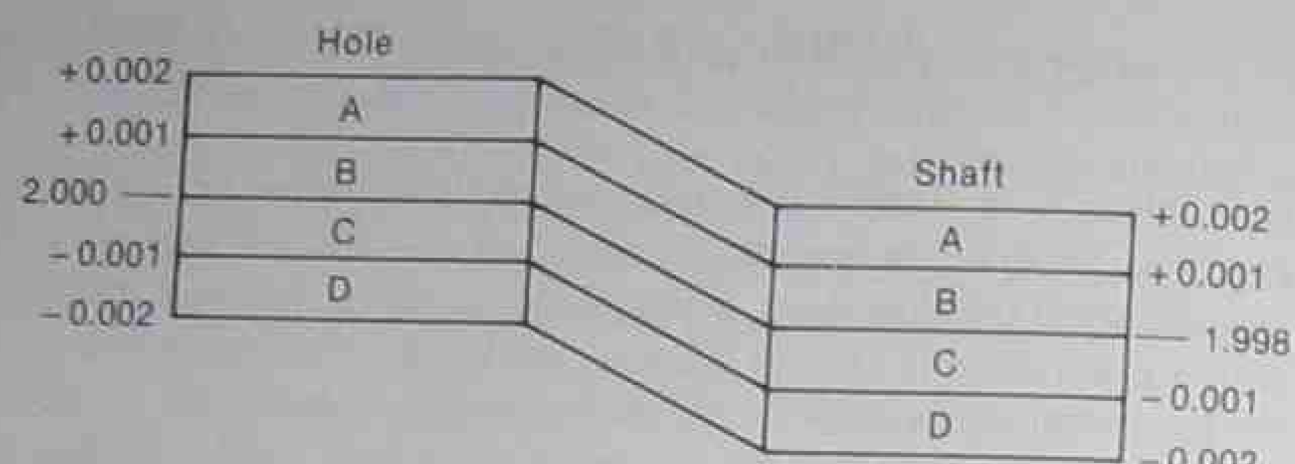


Figure 4–7 A selective assembly system for combining shafts and holes that insures a desired limit of clearance between any appropriate pair of components.

provide a maximum clearance of 2.002 − 1.999 or 0.003, whereas the minimum clearance is 2.001 − 2.000 or 0.001. Similar calculations pertain to the other three groups.

We note that the allowable tolerances on both parts could be increased in the above example and a greater number of groups, still having an interval of 0.001, would be used. It should also be realized that the added costs associated with the inspection of *all* individual parts, plus the need to segregate and store parts in specific groups, introduce expenses greater than those connected with statistical interchangeability. Therefore, a careful cost study should be made before it is decided to use selective assembly.

4.8 TRANSFER OF REFERENCE SURFACES

When a part is designed and dimensioned, the process by which the part is to be produced should be considered. For this discussion, assume that the various surfaces involved are to be *machined** to final sizes. The workpiece must be placed on particular surfaces of the machine tool and then clamped in place before operations can be performed. From these locating surfaces, the various tools are referenced; then the necessary cutting operations follow. If the reference surfaces selected for part location on the machine tool differ from those specified by the designer, then a *transfer* of references must take place. Inevitably, at least some of the tolerances available for the machining operations will be *less* than those specified by the designer. A cardinal rule to follow if such a transfer is necessary is that the sum of new tolerances that affect a dimension must not *exceed* the tolerance assigned to the original dimension. A few examples that illustrate this concept are now presented.

For the piece shown in Fig. 4–8(a), the original design specifications are shown. In locating the part on the machine, suppose surface *A* is first machined and surfaces *B* and *C* are to be finished by using *A* as the reference surface; thus dimensions *AB* and *AC* must be considered. To maintain the original tolerance of 0.002 in. between *B* and *C*, the *sum* of the tolerances on dimensions *AB* and *AC* cannot exceed that value when the transfer occurs. One solution is shown in Fig. 4–8(b); this will satisfy the original specifications, but notice the reduction of initial tolerances that results.

* Discussed in Chapter 9.

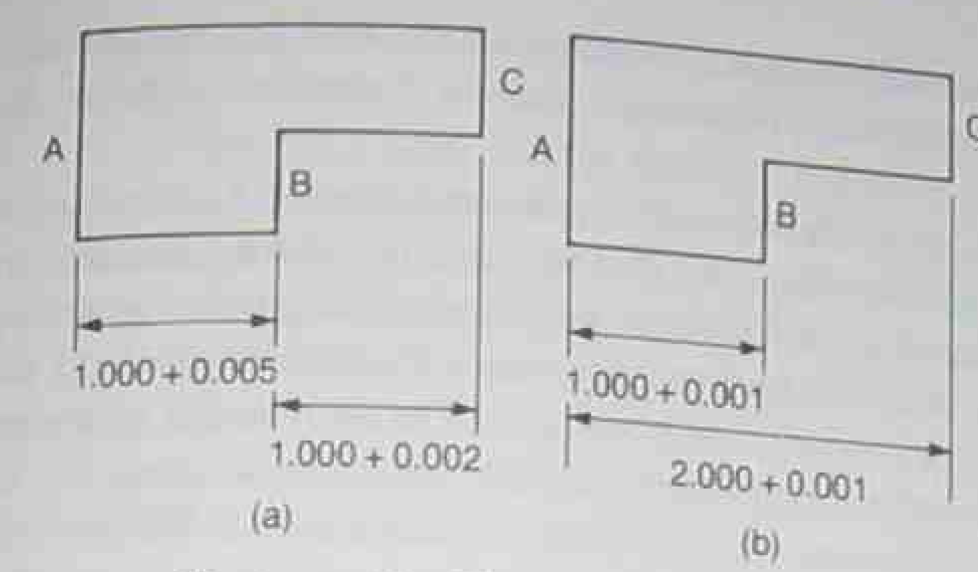


Figure 4–8 Transfer of references from initial dimensioning (a) where surface *A* is chosen as the reference surface for all dimensions (b).

Next consider Fig. 4–9(a), which shows the design specifications. If *A* is chosen as the reference surface, all new specifications are transferred with respect to *A*. The original tolerance on *BC* is the *smallest* and therefore is the one to consider *initially*. The transferred tolerances on new dimensions *AB* and *AC* cannot exceed the initial tolerance of ±0.002 on *BC*; the value of ±0.001 on each of the new dimensions satisfies that constraint as shown on Fig. 4–9(b). Now consider the next smallest initial tolerance of ±0.004 on *CD*. This dictates the sum of the new tolerances on *AC* and *AD*, and since *AC* carries a value of ±0.001 from above, the new maximum tolerance on *AD* is ±0.003. Finally the maximum tolerance on *AE* is given as ±0.002, since in combination with the

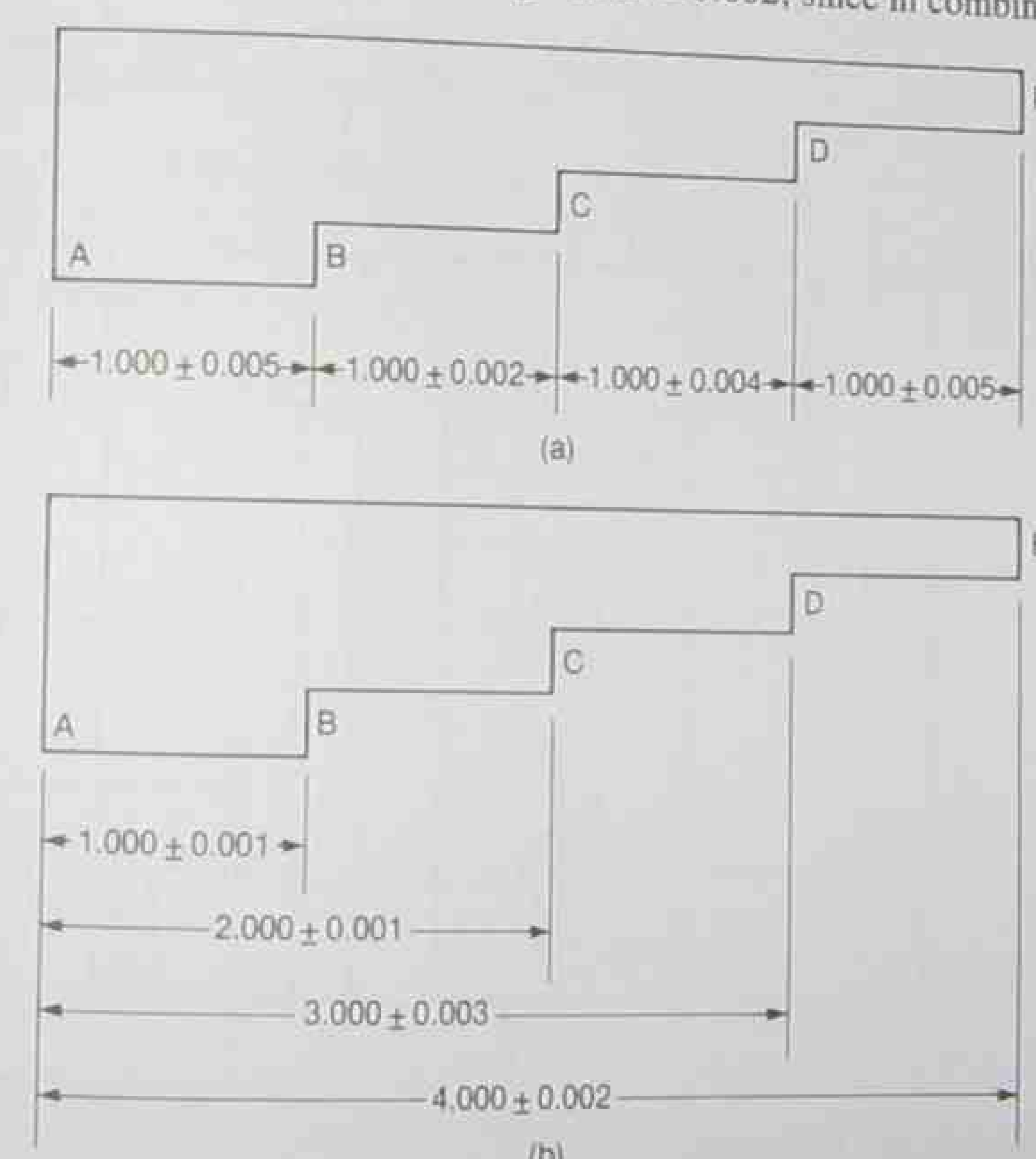


Figure 4–9 (a) Original dimensions and (b) new dimensions as transferred with *A* as reference surface.

value used on *AD*, it will not exceed the original tolerance of ±0.005 associated with *DE*. Note that the basic sizes of the transferred dimensions are in agreement with the design specifications on Fig. 4–9(a).

In many cases, a choice of reference surfaces is available, and each should be studied to provide an optimal situation. Fig. 4–10(a) shows initial dimensions and tolerances of a plate that is to contain three holes to be machined. Using hole centerlines as references for other dimensions is not the best practice, since the holes themselves do not exist originally. Suppose surface *B* is finished first and is then to serve as the reference for all other dimensions. Figure 4–10(b) shows one acceptable set of dimensions and tolerances. Since the smallest initial tolerance, ±0.002, was placed on the 0.500 dimension, this must be shared by new dimensions *C* and *D*; each carries a value of ±0.001 as shown. Now the maximum tolerance that could be applied to *E* is ±0.003 to meet the original value of ±0.004 on the 1.000 dimension, but this would then require a tolerance

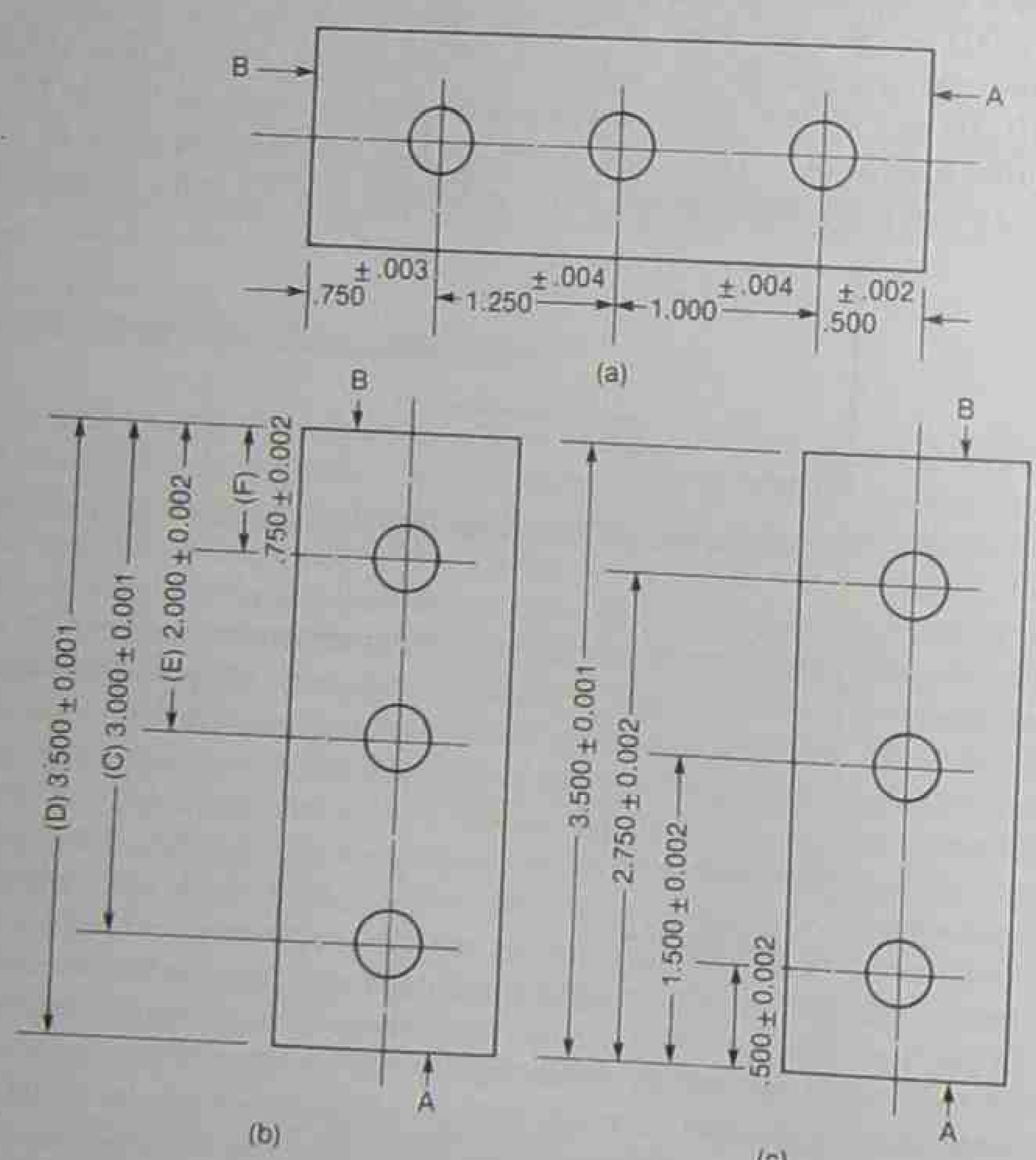


Figure 4–10 An illustration of transfer of references showing two possible results where (b) *B* is used as the reference surface and (c) where *A* is used as the reference; initial dimensioning is shown in (a).

of ±0.001 on *F*. It would be more sensible in general to share the available possibilities and use values of ±0.002 on both *F* and *E*, as shown in Fig. 4–10(b).

What if *A* were used as the basic reference surface? One combination of transferred dimensions and tolerances that satisfied the initial specifications is shown in Fig. 4–10(c). Several points are noted. If *A* is used as the reference, the *sum* of all individual tolerances is greater than that which results when *B* serves as the reference, that is ±0.007 versus ±0.006. Finally, if a particular individual machining operation requires a larger tolerance than others (that is, it is more difficult to produce the desired dimension within restricted limits), it is always possible to revise the tolerances on the transferred dimensions to handle such a situation as long as the *sum* of any two new tolerances does not lead to possible dimensional variations that would not satisfy initial specifications.

As a final comment, designers must avoid the use of redundant dimensions otherwise, problems are apt to occur. Fig. 4–11 illustrates such a situation. In Fig. 4–11(a) only

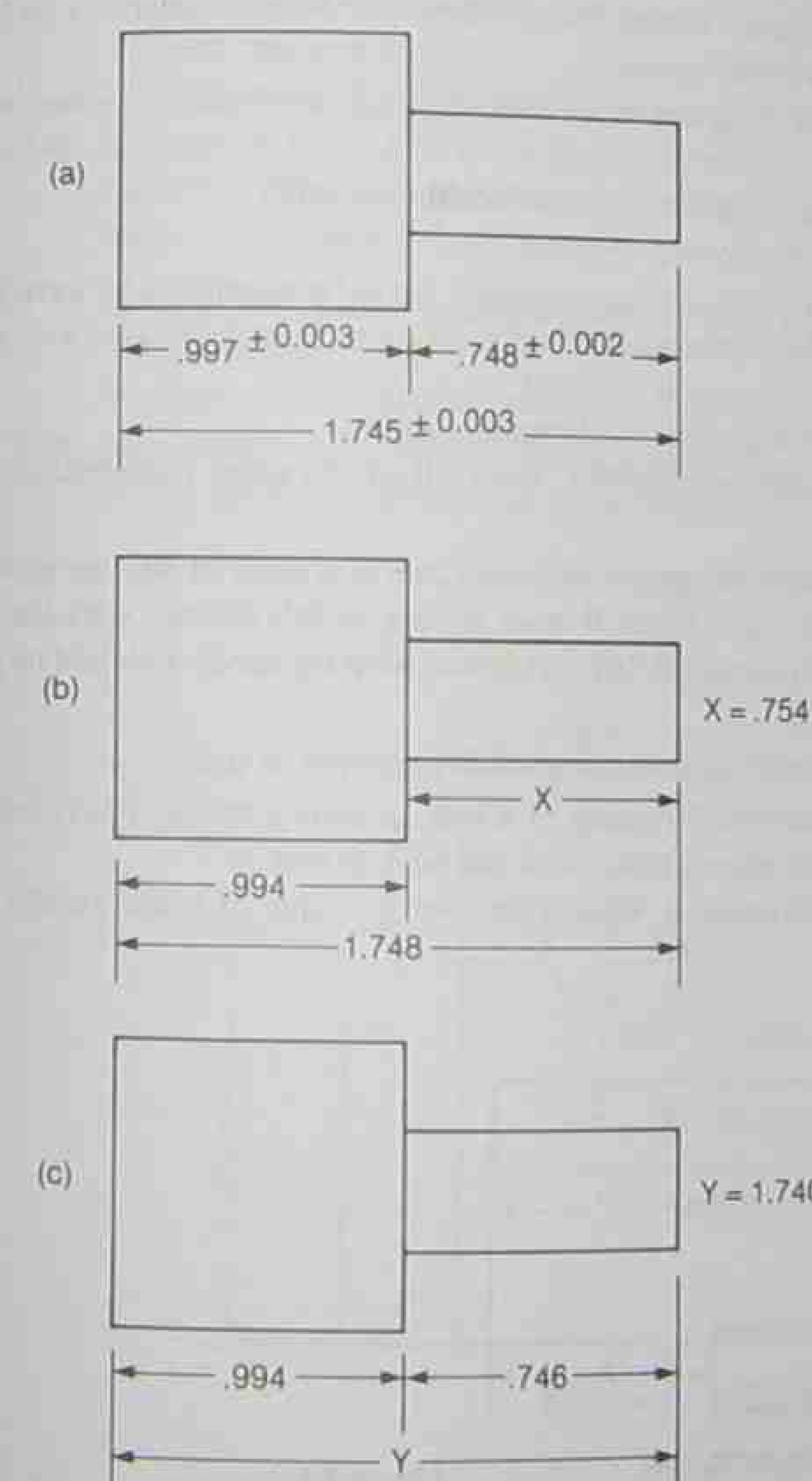


Figure 4–11 An illustration of redundant dimensioning (a), with two possible consequences shown in (b) and (c).

two basic sizes and tolerances should be specified, or the possible results in (b) or (c) could occur. The manufacturing people could certainly argue that they have correctly followed the original requirements in terms of two of the three dimensions, and, in fact, it is not always possible to satisfy all three dimensional specifications simultaneously.

PROBLEMS

4-1. Three blocks are to be assembled as shown in Fig. P4-1.
 (a) Determine the average size and the bilateral tolerance of dimension *A* that results, based upon the combinations of extreme possible conditions.
 (b) Using a statistical approach, determine the bilateral tolerances expected for dimension *A*.

4-2. Refer to Fig. P4-1. If blocks 1 and 3 are dimensioned as in Problem 4-1 and dimension *A* varies from 1.608 to 1.592, what tolerance must be placed on block 2 to satisfy any combination of worst conditions? Repeat this problem if the bilateral tolerance on block 2 is determined by using a statistical approach.

4-3. Refer to Fig. 4-2 in the text. Suppose dimensions *A*, *B*, and *C* remain as shown, while *D* is given as 2.899 ± 0.005.
 (a) Determine the extreme variations of *E* that could ever result.
 (b) Repeat part (a) from a statistical viewpoint.

4-4. Figure P4-4 shows a pair of typical components that are to be assembled by having the pins fit into the mating holes. The tolerances to be assigned are to be in the ratio of $x = y/4$, while the centerline tolerances are shown as ± 0.003.
 (a) For 100 percent interchangeability, determine *x* and *y*.
 (b) Repeat for statistical interchangeability, using all of the usual assumptions discussed earlier in the text.

4-5. Figure P4-5 shows three parts. For proper assembly, the pins must fit into the mating holes, and the bottom surfaces of parts *A* and *B* *must* remain in full contact with the reference surface. For 100 percent interchangeability, determine what tolerance *X* should be applied to the two-inch dimension.

4-6. Repeat Problem 4-5 if a statistical interchangeability analysis is used.

4-7. Parts *A* and *B* are to be assembled by means of a bolt, as shown in Fig. P4-7, and the right side of the bolt is to *always* make contact with the hole in part *A*.
 (a) Based upon *average* dimensions, what is the average value of *E* that results under the stated restrictions?

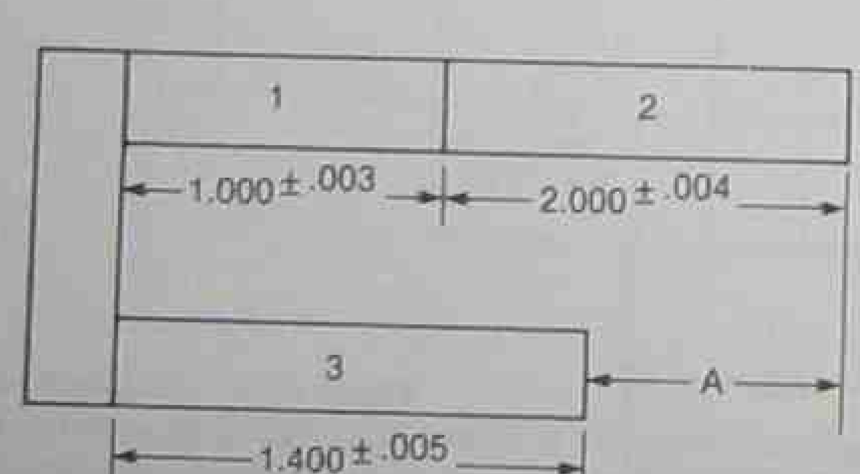


Figure P4-1

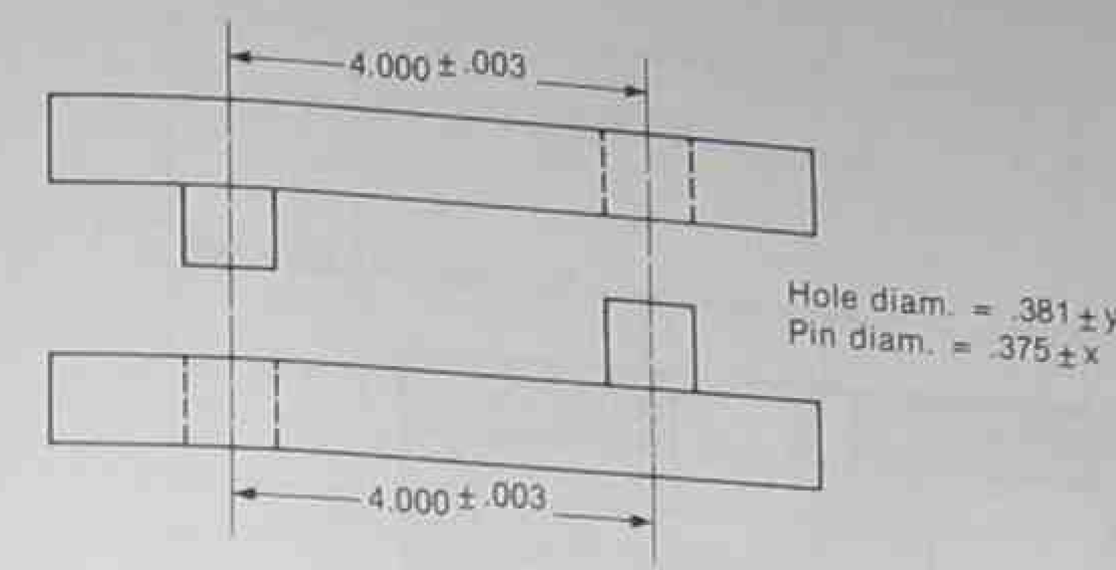


Figure P4-4

 (b) Using your answer in part a, what tolerance *X* must be assigned to the four-in. dimension so that the tolerance of 0.010 on *E* is maintained? This is to consider 100 percent interchangeability.
 (c) Repeat part b if statistical interchangeability is used.

4-8. A spacer *S*, bearing *B*, and a stepped shaft *C* are to be assembled in a housing as shown in Fig. P4-8. The minimum distance between surfaces *A* and *D* must not be *less* than 0.003, where *D* is always *below A*.
 (a) For 100 percent interchangeability, what tolerance *X* should be assigned to the 0.766 dimension?
 (b) Repeat a if statistical concepts are used.

4-9. Figure P4-9 shows original design dimensions. If for manufacturing purposes *A* is to be used as the reference surface for all operations, redimension the part accordingly so that all

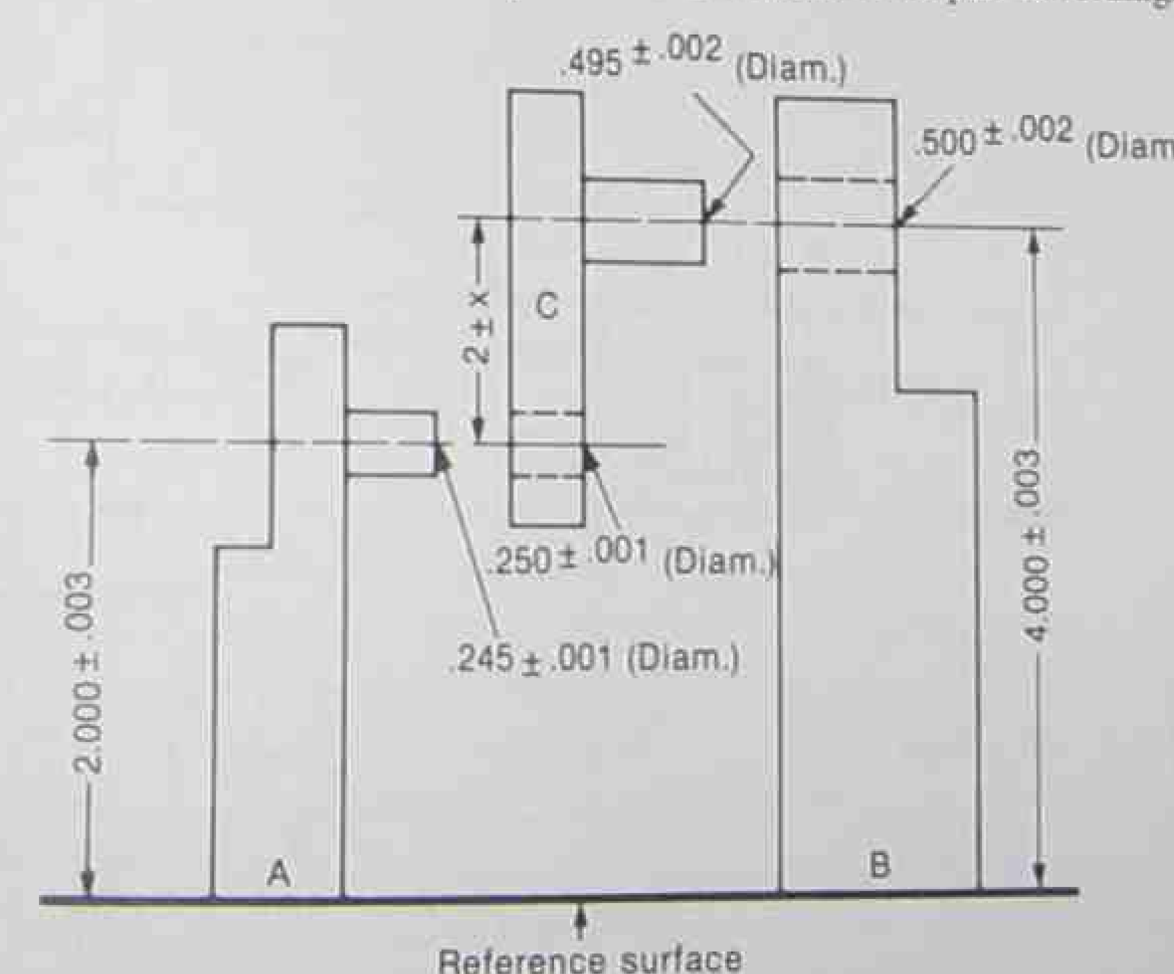


Figure P4-5

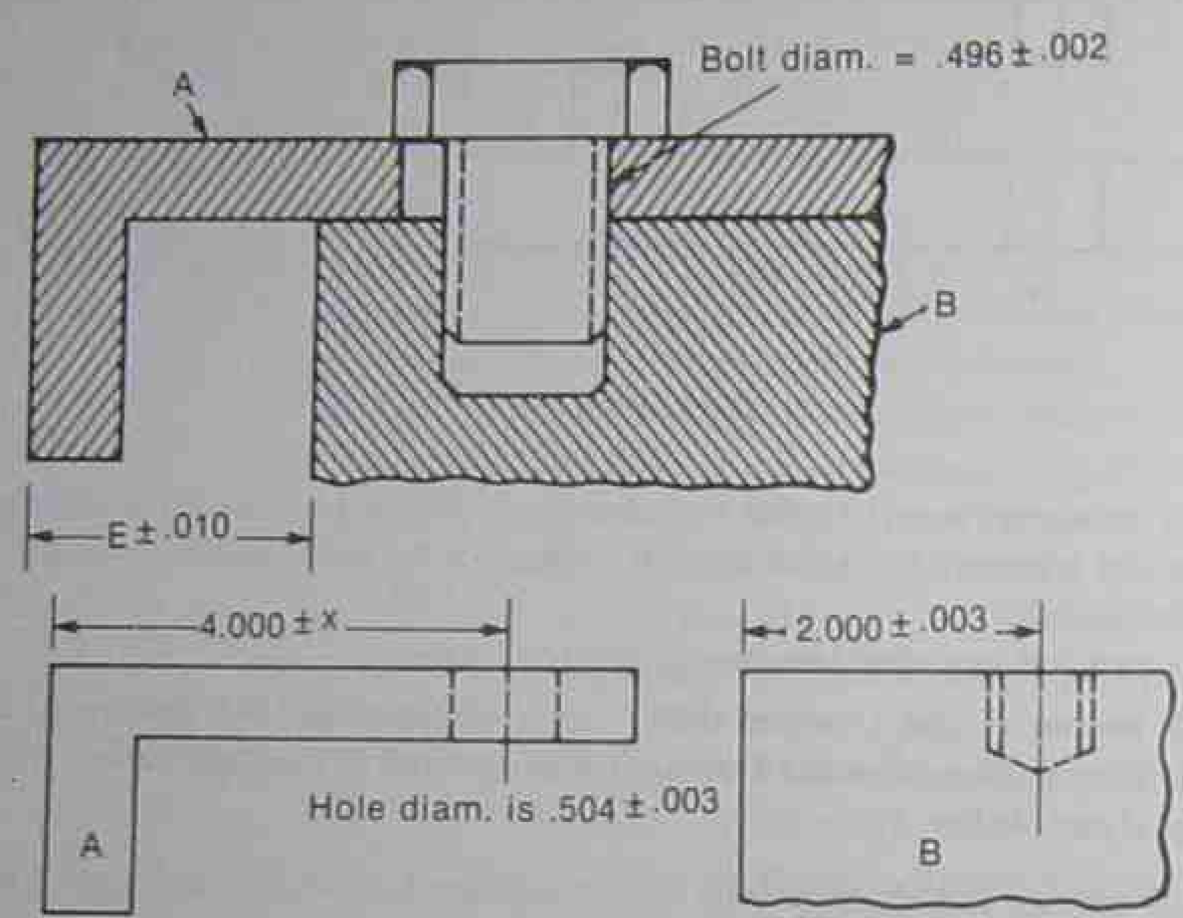


Figure P4–7

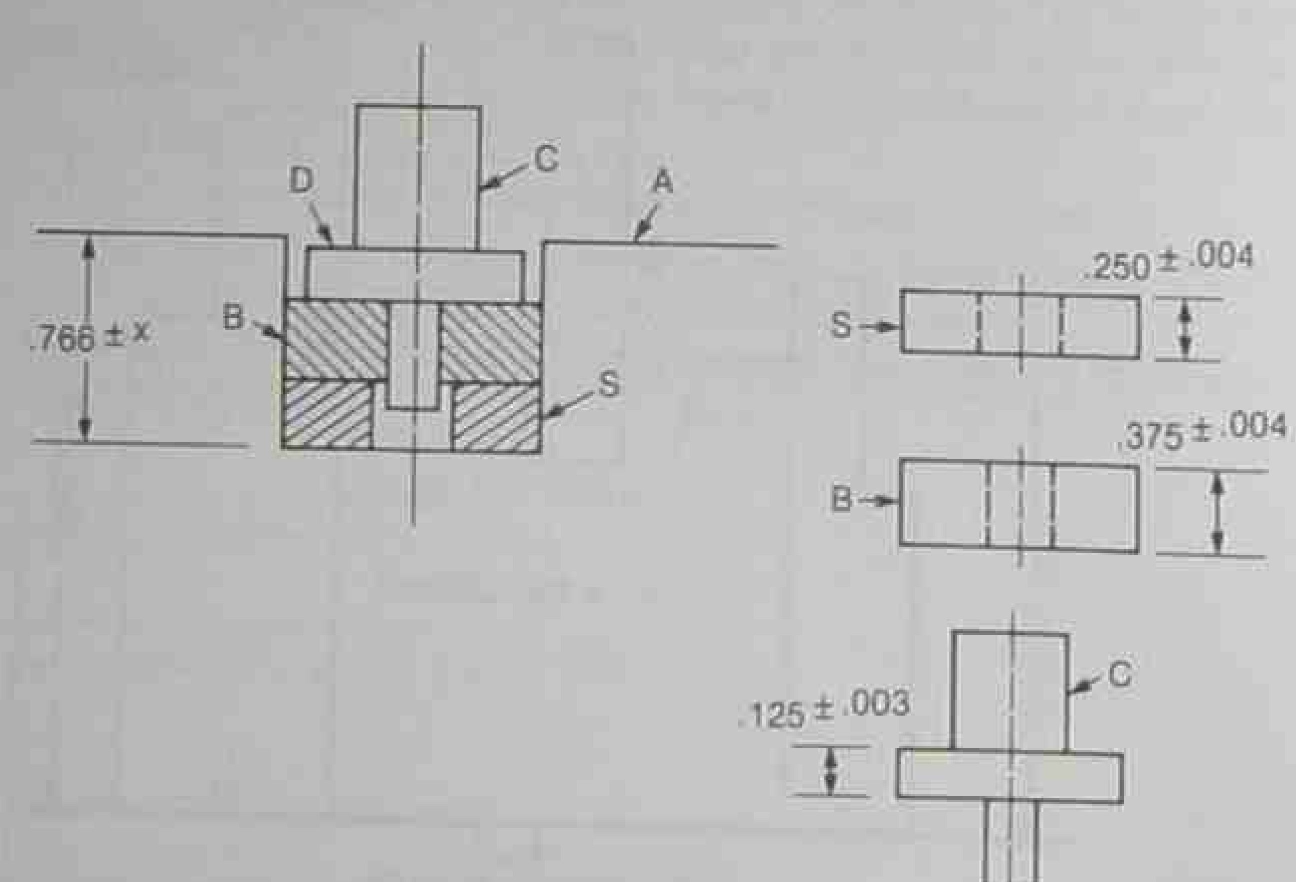


Figure P4–8

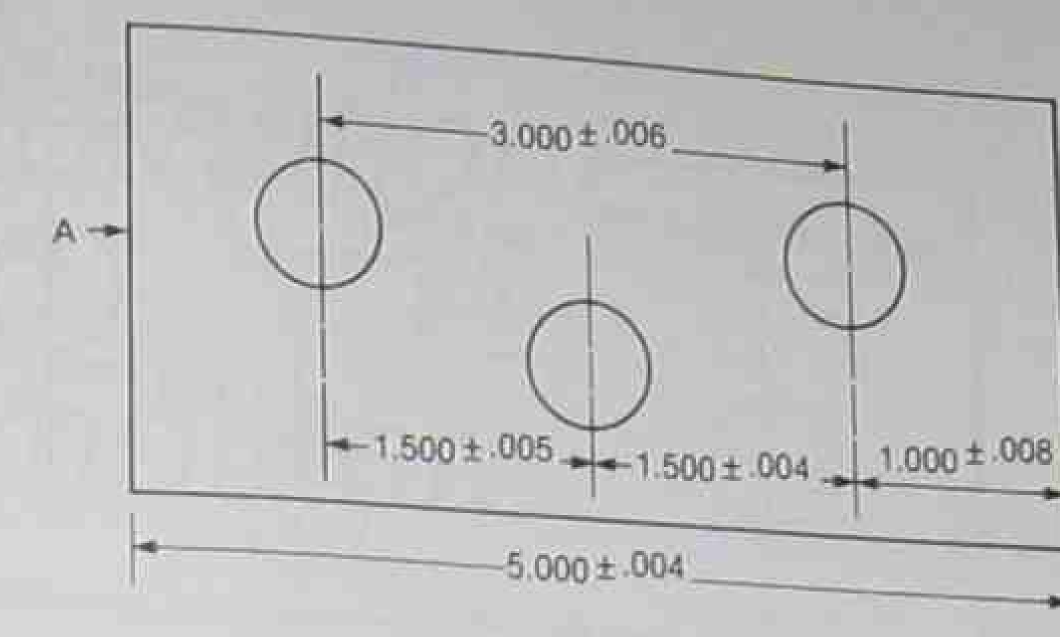


Figure P4–9

functional demands are satisfied and the maximum allowable tolerances are used. Keep the smallest single tolerance as large as possible.

4–10. Figure P4–10 is to be redimensioned by transferring dimensions with *A* as the reference surface.

4–11. Figure P4–11 shows original dimensioning. If *A* is to serve as a reference surface, transfer all dimensions so that all original specifications will be satisfied.

4–12. Redimension the part shown in Fig. P4–12, using *A* as the machining reference. In distrib-

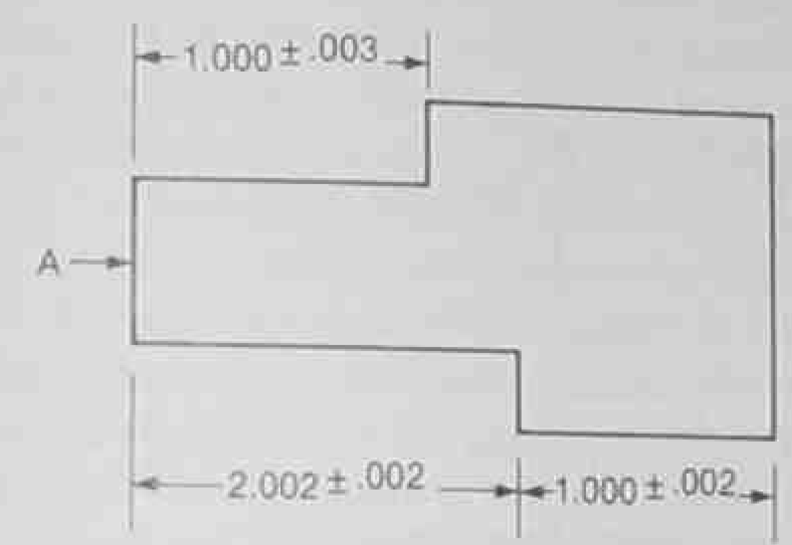


Figure P4–10

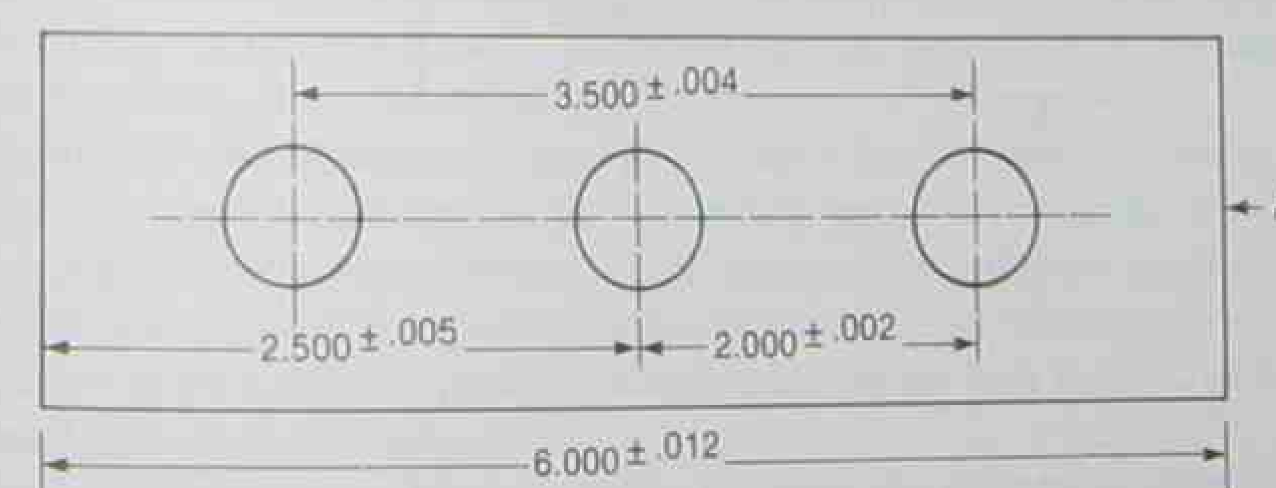


Figure P4–11

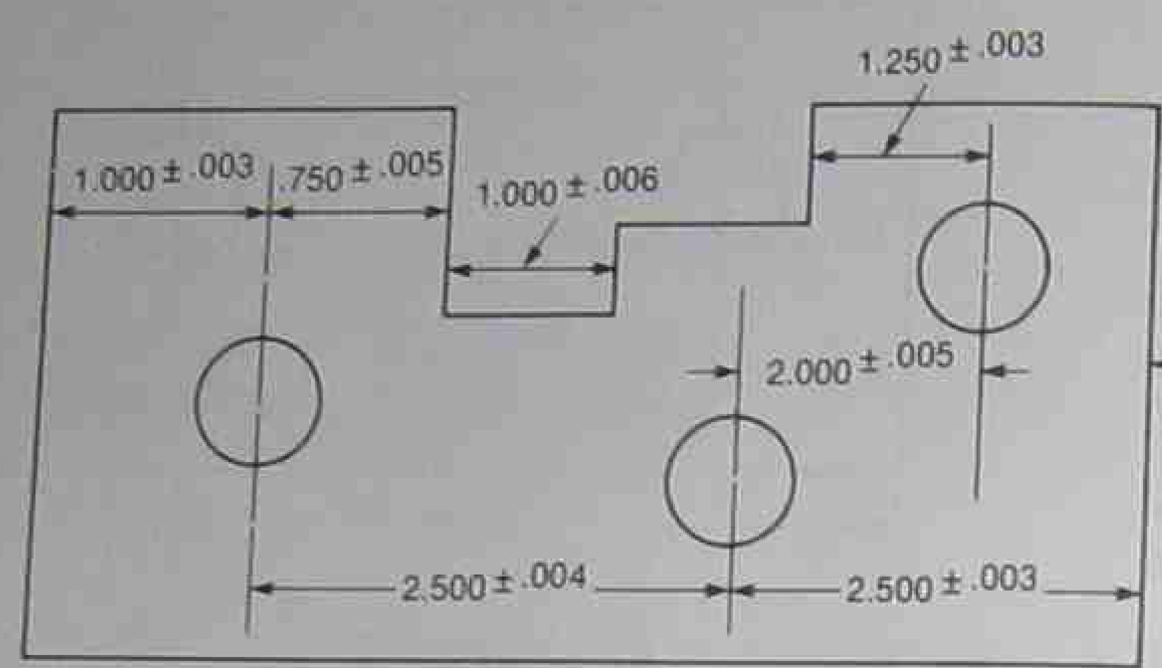


Figure P4–12

uting tolerances over any two new dimensions, use equal values for each individual dimension when possible. The maximum value of available tolerances must be used.

4–13. Consider the original dimensions of a part shown in Fig. P4–13. In transferring dimensions either surface *A* or *B* may be used.

(a) Complete a transfer for each surface, using the maximum tolerance available for each dimension.

(b) Which surface permits the greatest sum of individual tolerances?

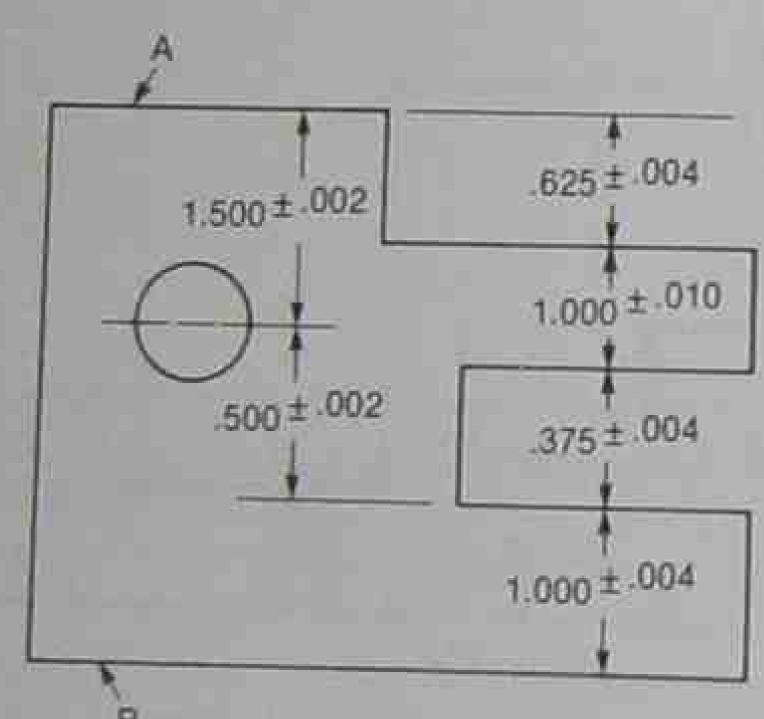


Figure P4–13

5 automation and computers

5.1 INTRODUCTION

One important consideration in a competitive industry is the degree to which manufacturing processes and human actions will be mechanized, automated, and computerized. There are clear benefits in doing so in many instances, but not in all. The deciding factor is whether a competitive advantage will be achieved by doing so. Recall that the economic survival of a manufacturer involves *delivering products, on time, with adequate quality, to sell in a well-informed market and gain a short-term profit.* A secondary objective is long-term economic survival, which is aided by offering new products, by making products good enough to assure repeat sales, by minimizing recall rates and warranty costs, and by a general reputation for fair commerce.

In the following sections, mechanization, automation, and computer technology will be defined and discussed. These topics have traditionally been linked in terms of mechanical progression; human controlled mechanical systems (machines) appeared first, after which the systems were made to *run themselves,* and recently computers became available to manage the systems more efficiently.

In parallel with mechanical innovations, there has been mechanization and automation of *information* transfer. At first there was paper and then the telephone, then photocopying and other conveniences. These systems speeded some aspects of information transfer, but they did not speed human comprehension of what all of the information means and where it should be going. Electronic devices are very helpful, particularly for transferring *appropriate intelligence.*

Until small computers became readily available, there appeared to be little connection between mechanical automation and the automation of information transfer. Lately, however, it has become very apparent that they are linked very strongly, and must be treated together. This may be seen in the following list of the tasks to be done in manufacturing.

5.2 THE TASKS TO BE DONE AND THE ASSOCIATED COSTS

The separate parts of the general industrial objective may be expressed under three headings, together with some of the major tasks in each, as follows. Many of these points were discussed in the previous chapters, and are listed here for emphasis.

1. The control of quality, by
   a. Inspecting incoming material.
   b. Routing the proper material to each process.
   c. Monitoring each process and machine, including handling, assembly, and packaging equipment, to ensure product quality.
   d. Keeping records of quality.

2. The control of scheduling or rate of production, by
   a. Allocation of machine use.
   b. Prediction and scheduling of downtime for failure, maintenance, and changeover.
   c. Effecting quick changeover of machines in order to process a variety of products on few machines.

3. The control of cost, by
   a. Maintaining minimum inventory.
   b. Balancing labor and machinery cost.
   c. Balancing production rate and tooling cost.
   d. Establishing a procedure to quantify the parasitic costs, such as
      (1) Defects in purchased material and the defects that develop in process, assembly, or handling. The cost of these defects increases if the defective material/product is allowed to continue through manufacturing, packaging, and marketing. *These costs may be minimized by complete product quality sensing, before processing and in process, but this costs money.*
      (2) Defects in machinery and tooling. These defects usually result in downtime of all or part of a product line, which is a loss of ability to recover investment. There may also be excess labor cost due to idle employees, and some loss of products that are damaged in a defective process. *These costs may be minimized by providing buffer storage between processes, by scheduled maintenance, and by adequate sensing of machine condition, all of which cost money.*

      (3) Inefficient use of the skills and physical labor of employees due to inadequate attention to the needs of the manufacturing system and perhaps due to inadequate attention to human relations.
      (4) Inadequate or inefficient communication of information. This is a problem throughout the entire factory; for example,
         (a) Design changes do not always reach the proper people to prevent the purchasing of the wrong materials or tooling.
         (b) Impending shutdown of a machine is not always known in time by the proper people to prevent severe unbalance in materials and labor in other parts of the plant.
         (c) Quick changes in production plans (in response to unanticipated market demands for more or fewer products, or for one product option over another) are not always communicated in sufficient time to purchase adequate materials and tooling to meet production needs, or to cancel orders for materials and tooling no longer needed, or to meet the changing needs in personnel or to change over machinery in proper time. *These costs can be minimized by a good organization, by thorough reporting procedures, and by the use of electronic communication, all of which cost money.*

The above list shows some of the challenges in industrial communication, or information flow. Some aspects of the challenge are very technical in nature, both in the content of the information and in the way information is transferred.

5.3 INFORMATION AND COMMUNICATION

It may be seen that for good management, information flow should be dynamic, responding to all changes in production rate, changes in machine condition, changes in the availability of labor, to name a few. Above all, specific people must be made accountable for every bit of communication, or for lack of it.

These are difficult requirements for a system of communication using paper or *word of mouth*. It is difficult for paper work to flow along unaccustomed channels, and it is even more difficult to grade recorded messages according to urgency and relevence. In the last three decades the telephone has been most helpful for some classes of information flow, but it lacks the ability to rank messages according to urgency. Telephones ring as loudly for trivial and bothersome messages as for urgent ones.

In the last decade the computer has been commercialized to the point where it provides a cost-effective aid to information flow. With computers it is easy to establish many different routes for information flow, beyond established mail routes or telephone extensions. It is possible to establish ways to rank the urgency of messages, and to establish accountability. Thus, if a production worker notifies the supervisor of an impending material shortage, the response of the supervisor can be recorded to establish responsibility, should there be an interruption of production. Or the supervisor could have previously sent messages on material needs to the material handlers or suppliers, and again responsibility could be fixed for any problems that develop. The importance of such an information network is that, in the short term, fewer material shortages are likely to occur, and in the longer term, deficiencies in the qualifications of personnel or organization will be highlighted rather quickly and definitively. It is apparent that information flow can strongly affect the cost of manufacturing. Thus the manufacturing engineer should be involved in the design of the communication system, both of the equipment and of the network of people involved.

Information transfer is a problem wherever there are people working together, whether that be in an airplane factory or in a fast-food restaurant. Using the latter example, we find that problems may begin when customers request the "product" in various degrees of processing in such terms as "rare, medium, or well done." They further request options in the product such as additions of catsup, pickles, mustard, and so on. The requests arrive at the "sales" department in random order, requiring complex communication, oral and nonoral, throughout the manufacturing facility (kitchen). Most often the system operates smoothly. But if there is a rush of orders for one or two products or options, and if the cook or other supplier of that product is either slow to receive the message or out of range of communication for a time, the system soon degrades into a state of confusion. Requests are shouted with increasing urgency, incomplete orders are set aside, which soon interfere with orderly traffic, and the number of errors increases. This sequence develops with varying frequency, hourly or daily, depending on many factors. The exact cause of the disorder is usually somewhat different each time, which taxes inexperienced workers and managers. And the search for the causes often is obscured by personal insecurities and blame shifting.

To prevent chaos it is vital to have good organization, which includes proper placement of people of the proper skills, with detailed assignment of tasks and responsibilities. In a restaurant there may be little need or opportunity to automate communication and systems of accountability. The need increases with the physical complexity of the system, with the sophistication of the product, and with the amount of competition in the market.

5.4 MECHANIZATION

We define mechanization as a state in which tools are applied to aid or replace animal power, but requiring human intervention for the purpose of control. Thus the hand shovel is a tool, since it allows the user to accomplish more than is possible without it, even though shoveling is hard work. The power shovel is also a tool. It moves large amounts of earth with little human effort, though it is guided in every motion by the human operator. The hand and power shovels are examples of the range over which tools may multiply human effort, and exemplify ranges or degrees of mechanization.

It is often surprising to see the great number of manual or unmechanized operations, even in a highly mechanized manufacturing facility. Assembly is often done manually, as is die-setting, machine loading and unloading, inspection, counting of inventory, and

even communication. A factory therefore is usually only partly mechanized, and very likely none will ever be completely mechanized. First of all, the cost to mechanize completely is prohibitive. Second, there are some human capabilities involving aesthetics and judgment that will not be profitable to mechanize. After all, consumer products must appeal to the highly varied subjective sense of consumers, and human judgment is required to determine when products are ready to risk in the marketplace.

5.5 AUTOMATION AND CONTROL

We have seen that mechanization consists in the use of tools to multiply the physical efforts of humans. The next step is to relieve humans of the task of guiding or controlling the tool. Tools that control their own motions within prescribed limits are said to be automatic; or by virtue of having had control devices connected to them, tools are automated. Few if any power shovels have been automated, probably because of the wide range of tasks required of them. Certain components of power shovels are automatic, such as the speed control of the engine, and the lift-limiting feature to prevent tipping over. These automatic features prolong engine life and prevent accidents, but without significantly affecting the quality of the work done.

In many instances a tool or machine is automated to maintain product quality. A hamburger grill may become too hot or too cool to produce quality hamburgers, unless controlled. Likewise, the chemistry of an uncontrolled electroplating bath may vary from the acceptable range, necessitating the stopping of production while adjustments are made.

Automatic machines *can* often produce parts with greater accuracy, uniformity, and speed than can manually operated machines; they also relieve humans of some tedious, dangerous, and hard jobs. But automation is cost-effective only for simple tasks. The majority of humans can readily perform fairly complex tasks, and for moderate pay. Some industries use cost guidelines in deciding whether or not to automate an operation. If, for example, a machine that replaces a person costs more than three times the annual salary plus benefits for the person, then the person is retained. Such reasoning explains why the dispensing of newspapers is well automated but the setting and clearing of tables in restaurants is not.

In the latter example one can discern two aspects of the complexity of a task. One is manual dexterity, or the skill in making complicated movements. The other is in the use of the *senses* (sight, pressure, sound, feeling of vibration, etc.) in guiding action. Several human senses are used to monitor even the simpler tasks. To integrate only one or two senses with complex motions in robots or other automated machies is very expensive.

Automation involves control, and controllers range from the very simple to the very complex. One of the simpler controllers is used on railroads. Two rails effectively "control" the direction of the railroad vehicles, without sophistication or instrumentation. A more complicated system is the system that controls an airplane along its intended course.

An airplane or a train may at one instant actually proceed in the proper direction all by itself, but both will soon proceed in a different direction and will need corrective input from the control system; the rail exerts pressure against a wheel, and the rudder directs the airplane. Both control systems must be designed to meet certain requirements. These may include limits on the amount of deviation allowed from the intended path before correction is effected, or limits on the abruptness of correction in order to prevent high stresses on vehicle parts or discomfort and damage to the payload. These requirements and limits are called *constraints* in the technical terminology of the field of control.

The train and the airplane are both controlled, but in the conventional meaning of the terms, the rails are fixed guides rather than controllers or control devices. A *true* controller provides the flexibility to follow a different course each day.

The principles of control are simple. A controller reacts to a signal sent to it from one or more sensors. One type of sensor in an airplane is a compass. It *feeds* a signal to the controller on the direction the airplane is heading. If this is not the *preset* or prescribed heading, the controller changes the direction of the airplane, which action is sensed by the compass, which *feeds* a new signal to the controller. This type of control system with the sensor is referred to as a *feedback control* system. Since action follows a sensed variable and that action changes the second variable, the total action is referred to as *closed loop* control. These terms may be compared with what is often referred to as *open loop* control, which is a contradiction in terms. In so-called open loop systems a mechanized or electrical action is effected without any measurement or sensing of the effect of that action. For example, a mechanical cam and follower, or a servomotor and ball-screw may be actuated by a given amount in order to move the table of a milling machine some desired amount. At the end of the actuating signal, if there is no sensor, there is no *feedback* to indicate that the table has been moved the proper amount. In such cases the cam and the servomotor are merely actuators or *effectors* and are not a part of a control loop.

Controllers can further be classified according to the manner by which corrections are made. An airplane that is off course could have the rudder turned abruptly some fixed amount until the airplane is on course. Or the abruptness of this action can be reduced by providing some ramping or gentle transitions at the ends of the control events. This introduces the need to detect the points at which ramping must be actuated. Alternatively, a corrective action could be applied that is proportional to the amount the airplane is off course. The airplane would steer more sharply when it is far off course than when it is near course, subject to some constraints. This proportional control has some advantages in that simple analog electronic circuitry can be used in the controllers instead of digital devices. Analog circuits are readily characterized by linear differential equations, which makes their design easy. The same differential equations characterize the forced vibration of a series spring-mass-dashpot combination, connected in a U-frame, and are of the form:

$$m\ddot{x} + C\dot{x} + Kx = F(t) \qquad (5\text{-}1)$$

where m is the mass, C is the damping coefficient, K is the spring constant, x is the displacement of the mass relative to the U-frame, and $F(t)$ is a time-varying force (forcing function) applied between the mass and the U-frame. The solution to the above equation depends on whether the quantity $\{C/m\}^2 - 4\{K/m\}$ is + or − or zero. (Beachley)

There is at least one mechanical system that behaves exactly like the example

above. The cab of a truck is a mass, and for isolation of the cab from the frame of the truck the cab is mounted on springs with dampers. The frame of the truck is equivalent to the U-frame mentioned above, and it bounces or oscillates at frequencies that vary with road and load conditions. We will describe the problem as a one-dimensional, or single degree of freedom, problem, even though truck cab motion is much more complex.

Some of the frequencies of frame oscillation can be near to the natural frequency of oscillation of the body. The latter can be readily adjusted, since it is simply proportional to $\sqrt{(K/m)}$. For driver comfort the natural frequency should be low, requiring a large m. But a large m reduces the amount of payload the truck can carry under limits imposed by governmental regulations. A small m is therefore required for economic purposes, but a small m will oscillate at larger amplitudes as well. One solution is to place an accelerometer on the cab, and a hydraulic cylinder between the cab and frame. The hydraulic cylinder imposes a force, which acts to decrease the acceleration of the cab when the frame oscillates. Such systems can respond as quickly as 30 Hertz (cps) over an amplitude of 5 cm, which is sufficient for driver comfort.

In the example of the truck, the physical system was similar to the simple model given earlier. However, the important design problem is that of the control circuit. It receives a signal from a sensor, the accelerometer, indicating something about the movement of the cab. The circuit will then calculate and effect an appropriate response, by opening or closing a hydraulic valve, provided that its own components are direct electrical analogs of those of the physical system. This is a straightforward or standard control system.

A control system designed for one particular application cannot, in general, be used for another. Furthermore, it will function poorly if the physical device it is designed to control changes its properties in some way. For example, should the truck cab be occupied by three very heavy people instead of the usual one, the control circuit will not provide the appropriate response to truck frame vibration. It is a simple matter to provide a means for changing the circuitry of the controller. There might be a knob to turn somewhere in the cab which can be adjusted to compensate for weight of the cab. Or there could be a device that senses the "at rest" length of the springs on which the cab is suspended, which could damage the electrical characteristics of the controller. Any number of schemes could be designed into the controller to vary the nature of the ride in the cab. But all of these are done by altering a part of the electrical circuitry that represents a physical quantity. These changes are called *gain scheduling*, because the changes are usually made by adjustments in the gain loop of an electronic amplifier in the controller.

As noted above, gain scheduling can be done manually, or it can be done automatically from some input taken from the *static* condition of the device or machine to be controlled. It can also be done *dynamically*. That is, the control circuitry could include provision for comparing the dynamic response of the physical system against some model behavior. For example, it could monitor the amount of overshoot or undershoot of corrective action taken by the hydraulic cylinder. If such a deviation from critical damping is larger than a prescribed amount, the auxiliary circuitry could then adjust the gain of the primary controller to compensate for the change in cab weight. This is called *adaptive control*. Adaptive control is distinguished from other types of control in that it is a method of control in which the controller adjusts its characteristics according to changes in the dynamic behavior of the machine.

Control, particularly adaptive control, is done most often against one or two constraints. These may include some safety consideration, or noise limit, or the like. The computer has brought the possibility to operate with several constraints simultaneously, and in fact it is reasonable to optimize performance around some predetermined *index*. For example, a lathe could be operated within simultaneous constraints to prevent tool breakage, prevent vibration, and produce a good surface finish. At the same time the machine can be made to operate at the overall most economic feed and speed, when the cost of labor, the cost of cutting tools, and the cost of chip removal are taken into account. It is then necessary to have all of these variables either expressed in mathematical form or available in tabular form for use in approximate equations in order for the control system to be able to arrive at an optimum. There might even be provision for automatic updating of the cost of cutting and labor costs each day so that the controller can ''find'' a new optimum each day. These possibilities are yet several years in the future.

5.6 COMPUTERS

The increasing availability of digital computers alters the economics of automation and control. A computer and simple electronics now replace complex control systems, and provide virtually any desired degree of sophistication to the control routine. The problem of designing control systems has shifted from the design of complex electronics to the programming of computers so that they will respond to sensors and continually alter the gain setting of the controller.

A second useful capability of computers is electronic communication. Several computers hooked into a network become an instantaneous and precise medium for transferring information. The information ranges from merely words to the display of complex designs. The designs are those of the company's products or perhaps of a complicated mold for plastics or a die for sheet metal forming; they are entered into electronic storage in the *graphics* mode in computers of a wide range of sophistication. Several people in several locations can then connect simultaneously to the same source and discuss possible revisions in a design from identical displays, or viewers could connect serially, as the need arises; these might be the material purchasing department, the plant engineers, and the manufacturing engineers. The buyer of the part may also have access to the design in order to see how it fits with some other parts to be supplied from elsewhere. Someone responsible for production scheduling could have access to the plant schedules and predict when the viewed part may be finished. A small but important advantage of a system is that parts will not be delayed in manufacturing because the part print (on paper) has not yet made the rounds, or because someone has scribbled ambiguous notes and revisions on a print. The major disadvantage is that computers malfunction, but paper does not.

Designers will find the use of computer data bases to be mandatory in a few years. The current procedures in designing objects involves calculations of various types, for which most designers have references and text books. If these references are very familiar, it is a simple matter to find the desired equations for use in calculation. However, if the new problem involves some new boundary condition, for example, if available equations are for conditions of small strains whereas the new problem involves larger strains, some considerable study may be required to find appropriate equations. A well-developed communication network in a design group could include access to an *expert system*, which is an interactive program whereby an inquirer at a keyboard can be led through a series of questions. The answers given will be used by the program to direct the inquirer to the best source of information for solving the new problem.

A parallel problem for designers is to locate materials, components, and supplies. Usually catalogs are in hand for this purpose. However, some effort is required to update the stock of catalogs and the price lists that may be attached. If all of the cataloged information is in electronic storage, it is a simple matter for the supplier to initiate sending updated information daily or weekly. This avoids contacting suppliers by telephone for prices, availability, discontinued products, new products, and other vital information.

5.7 THE MACHINERY OF MANUFACTURING

A major responsibility of manufacturing engineers is to specify, purchase, and arrange installation of machinery. The broadest considerations are the degree of mechanization, automation, and control that will affect the costs over the long term. In this section we will focus on the hardware and not consider economics.

Machinery is available for most types of manufacturing operations, but it may be particularly useful to describe the equipment available for metal cutting. Metal cutting is an old process and is being done on a scale that encourages the development of a very wide range of sophistication in automation and control.

The simplest metal cutting machine is the lathe. The first lathe may have been a V-branch of a tree, in which another tree branch was manually turned while bark or twigs were being removed. At some point in history, bearings were made to aid in turning the wood more smoothly, and perhaps chisels and other cutting instruments were developed at the same time.

In the eighteenth century, lathes were used to cut metal. The cutting tools were plain carbon steel, so cutting could not be done at a high speed by modern standards. (This is covered in more detail in Chapter 7.) In early years the cutting tool was mounted on a small table that was moved manually by screw-nut pairs. Later, the tool was moved by an arrangement of shafts and gears connected between the screw-nut pairs and the spindle on which the workpiece was held. One could shift gears to effect a different amount of tool motion for each revolution of the workpiece, but when the workpiece stopped, the tool stopped moving. In a later development, the tool could be moved separately from the rotation of the spindle, but each motion was controlled by the human operator of the lathe.

As described, the old lathes were indeed automated, but lacked closed-loop electrical or mechanical control of tool motion. Actually, the human operator is the controller, a far more sophisticated one than any electrical or mechanical controller. But for many operations the human system is underutilized and therefore probably too expensive. Thus there have been many innovations in control devices to aid or replace some or all of the involvement of humans in some lathe operations. In those systems involving feedback from sensors, the controller uses only one or two sensors, thus implying that the several other senses resident in humans will not be needed. This is true in some cases, but it leads to very big surprises and failures in other instances. For example, a problem in metal cutting is the occasional, unexpected failure of a tool. This type of failure is immediately evident, even to the untrained eye. In fact, a machinist can give a continuous appraisal of the wear of the tool and can predict failure in time to prevent it. Automatic detection of impending tool failure has not so far been achieved, despite many efforts to do so. Cutting force and noise analysis (acoustic emission) have been used without great success.

A number of mechanical aids to the machinist have been developed over the years. In the cutting of parts requiring several operations, the machinist must exchange a sequence of tools. This takes time, partly to loosen and tighten tools, but more to accurately adjust the location of the tool tip so that it can be moved the proper amount to produce a part of the correct dimension. This problem was alleviated by mounting several tools on a precisely made turret. The turret is rotated to bring up any desired tool, and each tool is preset. Some of the tools on the turret can *feed* faster than others, so some machines are equipped to advance the turret at the proper rate for each tool that is in the cutting position. These developments increased the productivity of machines up to fivefold, and reduced the skill required to operate the machine. A skilled technician is still required to set up the machine, however.

One popular development in lathes and other machinery is "numerical control," or NC. Most NC lathes use a one-inch-wide tape of paper or Mylar plastic in which eight columns of holes may be punched with as many lines as are required to perform an operation. The tape progresses through a reader where an array of eight sensors (photocells or pneumatic sensors) reads which column has a punched hole and which does not. This constitutes eight bits of binary information, or up to 256 (2^8) different "numbers" or codes per line of holes. A series of relays, with no memory, is arranged to interpret each code as an instruction to effect some action. These might include:

1. Set the tool feed direction either left or right.
2. Set the feed rate to be any of ten different values.
3. Start the feed motor.
4. Advance the feed motor 600 pulses.
5. Set the spindle speed to any of five different values.
6. Start the spindle.
7. Stop all motors.
8. Rotate the tool holder to bring up a new tool.

These signals must be directed to proper locations, and part of the code includes such information. But there must also be the proper interpretation of that code by a receiver. The instruction to advance the feed motors is received by a servo controller

which converts the instruction into a particular binary pulse stream to advance the servo 600 pulses. For this purpose the system contains many transistorized electronic elements called *chips*.

The tape reader and the associated electronics can accept code faster than most machine actions can take place. Thus the tape reader can sequentially read code for several separate servos of the machine. For example, both the feed and depth of cut can be varied at the same time, which is referred to as *two-axis control*. Some machines have even more motions actuated, perhaps as many as six. With multiaxis capability, a machine can cut complex contours and make very complex parts.

There are varying degrees of real control associated with NC. In the simplest case, servo motors that drive the screw-nut sets may receive a given number of pulses. There may be no method provided to assure that the motors have really moved. In this case the system should be called "numerically actuated." Better systems use servo motors that *feed* servo position signals *back* to its controller, and if the servo has not advanced in accord with the pulses sent to it, the controller may simply send some more pulses, or it may send a signal to the central controller, which may shut down the entire machine. In the older machines a human operator was then required to determine why the machine stopped. In newer machines a message is flashed, indicating the reason for stopping. This capability is a very elementary form of *artificial intelligence*.

The ultimate in NC is the system that uses sensors at various locations to verify that the tool or part has moved to its intended location. The sensor may be an interferometer, or a differential transformer, or other device. Interferometers are accurate to 300Å (0.00003 mm), and the differential transformers are accurate to 0.01 mm. The advantage of sensing the position of the tool or workpiece is that greater accuracy of machining can be done. Most machines, workpieces, and tools deflect more than the desired accuracy when the forces of cutting are applied. An accurate and well-placed sensor will signal the servo controllers to compensate for this deflection. This is *feedback* of the most useful nature. An additional useful feature on NC controllers is a readout of the command from the controller and a readout of the actual position of the driven member.

One of the very tedious tasks connected with NC is the punching of the tape. This involves looking up code in a large table of codes, and it requires absolute precision in punching the holes as well. In the early days, after punching, the tape was tried out on the machine. Immediate success was rare. Failure required a lengthy and tedious search for errors.

Computers have been very helpful for punching tape. Computer programs of various sophistication are available, ranging from simple table lookup of NC code to programs that can receive information on the dimensions of the part to be machined, the size and shapes of tool to be used, and the particular machine to be used. In the latter case the computer then proceeds to punch the tape without human intervention. This system is variously referred to as computer-aided NC or computerized NC, or CNC.

Obviously, if a computer can be programmed to punch tape for the NC machine, all of the code and characteristics of the NC machine are part of that program. It is a small step, then, for a computer to control the machine without the intermediary punched tape. Some refer to this type of system as CNC, and others call it direct NC or DNC. (DNC is also an acronym for distributed numerical control, where several computers may be involved in controlling several machines.)

There is yet one type of machine that receives much attention, and that is the robot. It is built to the standards and by the methods of the machine tool industry, and thus it must be regarded as one extension of machine tool technology. Robots usually have a wider range of motion than most machine tools, with corresponding lower accuracy of motion. They are controlled by the same devices as are machine tools, but in some instances the robot may have one or two more movements or degrees of freedom than do machine tools. The movements of the machine tools may consist of four or five sliders, but each usually acts separately. Robots often consist of one actuator mounted upon another, which adds versatility to its motions but also "stacks" errors due to the inherent looseness of fit in joints, which is exacerbated by wear. The motion of the "end" of the system then is the result of the motion of several joints, and may be designed to move in rectangular, cylindrical, or spherical coordinates. The calculation of the location and velocity of the "end" involves a great number of trigonometric calculations, which must be done at great speed in order to effect smooth motion. Thus computers of larger capacity are required to control robots than those required to control most machines.

So far we have discussed single, stand-alone machines. These may be sufficient for some products in some industries. But most products are made in several operations, requiring several machines. A major economic difficulty arises when only 50 or 500 identical parts must be made, as is the case with some 70 percent of the products made in the United States. Again it is useful to imagine what takes place in familiar surroundings such as a restaurant; few survive if they produce one set menu. However, given the availability of equipment to produce one simple product, this equipment could probably be revised slightly so that it can expand the range of products. The system is now *flexible* as to product. With further revision the system could also become *flexible* as to production rate, the discontinuance of old product lines and the addition of new products.

The design of manufacturing systems usually requires considerable knowledge of logical groupings of manufacturing processes: groupings based on skills required to set up and maintain the machines, or based on the materials being processed, or even groupings based on the specific types of controllers on the machines. These groupings are called *manufacturing cells*, and factory automation usually starts by attaching simple microprocessors to machines in the cells.

Products usually must pass through a succession of *cells*, possibly including inspection stations and packaging machines. One major source of cost is the handling of materials as they move into and out of the cells. Material handling is readily automated, at some cost. But automated material handling requires coordinated control of all of the *manufacturing cells* in a system. Another dimension of flexibility must now be provided in terms of routing among the cells to provide for product mix and machine breakdown. Material handling can be done by robots, by linear conveyors, or by loop conveyors.

Coordinated control of *flexible manufacturing systems* (FMS) usually requires a special information transfer system. The reason is that within a manufacturing system there will almost inevitably be computers and peripheral equipment from several manufacturers, each of which uses a different communication method or protocol. This was not done capriciously in the industry. Rather, the several tasks to be done by computers, (for example, analysis, graphics, interactive communication, machine control, and the like) can best be done in special ways. A popular method for solving this problem is for all industries to purchase machine (computer) linkage or bus systems that operate with (nearly) the same data transfer rates, using the same system of priorities of access of computers to the bus, and the same limit on the amount of information that can be put into or taken from the bus when connected. Each manufacturer of computer/control equipment then supplies an interface module between his equipment and the bus, which then makes their proprietary communication method invisible to the bus system.

The useful range of information transfer extends well beyond the actual machines making parts. For example, the design of the part could be stored electronically (on magnetic media), and some of that stored information on dimensions and tolerances could be "accessed" directly to set up the measuring devices at inspection stations. Again, the measurements taken at the inspection station could be analyzed periodically, with results provided to management on the progression of product quality. When several functions are interconnected with manufacturing system control, then *computer integrated manufacturing* (CIM) has been achieved.

Total CIM is a massive undertaking which few industries can afford. Furthermore, unless there is some major change in societal thinking, total CIM will be limited as much by the human dimension as by dollar cost. Engineers are often advocates of increased automation, and find societal and money constraints irksome.

REFERENCES

Beachley, N. H., and H. L. Harrison, *Introduction to Dynamic System Analysis*. (New York: Harper and Row, 1978) or equivalent.

PROBLEMS

5-1. Assess the amount of automation in the following areas: discuss the cost of fully automating and the inflexibility that would result from full automation:
  (a) A restaurant.
  (b) A car wash facility.
  (c) A bank.

5-2. Describe the type of sensors used in the following, and classify the type of control in terms of unramped on/off, ramped on/off, or proportional control:
  (a) Trains moving along tracks.
  (b) Thermostats in home heating systems.
  (c) Thermostats in the coolant system of an auto engine.

(d) Power brakes of automobiles.
(e) Gasoline filler nozzles.
(f) Servos on lathes, with and without sensors.
(g) Strain gages on cutting tools for force measurements.
(h) Sensors for control of surface finish in grinding.
(i) Lawn mower engine speed.

FOREWORD TO PART B

In the broadest and most traditional sense, there are four major processing categories. They may be called by different names; our preference is to use the terms *cutting* (machining), *forming*, *joining*, and *casting*. It would be a rare individual who would possess equal, in-depth knowledge and experience in all four areas. For example, one who has devoted his professional career to the subject of casting may possess little, if any, understanding about forming and vice versa. Yet is is not necessary to be an expert in all processes to understand and teach the most important aspects of each. Since entire texts have been written on an in-depth approach to each of these four processes, it must be realized that single chapters cannot cover, nor are they intended to, all that a specialist might desire. What we present here are the essential features and fundamental concepts of each process. If this much is understood then the reader should be able to continue with those specialized texts that go into greater depth and detail.

Chapter 6 is a minimal, but we think necessary, review of stress, strain, and mechanical behavior of solids. Of special importance is the coverage of strain hardening of metals and the material on nonmetals. This usually receives little if any coverage in other texts of this type, yet it is of decided importance if any quantitative analyses of certain processes are to be undertaken. This background finds major application in Chapter 8. Chapter 7 is a coverage of a number of the most important concepts of engineering materials; it is not a rehash of what is usually dealt with in a materials science course, nor is it intended to go into the many concepts of physical metallurgy. But with this background, the contents of Chapters 10 and 11 should be more meaningful. Chapter 12 introduces certain basics of the ''newer'' or ''nontraditional'' processing methods.

6
stress, strain, and mechanical properties

6.1 INTRODUCTION

Since the concepts of stress and strain are applied in later sections of this book, it seems sensible to discuss these terms in reasonable detail here, then use them directly whenever needed in chapters that follow. The coverage that follows is *sufficient* for the purposes of this text but is not put forth as an all-inclusive discussion of these two quantities. Both stress and strain are known as *tensors;* a tensor may be defined as a collection of components which may be *transformed* from one set of coordinate axes to another set by following certain mathematical rules. Because of the type of problems of major concern in this book, a background of tensor analysis is unncessary; however, it must be understood that stress and strain are not *vector* quantities in the usual meaning.

6.2 STRESS

This is usually defined as force per unit area. Although such a definition is adequate as far as it goes, neither the three-dimensional nature of stress nor the fact that it is a tensor quantity are made evident by this simple definition. Consider Fig. 6–1, where a force F acts normal to an area A as in *uniaxial tension,* with the x-y coordinate system shown. The stress S_y is simply F/A, and since no force acts *parallel* to A, no stress is associated with the x direction. Now consider a new coordinate system, x'-y', acting at some arbitrary angle Θ as indicated. What stresses act upon a plane at this angle due to force F? First, resolve this force into components F_x and F_y as shown; these act upon area A', which is simply $A/\cos\Theta$. Then since $F_{y'} = F\cos\Theta$ and $F_{x'} = F\sin\Theta$, we have

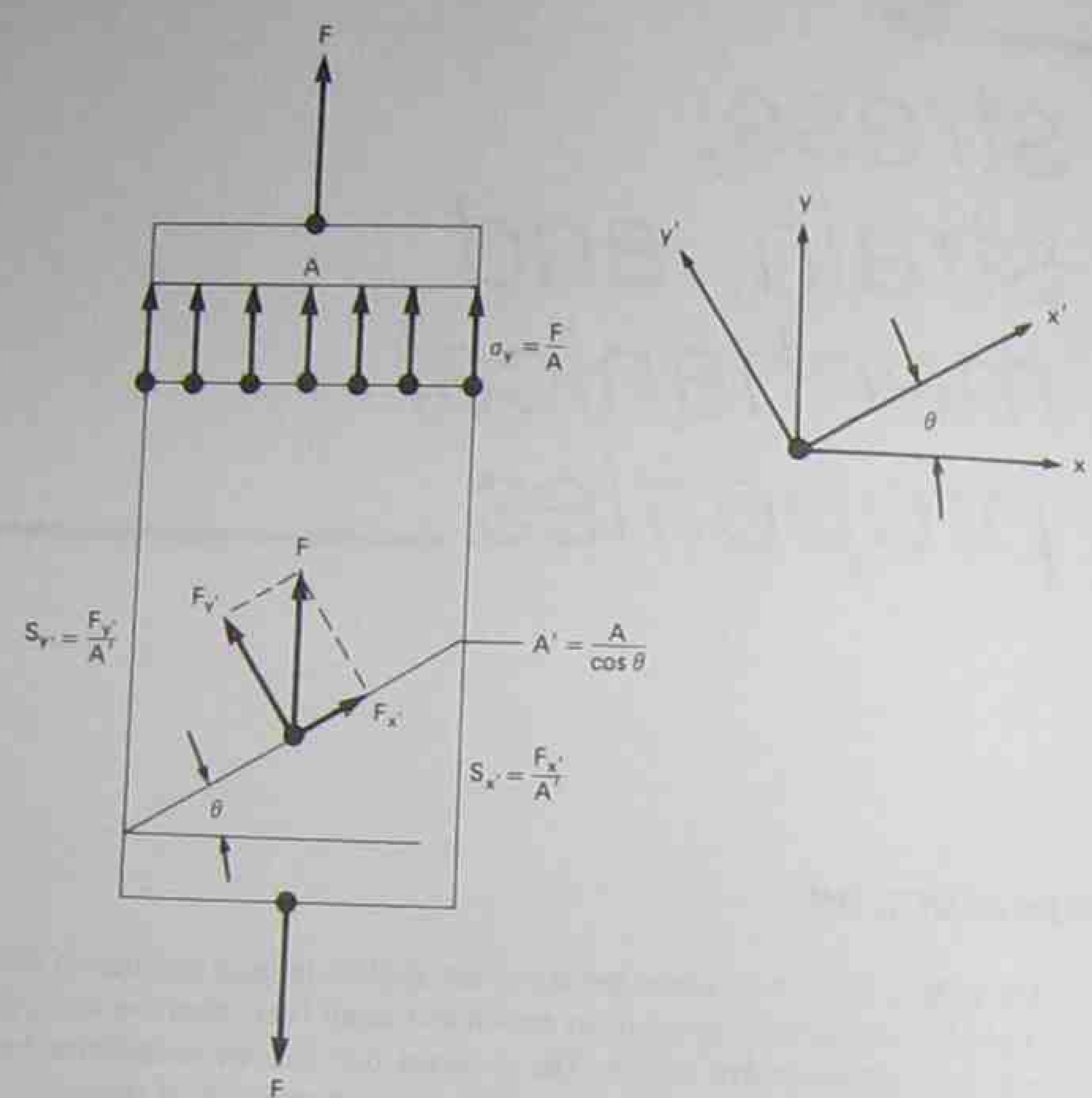


Figure 6–1 Forces and stresses related to different sets of axes.

$$S_{y'} = F_{y'}/A' = \frac{F \cos \Theta}{A/\cos \Theta} = \frac{F}{A} \cos^2 \Theta = S_y \cos^2 \Theta \qquad (6\text{–}1a)$$

and

$$S_{x'} = F_{x'}/A' = \frac{F \sin \Theta}{A/\cos \Theta} = S_y \sin \Theta \cos \Theta \qquad (6\text{–}1b)$$

This is the simplest illustration of stress transformation, and it is noted that to accomplish this requires the use of a *product* of two angular functions. Note that the equivalent force transformation required only one angular function; this illustrates one of the basic differences between vector (here force) and tensor (here stress) transformations.

By convention, *normal stresses* (such as S_y and $S_{y'}$) act perpendicular to planes, whereas *shear stresses* (such as $S_{x'}$) act parallel to planes. When a normal stress tends to cause extension of a body, it is considered *tensile* (or positive), whereas contraction is due to compressive (or negative) normal stresses. Shear stresses tend to distort the shape of the

body. We shall not be concerned here as to how positive and negative shear stresses are defined, but it is worth mentioning that in problems where a detailed stress analysis is demanded, such a definition is essential. Note that from this point on, the symbol τ will be used to denote a shear stress, whereas either S or σ will denote normal stresses.

Example 6–1

A uniaxial tensile load of 10,000 lbf (44.5 kN) is applied to the ends of a solid round bar whose diameter is one inch (25.4 mm). Determine the stresses acting on a plane oriented 70 deg from the axis of loading.

Solution With reference to Fig. 6–1, where Θ is 20 deg, from Eqs. (6–1a) and (6–1b),

$$S_{y'} = \frac{F}{A} \cos^2 \Theta = \frac{10{,}000}{\pi/4} \cos^2 20 = 11.24 \text{ ksi } (77.5 \text{ MPa})$$

$$\tau_{x'} = \frac{F}{A} \sin \Theta \cos \Theta = \frac{10{,}000}{\pi/4} \sin 20 \cos 20 = 4.09 \text{ ksi } (2.82 \text{ MPa})$$

Two points are noted here:

1. These answers do not describe the complete stress state of an element oriented as indicated, since, as shown shortly, there is a different normal stress acting 90 deg from $S_{y'}$. This does point out the danger in considering that stress is simply force divided by area.
2. To determine the full state of stress, the complete stress transformation equations should be used. Alternatively, this may be found by a graphical plot called *Mohr's circle*. Traditionally, normal stresses that cause extension are called *tensile* and carry a positive sign, whereas those causing contraction carry a negative sign and are called *compressive*.

6.3 MOHR'S CIRCLE AND PRINCIPAL STRESSES

For the majority of engineering problems, the magnitudes of the three so-called principal stresses are of greatest interest. Principal stresses are normal stresses acting on planes where shear stresses are zero. For *every* applied stress state there are three and *only* three principal stresses; they are mutually perpendicular. A circle plot, due to Mohr (Ref. [1]) is one way to determine the values of the two unknown principal stresses. For a uniaxial case, the applied stress is one of the principal stresses while the remaining two are both zero; thus there would be no need to plot a circle since all three of these stresses are noted directly. Thus, our main concern here is with so-called biaxial stress states.* Figure 6–2

* For those rare ''three-dimensional'' situations, the roots of a cubic equation must be found. These would be the three principal stresses. See, for example, reference [2].

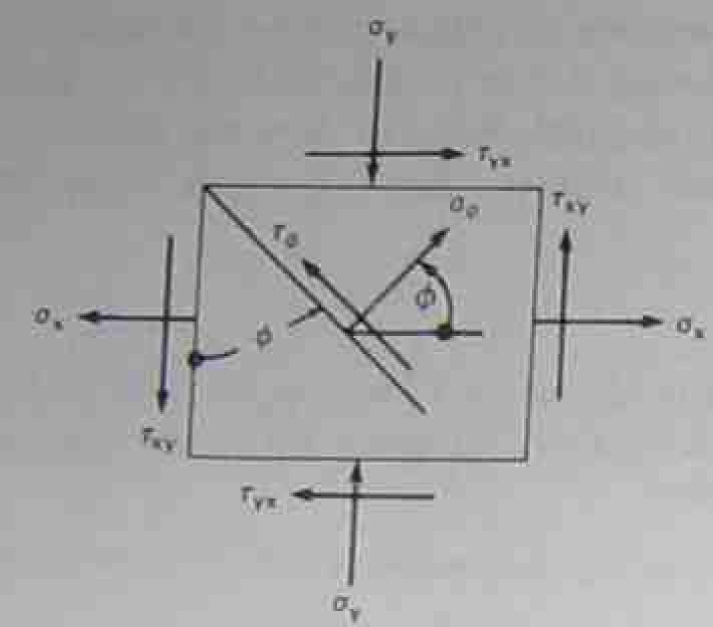


Figure 6–2 Stress element, in the physical plane, for a biaxial stress state.

illustrates such a situation, the known principal stress being zero. It is noted here that if this known normal stress were not *zero*, the stress state would still be *equivalent* to a biaxial condition in using Mohr's circle for determining the other two principal stresses.

Numerous *conventions* are used for plotting such a circle, and the one used here is the most sensible in our opinion. The convention employed is as follows:

1. Normal stresses are *plotted to scale* along the abscissa, tensile being positive and compression being negative.
2. Shear stresses are *plotted to the same scale* along the ordinate. If a shear stress would tend to cause clockwise rotation of the stress element, it is plotted as a positive value, whereas a tendency to cause counterclockwise rotation is considered negative.
3. Angles between two planes or directions on the circle plot are twice the corresponding angle on the physical plane containing the stress element. Figure 6–3 shows the correct circle plot for the general case given in Fig. 6–2, which indicates what is

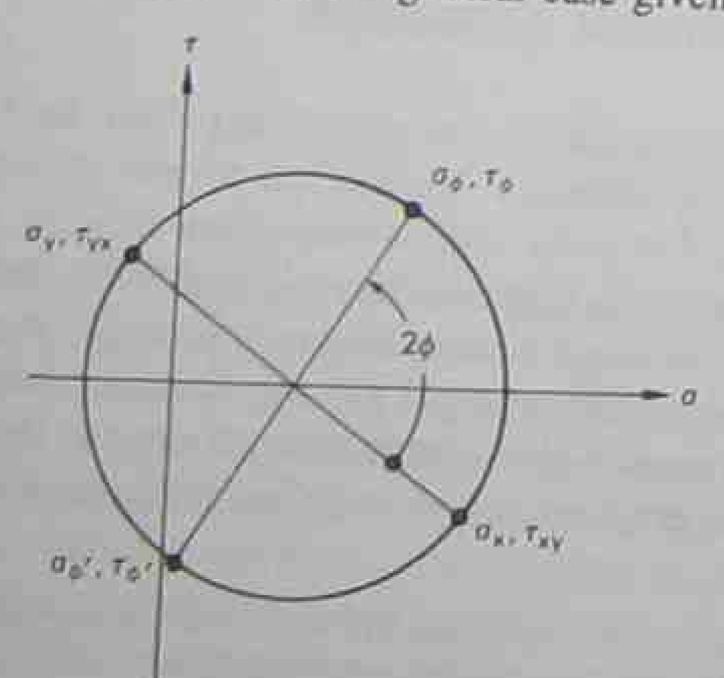


Figure 6–3 Plot of Mohr's circle for the stress state in Fig. 6–2.

meant by the physical plane. The *only* use of Mohr's circle in this text will be to define the *magnitudes* of the unknown principal stresses.

The equations that produce the values of the two unknown principal stresses and the maximum shear stress acting *in that plane* are

$$\sigma_{1,2} = \tfrac{1}{2}(\sigma_x + \sigma_y) \pm \tfrac{1}{2}\{(\sigma_x - \sigma_y)^2 + 4\tau_{xy}^2\}^{1/2} \qquad (6\text{–}2)$$

note that the third principal stress, σ_3, must be known at the outset. Then

$$\tau_{max} = \tfrac{1}{2}\{(\sigma_x - \sigma_y)^2 + 4\tau_{xy}^2\}^{1/2} \qquad (6\text{–}3)$$

gives the largest shear stress *in the x - y plane*. Note that the use of Eqs. (6–2) and (6–3) negates the need for using a plot of Mohr's circle if only the principal and maximum shear stresses *in the plane of concern* are required.

Example 6–2

1. A stress state is given by $\sigma_x = 10$, $\sigma_y = -5$, $\tau_{xy} = -6$ and $\sigma_z = \tau_{xz} = \tau_{zy} = 0$. Find the principal stresses from a plot of Mohr's circle.
2. A body is subjected to pure shear such that $\tau_{xy} = \tau_{yx} = 5$. Find the principal stresses. Pure shear implies that no normal stresses are applied.

Solution

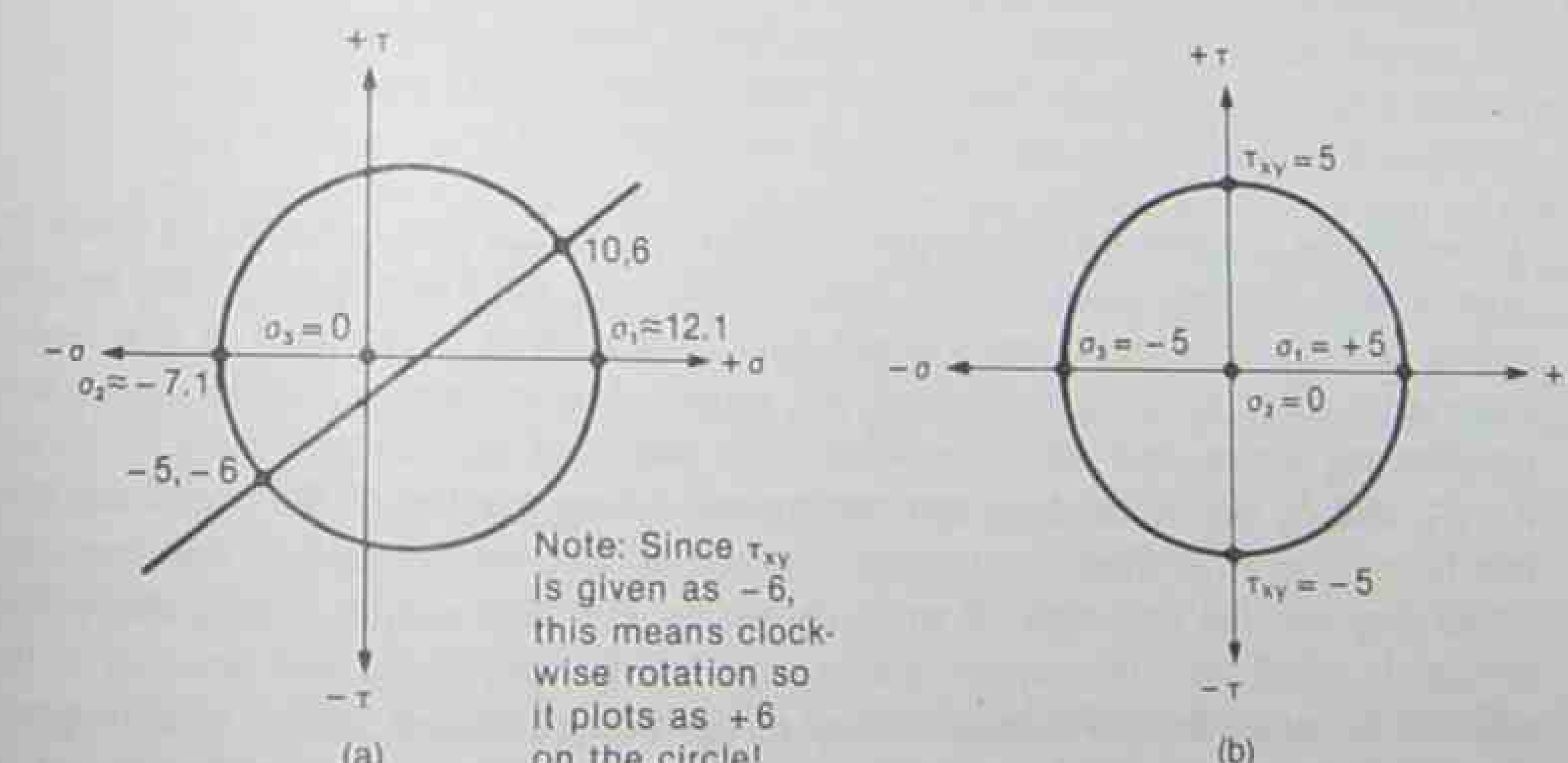


Figure E6-2

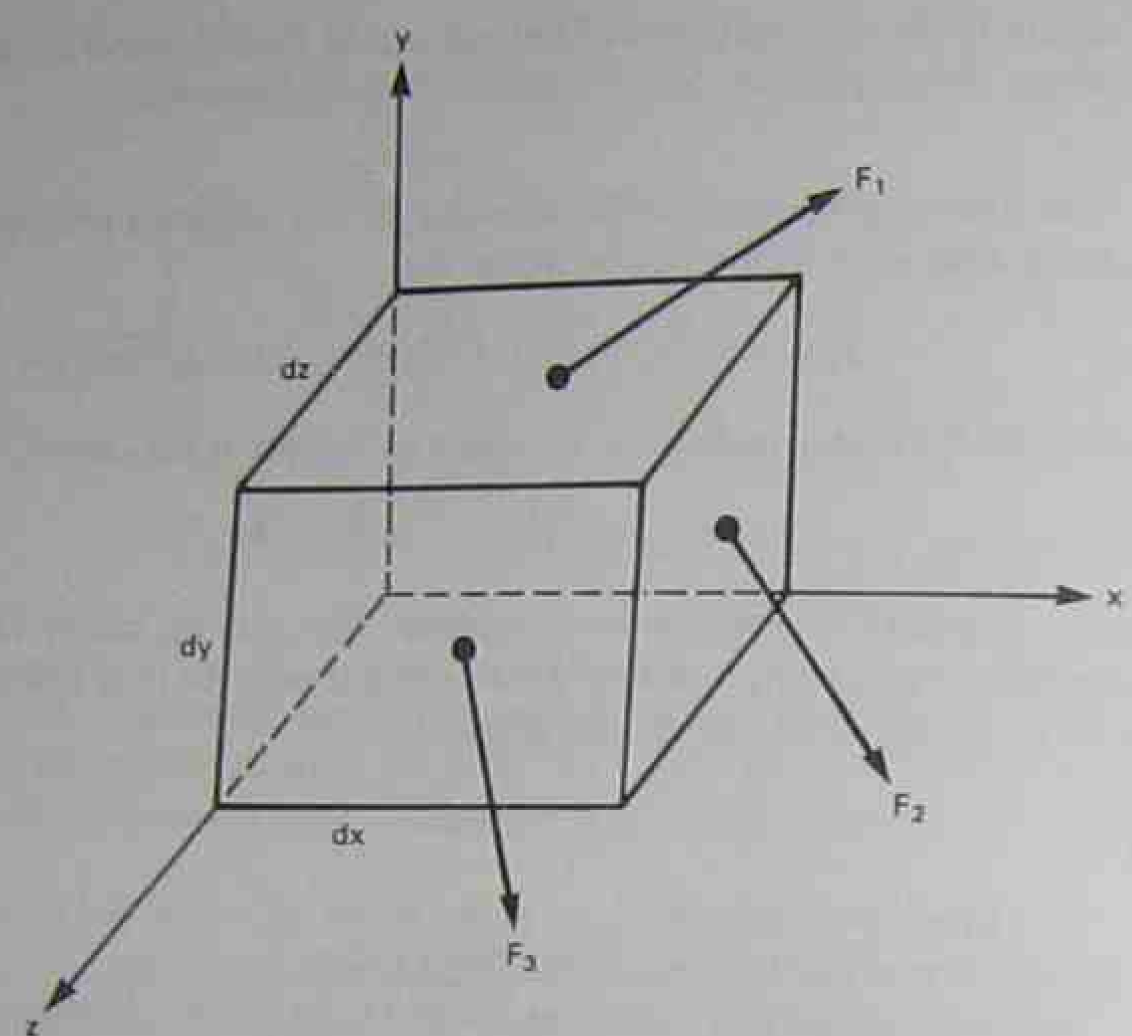


Figure 6–4 Generalized forces acting on a small body.

To portray the complete nature of stress, consider Fig. 6–4, where a number of forces act upon a small element which is under force equilibrium. All forces acting upon an individual face can be resolved into a single force; that is what is indicated in that figure. For clarity, only three such resultants are shown, but the reader should realize that equilibrating forces must act on the other three planes of the element. Now the forces F_1, F_2, and F_3 can be resolved into components acting parallel to the coordinate system $(x - y - z)$; if these components are then divided by the area of the face upon which they act, the total state of stress shown by Fig. 6–5 results. This collection of stresses is called the *stress tensor*, but in most real-life problems, and certainly in the majority discussed in later chapters, many of the stresses shown in Fig. 6–5 disappear (that is, they are zero). For this reason, further discussion of the stress tensor is unnecessary.

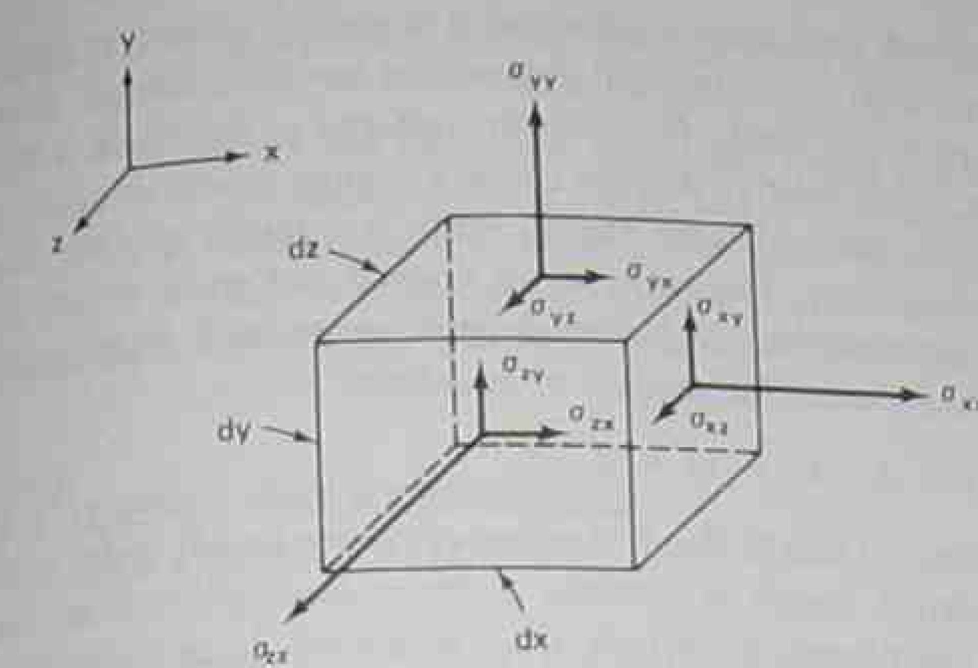


Figure 6–5 Stress element for a homogeneous state of stress. By convention, all stresses are considered positive as shown.

6.4 ALTERNATIVE DEFINITIONS OF NORMAL STRESS

Nominal (or engineering) normal stress carries the symbol S and is *defined* as

$$S \equiv F/A_0 \tag{6–4a}$$

where A_0 is the area of the body prior to the application of F.

True (or natural) normal stress carries the symbol σ and is defined as

$$\sigma \equiv F/A \tag{6–4b}$$

where A is the instantaneous or current area coincident with the load F. Shear stresses are not so distinguished. Although they are discussed in some detail later in this chpater, it is sensible here to indicate why these two definitions have been introduced. Students inevitably take courses in statics, strength of materials, and mechanical design *before* delving into manufacturing processes. Practically without exception, those earlier subjects are geared exclusively to elastic (small deformation) behavior of solids. Within such a constraint, any area *change* under a load F is so small that irrespective of whether one uses either A_0 or A, as in Eqs. (6–4a) and (6–4b), the computed values of S and σ are, for all *practical* purposes, identical. However, when large deformations of a body or workpiece are involved, as in plastic deformation, A becomes significantly less than A_0, and nominal stresses are no longer adequate to truly describe such a situation. Because *plasticity* is rarely ever discussed in traditional undergraduate courses, most students are never introduced to the concepts of true stress and true strain. In Sec. 6.6 we discuss their relationship.

6.5 STRAIN

When a solid is deformed under the application of external forces, points in the solid are displaced from their original positions or locations. Such displacements are used to *define* strain in such a way that excludes rigid body movements such as pure rotation or pure translation. For example, if a block rests upon a table and it is simply rotated or simply moved in some linear direction, none of the original dimensions has been altered; therefore, no strain has been induced in the block. If, however, one end of the block is held in place and external forces or, equivalently, stresses cause a change in any dimensions of the block, then strain has been induced. Like stress, strain is a tensor quantity, but nothing further need be said about that. The concepts of normal and shear strains are directly analogous to their stress counterparts if correct definitions are used. *Nominal* (or engineering) normal *strain*, which carries the symbol e, is defined as*

$$e \equiv \Delta \ell/\ell_o = (\ell - \ell_o)/\ell_o \tag{6–5a}$$

where $\Delta \ell$ represents some change in length due to loading, ℓ is the length under some applied load, and ℓ_0 is the original length of concern *prior* to load application. The corresponding *true* (also called logarithmic or natural) *strain* is defined, in an incremental way, as

$$d\epsilon \equiv d\ell/\ell \tag{6–5b}$$

which upon integrating between proper limits can be expressed as

$$\epsilon = \ell n(\ell/\ell_o) \tag{6–5c}$$

As with stress definitions, tensile strains are considered positive, while compressive are negative. In Sec. 6–6, we show how ϵ can be expressed by parameters other than length. The use of nominal values is again appropriate where small (e.g., elastic) deformations are involved, but the use of true strain proves of greater use and convenience where large (plastic) deformations are encountered. Note that both e and ϵ are dimensionless quantities.

Example 6–3

1. A bar of initial length of six inches (about 152 mm) is stretched to a length of eight inches (203 mm). Calculate the nominal and true strains induced in the direction of stretching.
2. If the eight-inch length is then further extended to twelve inches (304 mm), find the additional nominal and true strains induced during this step.
3. Compare the *sum* of both types of strain induced during the individual steps and then compare it with the strains that would have been computed if the six-inch dimension had been extended to 12 in. in a single extension. Assume that this uniform extension is possible for this problem.

Solution

1. $e_1 = 2/6$ [Eq. (6–5a)] $= 0.333$
   $\epsilon_1 = \ell n\ (8/6)$ [Eq. (6–5c)] $= 0.2877$
2. $e_2 = 4/8 = 0.500$
   $\epsilon_2 = \ell n\ (12/8) = 0.4055$

* Whenever e is used as the base of natural logarithms, it will be so stated to avoid confusion. The use of exp will also clarify matters.

3. $e_1 + e_2 = 0.833$ and $e_{total} = 6/6 = 1.0$
   $\epsilon_1 + \epsilon_2 = 0.6932$ and $\epsilon_t = \ell n(12/6) = 0.6932$.

Note that summation of nominal strains is not equivalent to the total strain whereas the summation of true strains is. This indicates what is called the *additive property* of true strains. In most real-life engineering problems, four-place decimal accuracy is to be questioned, even though calculators are capable of providing this. Young as well as older engineers must realize that such accuracy is seldom warranted in practice.

6.6 USEFUL RELATIONSHIPS AND OBSERVATIONS

As mentioned earlier, the concepts of nominal stress and strain are used extensively in undergraduate courses, since elastic deformation is of major concern. In situations where plastic deformation dominates, true stress and strain are more useful. Assuming that nominal values are available, the corresponding true values can be determined quite readily. This is now shown.

Although changes in volume occur during elastic deformation of most solids, these are very small. Volume changes during plastic deformation, especially with *metals*, are negligible. What this implies is that for a fixed *mass* of metal, the *density* remains essentially constant during such deformation. Since metallic structures have very high packing factors (that is, the atoms are very closely packed), it is difficult to pack them any closer; thus, the structure retains its density. This is not true for other solids such as *polymers*, since they have structural configurations much different than metals. Because of these different conditions, what follows is related to metals subjected to plastic deformation. If volume constancy prevails (that is, the solid is practically incompressible), then

$$V_0 = A_0\ell_0 = A\ell = V \tag{6–6}$$

where the subscript zero refers to initial conditions under no load and the other terms refer to instantaneous values under load. Considering Eqs. (6–4a) and (6–4b) and assuming some force F, we have

$$F = SA_0 = \sigma A \quad \text{or} \quad \sigma = S(A_0/A) \tag{6–7}$$

From Eq. (6–6), (A_0/A) equals (ℓ/ℓ_0) and with Eq. (6–5a), (ℓ/ℓ_0) equals $(1 + e)$; thus

$$\sigma = S\,(1 + e), \qquad \text{so } \sigma \text{ is always greater than } S. \tag{6–8}$$

Now by combining Eqs. (6–5a) and (6–5c), we obtain

$$e = (\ell/\ell_0) - 1 \quad \text{or} \quad \ell/\ell_0 = 1 + e \tag{6–9}$$

which leads to

$$\epsilon = \ell n\,(1 + e) \tag{6–10}$$

Hence, for uniform deformation, e is always $> \epsilon$.

Thus, if nominal values of $S - e$ coordinate points are available, they can be converted to equivalent $\sigma - \epsilon$ values from Eqs. (6–8) and (6–10). Several points are worth noting here:

1. To use Eqs. (6–8) and (6–10) in a meaningful way, the nominal strain e must be *uniform*.
2. From Eq. (6–6), it is a simple matter to express Eq. (6–5c) as

$$\epsilon = \ell n(A_0/A) \tag{6–11a}$$

which for round cross sections gives

$$\epsilon = 2\ell n(D_0/D) \quad \text{in terms of diameters.} \tag{6–11b}$$

3. If strains are small, say elastic strains where $e < 0.010$, then $e \approx \epsilon$ and $\sigma \approx S$. As an example, consider a nominal strain of 0.005; then

$$\epsilon = \ell n(1 + 0.005) = 0.004987 \approx 0.005 \tag{6–12}$$

which is practically identical to e. The corresponding σ would only be 0.5 percent greater than S. This is illustrates why the use of true stress and strain is never introduced when elastic deformations are involved; there is no need to do so.

4. Next consider a specimen of length ℓ_0 that is doubled in length. The comparable strains are

$$e = (2\ell_0 - \ell_0)/\ell_0 = 1 \tag{6–13}$$

and

$$\epsilon = \ell n(2\ell_0/\ell_0) = 0.693 \tag{6–14}$$

both being tensile. What is required to induce equivalent compressive strains by decreasing ℓ_0 to a necessary level?
There,

$$e = -1 = (\ell - \ell_0)/\ell_0 \tag{6–15}$$

so ℓ must approach zero; that is, the value of ℓ_0 must be reduced to *zero* thickness, which is physically impossible. Also, with this definition a value of e of -1.1 is impossible to attain. Now

$$\epsilon = 0.693 = \ell n(\ell/\ell_0) \tag{6–16}$$

so ℓ must be equal to $\ell_0/2$ to provide this compressive true strain. It is noted that doubling the length or halving it (that is, the true strains are identical except for sign) does produce quite similar changes in the deformed structure of the metal as indicated by property measurements. The same is not true if equivalent nominal strains are involved (that is, $e = \pm 1.0$).

5. For volume constancy,

$$V_0 = t_0 w_0 \ell_0 = V = tw\ell \tag{6–17}$$

where the symbols indicate thickness, width, and length, respectively, and any

instantaneous volume V under load is equal to the original volume V_0. Then

$$dV = dV_0 = 0 = twd\ell + w\ell\, dt + \ell t\, dw \tag{6–18}$$

or

$$dw/w + d\ell/\ell + dt/t = 0 \tag{6–19}$$

so, using Eq. (6–5b) in a general way, we obtain

$$d\epsilon_1 + d\epsilon_2 + d\epsilon_3 = 0 \tag{6–20}$$

Thus, for volume constancy, the sum of the three incremental normal true strains is zero.* This is a useful relationship in plasticity calculations, and it is noted that if nominal strains are used, the resulting expression is not so simple. If Eq. (6–20) is integrated by using proper limits, then

$$\epsilon_1 + \epsilon_2 + \epsilon_3 = 0 \tag{6–21}$$

where these are total strains. Whenever this relation is used in plasticity analyses, the *elastic* portion of the total strain will be neglected; that is, the plastic portion is taken as being equal to the total. In large deformation processes, this is a reasonable assumption and leads to great simplification.

Example 6–4

1. A force of 2000 lb (8.96 kN) is applied to the ends of a solid rod of 0.250 in. diam. (6.35 mm); under load, the diameter reduces to 0.200 in. (5.08 mm).
Assuming uniform deformation and volume constancy,
a. Find the nominal stress and strain.
b. Find the true stress and strain.
2. If the original bar had been subjected to a true stress of 50,000 psi (345 MPa) and the diameter was 0.220 in. (5.59 mm) under that stress, what is the nominal stress and strain for these conditions?

Solution

1. a. With Eq. (6–4a), $S = F/A_0 = (2000)/(\pi/4)(1/4)^2 = 40.74$ ksi (280.9 MPa)
With Eq. (6–5a), $e = (\ell - \ell_0)/\ell_0 = (\ell/\ell_0) - 1$ and with volume constancy, $A_0\ell_0 = A\ell$ or $\ell/\ell_0 = A_0/A$, so $e = (A_0/A) - 1 = [(0.25)^2/(0.20)^2] - 1 = 1.56 - 1 = 0.56$.
b. Since $\sigma = F/A$, $\sigma = (2000)/(\pi/4)(0.2)^2 = 63.66$ ksi (439 MPa). Check with Eq. (6–8). $\sigma = S\,(1 + e) = 40.74\,(1.56) = 63.6$ ksi.
Now $\epsilon = 2\ell n\,(D_0/D)$ from Eq. (6–11), so
$\epsilon = 2\ell n\,(0.25/0.20) = 0.446$
Check with Eq. (6-10). $\epsilon = \ell n\,(1 + e) = 0.446$.
2. The true strain is $\epsilon = 2\ell n\,(0.25/0.22) = 0.256$, so
$\epsilon = \ell n\,(1 + e) = 0.256$; therefore, $e = 0.292$. *Note* that
$e^{0.256} = (1 + e)$ can cause confusion since e on the left is the *base of natural*

* Only normal strains relate to volume changes. Shear strains cause a shape change only.

logarithms, whereas e on the right signifies nominal strain. Using exp (0.256) = $(1 + e)$ is far better.

Now $\sigma = S(1 + e)$, so $S = 50{,}000/1.292 = 38{,}700$ psi or 38.7 ksi.

Example 6–5

A metal specimen of 0.357 in. diameter (9.07 mm) is loaded in tension. When the applied force reaches 3000 lb (13.44 kN), elastic behavior ends; careful measurements show that at that instant, the diameter is 0.3566 in. Compare the nominal and true stresses.

Solution

$$S = (3000)/(\pi/4)(0.357)^2 = 29.97 \text{ ksi } (206.6 \text{ MPa})$$
$$\sigma = (3000)/(\pi/4)(0.356)^2 = 30.04 \text{ ksi } (207.1 \text{ MPa})$$

Note that the strain here is about 0.001 and that for all practical purposes, $S = \sigma$.

Example 6–6

A small circle of 0.500 in. diameter (12.7 mm) is printed on the face of a thin, flat sheet of steel. Under applied forces in the plane of the sheet, the circle changes into an ellipse having major and minor diameters of 0.600 and 0.520 in., respectively (call these the 1 and 2 directions). If direction 3 is normal to the surface, find ϵ_3 at this instant.

Solution

$$\epsilon_1 = 2\ell n\,(0.6/0.5) = 0.182 \text{ (positive)}$$
$$\epsilon_2 = 2\ell n(0.52/0.5) = 0.078 \text{ (positive)}$$

With volume constancy, Eq. (6–21) gives

$$\epsilon_3 = -(\epsilon_1 + \epsilon_2) = -0.260 \text{ (negative)}$$

6.7 MECHANICAL BEHAVIOR AND PROPERTIES OF SOLIDS

When any body or structure is subjected to external loads or forces, the body will elongate, shorten, distort, or undergo a combination of such shape changes; the overall effect will depend upon the type of loading. If the magnitudes of the induced stresses are less than some critical value, the deformation or behavior is called *elastic*, since upon removal of all loads, the body recovers its original shape or dimensions, much like the removal of a load from a spring.

If the induced stresses are large enough, then upon load removal, the body will display dimensions that differ from their original values. In addition a definite change in *shape* can occur. Such behavior is called *plastic*, and the body is said to have *yielded*. With many solids, elastic deformations are small, whereas plastic deformations can be much larger.

In many design considerations, plastic deformation is to be avoided, and *elastic properties* are used in calculations. However, in producing components such as an automobile fender or oil pan, it is essential to induce large plastic deformation in an initially flat sheet in order to *form* the desired part. There, the plastic properties of the material are of primary concern. What follows is a coverage of the major properties of interest; it discusses how they are determined and, in certain instances, points out some precautions that should be considered when one intends to use such properties selected from sources such as handbooks.

6.8 THE TENSILE TEST— ELASTIC PROPERTIES FOR ISOTROPIC SOLIDS*

Beyond question, this is the most widely used test in studies of mechanical behavior and, in all likelihood, most readers have been exposed to such a test. Although it is relatively straightforward, there are certain subtleties and restrictions that are often either overlooked or not pointed out to students when such tests are conducted. The measurements are generally obtained under controlled conditions, these being

1. The rate at which deformation is induced is quite low. In terms of *strain rates* (that is, the strain achieved divided by the time to achieve it), these are of the order of 10^{-3} per second. More is said about strain rate in Chap. 8.
2. Temperature of the surroundings is of the order of 25°C.
3. All measurements are restricted to a *gage section* that experiences a state of uniform, uniaxial stress during deformation.

For many materials, properties determined from a standard tensile test may be quite different if the magnitudes of strain rate and temperature are at noticeable variance with 1 and 2 above. Thus, care must be exercised if so-called *basic* property values, as usually listed in standard handbooks, are used in analyses where the actual conditions under which deformation will take place are greatly different from those under which that basic property was determined. A careful literature search will often provide more useful property values where, for example, strain rates and temperature are quite different from those listed above.

The basic information obtained from a tensile test is the magnitude of the force F required to cause a certain extension $\Delta\ell$. These findings are then converted to some type of stress-strain data; the definitions given earlier are used. Figure 6–6 is a schematic of a load-extension diagram typical of most *ductile metals*. Note that on the strain scale, which includes the total extension, the elastic region encompasses only a small portion of the full diagram. As such, it appears as a nearly *vertical* line if all data are plotted against a common scale. Figure 6–6(b) displays the elastic region where the strain scale is expanded; note that on Fig. 6–6 linear or cartesian coordinate scales are used. If the load-extension data are converted to nominal stress-strain data, by using Eqs. (6–4a) and (6–5a), the shape of the S-e curve is the same as the F-$\Delta\ell$ curve, since the constants A_0

* *Isotropy* means that a particular property is not affected by the *direction* in which the property is measured. *Homogeneity* means that a property does not vary with *location* in the body.

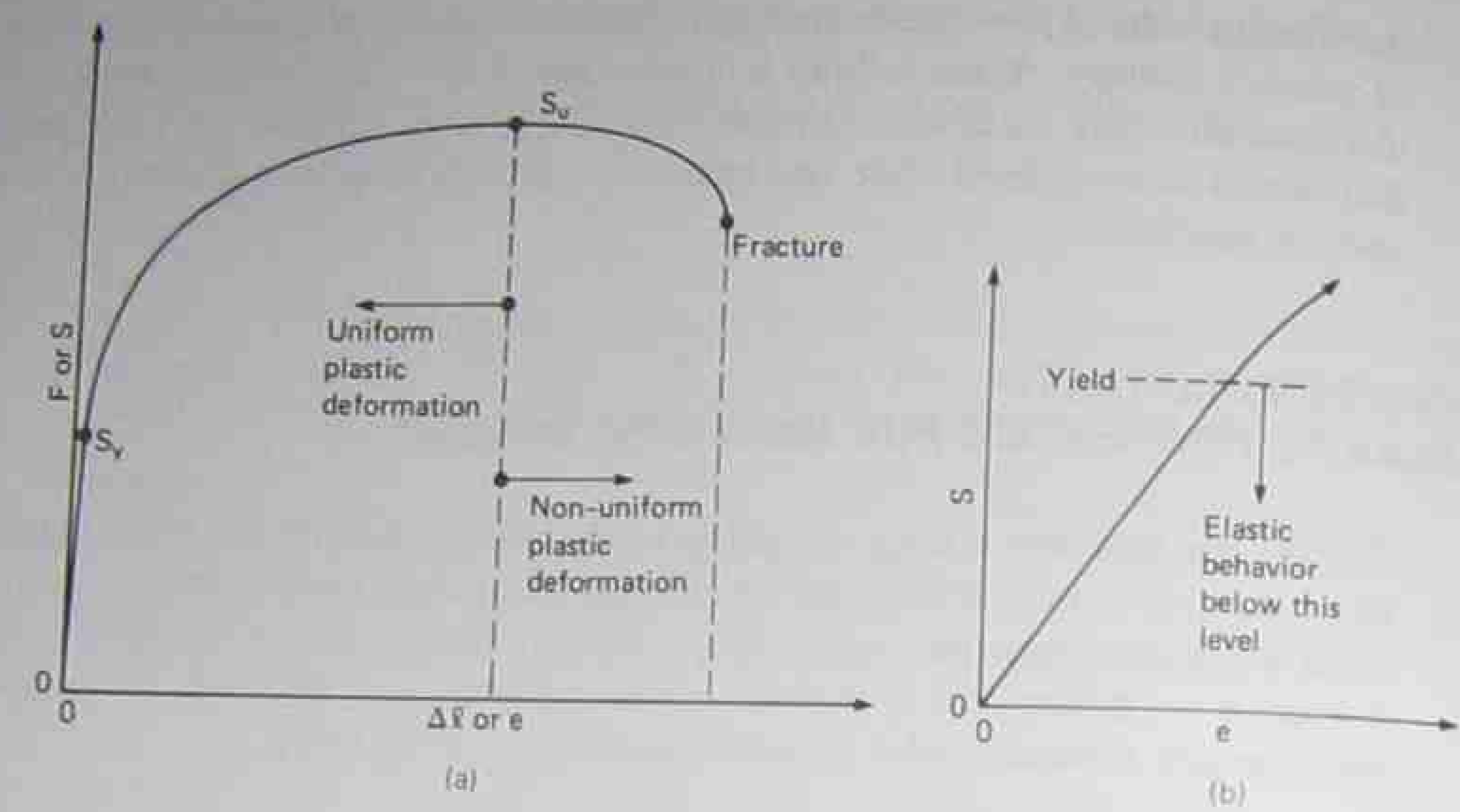


Figure 6–6 (a) Load-extension or nominal stress-strain plot for a ductile metal during a tension test, and (b) the expansion of the strain axis to clearly show the elastic region.

and ℓ_0 are used in the conversions; thus, Fig. 6–6(a) carries dual notations. Note that the load reaches some maximum or ultimate value and then proceeds to decrease.

One important elastic property is related to the initial part of the *S-e* curve, the slope of which is *linear* for almost all metals. This is the *elastic modulus E* (also called Young's modulus), which is

$$E = S/e \tag{6–22}$$

This relation, which describes elastic behavior, is often called Hooke's law.* Physically, *E* is a measure of *stiffness* or rigidity and carries units of stress, since strain is dimensionless.

A second elastic property that can be evaluated from a tensile test is called Poisson's ratio, ν. This requires the simultaneous measure of length and lateral (i.e., *across* the section) dimensional changes. Calling the loading direction 1 and the transverse direction 2 or 3, where these are perpendicular to each other, then

$$\nu = -e_2/e_1 = -e_3/e_1 \tag{6–23}†$$

and for *metals* the usual range of ν is from about 0.2 to 0.4.

We note that a distinction is sometimes made among the elastic limit, proportional limit (neither of which will be discussed here), and the yield strength but such a distinction is unnecessary for our purposes. Sole emphasis will be placed upon the *yield strength*, which is considered to be the level of stress that coincides with the onset of plastic

* Strictly speaking, Hooke discussed the relation between load and extension, since stress and strain were not defined at that time; see Ref. [3] for a discussion. Later in this chapter, "Hooke's law" will be cast in the more general, three-dimensional form.

† $\epsilon_2 = \epsilon_3$ due to the assumption of isotropy, and both are compressive here. Note that since e_2 and e_3 are compressive, ν is a positive number.

deformation. Technically this is not an elastic property, yet it is useful to define it here, using the symbol *Y*, so that

$$Y \equiv F_y/A_0 \tag{6–24}$$

For most ductile metals, the actual onset of yielding is difficult to measure, and the simplest way to determine this property is by some definition, as is discussed shortly.

As with *Y*, the property called *tensile* (or ultimate) *strength*, which is certainly not an elastic property, is also included here; its symbol is S_u and it is defined as

$$S_u \equiv F_u/A_0 \tag{6–25}$$

where the F_u implies the ultimate or maximum load experienced by the specimen. Note that S_u is the largest *nominal* stress measured during a test. What Eqs. (6–24) and (6–25) indicate are that *strength* properties of the test material are particular stress levels of importance. Do not *confuse* strength and stress. Depending upon the applied load, stress can have many values, whereas strength indicates a particular property of the material.

Attention is called to the fact that from the onset of yielding *Y* up to the tensile strength S_u the tensile specimen undergoes essentially *uniform* plastic deformation such that changes in area along the gage section decrease uniformly during loading.

Other plastic properties, of which little use is made in this text, are simply defined here. The *bulk modulus B* is

$$B = E/3(1 - 2\nu) \tag{6–26}$$

while the *shear modulus G* is

$$G = E/2(1 + \nu) \tag{6–27}$$

and the full derivation of these relations may be found elsewhere [2]. What is implied by Eqs. (6–26) and (6–27) is that of the four elastic *constants*, *E*, ν, *B*, and *G*, only two are independent. Thus, if *E* and ν are determined experimentally, then *G* and *B* can be calculated directly. If a solid exhibits anisotropic behavior, then these *constants* vary with direction and the simplified form of, say, Eq. (6–22) is no longer valid. Anisotropic elasticity is beyond the intent of this text, but coverage can be found elsewhere [4].

Returning to Eqs. (6–22) and (6–23), consider the case where three normal stresses are applied simultaneously to an isotropic body. Each not only induces a normal strain but also causes "Poisson strains" in the other two directions. Thus, in say the one direction, the induced strains are

$$e_1 = S_1/E, \quad e_1 = -\nu S_2/E, \quad e_1 = -\nu S_3/E \tag{6–28}$$

Since the elastic equations are all linear, the concept of superposition can be applied so that the total strain in that direction is

$$e_1 = (1/E)[S_1 - \nu(S_2 + S_3)] \tag{6–29}*$$

* If the three normal strains e_1, e_2, e_3 are added together, they equal zero *only* if $\nu = 0.5$; this would indicate constancy of volume.

where the signs of stresses [that is, positive (tensile) or negative (compression)] must be adhered to. This is the three-dimensional form of Hooke's law and includes the special case for uniaxial tension where $S_2 = S_3 = 0$. Strains in the other two directions result by altering the subscripts in Eq. (6–29).

Since $\nu < 0.5$ for most solids, there is a volume *increase* during tensile deformation or a *decrease* during compression. Since, however, these volume changes are very small, and since the primary interest in this book is not concerned with elasticity theory, we shall always *assume* volume constancy as a simplification.

Example 6–7

A specimen of steel, having properties of $E = 30 \times 10^6$ psi (207 GPa) and $\nu = 0.3$, is subjected to stresses $S_1 = 10$ ksi, $S_2 = 5$ ksi, and $S_3 = -7$ ksi. What is the strain in direction 3?

Solution Noting that $E = 30 \times 10^3$ ksi, recast Eq. (6–29) as

$$e_3 = (1/E)\,[S_3 - \nu(S_1 + S_2)] =$$
$$1/(30 \times 10^3)\,\{-7 - 0.3(10 + 5)\} = -0.383 \times 10^{-3}$$

6.9 THE TENSILE TEST— PLASTIC PROPERTIES FOR ISOTROPIC METALS

The ability of many metals to undergo a large degree of plastic deformation prior to fracture is referred to as *ductility*. To quantify this property, two commonly used measurements are developed from a tensile test; note that both are made *after* the specimen has fractured into two pieces. These properties are defined as

$$\%\text{ Reduction of area} = A_r = 100\,(A_0 - A_f)/A_0 \qquad (6\text{–}30)$$

and

$$\%\text{ Elongation} = \mathscr{E}\ell = 100\,(\ell_f - \ell_0)/\ell_0 \qquad (6\text{–}31)$$

where the subscript f refers to the *minimum* specimen area and the gage length after fracture. Figure 6–7 illustrates these points, noting that the $\mathscr{E}\ell$ is simply the nominal strain at fracture times 100. As seen on that figure, as extension is continued beyond the maximum load F_u, deformation becomes highly localized, and the phenomenon called *necking* begins. Most of the further plastic deformation, from this point until fracture, occurs in the region of the neck; as a consequence, the overall deformation between the initial gage marks becomes more and more nonuniform, so any measure of nominal strain is really a meaningless average. In situations where necking does occur, $\mathscr{E}\ell$ in Eq. (6–31) does not imply uniform extension. With regard to A_r, which is based upon the *minimum* area of the neck after fracture, this is indicative of the single *largest* deformation or strain induced in the specimen.

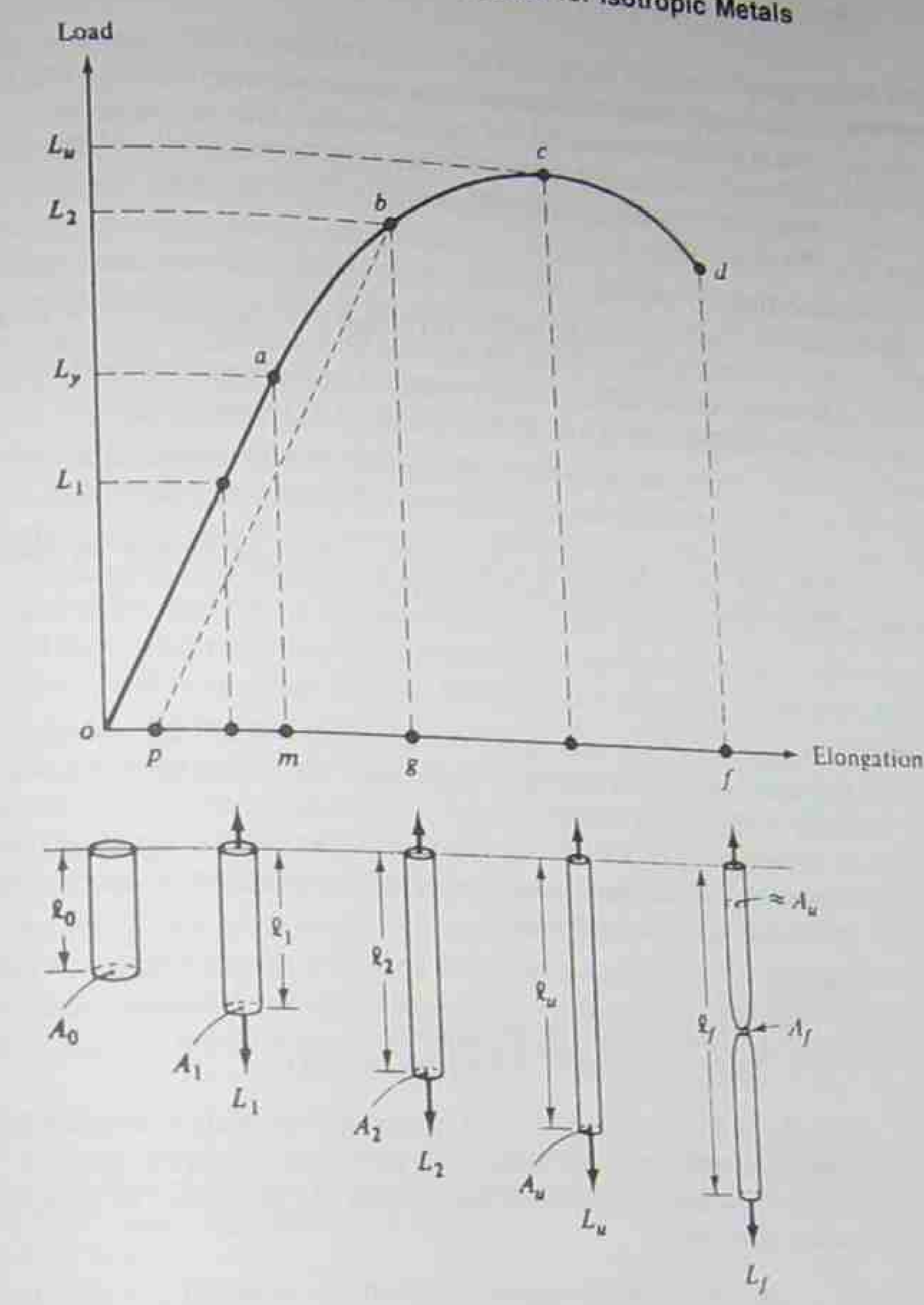


Figure 6–7 Tensile test results indicating changes in gage length at different loads. Note necking after L_u.

Although this fact is often overlooked, it should be realized that these indicators of ductility are influenced by whatever values of ℓ_0 and A_0 are chosen initially. For instance, with necking occurring, smaller values of ℓ_0 will lead to larger values of $\mathscr{E}\ell$ and, in a similar way, different starting values of A_0 can affect A_r. (The affect here is much less pronounced than is the effect of ℓ_0 on $\mathscr{E}\ell$.) Because of these possibilities, *standard* values of ℓ_0 and A_0 are used so that comparisons of these two properties, when obtained from standard specimens, will have some rational meaning. Yet, because the numerical values of both are geometry dependent, they are not truly basic in that context.

Example 6–8

Under tensile deformation, a specimen fractures when the *maximum* load is reached; the gage length was 60 mm and the area was 83.33 mm². Prior to loading, ℓ_0 = 50 mm and A_0 = 100 mm². Find the true strain at fracture, ϵ_f, using both length and area changes, and explain why these results occur. Also find A_r and $\mathscr{E}\ell$ and show that the volume did not change during this deformation.

Solution

$$\epsilon_f = \ell n(60/50) = 0.182 \text{ and } \epsilon_f = \ell n(100/83.33) = 0.182$$

Since the deformation was *uniform* up to F_u and fracture occurred at maximum load, the fracture strains must be the same since $(\ell/\ell_0) = (A_0/A)$.

$$A_r = (100)(100 - 83.33)/100 = 16.67 \text{ percent from Eq. (6–30)}$$
$$\mathscr{E}\ell = (100)(60 - 50)/50 = 20 \text{ percent from Eq. (6–31)}.$$

Note that $V_f = A_f\ell_f = (1.2)\ell_0(0.833)A_0$, which agrees with the statement of uniform deformation (and constancy of volume) above.

Example 6–9

1. A tensile specimen, with ℓ_0 = 2.000 in. and D_0 = 0.505 in., is pulled to fracture; necking preceded fracture. The final gage length was 2.780 in. and the *minimum* neck diameter was 0.321 in. Repeat Ex. 6–8.
2. A second value of ℓ_0 = 1.400 in., placed on the same specimen, showed a length at fracture of 2.05 in. Find the value of $\mathscr{E}\ell$ for these numbers, compare with 1, and comment on the reason for any difference.

Solution

1. $\epsilon_f = \ell n(2.78/2) = 0.329$ (length changes)
$\epsilon_f = 2\ell n(0.505/0.321) = 0.906$ (area changes)

Equivalence *does not* result here due to necking. The strain based upon length changes is a meaningless value, since the strain *along* the length is highly nonuniform. The value based upon the minimum neck diameter does indicate the *maximum* true strain induced; as such it is often more useful.

$$A_r = (100)(A_0 - A_f)/A_0 = 100\,(0.2 - 0.081)/0.2 = 59.5 \text{ percent}$$
$$\mathscr{E}\ell = (100)(2.78 - 2)/\,2 = 39 \text{ percent}$$

Here, $V_f = (1.39)\ell_0(0.405)A_0 = 0.563A_0\ell_0$. This does *not* mean that volume constancy did not occur; rather, the nonuniform deformation does not permit this type of calculation as in Ex. 6–8, where deformation was uniform.

2. $\mathscr{E}\ell = 100(2.05 - 1.4)/1.4 = 46.4$ percent. After necking, further extension is concentrated in the region of the neck, and shorter starting gage lengths are more proportionally affected; this leads to a larger $\mathscr{E}\ell$.

The end of fully elastic behavior and the onset of yielding is not usually displayed by an abrupt or obvious change in the $F - \Delta\ell$ plot; rather, there exists a gradual region of transition. The most likely cause of this behavior is the variation of orientation of grains

within the specimen. Due to differences in the alignment of crystallographic planes with respect to the applied load, slip occurs more readily in some grains than others.* In succession, as the load gradually increases, other less *favorably oriented* grains then slip until finally the entire structure has been plastically deformed. Beyond this point, as loading continues to increase, the entire specimen displays macroscopic uniform plastic deformation. This overall behavior is shown schematically in Fig. 6–8 (note that the elastic region is greatly exaggerated here to assist this explanation) for a specimen of aluminum and is *typical* of most ductile metals. Because of the uncertainty of exactly where yielding begins, it is common practice to *define the yield strength* by using an *offset* method. For a 0.2 percent offset, which is often used, a line is drawn parallel to the initial elastic region but offset by a *strain of 0.002* as indicated. The intersection of this line with the curve gives a stress that is taken as the yield strength (this is also called the *proof* stress), and it is obvious that this stress level will induce a small amount of permanent strain (0.002) in the specimen. A few metals, notably annealed, low-carbon steel, display a true yield point; see Fig. 6–8. In fact, one usually sees an upper *A* and lower *B* yield point for this material.† Since *B* is much less influenced than *A* by variations in loading rate, specimen alignment, and the stiffness of the testing machine, the yield strength in such cases is based upon *B*, and there is no need to employ an offset method. Note also that Fig. 6–8 shows plots on cartesian coordinates, and the higher elastic modulus of steel is reflected by a steeper slope.

With reference to Fig. 6–6(a) it is seen that as elongation continues beyond initial yield, an increase in load is demanded; this occurs at a continuously decreasing *rate*. As mentioned earlier, the deformation from the onset of yielding up to this maximum load is uniform. The reason for this $F - \Delta\ell$ behavior is attributed to the phenomenon called *work hardening* (or *strain hardening*). As explained by dislocation theory,‡ initial plastic deformation requires increased loading for the next increment of deformation to occur. As a consequence, plastic deformation increases the current yield strength of the specimen to such an extent that it offsets the accompanying decrease in cross-sectional area; therefore,

* See Chapter 7 for further discussion.

† At strains beyond *B* a horizontal sawtooth region follows. These small oscillations are caused by micro-neck formation that produces localized necks along the gage section; they are called Lüders bands, lines, or strains. See Chapter 9 regarding stretcher strains.

‡ See Ref. [2].

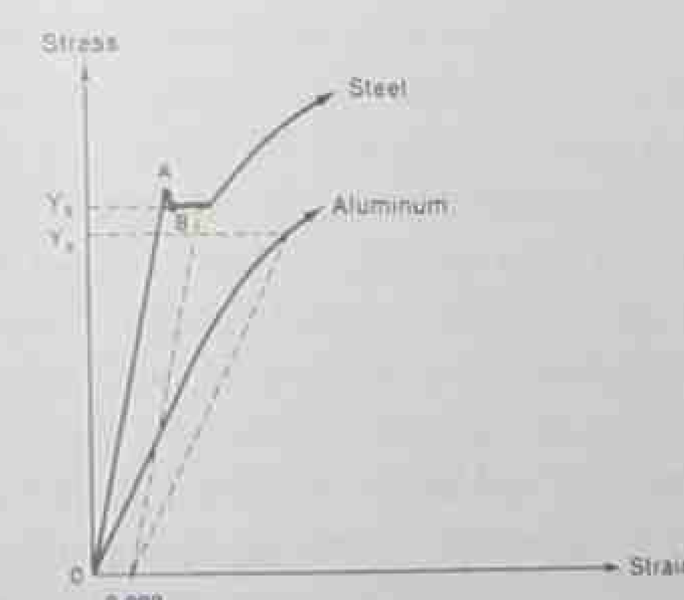


Figure 6–8 Offset method used to define yield strength.

the *load-carrying* capacity increases (that is, an increase in load is necessary to cause further elongation). Once F_u is reached, a condition of *tensile instability* occurs; note that $dF = 0$ at this point, and *necking* begins somewhere along the specimen. Although the metal in the neck *continues* to strengthen, it can no longer compensate for the faster-paced reduction of the minimum neck area, which now governs the load necessary to cause further extension. Thus, although the neck possesses the *current* highest yield strength throughout the specimen, it has the *lowest* load-carrying capacity, and, as a consequence, a drop in load results as extension continues. Testing machines are designed to apply only that load needed to cause further extension, and this is why the load drop from necking to fracture is observed. If static loads (for example, dead weights) were applied, the specimen would fracture at F_u, since the necessary load reduction would not be accomplished. Thus the major significance of the *tensile strength* S_u is that it indicates the maximum load-carrying capacity that a part of starting area A_0 can withstand.

The principal *plastic properties* determined from a tensile test are most readily found by plotting true stress versus true strain; that is, σ versus ϵ as defined by Eqs. (6–4b) and (6–5c). When plotted on cartesian coordinates along with their nominal counterparts, such results are shown schematically in Fig. 6–9.

From that figure it is seen that at initial yielding, S_y and σ_y are equivalent (see Example 6–5), so no distinction is made between the nominal and true yield strength; this is simply referred to as Y. Using the definitions from Eqs. (6–8) and (6–10) we see that it is obvious that as extension proceeds beyond Y, the $\sigma - \epsilon$ curve must be displaced above the $S - e$ curve, where, up to ultimate, for any $S - e$ combination the resulting σ is $>S$, whereas $\epsilon < e$. Beyond necking, σ continues to rise until fracture and, if careful

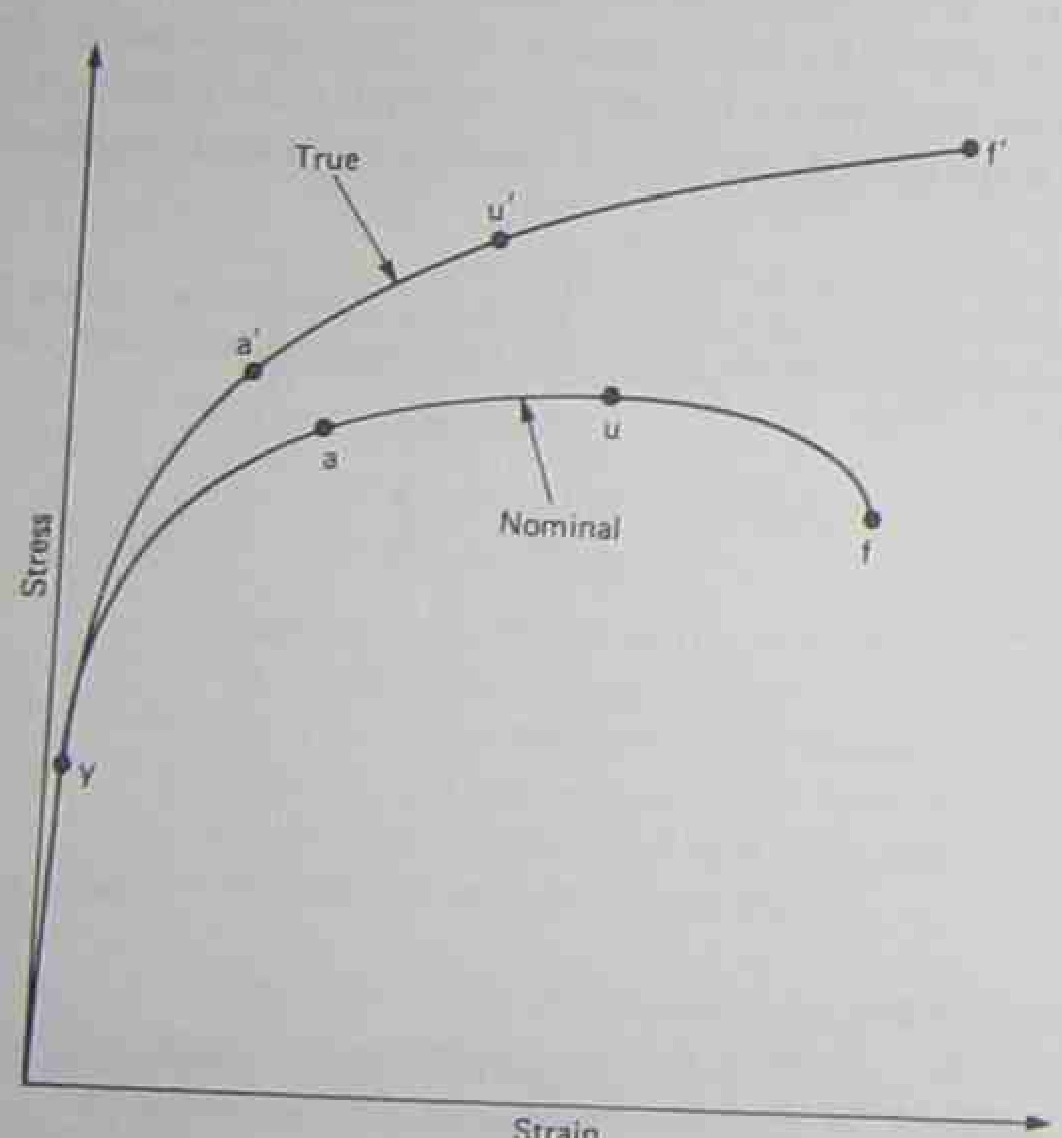


Figure 6–9 Comparison of nominal and true stress-strain curves.

measurements are made, and necking is relatively large, the values of ϵ based upon the minimum neck diameter can exceed e. For this reason, ϵ_f is shown to be larger than e_f on Fig. 6–9.

With many ductile metals that contain no effects of prior work hardening (that is, they are fully annealed) *prior* to tensile testing, the $\sigma - \epsilon$ behavior can be well described by a power-law expression,

$$\sigma = K\epsilon^n \tag{6–32}$$

where K is called the *strength coefficient*, having the same units of stress as σ and n is the *strain hardening exponent*. Equation (6–32) is called the *strain hardening equation*. As will be shown shortly, n indicates the extent of uniform true strain a metal can withstand under uniaxial tension. Other forms of strain-hardening equations have been proposed (see, for example, Ref. [5]), but Eq. (6–32) will be used in this text. Manipulations of Eq. (6–32) as given below can be applied to other forms of $\sigma - \epsilon$ behavior.

If we assume that Eq. (6–32) is applicable, the most direct way to determine K and n is to plot the $\sigma - \epsilon$ data on logarithmic coordinates, since a straight line must result. Figures 6–10 through 6–12 show actual test results using commercially pure aluminum, a plain low-carbon steel, and an alpha brass.

To determine K, each line is extrapolated to intersect a strain of unity (note that although ϵ_f is <1 in each case, we are only interested in the *equation* of the line itself) so that the value of K is equivalent to the stress corresponding with $\epsilon = 1$. By measuring the *slope* of the line, or by selecting two pairs of $\sigma - \epsilon$ values that lie *on the line* and using Eq. (6–32) with K defined numerically, the value of n is determined.

With reference to Fig. 6–10, the following are noted:

1. Up to a strain of about 0.0005 (that is, Zone 1) elastic behavior results. Here $\sigma = E\epsilon$ (or $S = Ee$) results, and if that line is extrapolated to a strain of unity, the corresponding stress level is the value of E. This provides a most practical way to determine E for solids, if, of course, Hooke's law prevails, since once the points are plotted they *must* lie on a line at 45 deg to either axis, so a best fit is easy to obtain without recourse to mathematical curve fitting.
2. For strains between 0.0005 and slightly less than 0.01 (Zone 2) the transition behavior can be seen. Some workers have fitted a straight line to such points and consider the material to show a *double n* value (that is, one at low plastic strains and another for higher levels). What seems to be overlooked here, and would be made perfectly clear if a *full* plot of $\sigma - \epsilon$ data were produced on logarithmic coordinates, is that seemingly *all* ductile metals would show such a double n value. Consider both Figs. 6–11 and 6–12 in this context where, for the plain carbon steel, the *initial* n is zero (because of the pronounced yield point). Use of double n values can cause confusion, and its practice is discouraged. In most forming operations where Eq. (6–32) finds greatest application, the strains are generally well beyond the transition region (that is, in Zone 3) and the behavior in Zone 2 is irrelevant. The use of n by the vast majority of persons who apply it in forming operations is understood to be related to the behavior in Zone 3.

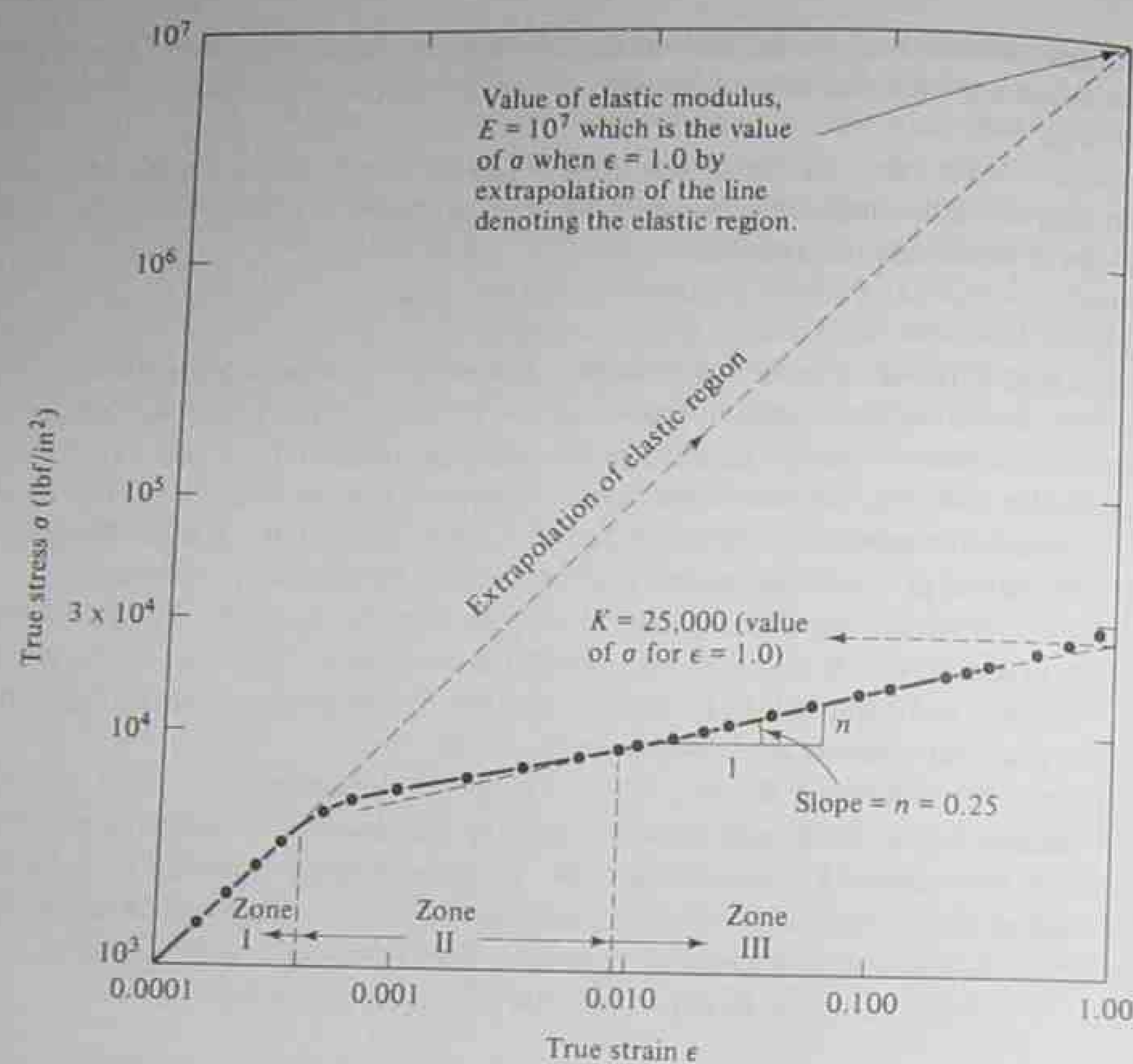


Figure 6–10 True stress-strain behavior of 1100-0 aluminum tested in tension and plotted on logarithmic coordinates.

3. Once the transition region is exceeded, the behavior can be viewed as being fully plastic, and it is those data points that should be used to define K and n; note that one should *not* give equal weight to any test points at strain levels less than those in Zone 3. We have seen instances where such an approach is overlooked and, in essence,points in Zone 2 are handled together with those in Zone 3, the result being that *incorrect* values of K and n are defined.

It is crucial, therefore, to note several points with regard to Eq. (6–32); these are

1. It is based upon true stress-true strain data beyond the initial transition region; usually this will be at strains of the order of 0.04 to 0.08.
2. To compute an original yield strength via $K(0.002)^n$ can be quite inaccurate; instead, a definition such as one based upon offset is more correct.
3. After necking, that region is subjected to a state of triaxial rather than uniaxial stress, and calculations of $\sigma = F/A$ and $\epsilon = \ell n(A_0/A)$, as based upon the minimum

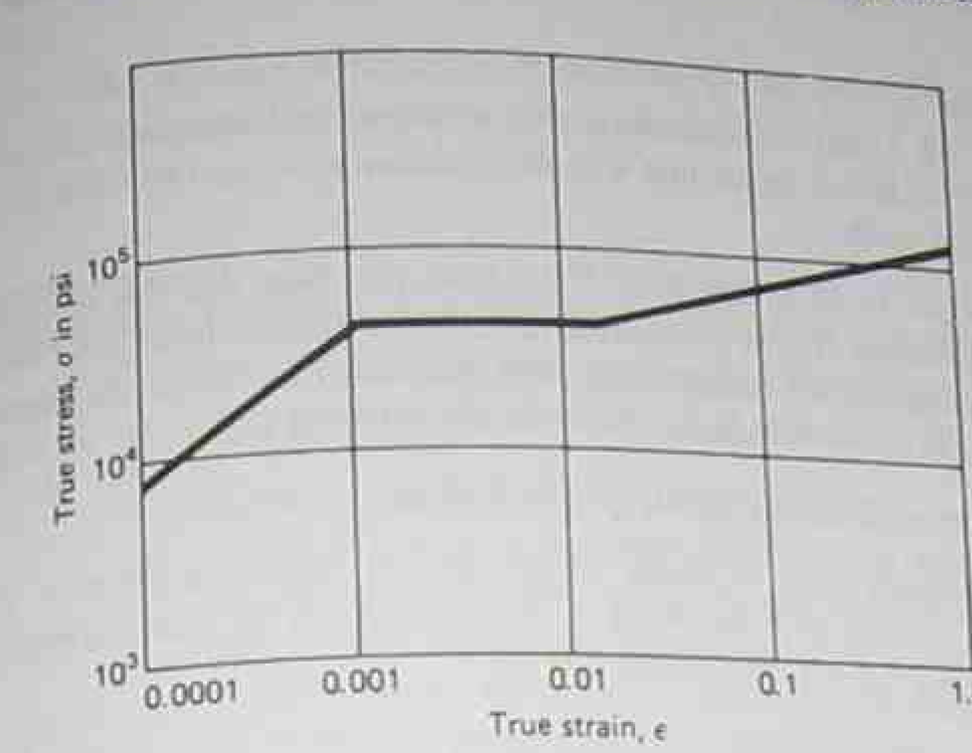


Figure 6–11 True stress-strain behavior of an annealed, low-carbon steel tested in uniaxial tension and plotted on logarithmic coordinates.

neck area, will produce points that begin to fall above the line described by Eq. (6–32). If careful corrections [6] are made, these corrected values will show quite a reasonable fit with Eq. (6–32).

4. With isotropic materials, the $\sigma - \epsilon$ data obtained from carefully conducted direct compression tests (other tests are also used) often display behavior equivalent to tensile results and can be conducted out to strains well in excess of that which coincides with necking in tensile testing. As a consequence, we do not *limit* the range of applicability of Eq. (6–32) to strains equivalent to the necking strain in tension. Indeed, the major stresses induced in many forming operations, where the use of a form such as Eq. (6–32) finds major application, are compressive and necking does not occur.

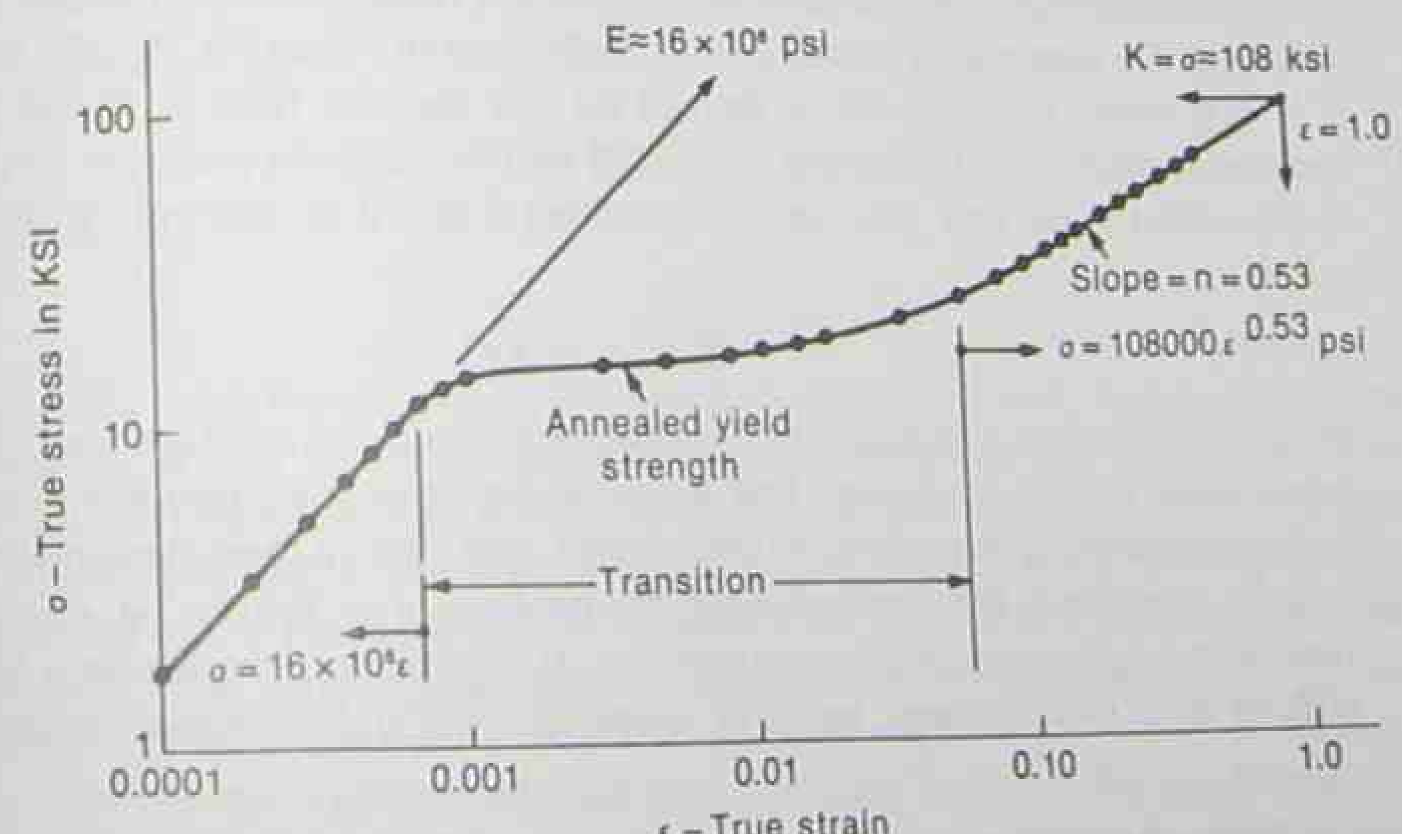


Figure 6–12 Same as Fig. 6–11 for an annealed alpha brass.

Example 6–10

Refer to Figs. 6–10 and 6–12. Using the strain hardening equations indicated, calculate σ for a strain of 0.002 in each case, and compare this with the probable yield strengths that would be found by using the offset method.

Solution From Fig. 6–10, $\sigma = 25{,}000(0.002)^{0.25} = 5290$ psi. Note that this is slightly below the stresses in the transition region for a strain of 0.002, where an offset yield would be about 6000 psi. (This *cannot* be determined accurately from this plot.) Thus, this method for finding Y (annealed) might seem reasonable. Now, however, using Fig. 6–12, we obtain

$$\sigma = 108{,}000(0.002)^{0.53} = 4010 \text{ psi}$$

whereas the actual yield strength must be about 14,000 psi, as shown on the plot. This illustrates the error that can result if Y annealed is computed from the strain hardening equation. As indicated on both figures, K has units of psi.

Here, although *not* a property, an expression for the degree of cold work is introduced; symbolically

$$r \equiv (A_o - A)/A_o \tag{6–33}$$

where r is the cold work induced during a reduction of area from A_0 to some smaller A. If multiplied by 100 it is often called percent cold work. Note that this is *not* the same as the *property* A_r defined by Eq. (6–30). In essence, A_r is indicative of the maximum r that can be induced at fracture under *uniaxial tension*. We note that it is not always possible to define strain as a measure of r (such as in hardness testing), but whenever work hardening can be related to an area change, then the *homogeneous** strain induced will be calculated by using Eqs. (6–11) and (6–33) as,

$$\epsilon = \ell n(1/[1 - r]) \tag{6–34}$$

Thus, if a specimen is subjected to, say, 10 percent reduction of area, called *cold working* below certain temperature levels, then the *uniform* strain, from Eq. (6–34), is 0.105, since $r = 0.1$. Now, if Eq. (6–32) is defined for this material (that is, K and n have been previously determined), inserting $\epsilon = 0.105$ into the equation produces a value of σ that is equivalent to the new *yield strength* resulting from 10 percent cold working. In other words,

$$Y = \sigma = K\epsilon^n \tag{6–35}$$

provides a means for calculating the yield strength of a metal as a function of plastic strain induced due to work hardening. One major limitation to this concept is that most deformation processes do not induce uniform strains, and even in those cases where a value of r can be determined from area changes, the predictions of Y from Eq. (6–35) should be viewed as a *reasonable first approximation;* if anything, they provide a lower bound. With all of its qualifications, however, using Eq. (6–35) does at least *attempt* to include

* In most operations, the induced strain is not homogenous or uniform across the section; this is discussed in greater detail in Chapter 8.

the actual effects of work hardening into particular analyses rather than assuming that Y is constant as has often been done.

Example 6–11

The work (strain) hardening behavior of a metal is described by $\sigma = 690\ \epsilon^{0.2}$ (MPa). If a piece of the metal were plastically deformed in a uniform manner by 40 percent reduction of area, calculate the expected yield strength of the cold-worked metal.

Solution Since $r = 0.4$, using Eq. (6–34) gives $\epsilon = \ell n(1/[1 - 0.4]) = 0.51$; thus, from Eq. (6–35), $Y = \sigma = 690(0.51)^{0.2} = 603$ MPa. As discussed in relation to Eq. (6–40), the induced strain here (0.51) is much greater than n (that is, 0.2). Although a maximum uniform strain of only 0.2 (that is, n) can be induced due to *pure tension*, much larger strains can be induced by many deformation processes. Thus, the strain hardening equation is *not restricted* to use where the induced strains are limited by n or ϵ_u in a tensile test.

Two useful relationships are now developed. Recalling that $F = \sigma A$ and that at ultimate load F_u the slope of the $F - \Delta\ell$ curve goes to zero, it follows that

$$dF = \sigma\, dA + A\, d\sigma = 0 \tag{6-36}$$

or

$$d\sigma/\sigma = -dA/A \tag{6-37}$$

Considering that $-dA/A$, when integrated between limits of A_0 (lower) and A (upper), becomes $-\ell nA/A_0) = \ell n(A_0/A)$, it follows from Eq. (6–11) that $-dA/A = d\epsilon$. Thus,

$$d\sigma/\sigma = d\epsilon \quad \text{or} \quad d\sigma/d\epsilon = \sigma \tag{6–38}$$

with $\sigma = K\epsilon^n$,

$$d\sigma/d\epsilon = nK\epsilon^{n-1} = \sigma = K\epsilon^n \tag{6-39}$$

therefore,

$$\epsilon_u = n \tag{6–40}$$

because the analysis referred to the condition at ultimate load, hence ϵ_u. In words, the numerical value of the strain-hardening exponent equals the true strain at ultimate load, especially for most annealed metals. In practice, because of testing machine insensitivity it is truly impossible to decide the *exact* value of F_u (as seen in Fig. 6–6(a), the load curve is quite flat over a broad change in extension at the upper load levels) and hence the exact value of ϵ_u from which n could then be deduced. The most accurate method yet found to define n is from a plot such as that shown on Fig. 6–10.

Now the true stress at ultimate load is σ_u (*don't* call this the true tensile strength). Historically, tensile strength is defined as the largest *nominal* stress S_u, whereas the largest true stress coincides with fracture. Referring to σ_u in that way could cause confusion and is to be avoided for this reason. Now from Eq. (6–32) we have

$$\sigma_u = K\epsilon_u^{\,n} = K(n)^n \tag{6-41}$$

using Eq. (6–40). Considering the condition at ultimate load, we have

$$F_u = S_u A_0 = \sigma_u A_u = K(n)^n A_u \tag{6-42}$$

using Eq. (6–41). So

$$S_u = K(n)^n A_u / A_o \tag{6-43}$$

and from Eq. (6–11)

$$\epsilon_u = n = \ell n(A_o/A_u) \quad \text{or} \quad A_u/A_o = e^{-n} \tag{6-44}$$

so that

$$S_u = K(n/e)^n \tag{6-45}$$

where the symbol e in this derivation is the base of *natural logarithms* and *not nominal* strain. The value of S_u in this equation is the value for the metal containing *no prior cold work* before tested in tension.

There is one very practical use of Eq. (6–45), and this is in regard to the comments made earlier about properly determining K and n from experimental data. Suppose a number of $\sigma - \epsilon$ values have been determined along with a measured value of S_u (this is quite accurately determined, since any errors arising from an initial measure of A_0 and the value chosen as F_u are truly minimal), and the data are plotted as in Fig. 6–10. If improper weight is given to any points in Zone 2, the resulting line would describe incorrect values of K and n. But if those values were used in Eq. (6–45) and the *calculated* value of S_u was different from the correctly measured value (say 2 percent or so), this should indicate to the experimenter that an adjustment of the initial line is needed and that Zone 2 points should be ignored.

Example 6–12

From a tensile test, the measured value of the tensile strength S_u is 28 ksi (193 MPa). When the $\sigma - \epsilon$ data are plotted on logarithmic coordinates, an experimenter fits a straight line to the large-strain data and concludes that $\sigma = 50\epsilon^{0.25}$ (where $K = 50$ has units of ksi). Comment on the appropriateness of the K and n values.

Solution Using Eq. (6–45), we find that the *computed* value of S_u is

$$S_u = 50(0.25/e)^{0.25} = 27.54 \text{ ksi}$$

Since this is less than two percent from the measured value of 28 ksi, the values of $K = 50$ ksi and $n = 0.25$ are quite reasonable.

It is noted that cold working of metals increases the tensile strength as well as the yield strength. Details on this point can be found in Ref. [2].

6.10 MECHANICAL BEHAVIOR OF NONMETALS

6.10.1 Introduction

Although many of the basic definitions and concepts covered earlier in this chapter are applicable where ceramics or polymers are involved, there are certain differences that may be observed, and it is these that are discussed briefly here.

6.10.2 Ceramics

One of the characteristics of most ceramics is their inherent brittleness, which leads to higher compressive strength prior to fracture as compared to tensile strength. By and large, the stress-strain behavior to fracture is primarily elastic, so in applications where ductility is needed, these solids are not used. Yet the so-called traditional ceramics, and many newer ones that are man-made, do possess special properties that satisfy certain engineering applications. The making or building of chinaware, glasses, and structures such as bridges and cathedrals (which primarily support compressive loads) have utilized traditional ceramics for centuries. Applications of newer ceramics include refractory solids used in industrial furnaces, catalytic converters on cars, gas turbines, jet engines, and the like. In such high-temperature applications, ceramics retain their strength and thermal stability (that is, they do not *creep** very much) at temperature levels for which most metals and polymers would *not* be satisfactory.† It is noted that ceramics are also used as cutting tools for particular machining operations. In essence, traditional ceramics find use where the loading is primarily compressive,‡ whereas the more recent man-made types are used where the retention of strength and limited creep at elevated temperatures are needed.

6.10.3 Polymers

Some of the long-standing and traditional definitions of certain mechanical properties associated with metallic behavior have not been followed exclusively by those measuring and reporting the properties of polymers. This causes confusion and may lead to problems where designers are concerned.

One obvious difference between the behavior of ductile metals and ductile polymers is seen in a typical tensile test. Consider Fig. 6–13, which shows an annealed low-carbon steel specimen taken to fracture under tensile loading. As indicated by Eq. (6–40), the maximum *uniform* strain occurs at the ultimate load after which necking occurs. This is followed by a localized region in which most further plastic deformation occurs in the

* See Chapter 7 for a discussion on creep.

† Only so-called *superalloys*, which utilize, for example, cobalt, nickel, and chromium, approach the behavior of ceramics at temperatures of the order of 1000°C; however, ceramics are used at even higher temperatures.

‡ See the excellent texts by J. E. Gordon entitled *The New Science of Strong Materials—or Why You Don't Fall Through the Floor*, 1968, and *Structures—or Why Things Don't Fall Down*, 1978 (London: Penguin Books.)

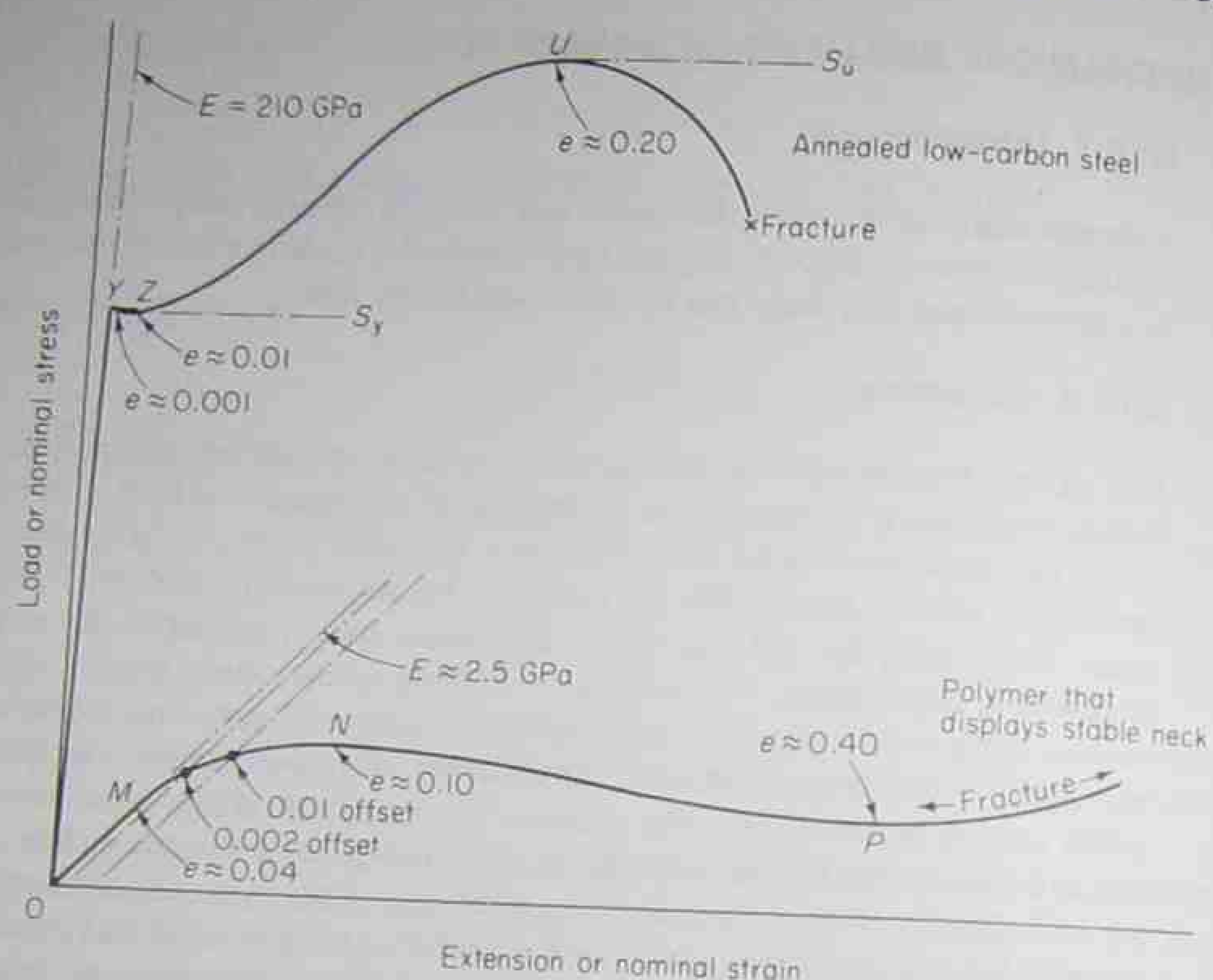


Figure 6–13 Nominal stress-strain curve for a polymer that displays stable neck propagation and an annealed low-carbon steel that displays a pronounced yield point.

vicinity of the neck. Further strain, as evidenced by a continued reduction of the diameter of the neck, follows until fracture occurs.

Compare that behavior with the specimen of high-density polyethylene (HDPE) shown in Fig. 6–13. The behavior up to the onset of necking is similar; that is, the area of the gage section decreases as the load increases. As a local neck begins to form, the area in that region continues to decrease under a decreasing load (just as with metals), but fracture does *not* occur. Instead, the neck reaches some value characteristic of the polymer, and a *stable* neck begins to propagate along the specimen towards the shoulders at the ends. This is called *cold drawing* in polymer terminology. During this stable neck growth, the *maximum strain* induced remains constant, since the minimum *area of the necked region does not essentially change during this* process. Note too that the period of stable neck growth proceeds under *constant* load. In essence the magnitudes of true stress and true strain remain fixed. This behavior is quite different from that observed with common ductile metals and probably accounts for some of the confusion found in the literature. For example, instead of considering the true strains involved, terms such as *draw ratio* (that is, ℓ/ℓ_0) or the use of nominal strain $(\ell - \ell_0)/\ell_0$ are most often used in the polymer literature. Both of these are practically meaningless, since they are a function of the initial gage length ℓ_0, which can take on numerous values. As shown by Ex. 6–9, that is why such values must be standardized. In addition, using nominal strains after

necking occurs simply gives a type of average value due to the nonuniform strains occurring over the full gage length. If the definitions of the true stress and strain were used, confusion would be avoided, since, as mentioned above, those values are basically constant during stable neck propagation. It's unfortunate that this has been so misunderstood. Another misconception has also been apparent in the literature on the behavior of polymers; that has to do with what is called the *fracture stress*. We note that the necked region, probably due to alignment of backbone chains, must possess a greater *load-*carrying capacity than the regions adjacent to the neck. As the stable neck grows along the specimen, it eventually runs into the shoulders of the specimen. The much larger area of the shoulders requires the load to increase as they come into play, and eventually fracture occurs somewhere in the stable neck. To call this fracture *load* is pure nonsense. A more fundamental approach would be to produce a new test specimen from the region containing the stable neck and *continue* a test using that specimen. If measurements of true stress and strain were made for the entire test, a sounder description of such behavior would result. Figures 6–14 and 6–15 show such results for several polymers. Note that the test points at the higher strains under uniaxial tension were obtained with a specimen made from the region of stable neck growth.

Figures 6–13 and 6–16 show *nominal* stress-strain curves; the first concerns a polymer that displays stable neck growth and an annealed plain carbon steel showing

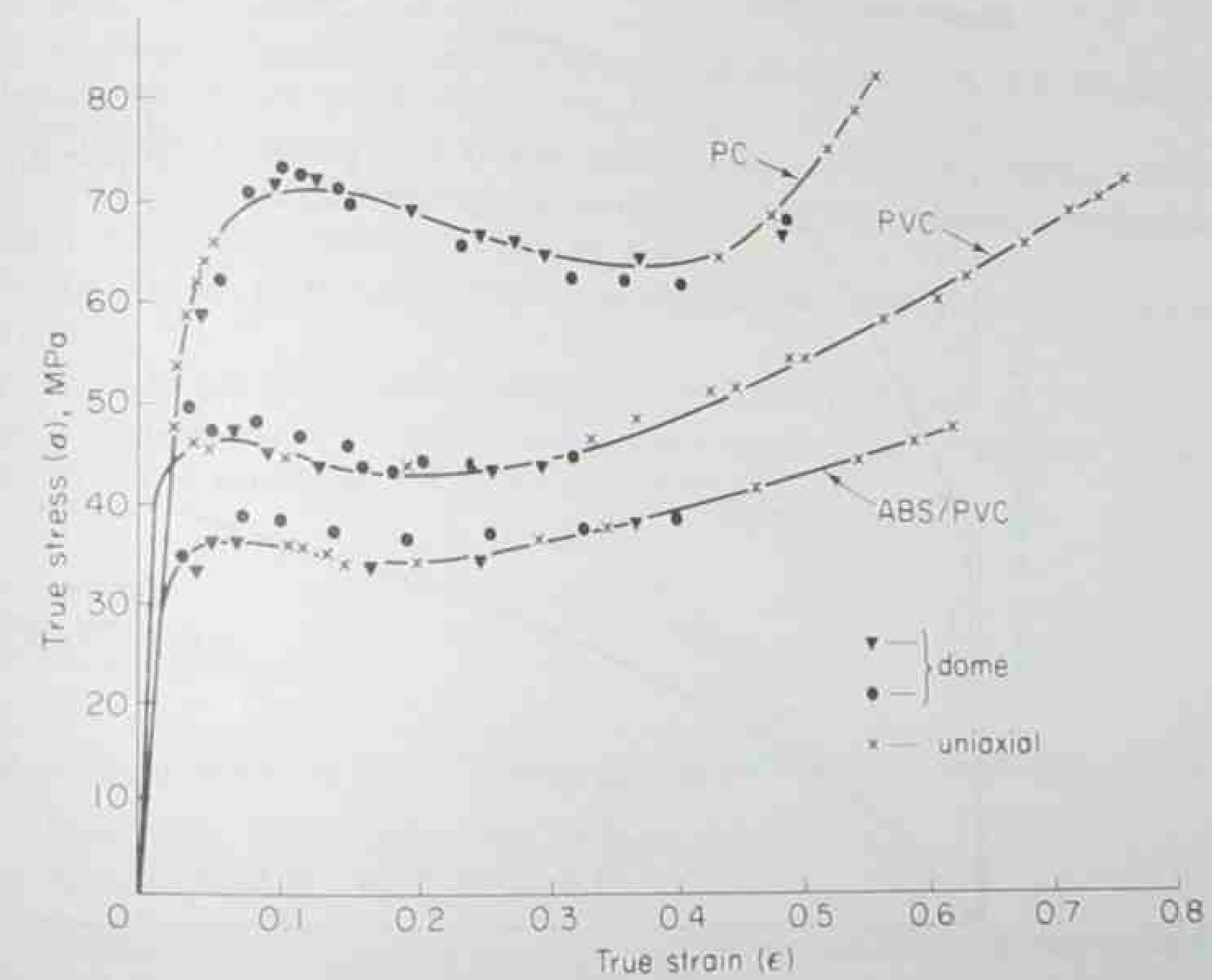


Figure 6–14 True stress-strain behavior of three polymers subjected to uniaxial and balanced biaxial tension.

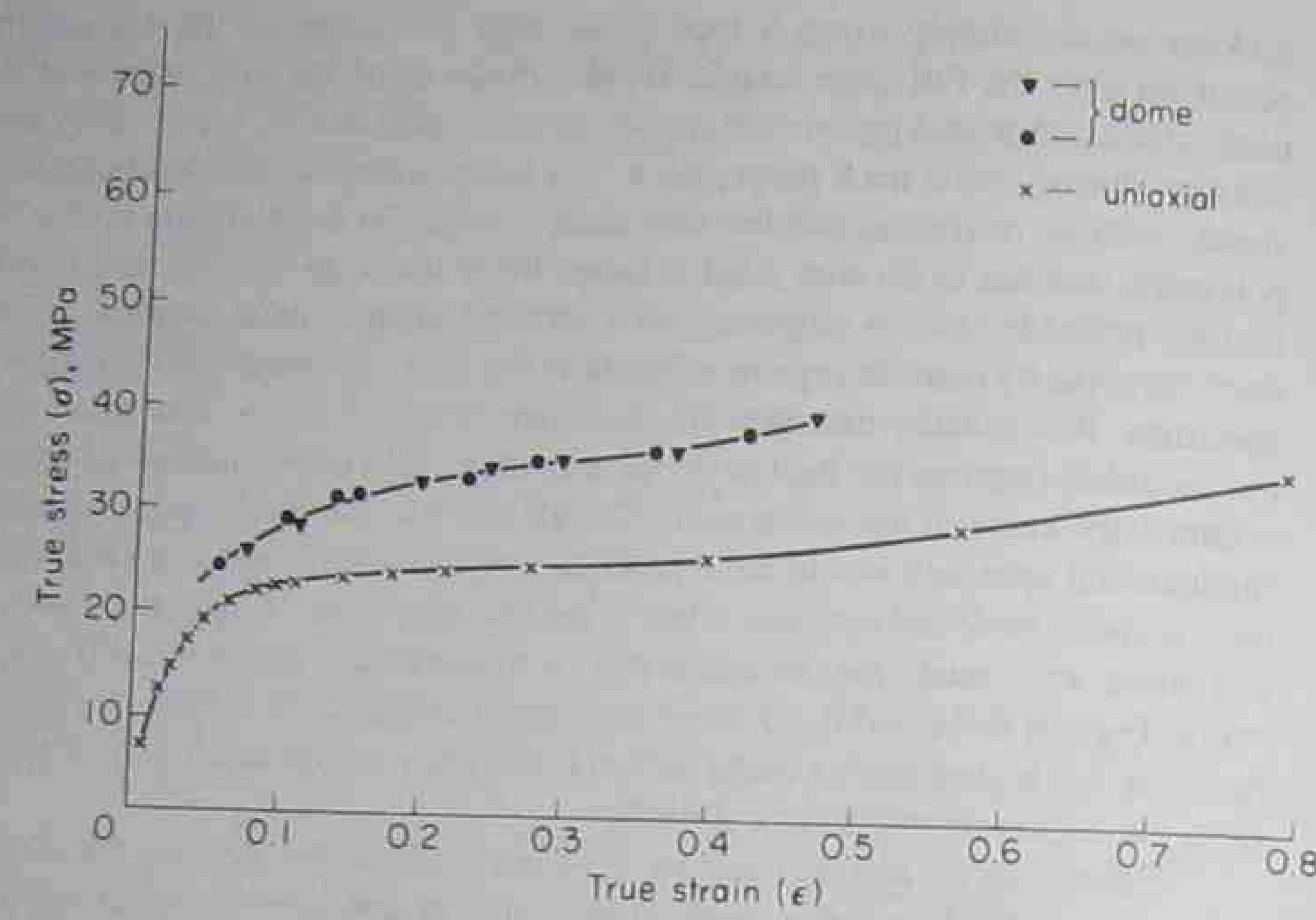


Figure 6–15 True stress-strain of high-density polyethylene subjected to uniaxial and balanced biaxial tension.

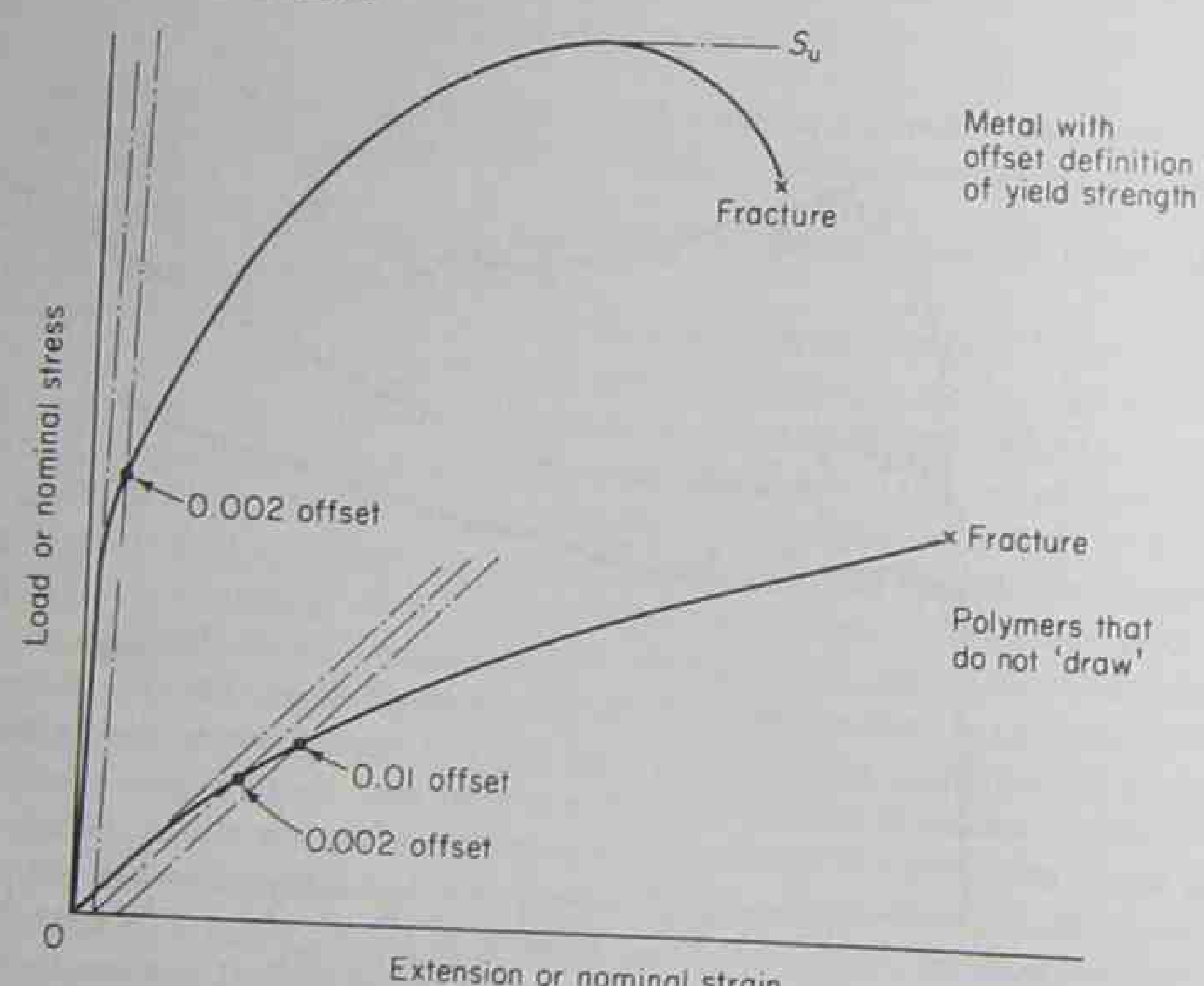


Figure 6–16 Nominal stress-strain curves for a polymer that does not display localized necking and a ductile metal that does not display a pronounced yield point.

a pronounced yield point, whereas the second involves a polymer that does *not* display a stable neck and a metal having no pronounced yield point. One reason for including these figures is to indicate why confusion often exists in defining the yield strength for polymers. Those that show stable neck growth display a load drop as necking begins, and the load associated with the start of necking is often taken as the load at yielding. The nominal stress at point N on Fig. 6–13 is called the *yield stress*. It is obvious that such a definition cannot be followed in Fig. 6–16, since no similar load drop occurs. We prefer to use an offset definition in any case; this at least provides consistency although the choice of percentage of offset is certainly arbitrary.

From these brief comments it is obvious that certain traditional methods used to define particular properties of ductile metals have not really carried over in a one-to-one correspondence for polymers. Thus, those who must select polymer properties in design or manufacturing studies should be aware of this. Reference [7] contains a more thorough discussion on this topic. Details of the structure of materials alluded to in this chapter are discussed in greater detail in Chapter 7 of this text.

REFERENCES

1. O. Mohr, Zivilingeneur, 1882, p. 113.
2. R. M. Caddell, *Deformation and Fracture of Solids* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1980), pp. 5–9, 50–52, Chap. 7, 105, 113.
3. J. E. Gordon, *Structures* (New York and London: Plenum Press, 1978), Chapters 2 and 3.
4. N. H. Polakowski and E. J. Ripling, *Strength and Structure of Engineering Materials* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1966), pp. 116–18.
5. W. Johnson and P. B. Mellor, *Engineering Plasticity* (New York: Van Nostrand Reinhold, 1973), pp. 15–17.
6. P. W. Bridgman, *Transactions*, American Society of Metals, **32** (1944), pp. 553–74.
7. D. K. Felbeck and A. G. Atkins, *Strength and Fracture of Engineering Solids* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1984), pp. 276–80, 295–20.

PROBLEMS

6–1. A bar of area $A_0 = 0.2$ in.^{2} is subjected to tensile loading with the following results:

Force (lbf):	1000	2000	3000	4000
Area (in.^{2}):	0.19	0.185	0.181	0.17

Find the true and nominal stresses at each load and comment on the difference in magnitudes as F increases.

6–2. Calculate the ratio of true to engineering strains (ϵ/e) for values of e of 0.001, 0.01, 0.05, 0.10, and 0.50. Comment on the trend that is shown.

6–3. With the data in Problem 6–1, find the engineering strain e at each load, then compute σ using Eq. (6–8). If Problem 6–1 has been completed first, compare the values of σ for both approaches.

6–4. With the data in Problem 6–1 calculate e and ϵ for each load independently. Then use Eq. (6–10) to find ϵ from e and compare the results.

6–5. A metal bar of dimensions $\ell_0 = 76$ mm, $w_0 = 12.7$ mm, and $t_0 = 7.6$ mm is loaded in tension to 22 kN. Using a ruler (not the most accurate device), you eyeball the length under the load to be 89 mm, the width to be 11.9 mm, and the thickness to be 7.1 mm. Find the true strains in the three directions and comment on the accuracy of your measurements.

6–6. Find S and σ in Problem 6–5 and discuss any calculation that might be questionable.

6–7. A rectangle of dimensions one inch by 0.75 in. is printed on the surface of a metal sheet. In-plane loading causes these dimensions to change to 1.25 in. and 0.68 in. respectively. If the unloaded *thickness* of the sheet was 0.125 in., what is the thickness after the loads are applied?

6–8. A metal has an elastic modulus of 68.9 GPa and Poisson's ratio of 0.3. If elastic deformation results due to three normal stresses,

$$\sigma_1 = +20.7,\ \sigma_2 = +10.3,\ \text{and}\ \sigma_3 = -6.89\ \text{(all in MPa)},$$

(a) Find the three normal strains.
(b) Determine if there is any volume change.

6–9. Repeat Problem 6–8 if $\nu = 0.5$ instead of 0.3.

6–10. Handbook data for a certain metal gives $\mathscr{E}\ell = 30$ percent and $A_r = 45$ percent. Did this metal neck when tested in tension? Explain.

6–11. The strain-hardening behavior of commercially pure aluminum is given by $\sigma = 22{,}000^{0.23}$. If a bar of this metal were cold-worked 20 percent, calculate the resulting yield strength. Note that K has units of psi.

6–12. During a tensile test of a brass specimen ($d_0 = 12.8$ mm) a *maximum* load of 53.4 kN is reached; at this point the area is 60 percent of the starting value. If a second identical specimen were loaded until the induced strain was $n/2$, what *load* is required to cause this condition?

6–13. A metal follows $\sigma = K\epsilon^n$ and displays a tensile strength of 300 MPa. To reach the maximum load required an elongation of 35 percent. From this information find K and n.

6–14. A tensile specimen having $\ell_0 = 2.000$ in. and $d_0 = 0.357$ in. is subjected to loading, and the following are found:
yield load = 2500 lbf.
diameter at ultimate load = 0.305 in.
diameter at fracture = 0.280 in.
elastic modulus = 15×10^6 psi
Upon completing this test you are informed that n for this material is 0.45 and in the annealed condition it obeys $\sigma = K\epsilon^n$. In addition, the specimen you started with contained some cold work *before* you pulled it in tension.
(a) How much strain had been induced *before* the specimen was tested?
(b) What maximum *load* was reached during the tensile test?

6–15. When an annealed metal is pulled to fracture, the maximum tensile strain is found to be 0.8. Suppose a specimen of this annealed metal has a $d_0 = 10.2$ mm and is pulled until the diameter is 8.9 mm (assume uniform reduction at this point). *Starting* with the 8.9-mm specimen that has been cold worked, what reduction of area at fracture would you expect?

6–16. At the end of this chapter it was stated that the *tensile strength*, S_u, increases with cold working. Develop a relationship between the tensile strength due to cold working (call this S_w), the annealed tensile strength S_u and cold working r. *Hint:* Consider a general load-extension curve and think about what happens if you induce a load greater than initial yielding but less than the maximum load and then unload to zero. Now, starting in this condition, what happens if you carry the test to fracture?

7 engineering materials

7.1 INTRODUCTION

The word *materials* covers a spectrum of different atomic and molecular structures which find many applications. In some cases, ready availability is the reason why certain materials are used. For example, eskimos build igloos out of ice, whereas Indians of the southwest used clay to build shelters. Fabrics are used in large quantities in the furniture and automotive industries; wood finds extensive use in the housing industry, while paint is used in cases too numerous to mention. Depending upon one's viewpoint, materials can be categorized in various ways. Solid, liquid, and gaseous are one such category whereas organic and nonorganic are another. Composites are often considered to be natural or man-made, whereas metals are often considered to be ferrous or nonferrous.

In this text, *engineering materials* are intended to include metals, polymers, and ceramics, since, by and large, it is these three categories that are most often subjected to the major *manufacturing processes* discussed in chapters that follow. With much of today's publicity being devoted to topics such as robotics, computers, and processes that find important but rather limited application, the traditional and most widely used industrial processes seem to have been relegated to a position of limited importance. Product design, inventory control, scheduling, and the like are all important aspects in the broad field of manufacturing, yet materials must be *selected* and *processed* or parts would never be produced. It is our contention that a reasonable knowledge of the interaction among the material being processed, the design, and the processing method itself should be of definite concern to manufacturing engineers. They should not be expected to be experts in materials science, but an understanding of certain broad, basic principles, and a rea-

sonable awareness and vocabulary connected with materials will permit a better interaction with specialists, which is often necessary.

Many other texts that address manufacturing processes begin the coverage of engineering materials with sketches of unit cells, dislocation models, polymeric chain structure, Miller indices, and so forth, but little if *any* use is then made of aspects of materials science. Except for certain terms or definitions that are needed for explanatory purposes, we see no need to present a detailed review of such topics since the reader (as mentioned in the preface) is expected to have covered that material in an introductory course in materials science.

One last point pertains to the depth of coverage to be presented. Consider the topic of equilibrium diagrams of metals. Since the success of various *strengthening mechanisms* is limited by phase changes, we shall discuss a few in detail in order to demonstrate their importance. The reader can then extend these ideas to the myriad of other equilibrium diagrams without the need of specific discussion.

7.2 METALS

The basic differences among the three groups of engineering materials, which leads to their different properties, arise from the manner in which atoms are arranged and held together by bonding forces. For our purposes, metals are considered to form a *crystalline* structure.* A small number of atoms form a single *unit cell*, where the combination of the type and size of cell varies from one element to the next. Each cell attaches to adjacent cells, and this repetition in three dimensions is called the *space lattice*. The most common cell structures are the body-centered cubic (BCC), face-centered cubic (FCC), and hexagonal close-packed (HCP), and the balance of forces between positively charged nuclei and negative electrons is referred to as the *metallic bond*. One important characteristic of metallic structures is their relatively high *packing factor;* that is, the individual atoms are packed tightly together because of high bonding forces. This leads, for example, to the high density and, generally, high strength of metals among other properties. To cause metals to deform under applied loads or forces, one of two major mechanisms is required.† They are called *slip*, where a relative sliding displacement between atoms on adjacent planes is caused by *shear* stresses, and *cleavage*, where adjacent planes of atoms are literally split due to *tensile* stresses. In either case, the applied forces must overcome the bonding forces for such results to occur.

From a crystallographic point of view, a *perfect* structure is one in which every atomic site is occupied by an atom, and the space lattice would range undisturbed from one free surface to another. Any type of interruption of such a perfect structure is deemed a *defect* by the crystallographer and would appear to be undesirable. To the engineer, however, such defects can have important positive benefits, as we will discuss shortly.

* Metals can be produced in an amorphous form, but the total industrial application of such metals is highly limited at the present time.

† Twinning can also occur but is not discussed here; see, e.g., R. A. Flinn and P. K. Trojan, *Engineering Materials and Their Applications* (New York: Houghton Mifflin Co., 1981), p. 64.

The major defects are called *point, line,* or *surface* (area). A point defect is called a *vacancy,* which is simply a missing atom; a line defect is called a *dislocation,* which may be a row of missing atoms, while an area defect is a *grain boundary,* where a mismatch of atoms occurs between adjacent grains. In the main, the latter two types of defects are of the greatest importance to us as they provide certain benefits. First, except for very small metallic *whiskers,* all metals contain dislocations in exceedingly large numbers. The presence of dislocations explains why real metals do not possess the theoretical shear strengths that should prevail if the structure were perfect.* Also, any aspect of the structure that acts to inhibit the *easy glide* or movement of dislocations is considered to be a *strengthening mechanism.* The only way for slip to occur in a perfect metal would be for entire planes of adjacent atoms to displace simultaneously; of course, if that were to happen, immediate fracture would most likely occur. Although extremely high stresses would be needed to cause fracture, no significant degree of *plastic* deformation would result, and those processes that require such deformation to produce a part would be basically nonexistent. With the presence of dislocations, the stress needed to cause slip is about 0.1 percent of the theoretical value based upon perfect structure analyses, but because of dislocations, plastic deformation can occur before fracture is reached. This ability of *ductile* metals permits many parts to be formed to shape by the application of appropriate forces. Thus, the crystallographic defects can be a benefit to the engineer concerned with plastic deformation.

Practically all commercially used metals are *polycrystalline;* that is, a large number of separate crystals or grains make up the overall structure. The interface between adjacent grains is called a *grain boundary.* Figure 7–1 is a photomicrograph of a steel containing 0.02 percent carbon, showing distinct grain boundaries. Why and how grains form can be explained with the schematics shown in Fig. 7–2.

Initially, solidification will begin in regions where the rate of heat transfer is greatest. If it is assumed that the molten metal has been poured into some kind of container, this rate is usually highest at the interface between the melt and the walls. Solid nuclei begin to form, much like the freezing of water in an ice-cube tray, as in Fig. 7–2(a). Solidification towards the center continues and is shown as rows of atoms in Fig. 7–2(b), where the atomic spacing is uniform. Eventually, rows growing from different directions reach a point where a mismatch or interference occurs, as shown in Fig. 7–2(c). The atom shown solid cannot meet the spacing demands of the two competing rows, so it is forced away from the equilibrium position of either and is now considered to be at a position of *higher energy level.* The continuation of this process finds a number of adjacent atoms undergoing this same process, and it is these atoms that form the grain boundary. For this reason, the final grain boundary displays a *higher* energy level as compared with the grain interiors; this is shown in Fig. 7–2(d). As will be discussed later, the average *grain size* can be altered. Again, this is considered to be a defect, even though the strength of the structure can be altered with grain size. The engineering viewpoint is to take advantage of grain size to modify the strength or ductility accordingly; this is discussed in some

* See R. M. Caddell, *Deformation and Fracture of Solids* (Englewood Cliffs, N.J.: Prentice-Hall Inc., 1980), Chapter 7.

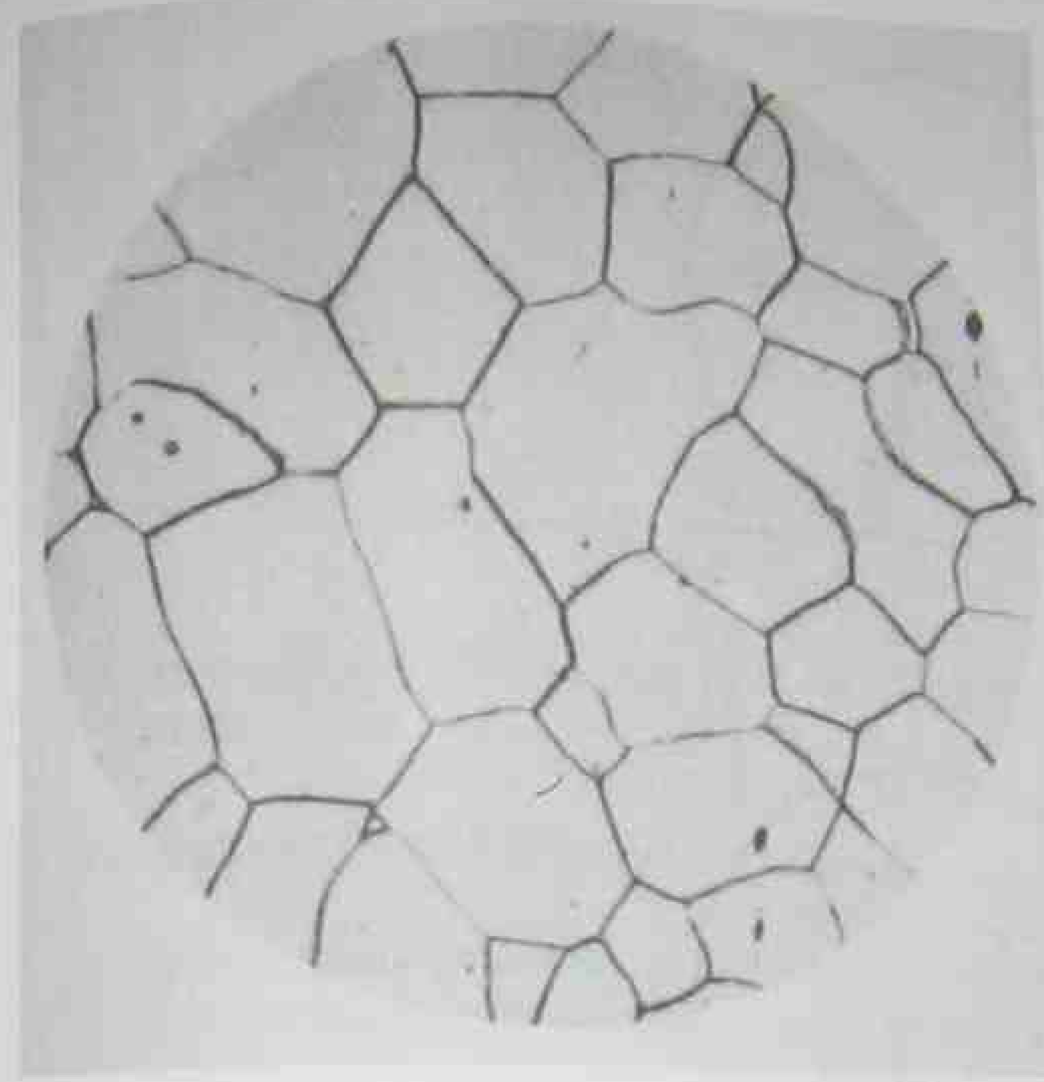


Figure 7–1 Photomicrograph of a low-carbon steel showing distinct grain boundaries.

detail later. With this brief introduction we have indicated many terms and a few basic concepts that will be used in much of this chapter.

7.3 SOME GENERAL OBSERVATIONS

The word *equilibrium* requires understanding, and we introduce a simple but useful explanation that will serve to explain various phenomena. Essentially, the concept of energy levels is involved, and we can refer to a book lying on a desk as an illustration. Since the book is several feet above the floor, it possesses potential energy; yet unless a force is applied to move the book to the edge of the desk and cause it to fall to the floor, it will remain in place. One might consider the book to be in a state of equilibrium, yet it is not at its lowest energy level with respect to the floor. To bring this about requires an input of energy, such as pushing the book to the table's edge, thereby causing it to fall; the energy supplied can be called the *activation energy* and is needed to cause a lowering of the potential energy that the book possessed prior to falling. Next, consider Fig. 7–3, which shows a marble in a trough. The datum plane shown is assumed to be at a position of lower energy level; let us consider this to be loosely the level of equilibrium. To reach this level, the marble must first be *lifted* to the top of the trough, then under the force of gravity it will fall to the reference plane. In order to cause this lowering of the energy of the marble, an input of energy must be first induced; again this is viewed as activation energy. One other observation is important here. If the marble rested in a trough that was higher from the datum, it would be further from equilibrium than in the first case. Then to drive the marble to the datum requires *less* activation energy than before. Thus, the

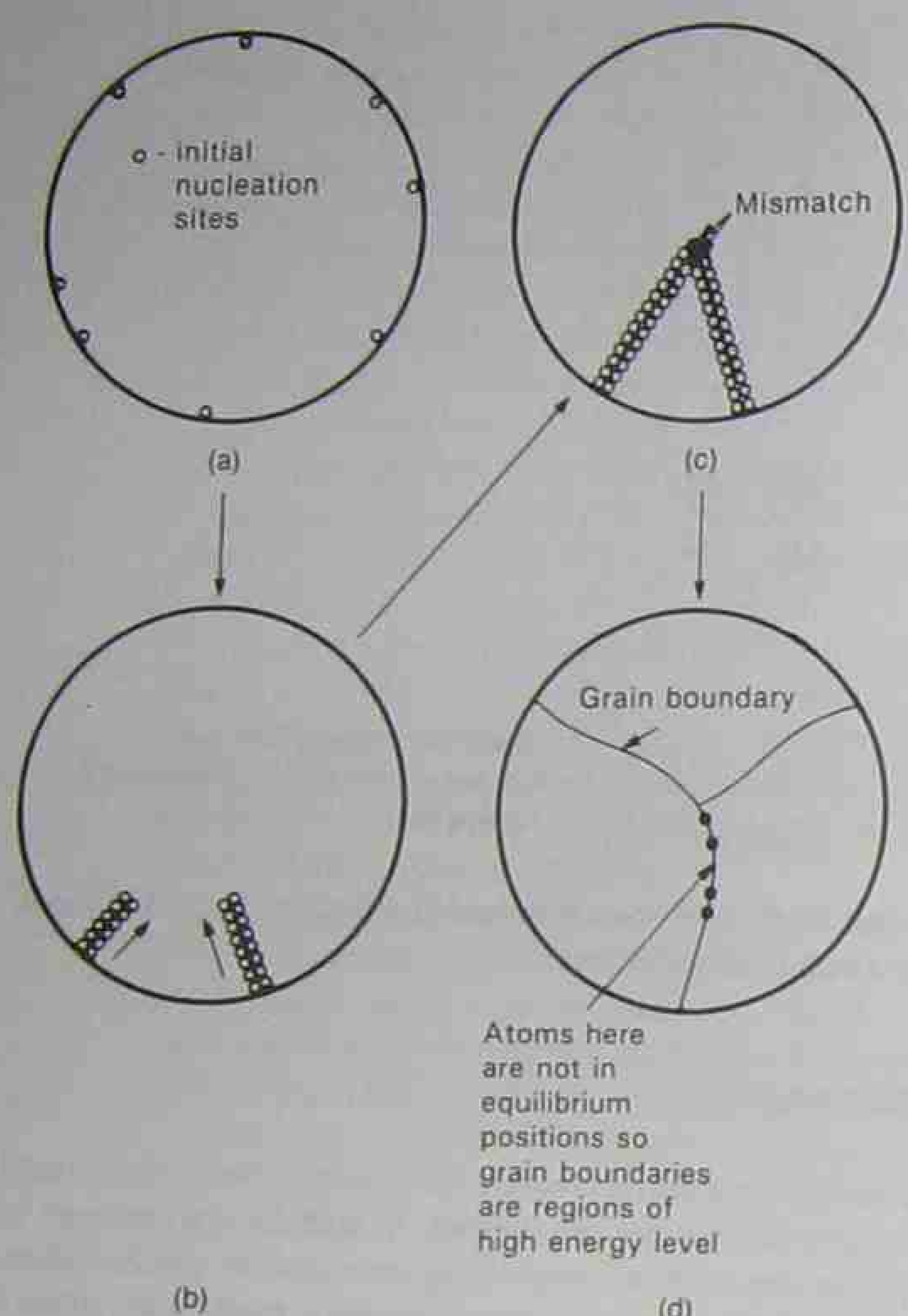


Figure 7–2 Schematic showing how grain boundaries form.

further from equilibrium is the condition of a structure (we might say the more unstable is the condition), the lower is the activation energy needed to drive the structure towards a lower energy level. Although a rigorous analysis of this topic involves certain concepts of thermodynamics, this simple physical explanation will prove adequate for our needs.

Bulk metals are initially cast to shape; that is, molten metal is poured into a container of some sort and permitted to solidify. Subsequent operations are then used to change the initial shape into the many types that find industrial use, for example, plates, sheets, tubes, and the like.* It is important to realize that no piece of bulk metal exists in a fully unflawed condition. Small voids or holes are *always* present; sound processing techniques can usually keep such flaws to a tiny and acceptable level, *but* they cannot be

* See Chapters 8 and 11 for greater detail.

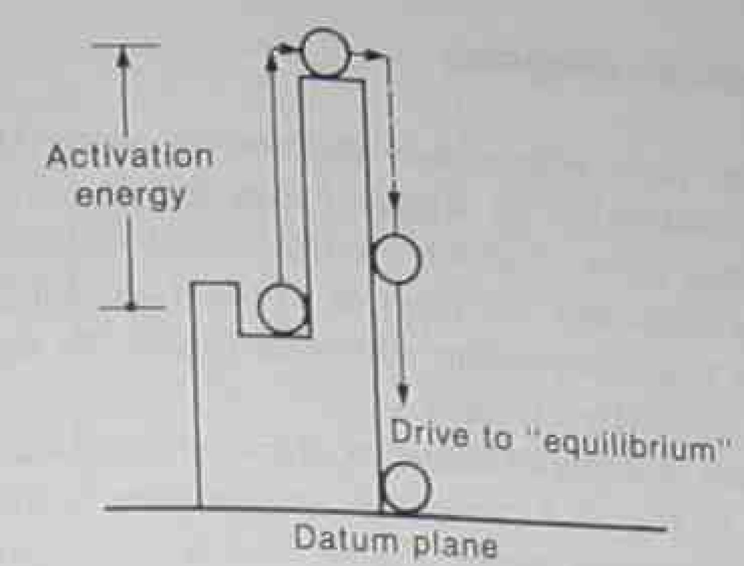


Figure 7–3 Conceptual model to illustrate activation energy and energy levels.

wholly eliminated. Note that these are in addition to the atomic level defects discussed earlier.

7.4 EQUILIBRIUM DIAGRAMS

To understand how and why the properties of metallic alloys can be altered by various *heat treatments* it is essential to have at least a minimal understanding of *equilibrium diagrams* as a starting point. What follows is not intended to cover the many details of physical metallurgy; rather, a few of the more commonly used types of alloys will be addressed to illustrate some of the *strengthening mechanisms* that can have a significant influence on the mechanical properties of these alloys. An *alloy* can be considered to be a mixture of elements that displays various levels of solubility; for example, nickel and copper will mix together in any percentage (by mass) of either element, and thus they are said to be completely soluble and can form an alloy of any composition. This is an exception rather than the general rule, as we shall see.

Certain words require definition, as they will be used extensively in discussing metals. A *phase*, which may be liquid or solid, is a homogeneous mass of matter. For particular alloys and levels of temperature, a *phase transformation* may occur. Such transformations may be from liquid to solid (such as the freezing of water) or from one solid phase to a different solid phase. Often the term *solid solution* is used; this is nothing more than a solid phase.

In general, the mechanical properties of alloys cannot be fully explained by considering the amounts of phases alone; rather it is the size, shape, and distribution of phases that are important. Here the words *microconstituent* and *microstructure* are introduced. A microconstituent may be defined as a repeatable portion of the overall structure as viewed with an *optical* microscope; it may be a single phase or may be composed of two phases that are insoluble in each other.* The microstructure is the combination of all microconstituents. Discussing a few equilibrium diagrams should make these definitions clear.

* This restricts considerations to features larger than the resolution of light microscopes, i.e. of the order of about 100 μm. Electron microscopes are capable of much greater magnifications, but the use of only traditional optical microscopes is implied here.

7.4.1 The Iron-Iron Carbide Diagram

Because they are one of the most widely used engineering materials, steels will be discussed first. Although considered to be alloys of iron and carbon, in practice they *always* include other elements, some of which are introduced intentionally to produce *alloy steels* whose particular properties are superior to *plain-carbon steels*. Other elements, called impurities, are controlled within allowable limits. To remove them completely is uneconomical and not really essential.

Ignoring the minor effects of impurity elements, the iron-iron carbide (or simply the iron-carbon) diagram shown in Fig. 7–4 can be used effectively to discuss the various equilibrium structures of *plain-carbon steels*. Figure 7–5 is an enlarged view of the left-hand end of the full diagram; since few steels contain more than 1 percent carbon, this figure will receive major attention. To simplify this discussion we have rounded off

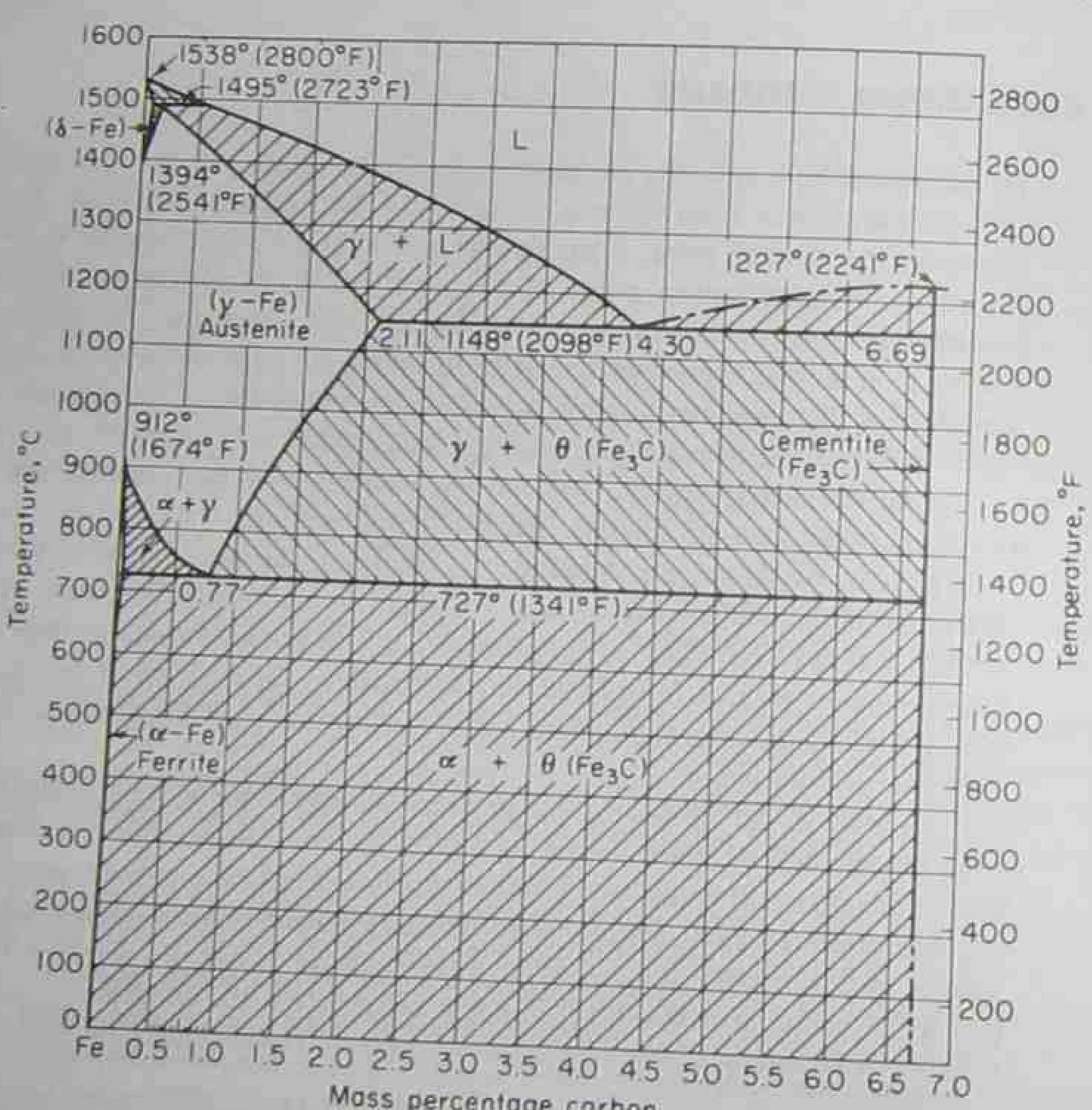


Figure 7–4 Iron-carbon equilibrium phase diagram in absence of free graphite.

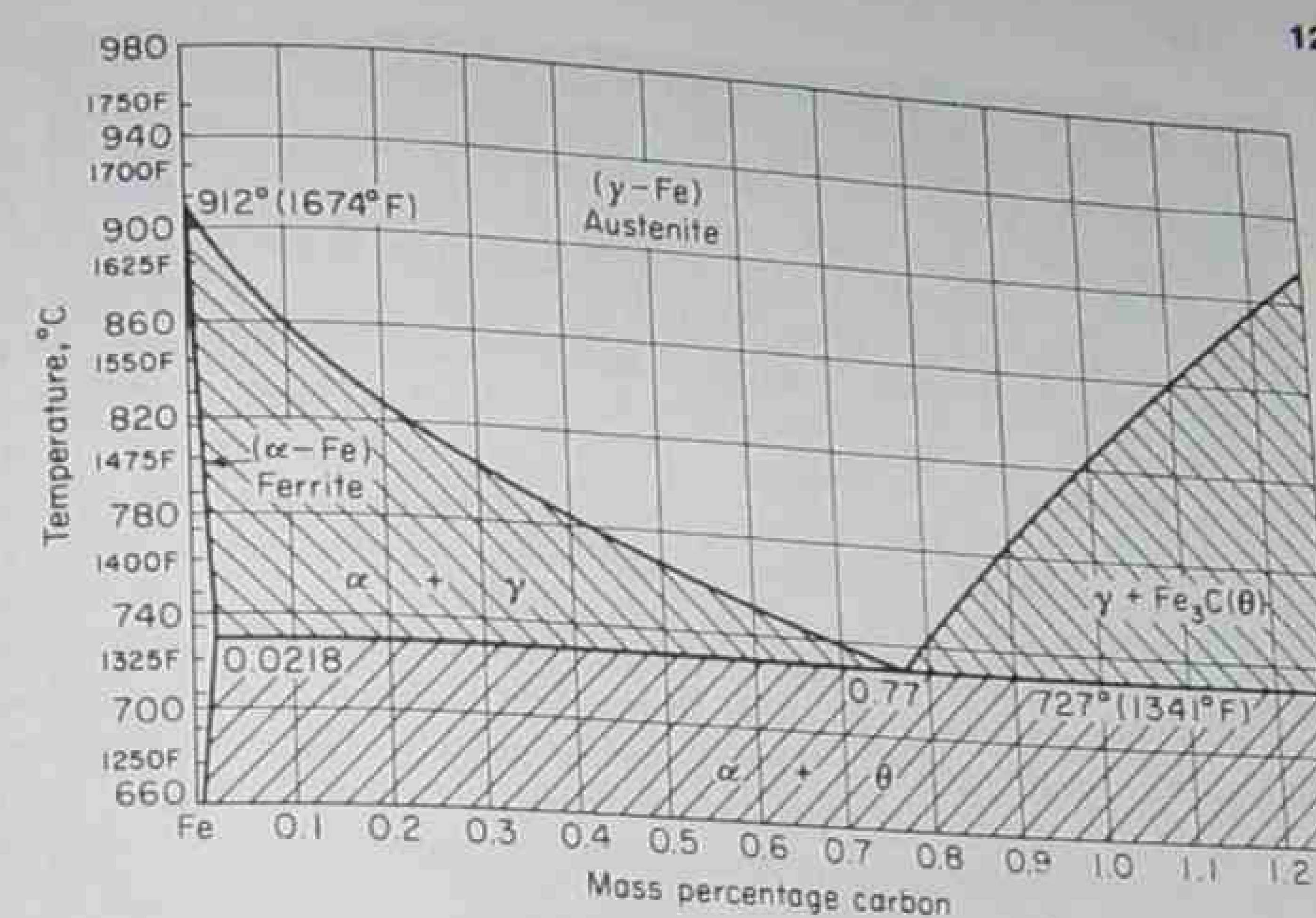


Figure 7–5 Iron-carbon equilibrium phase diagram to 1.2 percent carbon.

certain compositions and temperatures. For example, what is called the *eutectoid* is indicated as having a composition of 0.8 percent carbon, whereas the most recent measurements indicate 0.77 percent. The approximations we make have no real affect on the conclusions we shall draw. It is important to remember that Figs. 7–4 and 7–5 are equilibrium phase diagrams, and the various microstructures that result are due to a cooling rate that approximates equilibrium cooling. In practice this does not mean an infinitely slow cooling rate. Of the commonly available means of cooling, *furnace* cooling is the slowest and, therefore, closest to equilibrium. There, the part is heated to an appropriate temperature and held for an adequate time period to produce a homogeneous structure (that is, all grains are identical as to chemical composition); then the furnace is turned off. With well-insulated furnaces it may take up to half a day before the inside reaches room temperature. *Air* cooling is the next slowest method used; there, after homogenization, the part is removed from the furnace and simply cooled by the surrounding air. In many cases, but not in all, removing the heated part and plunging it into special *quenching oils* rather than air will also produce the phases shown on the diagrams. Thus, as long as such phases result, the cooling rate is considered *close* enough to equilibrium that the diagram can be used for making *reasonable* predictions of the final structure. Now there are three distinct *phases* indicated on these diagrams. Their symbols and other comments are given below.

1. Austenite, γ, is an elevated temperature phase. It is *extremely* soft and cannot exist below the *lower critical* temperature of 727°C (1341°F) if true equilibrium cooling

prevailed.* Note that the carbon content of γ can vary over a wide range, with the maximum being about 2 percent.

2. Ferrite, α, results when γ of *less* than 0.8 percent C transforms as the temperature is lowered. The γ is a face-centered cubic structure (FCC), whereas the α is a body-centered cubic (BCC). Although harder than γ, the α is also soft, having a Brinell Hardness Number (BHN) of 80 at room temperature.† Note the α is almost pure iron, containing a maximum amount of carbon of a little over 0.02 percent at 1341°F (727°C) and decreasing to practically zero at room temperature.
3. Cementite, Θ, results when γ of *more* than 0.8 percent C transforms as the temperature is lowered. It is iron carbide, Fe_3C, and has a fixed composition of about 6.7 percent carbon. The unit cell of Θ is more complex than either α or γ (actually orthorhombic) and, unlike the solid structure of those two, is an *intermetallic compound*. One characteristic of such structures is their extreme hardness as compared with solid solutions; the BHN of Θ is of the order of 1000. Because of this high hardness, such compounds are quite brittle; however, microstructures containing free Θ display good wear resistance and are most effective when the Θ exists as relatively small particles dispersed uniformly throughout the microstructure.

To this point, equilibrium phases have been discussed. Now consider what results if a steel containing 0.2 percent C is heated to 1650°F (899°C), homogenized, and then *furnace* cooled (FC) to room temperature. The resulting microstructure is shown in Fig. 7–6. The light grains are ferrite, and those that show a dark, lamellar pattern (alternating plates of α and Θ) are called *pearlite* (P). Although the phases are α and Θ, the *microconstituents* are α and P. Pearlite forms whenever γ of 0.8 percent carbon transforms at the lower critical temperature (727°C), so the average carbon content of P is also 0.8 percent. To estimate the relative *amounts* of α and P, the *inverse lever rule* can be used. The first step is to locate the fulcrum point of the lever. Here, it is the average carbon content of 0.2 percent, as seen on Fig. 7–5. Since our interest is with microconstituents and not phases per se, the length of the lever of concern is determined from the compositions of the microconstituents at the temperature of transformation (1341°F). To *simplify* this type of calculation (little error results) consider the α to have zero carbon while the P has 0.8 percent carbon; thus, the *length* of the line of concern is 0.8 unit, and this is to be divided into two portions to give the relative amounts of α and P. Physically, the fulcrum point is much closer to the α composition than that of P, so a greater amount of α results in the end structure. The distance from the fulcrum (0.2 percent C) to the two ends of the line (from 0 to 0.8 percent C) gives ratios of ¼ and ¾. Therefore, the final structure contains 75 percent α and 25 percent P. Note that to arrive at the 75 percent value for α, the distance from the fulcrum to the P composition must be used and vice versa to arrive at 25 percent P. This demonstrates the *inverse* nature of the lever rule.

* This γ is for plain-carbon steels. *Austenitic stainless steels* are substantially harder than this type of austenite.

† BHN is one of various hardness designations used.

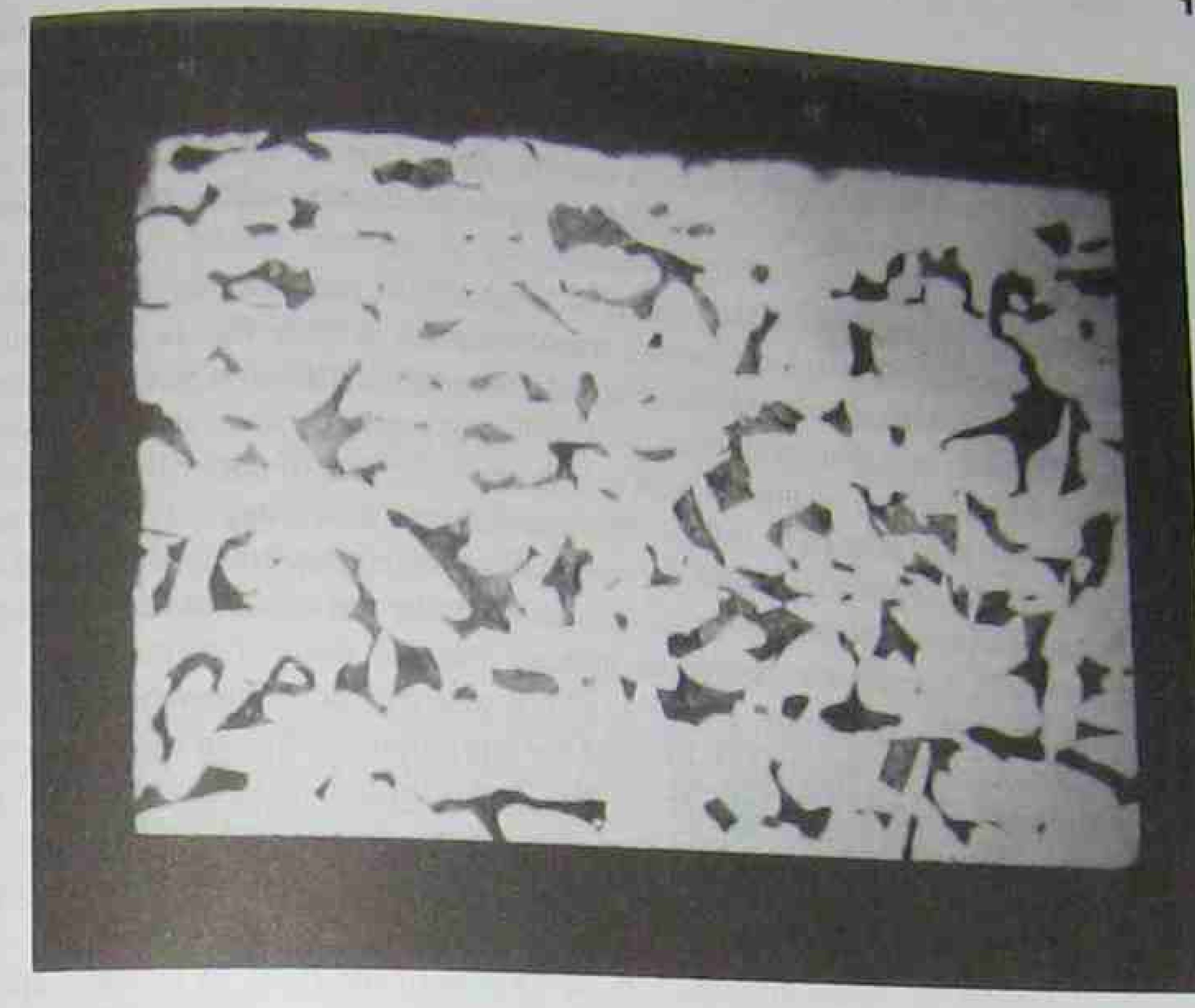


Figure 7–6 Photomicrograph of 1020 steel furnace cooled from the austenite region.

If oil quenching rather than furnace cooling had been used and if equilibrium phases resulted, a structure of about 75 percent α and 25 percent P would again result. However, the properties of the FC and OQ specimens would differ; this is readily indicated via hardness tests. How can this be when we have not only the same amounts of the two phases but also the same microconstituents? The answer is we do *not* have the same microconstituents in the grains of P. Careful observation of such grains made at the same magnification would indicate that the platelets of Θ in the furnace-cooled specimen are *thicker* and *less numerous* than those in the OQ specimen. In these cases the platelets of Θ in the FC specimen are about 10 times thicker than those in the OQ specimen. Because the *size* and *distribution* of the Θ phase differs, the pearlite in these specimens is considered to be a different microconstituent. With furnace cooling it will be called coarse pearlite, P_c, while the oil quench produces fine pearlite, P_f. Note that if air cooling had been used, the resulting structure is called medium pearlite, P_m, since the thickness and number of platelets of Θ would be between the others; this occurs because air cooling is faster than FC but slower than OQ.

To explain in a simplified manner why these differences occur, the terms *nucleation*, *growth*, and *diffusion* must be introduced and defined. Recall that for pearlite to form we must start with grains of γ that contain 0.8 percent C and that these are in the solid state. At the transformation temperature a group of iron and carbon atoms must find themselves positioned to form a unit cell of Θ. This will occur most likely at γ grain-boundaries, since they are at higher energy levels than the grain interiors and the structure desires a lower energy level. Thus the nucleation of Θ begins in those regions. Since Θ has a fixed composition, 6.7 percent carbon, atoms of carbon must diffuse from adjacent regions in order to produce this high carbon phase from γ, which contained 0.8 percent carbon. Thus those regions are depleted of practically all carbon, thereby becoming α. Figure 7–7 is a schematic of this process.

Diffusion is rapid at this elevated temperature, and the original nuclei begin a growth process. Although the overall transformation is quite fast (a matter of seconds), this *pearlite reaction* is *time dependent* and does require diffusion of carbon atoms. Now, the differences in P that result because of different cooling rates can be explained. When such a transformation occurs at a high temperature, the rate of nucleation is relatively low, whereas the rate of growth is high because the rate of diffusion is greater at elevated temperatures. Conversely, where the transformation temperature is lower, the nucleation rate is greater *but* the rate of growth is lower, since the rate of diffusion decreases at lower temperatures. Careful measurements indicate that the transformation of γ for furnace cooling occurs slightly below 1341°F (727°C), whereas for oil quenching it takes place at lower temperatures, say, 1000°F (538°C) for our purposes. So we should expect that the P for furnace cooling will have fewer but thicker plates of Θ than for oil quenching. Remember that *neither* of these are true equilibrium cooling and oil quenching is further from that condition than is furnace cooling. So, although both rates lead to *equilibrium* phases, the end microstructure differs.

To assess the actual difference in the harness of the *types* of P, three specimens of *eutectoid* composition (0.8 percent C) could be heated to, say, 1400°F (760°C). By using furnace, air, and oil quenching on one of each specimen, we would end up with pieces having 100 percent coarse, medium, and fine P respectively. The BHN of each have been found to be 240 (P_c), 280 (P_m), and 380 (P_f); these are reliable average values. Note that P_f is almost 60 percent harder than P_c; this could never be explained on the basis of phases alone and is an obvious example of why the use of microconstituents is more practical for explaining differences in properties. *Caution:* Oil quenching eutectoid steel of small size or cross section can produce a substantial amount of martensite (see Sec. 7.4.2). This

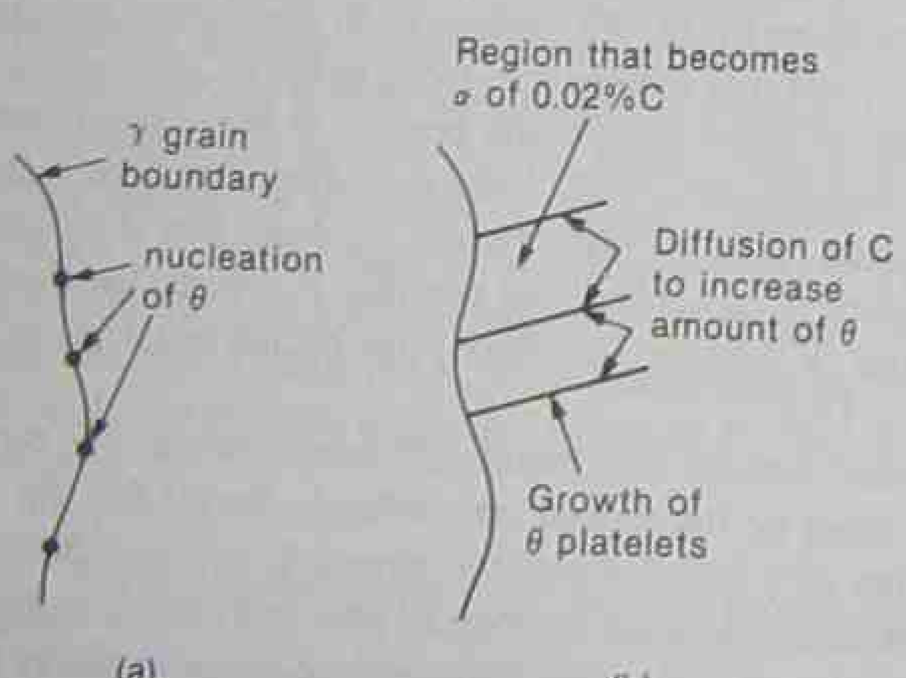


Figure 7–7 The pearlite reaction illustrated schematically.

makes welding of high carbon steel pressure vessels difficult and that is why the carbon content in boiler plate is kept so low by most state codes.

The approximate hardnesses of the FC and OQ pieces of 0.2 percent C (this is a *1020* steel using the AISI designation) can be computed by using a rule-of-mixtures type of calculation. The relative fraction of each microconstituent is multiplied by its BHN; then they are added. For the FC specimen, we had ¾ α (80 BHN) and ¼P_c (240 BHN), so the overall hardness is ¾(80) + ¼(240) = 120 BHN, while the OQ would give ¾(80) + ¼(380) = 155 BHN.

A useful relationship between the BHN and the *tensile* strength S_u of many *steels* is

$$S_u = 500 \text{ BHN, psi} \quad \text{or} \quad S_u = 3.45 \text{ BHN, MP}_a$$

Based upon the predicted hardnesses, S_u for the FC 1020 steel should be 60,000 psi (414 MPa), while that for the OQ is 77,500 psi (534 MPa). This is just about exactly what would be measured from a tensile test. Three points are worth noting here.

1. The method used above could be questioned on purely metallurgical grounds and may be due to a type of averaging affect in these polycrystalline structures. We note a similarity in connection with the elastic modulus of iron or steels. Tests on single crystals of iron indicate that the elastic modulus varies between 18 and 42 × 10^6 psi (124 to 289 GPa) depending upon the direction of loading. An average of 30 × 10^6 psi (207 GPa) results when these extreme values are used. Since bulk metals are almost always polycrystalline, it would seem that an averaging must occur, since these values typify the modulus of those metals.
2. The important point to see with such calculations is that it is the microconstituents and *not* the phases per se that govern properties. Our real concern is not to expect exact correlation between prediction and measurement with such a technique. Rather its use is to indicate the reason behind the noticeably different hardnesses and strengths of the two different structures that contain the same phases.
3. We know of no other alloy systems where the above averaging approach gives adequate predictions. Perhaps it is fortuitous that even one such system provides the means to indicate the importance of microconstituents as compared with phases.

For all plain-carbon steels up to say 1 percent carbon, the weighted averaging of the amounts of microconstituents and their appropriate hardnesses will give similar, reasonable predictions.* With low alloy, superalloys, and stainless steels, no such predictions should be attempted, *yet* the general affect on properties as a function of microstructure is still observed.†

* For steels with carbon less than 0.1 percent, lamellar pearlite may not form, so this averaging technique becomes questionable.

† In these metals, the alloying elements can have a great influence on the hardness of α, and the BHN of 80 is no longer appropriate. Also, the relative amounts of each microconstituent are not readily calculable.

Example 7–1

Estimate the tensile strength of a 10-mm-diameter rod of 1040 steel that is air-cooled from the γ region.

Solution The structure would be about 50 percent α and 50 percent P_m, so the BHN would be (½)80 + (½) 280 = 180, or S_u = 180 (500) = 90,000 psi.

7.4.2 Nonequilibrium Microconstituents

Compared with practically all other commercial alloys, steels possess a unique ability. If they are quenched at a *fast* enough rate, a nonequilibrium microconstituent called *martensite** can form upon the transformation of γ. An entire text could be written by experts in the many subtleties of martensite, M. For our purposes it is sufficient to cover only the major aspects of the formation of M and its importance in relation to mechanical properties.

Martensite is a body-centered tetragonal (BCT) structure of iron that is supersaturated with carbon; that is, it contains much more carbon than equilibrium cooling would allow. It is useful to realize that if all of the excess carbon were removed (this is discussed shortly), the BCT structure would revert to the BCC structure of α. There is no single hardness of M and for plain-carbon and low-alloy steels, *maximum* hardness is solely a function of the carbon content. Figure 7–8 shows a relationship of hardness as a function of percentage of carbon. We emphasize that this plot shows the *maximum* hardness attainable; in practical situations, lower values result, and the shaded region illustrates this. Note that the divergence from maximum becomes greater as the carbon content is decreased. When M is first produced, it is called *primary* or *as-quenched* martensite and is often defined as a diffusionless, shear-induced transformation. Unlike the P transformation that is time-dependent (since diffusion is needed), the M transformation occurs in

* Martensitic transformations do occur in other alloy systems, such as Cu-Mn and In-Th, but such alloys do not find *widespread* industrial use.

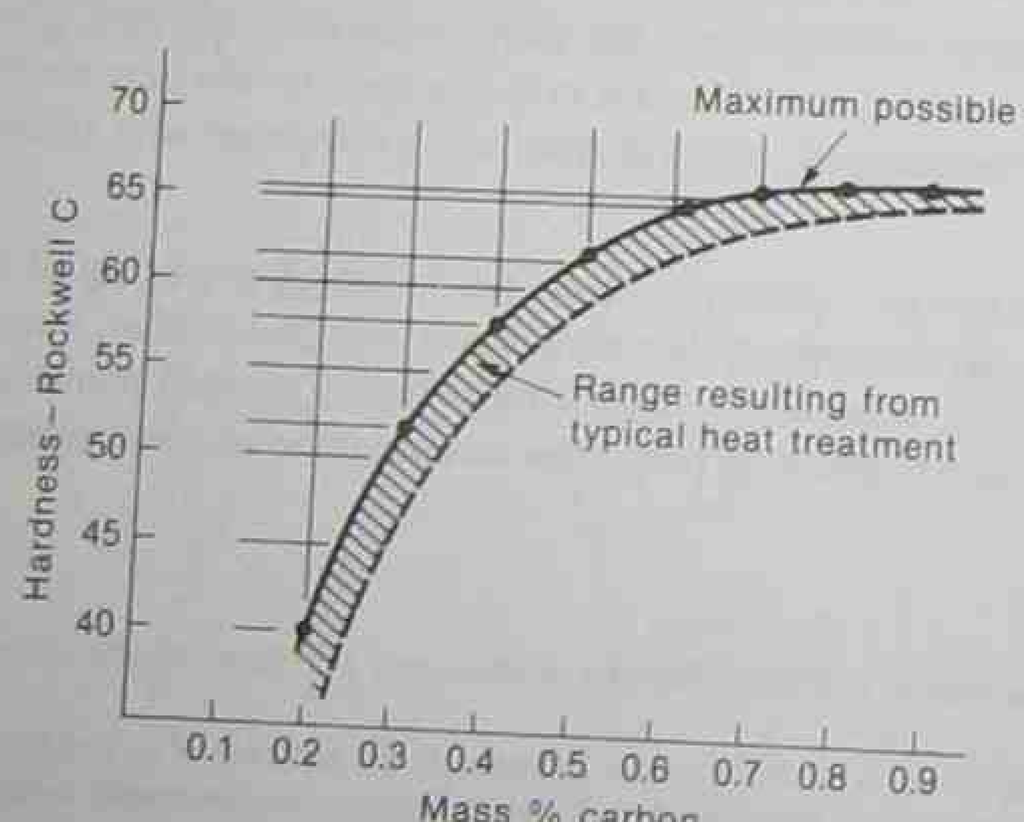


Figure 7–8 Hardness of martensite as a function of carbon content.

about 10^{-7} seconds, which is considered to be instantaneous. Although much harder than α and P, M is seldom used in the primary condition. The reason for this has to do with residual stresses that develop across the section of the quenched piece. The surface is the first region to transform from γ to M, and this requires a volume *expansion* of the γ; the center of the piece is still γ. As the center transforms to M, α, or P (depending upon the carbon content), a volume expansion of γ must occur. This causes one of two consequences, since the surface M is strained in tension due to the volume expansion of the center. First, the surface may crack due to the tensile strains induced, and cracked pieces are not too useful. If the piece does not crack, the surface is left in tension. Now as the center cools, say, to room temperature, it undergoes thermal contraction, thereby forcing the surface to also contract. This complex sequence of events locks in strains; the net result produces *residual* stresses which can reach a fairly high level. The expected tensile strength, based upon the hardness of the surface (via S_u = 500 BHN psi) would not be reached in a tensile test; rather a lower brittle fracture stress would be measured. Figure 7–9 illustrates this, noting that the applied tensile stress is added on to residual tensile stresses. Thus the measured tensile *strength* would fall below that which would result *if* no residual tensile stresses existed. To overcome this problem, the quenched part should, at the very least, be given a thermal stress-relief where diffusion effects tend to remove the locked-in strains. Heating to 200–300°F (93–149°C) for one hour per inch (25 mm) of cross section is usually adequate. Although there is little if any change in the hardness of the piece, the removal of the residual stresses leads to an increase in strength.* In fact, the relation between tensile strength and BHN used earlier is now appropriate. In the as-quenched condition, however, large deviations between measured hardness and predicted S_u can occur. Figure 7–10 illustrates this point. Even stress-relieved martensite (M_{sr}) has certain limitations in applications. Although it possesses high strength and wear resistance, its ductility and *impact resistance* are minimal. For that reason, it is usually heated

* Hardness involves plastic flow under high hydrostatic compressive stress. This inhibits fracture of primary M. Such is not the case under tension, as shown by the different values of S_u.

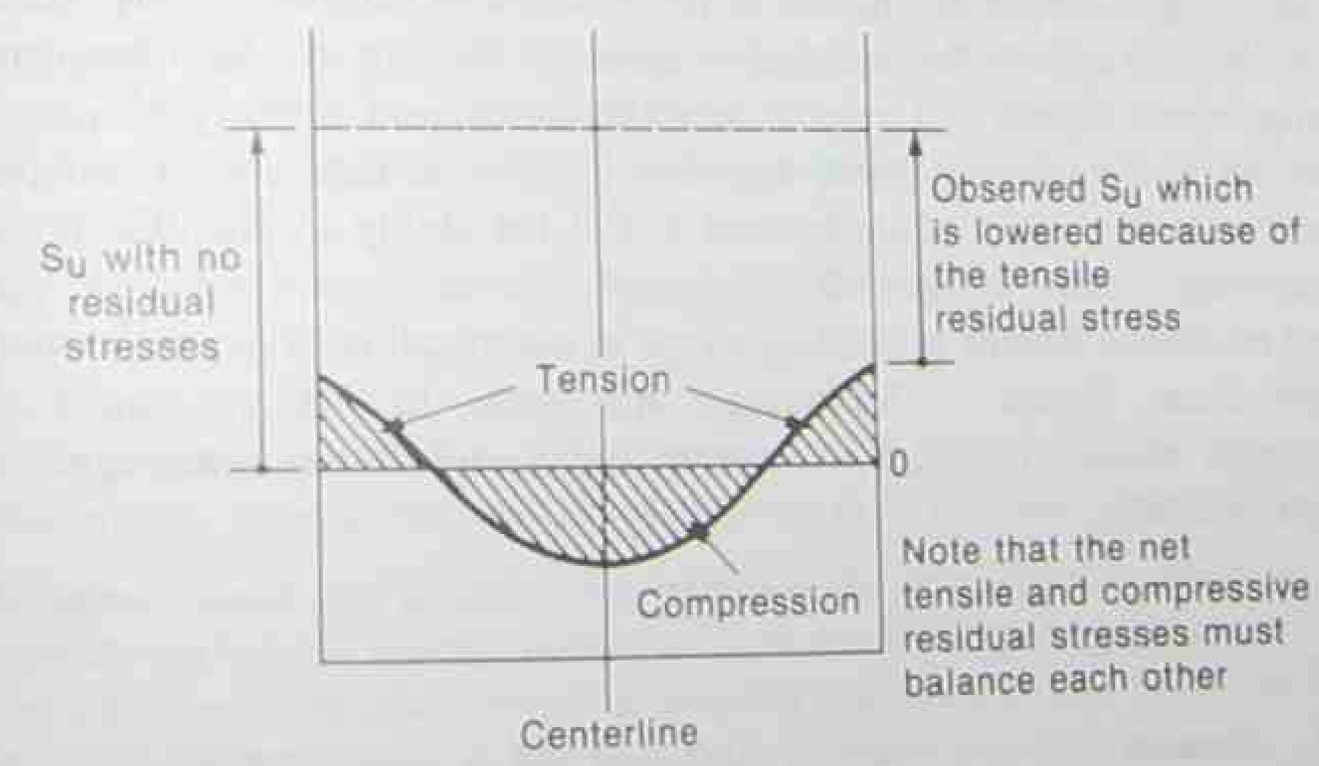


Figure 7–9 Illustration of a residual stress distribution resulting in primary martensite.

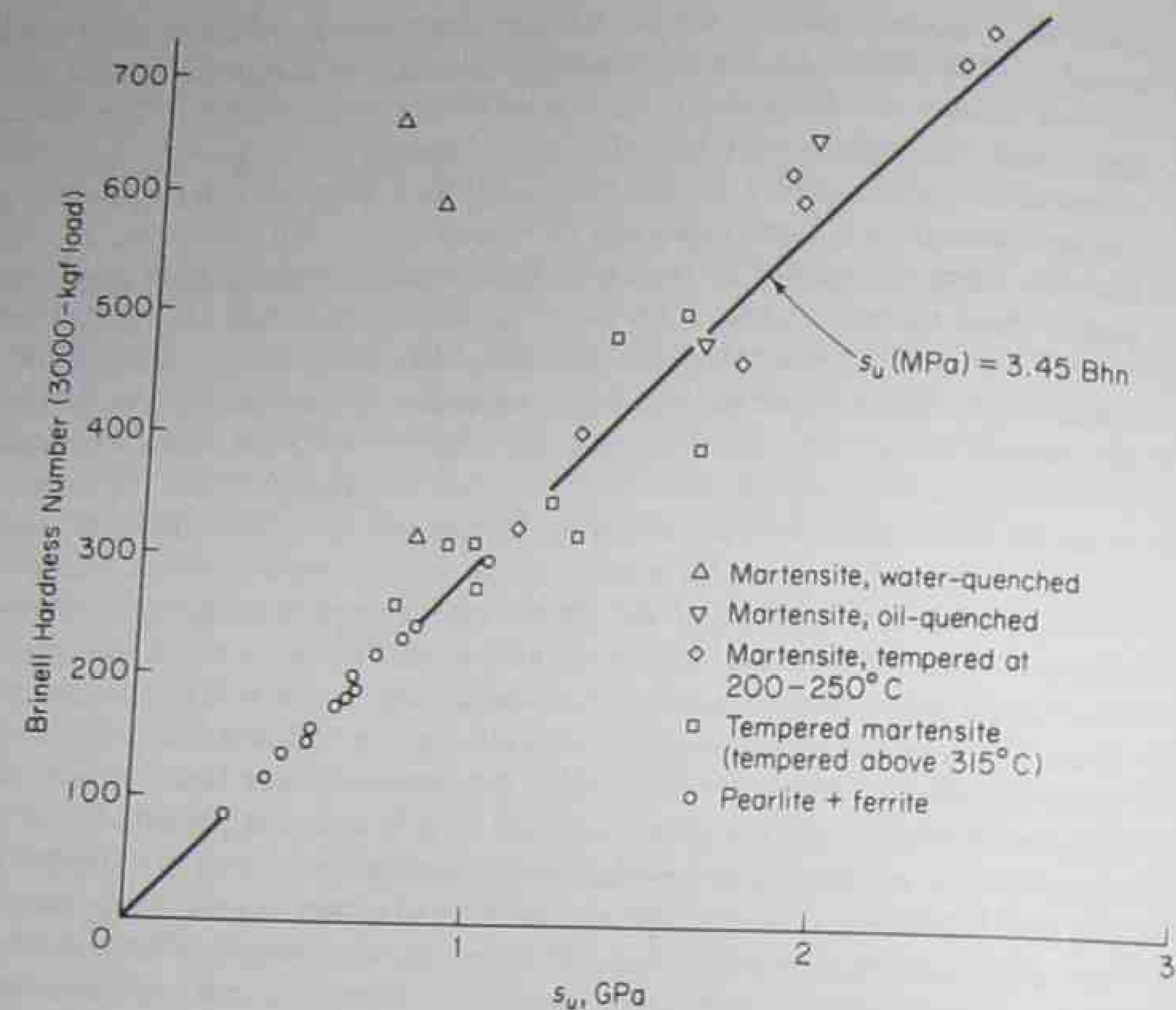


Figure 7–10 Relation between BHN and tensile strength for a number of steels. Note that the water-quenched steels, giving primary martensite, do not follow the usual trend.

(that is, *tempered*) to form what is called tempered martensite (M_T). There is no way to give a single simple description of the structure or hardness of M_T, since the end result is a function of both the temperature used and the time at which the piece is held at that temperature. Figure 7–11 shows how the hardness of primary M varies with temperature for some selected steels; note that the time is constant on that plot. Perhaps the most important point to understand is that the ability to alter the properties of M by tempering allows a trade-off of strength for an improvement of ductility and impact resistance thereby permitting a type of optimization of overall properties for specific applications. Figure 7–12 illustrates this feature in that one cannot attain maximum strength, ductility, impact resistance, and so on at the same time; some sacrifices must be made.

Example 7–2

Consider a ¾-inch round of 1050 steel that is water-quenched from just above 727°C. Estimate the probable hardness of the quenched piece.

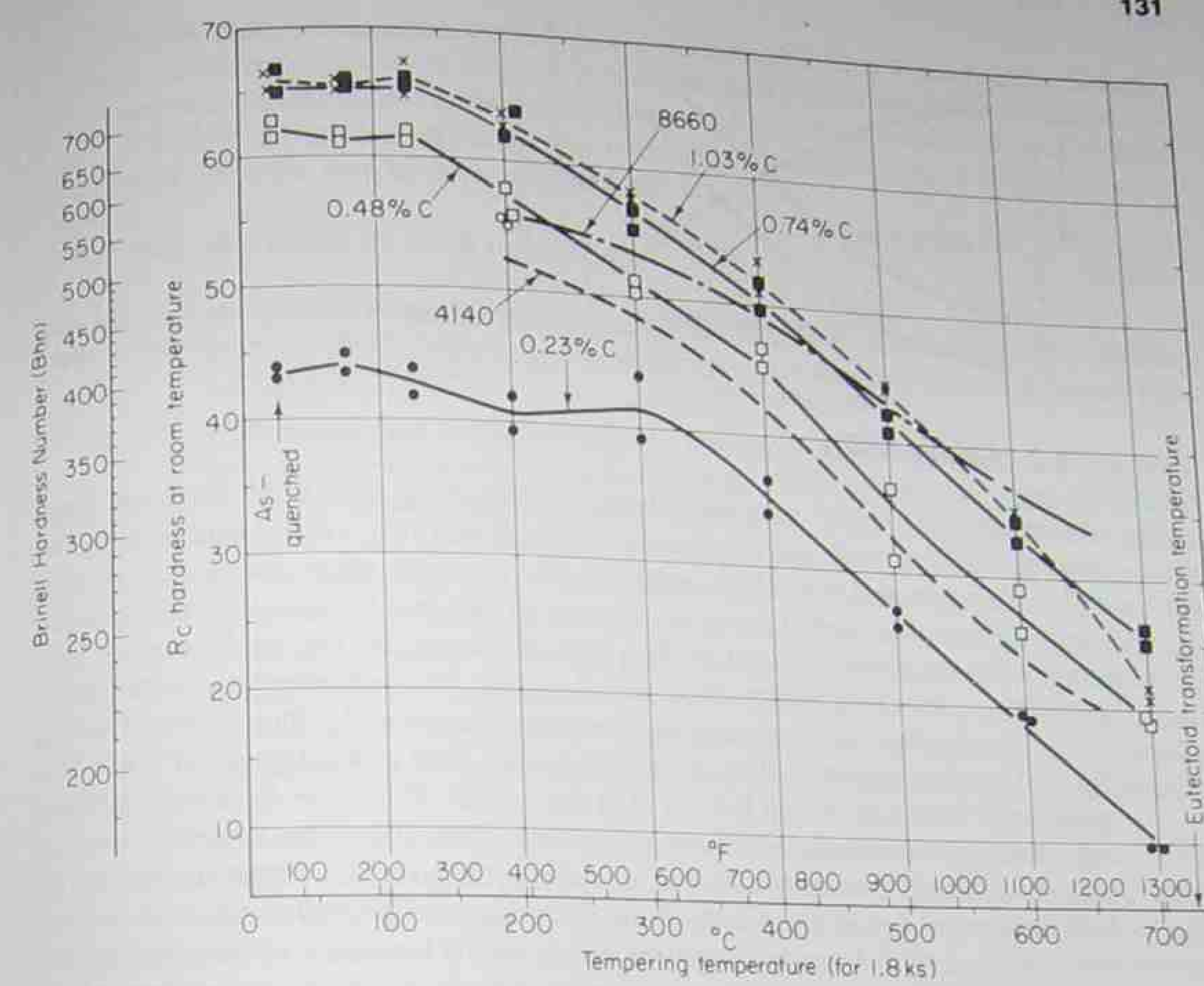


Figure 7–11 Hardness at 20°C as a function of tempering temperature for a number of steels, all tempered for 1.8 ks.

Solution Since the structure before quenching is about ⅜ α and ⅝ γ, the structure after quenching will be ⅜ α and ⅝ M of 0.8 carbon. From Fig. 7–8, the M will have a hardness of about 65 R_c or about 700 BHN. So the approximate hardness will be (⅜)80 + (⅝)700, or 467 BHN.

Example 7–3

If the quenched piece in Example 7–2 were heated to 400°C for 1.8 ks, then cooled to room temperature, estimate the tensile strength of the piece after this tempering operation.

Solution Since only the martensitic portion of the structure will temper, use Fig. 7–11 to estimate the hardness of that portion as about 500 BHN. Note the plots for 0.74 and 1.03 percent carbon. Thus the BHN of the tempered structure is (⅜) 80 + (⅝) 500 = 342, so S_u = 500(342) = 171 ksi.

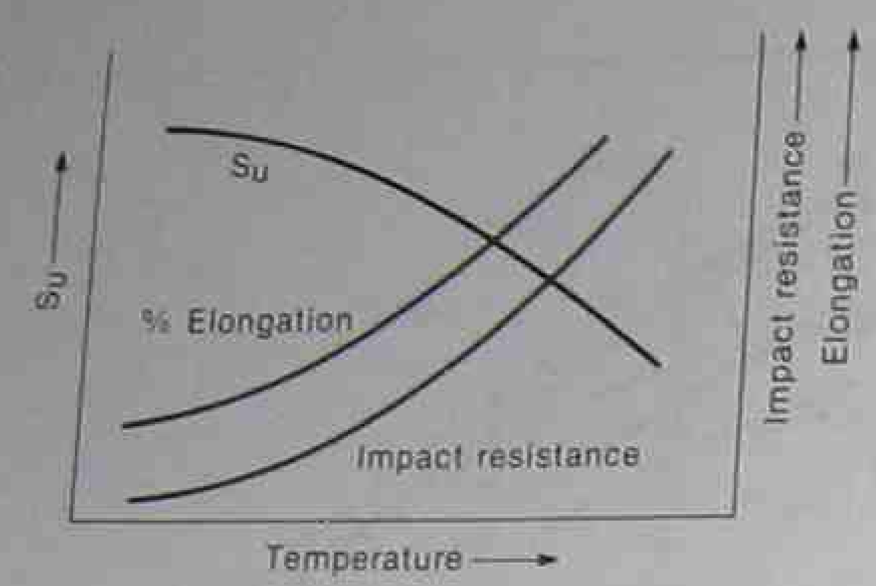


Figure 7–12 Schematic illustration of the effect of tempering temperature on certain mechanical properties.

7.4.3 Hardenability

An important property of many steels, both in understanding some of the rudiments of heat treating as well as in the process of welding (see Chapter 10), is called *hardenability*. We define this as the ease with which martensite forms to depth across a section. There is no quantitative method used to assess this property in the sense of saying that one steel possesses twice the hardenability of another. Instead, we say that one steel has better hardenability compared with another. The reader should *not* confuse hardness and hardenability. As discussed earlier, the hardness of M is basically a function of carbon content. Consider Fig. 7–13, where pieces of 1040 (plain-carbon) and 4340 (low-alloy) steels* have been quenched at rates fast enough to form M at the surface. The pieces have been sectioned and a hardness traverse made from surface to center. Since the carbon content is identical, both pieces show the same hardness at the surface, but a rapid decrease is seen with the 1040 steel as one moves towards the center. No such rapid drop-off occurs with the 4340 steel, so this steel has better hardenability; that is, it fully hardens at a slower cooling rate. As section sizes become larger, a requirement to form M across the section demands the use of a steel possessing good hardenability. This property is influenced by both higher carbon content and larger γ grain size prior to quenching, but the most effective means for improving hardenability is the addition of alloying elements such as molybdenum, chromium, nickel, manganese, and boron. A useful concept to keep in mind can be illustrated by considering a eutectoid steel. If γ is to transform, it must choose between P and M, and any action which diminishes the chance of forming P will enhance the formation of M. Since P requires diffusion of carbon (M does not), the presence of alloying elements must retard such diffusion, thereby improving hardenability.

* The numbers refer to the AISI designation for steels.

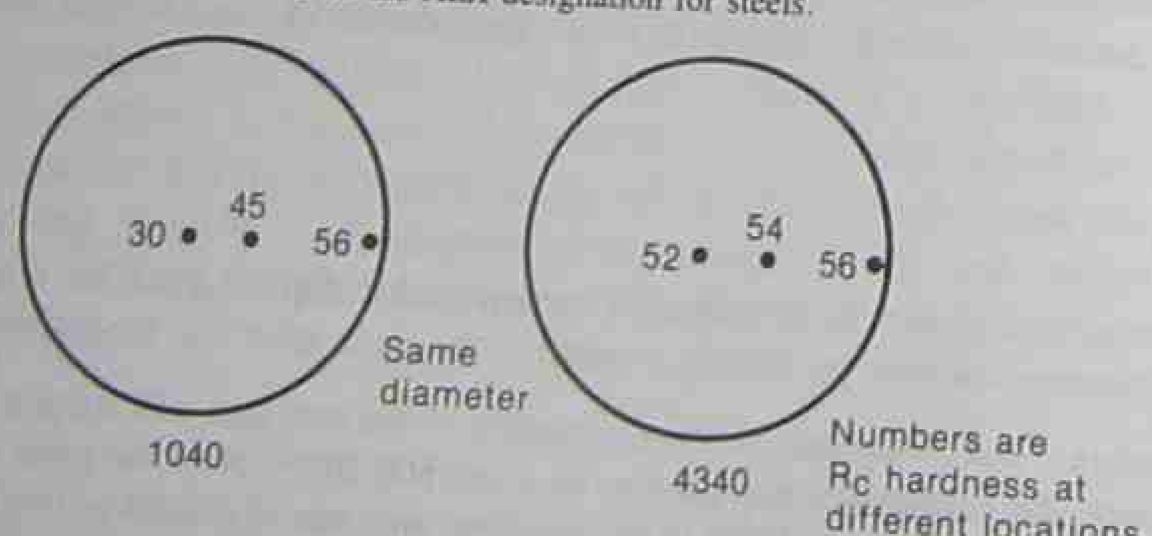


Figure 7–13 Surface to center hardness variation of 1040 and 4340 steels to show the superior hardenability of the low-alloy steel.

Two other points are worth mentioning.

1. The concept of hardenability pertains *only* if an alloy has the capability of forming M. This term is never used with the many commercial alloys that cannot form martensite.
2. With metals such as plain-carbon and low-alloy steels, both the section size of the part and the rate at which it is cooled from γ will influence the depth at which M can form. Because plain-carbon steels have poor hardenability, it is almost always essential to use water quenching to form any reasonable depth of M, but even with this fast quenching rate, it is difficult to form a significant amount of M with such steels if the carbon content is low (say less than 0.3 percent). The better hardenability of low-alloy steels permits them to be cooled at slower rates, and still form M to reasonable depth; oil quenching is often used. It is always sensible to use as slow a cooling rate as possible to form M, since that reduces any tendency to cause cracking. For that reason, water quenching low-alloy steels is usually avoided if at all possible.

In closing the discussion about hardenability, consider two additional figures. Figure 7–14(a) is a so-called time-temperature-transformation (TTT) curve (or IT curve, which is short for isothermal transformation) for a plain-carbon steel of eutectoid composition. The diagram is obtained in the following way. Consider a series of identical specimens of eutectoid composition, all homogenized at 1450°F (788°C); they all are composed of grains of γ containing 0.8 percent C. If the temperature of one specimen were suddenly reduced to 1200°F (649°C), held at that temperature for a time period denoted by point 1, and then rapidly quenched in water, a structure of martensite, M, would result. In essence, since M can only form from γ, there was *no transformation* of γ during time interval 1 at 1200°F. By repeating this procedure with longer and longer time intervals, we eventually determine the time at this temperature where a small amount of γ transforms to pearlite, P, before water quenching; the final structure in this case would be M plus a small amount of P. (Again recall that M forms from γ, and if some of the γ→P, we can't end up with 100 percent M.) Thus point 2 indicates the time needed to cause the *start* of transformation of γ at 1200°F. Continuing this procedure, we then find the time to produce complete transformation of γ→P at 1200°F; this is designated by point 3. Note that between points 2 and 3 the structure after water quenching would contain differing amounts of P and M; ratios of these two microconstituents would depend upon the amount of γ→P *prior* to water quenching, and the degree of γ→P itself depends upon the time interval. Repeating the procedure for a number of temperature levels produces points comparable to 1 and 3 with the 1200°F temperature; in essence, the time intervals to begin the γ→P transformation are determined. Connection of all points for

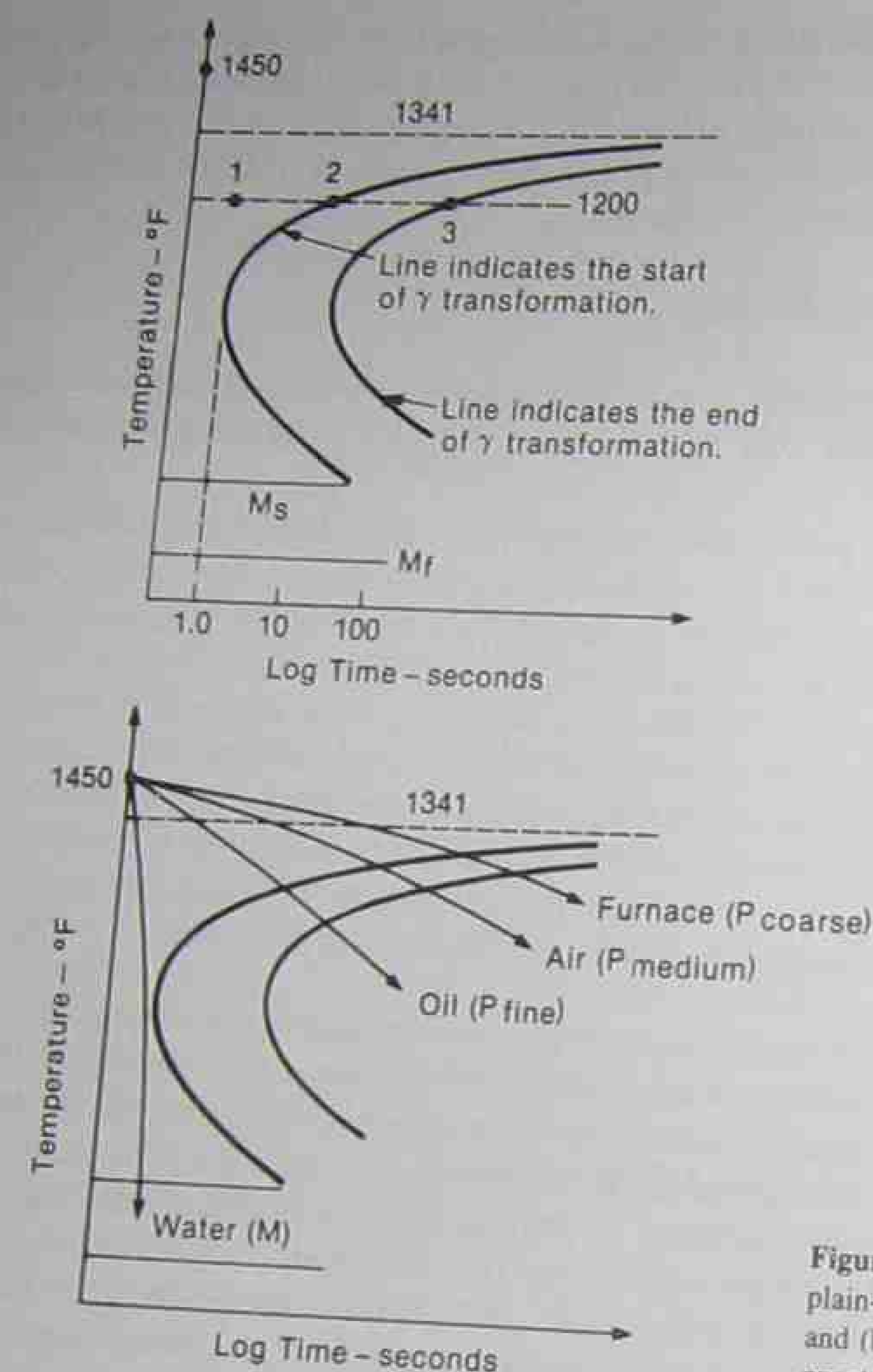


Figure 7–14 (a) T-T-T curve for a plain-carbon steel of eutectoid composition and (b) superposition of curves of different constant cooling rates on the curve in (a).

each of these conditions produces the boundary lines shown on Fig. 7–14(a). To complete such a diagram, it is necessary to determine the temperatures at which γ begins to transform to M and at which that transformation is completed; these are shown as the M_s and M_f lines on that same figure.

Of course, most practical heat treatments do not involve transformation of γ under isothermal conditions. Instead, parts are quenched in media such as water, oil, or air, or in a furnace, all of which produce continuous cooling transformations, or CT for short. To illustrate this point, and to avoid undue clutter on a single figure, Fig. 7–14(a) is reproduced as Fig. 7–14(b), and curves of *constant* cooling rates are superimposed to reflect the approximate rates of the four widely used quench media. We *emphasize* that this approach is taken to illustrate concepts and is not intended to provide wholly accurate predictions of the final microstructures. With water quenching, the cooling rate is considered to be

high enough that no γ transforms until the M_s temperature is reached (that is, the *nose* of the curve has been missed, so no P forms). At the other extreme, furnace cooling (FC), all of the γ transforms at a high temperature and coarse P results. For the intermediate cooling rates, medium P results with air cooling, whereas fine P is obtained with oil quenching. The following comments are important here:

1. Since Fig. 7–14(a) was the result of isothermal transformations, it is not technically correct to superimpose cooling curves on such a diagram and expect to make accurate predictions of the resulting microstructures. In Sections 7.4.1 and 7.4.2 we have alluded to the influence of cooling rate on the resulting microstructure, and with Fig. 7–14(b) an attempt is made to tie in the effect of cooling rate on the structures likely to result.
2. Both the carbon content and the percentage and type of alloying elements lead to IT diagrams that look quite different from Fig. 7–14(a).* In addition, the size of the piece and the quenching medium used can lead to variations in microstructure from surface to center, since the cooling rates at these two regions can be quite different. Thus the use of appropriate CT diagrams and associated data are far more reliable in predicting resulting microstructures than one would expect by using figures such as Fig. 7–14(b).† Once again, the technical details and decisions related to hardenability considerations are best left to the specialists in metallurgy and not to the typical manufacturing engineer.
3. The most widely used manner for presenting hardenability data directly is with the use of *Jominy* curves. Such results are obtained by using the so-called Jominy test.‡ Figure 7–15 compares typical findings for 1040 and 4340 steels. Note that the maximum hardnesses, obtained at a high cooling rate, are the same, since they both possess the same carbon content. However, due to the presence of alloying elements, the 4340 steel *maintains* a much higher hardness as the cooling rate is decreased. In essence, thicker sections of 4340 could be hardened, so its hardenability is superior to the 1040.

7.4.4 The Copper-Aluminum System

The reasons for discussing these alloys in some detail are twofold. First, they find widespread commercial use, and second, they involve the important ability to undergo *precipitation hardening* (also called *age hardening*). Figure 7–16 shows the portion of the equilibrium diagram of interest. Other than the liquid phase, two equilibrium phases are indicated, these being an α solid solution and an intermetallic compound Θ. Note that these *symbols* were used with the diagram in Fig. 7–4, but they are not to be confused with

* See, for example, *Isothermal Transformation Diagrams*, 3rd ed. (United States Steel, Pittsburgh, Pa., 1963).

† See, for example, M. Atkins, *Atlas of Continuous Transformation Diagrams for Engineering Steels* (British Steel Corporation, Sheffield, England, 1980).

‡ See, for example, D. K. Felbeck and A. G. Atkins, *Strength and Fracture of Engineering Solids* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1984), pp. 226–28.

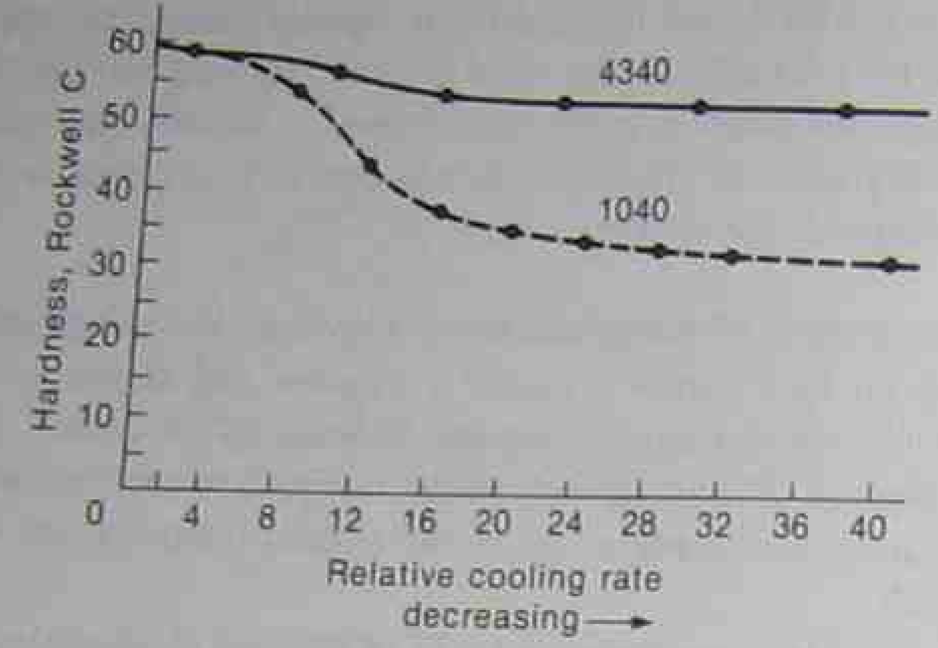


Figure 7–15 Typical Jominy curves for 1040 and 4340 steels.

ferrite and cementite. Here, the compound Θ is copper aluminide ($CuAl_2$), while the α is a solid solution of copper and aluminum where the maximum solubility of copper in aluminum is 5.65 percent.

First consider an alloy containing 4 percent copper that is heated to about 500°C, as shown by the dashed line on Fig. 7–16. When it is homogenized—that is, when Cu is uniformly dispersed in Al—we have a solid solution containing 96 percent Al and 4 percent Cu. Upon slow cooling, say by shutting off a furnace, the phase structure remains constant until a temperature just below point A is reached. From there down to about 200°C the diagram indicates that there are now two phases (α and Θ) that exist and that the amount of Θ increases as the temperature decreases. The phase boundary line beginning at 548°C and showing 5.65 percent copper drops downward to the left to about 05 percent copper at 200°C. This is an indication of *decreasing solid solubility* with tem-

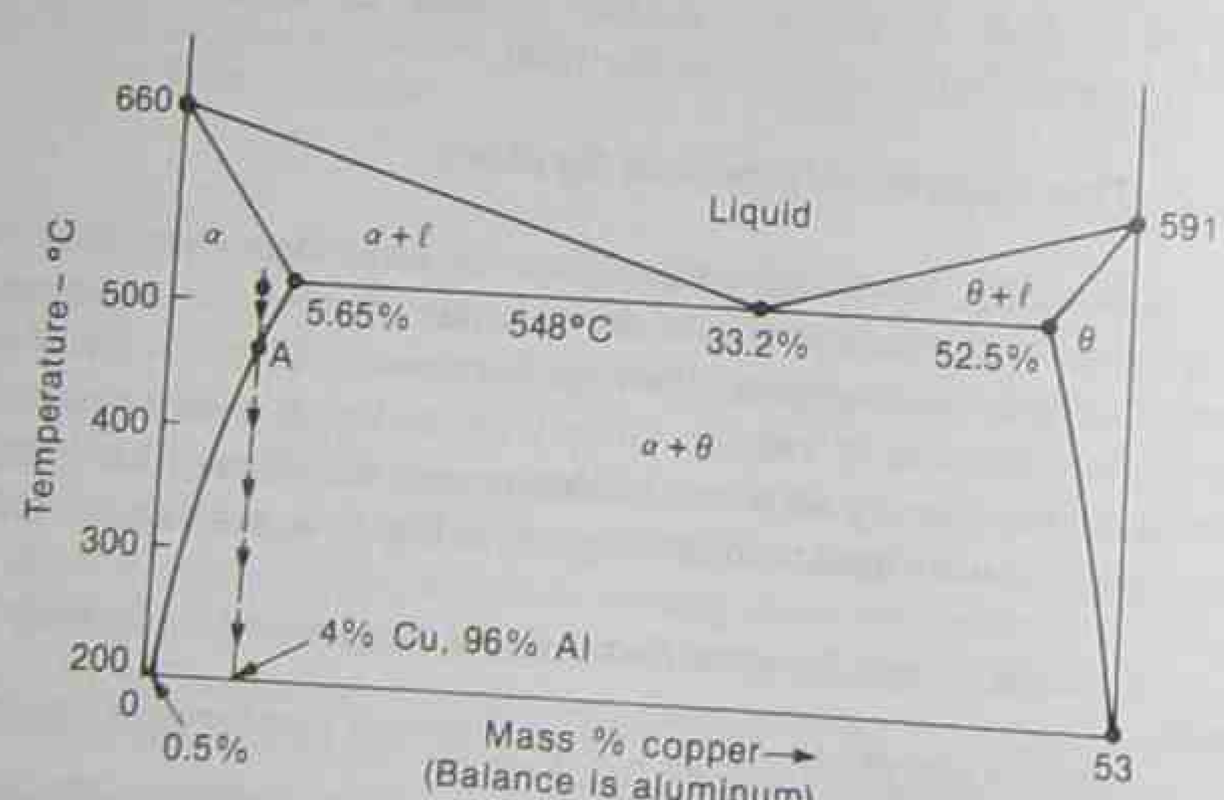


Figure 7–16 Sketch of aluminum-copper phase diagram to 53 percent copper.

perature; that is, the amount of copper that can be retained in solution with aluminum decreases with temperature under equilibrium cooling. As the copper content of the α changes from 4 percent at about 450°C to 0.5 percent at 200°C, the copper atoms that comprise this 3.5 percent difference combine with some aluminum atoms to form the Θ phase. During this temperature drop, the grain boundaries are at a higher energy level than are the grain interiors, so as the Θ forms, it will tend to precipitate principally in the grain boundaries (again tending to decrease the overall energy level of the structure). Figure 7–17 is a *schematic* microstructure indicating the end result. There are two practical disadvantages to such a structure. First, it would possess relatively low strength, since its major microconstituent α is almost pure aluminum. Second, the hard, brittle Θ phase, being confined to grain boundaries, would produce a more brittle structure than expected for this low strength and ductile matrix (α). Whenever intermetallic compounds (Θ here) are located primarily in grain boundaries, fracture often begins in those regions and proceeds as intergranular cracking, so the tougher, more ductile matrix does not contribute as much to fracture resistance; the end result is a lower ductility than would be expected. As mentioned regarding cementite in steels, hard phases are generally most effective if they are dispersed as small particles throughout the matrix. Precipitation hardening does that very thing, as now described.

Example 7–4

Refer to Fig. 7–16. If an alloy containing 15 percent copper were slowly cooled to room temperature, describe the resulting microstructure.

Solution After full solidification at just under 548°C, a structure of α solid solution and eutectic exists. As further cooling occurs, Θ will precipitate out of the α solid, primarily in α grain boundaries; the end result will consist of grains of α solid solution, Θ precipitate, and grains of eutectic.

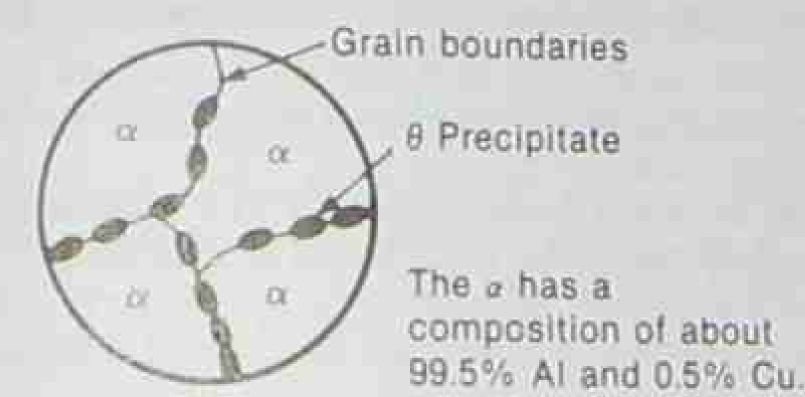


Figure 7–17 Schematic microstructure of a 96 percent aluminum-4 percent copper alloy furnace cooled from the α region to 200°C. Note that the Θ phase precipitates primarily in the grain boundaries.

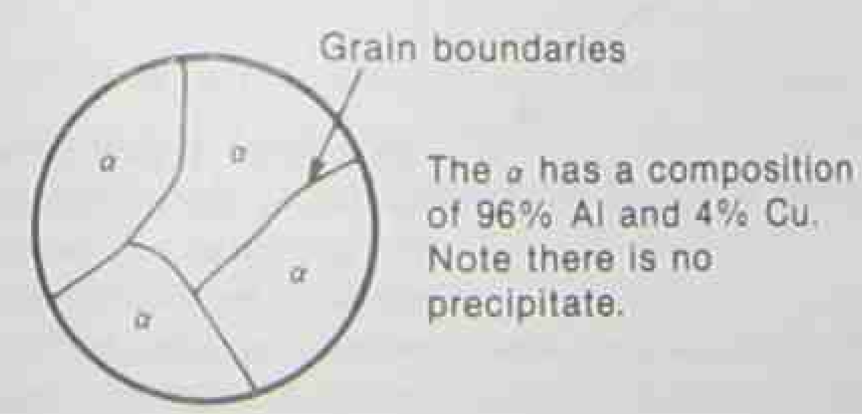


Figure 7–18 Schematic of the alloy in Fig. 7–17 that has been water-quenched from the α region. A supersaturated α solid solution results.

If this 4 percent copper alloy were quenched in water from 550°C, there would be inadequate time to form the Θ phase as earlier, since diffusion could not occur, and that is essential if Θ is to form upon the lowering of temperature. A microstructure of supersaturated α would result as shown in Fig. 7–18 and in that condition, it would be stronger and usually more ductile than the structure indicated in Fig. 7–17; note that *no precipitate* exists after the initial water quench, and all the copper has been used to form the supersaturated solid solution of α. This α is not near equilibrium and would like to lower its copper content as it did when furnace cooled. What follows will not describe certain metallurgical subtleties, since it is the end results that are of major interest here. Both time and temperature are of importance, but for simplicity, consider that temperature is held constant for now. Suppose we conduct a series of tests where we heat the supersaturated α to, say, 400°C for a few minutes, quench it in water, measure its hardness, then repeat this procedure for ever-increasing time intervals. The sketch in Fig. 7–19 shows a typical result of hardness versus time, and we note the following:

1. The hardness begins to increase.
2. The hardness reaches some maximum value.
3. The hardness begins to decrease and eventually becomes *lower* than that of the *initial* α.

A series of schematic microstructures, shown in Fig. 7–20, will assist in explaining this sequence of events. First there occurs a precipitation of relatively *few, fine* particles.* These particles tend to interfere with the easy movement of dislocations, thereby increasing hardness and strength. To form the precipitate, some of the copper atoms from the supersaturated α combine with aluminum atoms; this *reduces* the degree of supersaturation of the α, thereby lowering the energy level and causing the α to be closer to equilibrium. As this process is continued at longer times, more and more precipitate forms and the hardness continues to increase. Note that two effects† are counteracting here. To form

* These will always be indicated by dots, although they have a small platelet-like form during the early time stages and are of the order of 10 nm in size.

† The degree of *coherency* of the structure is also important but is not included here. See D. K. Felbeck and A. G. Atkins, *Strength and Fracture of Engineering Solids*, p. 194 for a discussion.

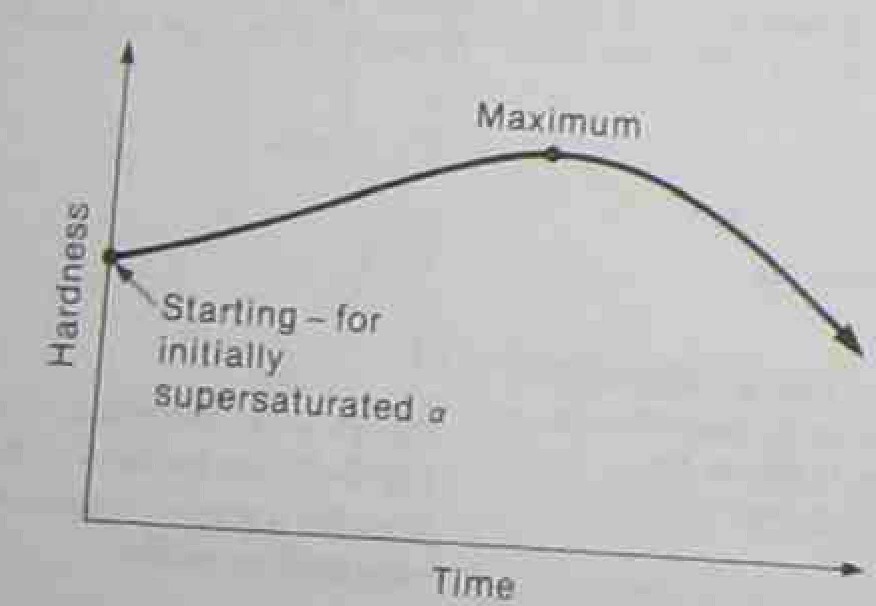


Figure 7–19 Typical hardness versus time plot as an initially supersaturated α solid solution undergoes precipitation hardening. Beyond the point of maximum hardness, overaging results and the hardness begins to decrease.

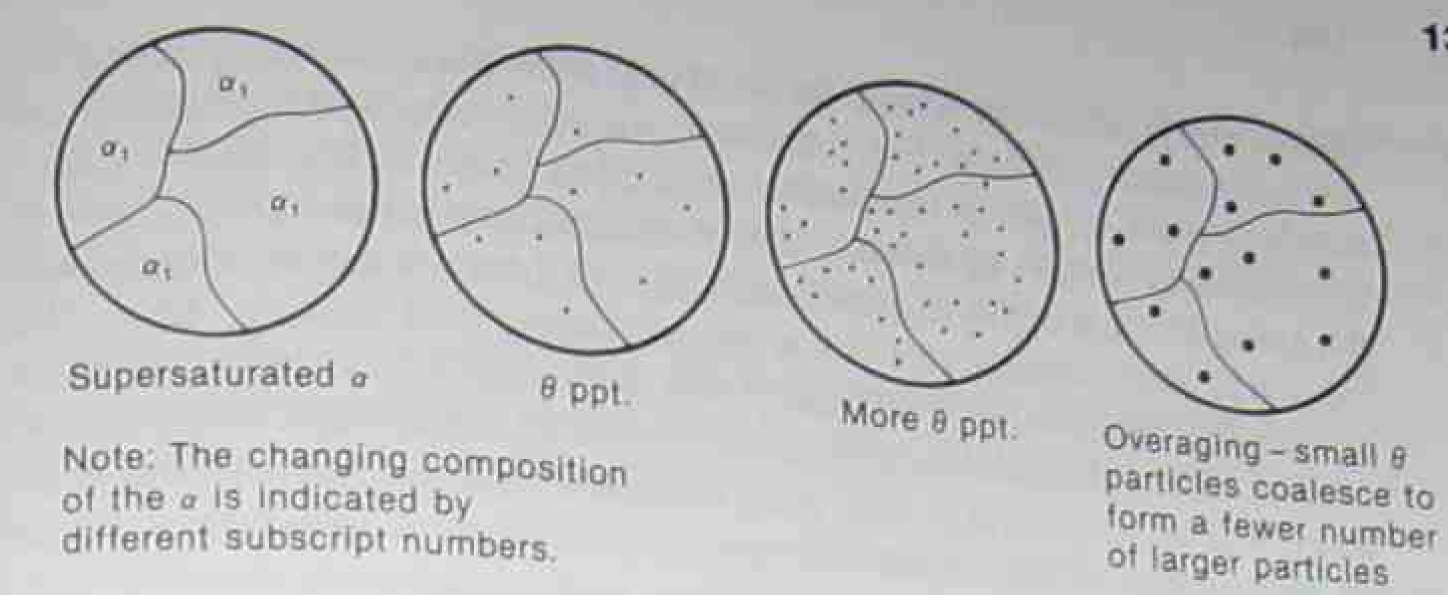


Figure 7–20 Series of schematics to illustrate the effects of precipitation hardening and subsequent overaging.

more Θ, a greater number of copper atoms must *come out* of the α solid solution, and the degree of supersaturation is continually lessened. The overall hardness at any time is, therefore, the result of two competing factors. If Θ is to form, it must do so at the expense of the supersaturated α, so at longer times, the α itself becomes softer, since the *solution hardening* effect of the excess copper diminishes. When the influence of the precipitate more than offsets the diminishing solution hardening effect, the *net* result is an increase of hardness and illustrates what is meant by precipitation hardening. This important strengthening mechanism permits one to attain a variety of combinations of strength and ductility for the many aluminum alloys used commercially. Note that some of these alloys will age-harden fairly quickly at room temperature (called *natural aging*), but most require the use of an elevated temperature for specific time periods (called *artificial aging*) to attain the desired structure in a reasonable time period. It is important to consider the possibility of *overaging* when such alloys are used in practice. If they are subjected to elevated temperatures for adequate time, they will begin to soften, and the resulting decrease in strength could lead to serious consequences. We note that various time-temperature combinations can produce similar results where higher temperatures lead to an increase in hardness in shorter time.* Designers must be aware of this and other potential problems when they specify the use of such alloys. For example, to produce what is known as a T-6 (artificially aged) condition can lead to residual stresses, while welding such a structure can cause a serious decrease in the original strength.

7.4.5 Cold Work, Recrystallization, and Grain Size

The effect of *cold working* (also called work hardening or strain hardening) was discussed in a quantitative way in Chapter 6. Here it is covered qualitatively as a strengthening mechanism and may be defined as plastic deformation carried out at temperatures below

* An excellent reference on this topic is John E. Hatch, ed., *Aluminum: Properties and Physical Metallurgy* (Metals Park, Ohio: American Society for Metals, 1984).

which *recrystallization* would occur. To explain these words, a series of schematic microstructures will help. For simplicity, consider a *single phase* material such as commercially pure copper or aluminum, or an alpha brass. In Fig. 7–21(a), the structure is considered to contain *equiaxed* grains of a particular average size. Due to cold working, the grains become highly distorted and although there is a definite change in the *shape* of the grains, their average size is not altered. During cold working, the number of initial dislocations increases tremendously,* and the energy of the grain interiors thereby increases. If the structure is then heated at an elevated temperature, nuclei of new strain-free grains will begin to form at *both* grain boundaries and interiors, and this begins to lower the energy level of the overall structure. With time, these initial grains begin to enlarge as other new nuclei form. During this time, many dislocations disappear and the energy level continues to decrease. When this process is first completed, we have a strain-free structure of relatively small grain size and an overall dislocation density about the same as existed prior to cold working. Figures 7–21(b) and (c) illustrate these comments. If the process is not stopped at this point, the phenomenon called *grain growth* will occur wherein smaller grains of severe boundary curvature (that is, smaller radius) will tend to disappear and a smaller number of larger grains results; see Fig. 7–21(d). In essence, the coarse grain structure is closer to equilibrium than the finer grain structure, since the total amount of grain boundary surface area is smaller. The above procedure is used if we wish to *refine* grain size, that is, cold work the metal then cause recrystallization to be completed throughout. It is noted here that the final recrystallized grain size, as in Fig.

* In the annealed condition, the dislocation density of bulk metals is about $10^6 cm^2$ whereas after severe cold working it increases to the order of $10^{12} cm^2$ (planar density). For the model used to explain this increase see, for example, R. M. Caddell, *Deformation and Fracture of Solids*, p. 172.

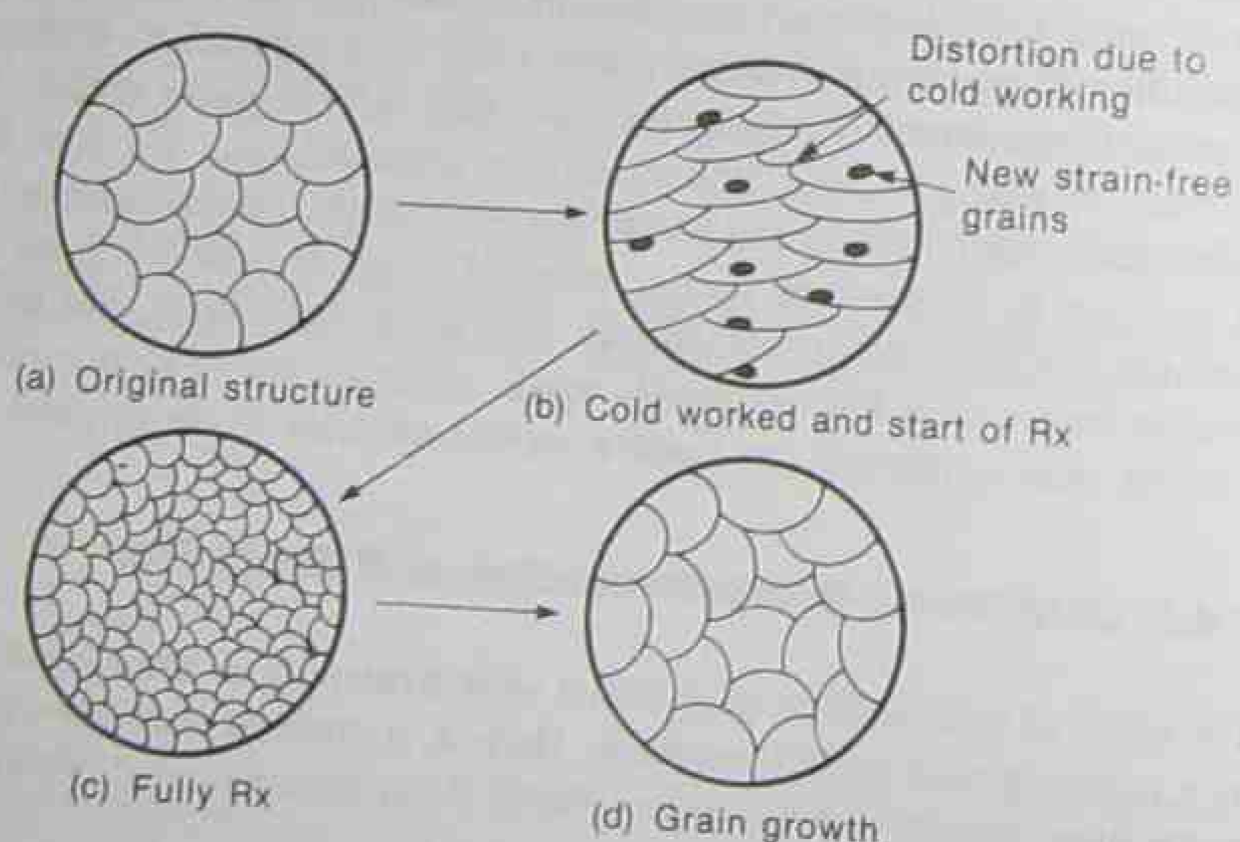


Figure 7–21 Series of schematics to illustrate the sequence of cold-working and annealing a single-phase structure. Note that a strain-free, fine-grained structure undergoes grain growth if heating is continued.

7–21(c), will be influenced by the degree of initial cold work and the time-temperature combination used in this process.

With single-phase metals, whose microstructure cannot be influenced by methods discussed in previous sections, their properties can only be altered by cold working or by changing grain size. Taken individually, this occurs as follows:

1. Cold working by itself causes a drastic increase in the *total number* of dislocations, but the number capable of *easy glide decreases*. As a general consequence, strength increases and ductility decreases.
2. If we compare a coarse versus a fine-grained structure, where grain boundaries provide the major impediment to dislocation motion, we see that the *average* glide distance is smaller with the finer grain size. That is, as dislocations move under the application of stress, they travel a shorter distance before reaching a grain boundary in the fine-grained structure; this ties in with an increase in strength.*

Figure 7–22 illustrates the typical behavior of the cold work-annealing cycle that is observed with single-phase structures.

In forming a part to its final desired shape, it is sometimes necessary to induce strains that would cause fracture *before* the end shape is reached. To avoid this, cold working is carried out to a level short of causing fracture, and the piece is then annealed to remove the effects of this initial cold work. Now that ductility is restored, the piece is further cold worked to its final shape.

7.5 POLYMERS

7.5.1 Introduction

Words such as *plastics* or, more recently, *engineering* plastics are now commonly used to denote those types of *polymers* that are usually man-made and used to make engineering components. As such, they are often called *synthetic* polymers to distinguish them from *natural* polymers such as wood, bone, and the like. The polymers of major importance in this context are comprised of organic molecules that are bonded together by various means to form a solid structure. Just as the unit cell formed the basic building block for crystalline metals, what is called the *mer* or monomer is the smallest repetitive unit from which polymers (that is, many mers) are developed. However, the fundamental forces that hold a polymer together are very different from metallic bonding, and this is why the properties of polymers are so different from metals. It is this difference in properties that usually leads to the use of polymers in numerous industrial applications, since they often fulfill a need that other solids cannot provide, or they do so more economically. To give a reasonable understanding of the behavior of polymers, it is necessary to delve into a brief coverage of a few of the principal aspects of structure as was done with metals.

* At temperatures above what is called the *equicohesive* temperature, the coarser-grained structure can be viewed as being stronger. This is where the phenomenon called creep becomes important.

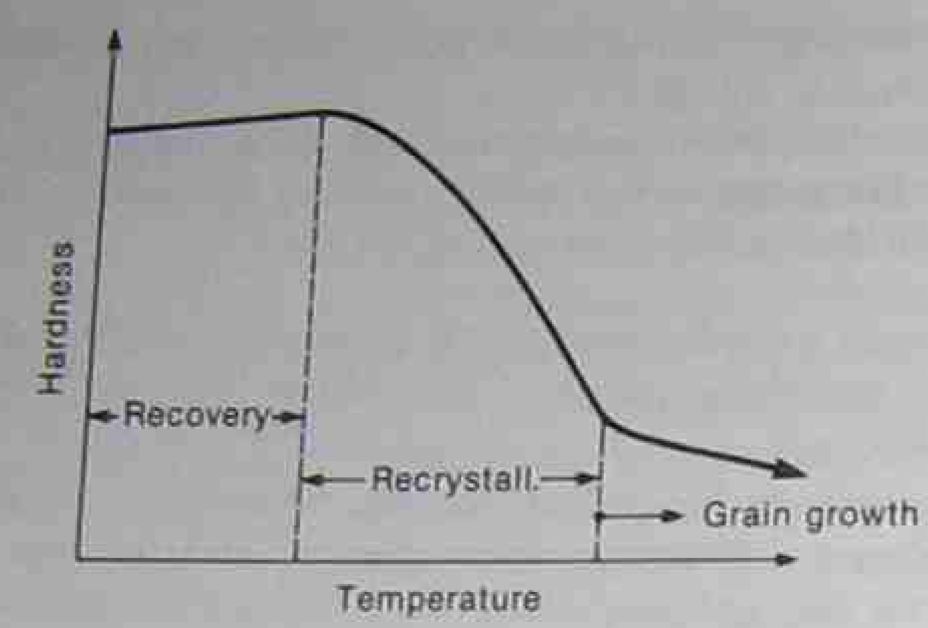


Figure 7–22 Sketch of the change in hardness of an initially cold-worked single phase metal that is heated to various temperature levels; time at temperature is constant here.

Again we stress that this coverage pretends no attempt of completeness; in fact, the numerous aspects involved with altering such structures are in many ways far more complicated than was the case with metals. Consequently, detailed coverage is best left to specialized books in the field of polymer science or polymer chemistry.*

7.5.2 Structure of Polymers

Carbon has been called the backbone of organic materials, and is the element of greatest single importance in synthetic polymers. Because of its chemical valency, carbon can provide anywhere from one to three bonds with other atoms, thereby forming molecules. As the length of the molecules increases, the *total* bonding force that holds the molecules together also increases. In most cases, the polymers of major interest to us consist of molecules that contain carbon plus elements such as hydrogen, nitrogen, fluorine, oxygen, sulfur, and chlorine. This rather large number of possible combinations leads to the variety of properties displayed by such polymers.

The atoms which comprise the molecules that constitute the backbone chain are held together by covalent bonding; this is the primary bond. It is a useful analogy to view a chain of such molecules as resembling a string of beads. Secondary bonds, such as Van der Waals forces, hold adjacent strings together.

It is important to realize that the molecular chains usually occur in a random orientation and the *packing* of atoms is nowhere near as close as in metals. As a result, the overall bonding forces are lower; thus, polymers are, on the whole, not as strong as metals nor as dense. When the backbone chains are held together only by secondary bonds, this is called a *linear* polymer, but this does not mean that all chains are parallel to each other. A loose analogy is to view the chains as the individual strands in a plate of spaghetti; in such a condition they are quite flexible to alter their positions under loading. There are several actions that can be taken to alter the properties of linear polymers. These include:

* An excellent *introduction* to this topic is the text by G. M. Moore and D. E. Kline, *Properties and Processing of Polymers for Engineers* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1984).

1. Forming a *copolymer*. Here the backbone chain contains more than one type of repeating mer.
2. *Crosslinking*, wherein covalent bonding between chains is used. This leads to an increase in strength and rigidity.
3. The addition of *plasticizers*. These are solvents that act between the chains and tend to make the end product softer. Note that, depending upon the relative *amount* of such additives, end properties can be quite different.
4. The addition of *fibers* that do not dissolve in the polymer. Such materials are usually called either reinforced plastics or, preferably, *composites*, and they display strength and stiffness much higher than the polymer itself.

7.5.3 Manner of Classification

7.5.3.1 Thermoplastic and Thermoset Of the several ways to categorize polymers as to groups, the use of *thermoplastics* and *thermosets* is an important one and is distinguished chemically by the degree of crosslinking. *Thermoplastics* utilize secondary bonds to hold the backbone chains together; upon being heated, such bonds are easily broken, and the polymer becomes quite pliable. It can then be formed to a desired shape and cooled to retain that shape. Upon reheating, this process can be repeated time after time. Note, however, that although the desired shape can be produced quite easily, if it is subjected to even modest temperature levels, a shape reversion can occur. We also note that thermoplastics can be recycled in a manner somewhat like that used with aluminum cans. With *thermosets*, the backbone chains are crosslinked in their final form. Upon initial heating and shaping to form, the crosslinks occur; upon cooling to a final shape, the end result is, essentially, one giant molecule. If reheating takes place, there can be no reshaping; instead, such a polymer simply chars and degrades. These polymers are more rigid and stronger than the thermoplastics but are incapable of recycling. In essence the heating, shaping, and cooling of thermoplastics is reversible, since the process can be repeated, whereas the heating, forming, and cooling of a thermoset is irreversible. Commonly used polymers such as polyethylene (PE), polyvinyl chloride (PVC), polymethyl methacrylate (PMMA), and polycarbonate (PC) are thermoplastics, while epoxies and phenolics are thermosets.

7.5.3.2 Amorphous and Crystalline No polymer has yet been produced commercially with a structure that is fully *crystalline* where the ordered repetition of the basic building block is as complete as in metals. Nevertheless, in a general sense, the simpler the mer, the greater the possibility of obtaining some crystallinity. The rest of the structure is *amorphous*, where the atomic arrangement is random in nature. The highest degree of crystallinity (about 90 percent) attainable with any of the commercially used polymers occurs in high-density polyethylene (HDPE); the remainder of the structure is amorphous. Polymers such as polycarbonate (PC) and polymethylmethacrylate (PMMA) are completely amorphous, while others can be produced with varying degrees of crystallinity. Various models* have been proposed to explain how the large molecules form crystals, but we shall not be concerned with such detail. It is sufficient to state that in the amorphous structure the atoms are less densely packed than in the crystalline structure. Let's consider a polymer that is *capable* of being produced in at least a partially crystalline form. Reference to Fig. 7–23 will assist here, where T_m refers to the melting temperature. If the cooling rate at T_m is *slow*, the tendency for crystallization is enhanced, tight packing of atoms occurs, and the specific volume decreases abruptly. Further cooling to room temperature shows a further drop in the specific volume. If, however, the rate of cooling at T_m is much faster, the time to produce crystallization is inadequate, and the structure that results is basically amorphous. Owing to the looser packing of atoms, the drop in specific volume is not as great compared with the crystalline structure. Further cooling shows a break in the curve, after which the decrease in specific volume occurs at a lesser rate. The temperature at this break is called the *glass transition temperature* T_g and has important consequences. Above T_g, the amorphous structure is usually called *rubbery*, since it is pliable and lacks any degree of stiffness. Below T_g the structure is much stiffer and is called *glassy*, probably because glass, which is amorphous, behaves in a similar manner. The strength and elastic modulus of such solids are greater when below rather than above T_g. Figure 7–24 illustrates the variation of modulus with temperature.

Another important property is the damping of capacity. This can be explained with the use of a rubber ball, where we desire the ball to bounce after it is dropped. Initially, the ball possesses potential energy; as it drops and strikes a surface such as a floor, some energy will be absorbed by the ball and the remainder will cause the ball to bounce. If the ball has a large damping capacity, the rebound height will be small, whereas the reverse occurs if the damping capacity is low. A schematic of damping capacity versus temperature is shown in Fig. 7–25; we note that at or near T_g the damping capacity is greatest. Returning to the rubber ball, we address its purpose. Assuming a child is to use the ball, we would want a low damping capacity (that is, high bounce) and a soft (rubbery) rather than a hard (glassy) structure. From Fig. 7–25, a polymer (or rubber) whose T_g is well

* See, for example, G. M. Moore and D. Kline, *Properties and Processing of Polymers for Engineers.*

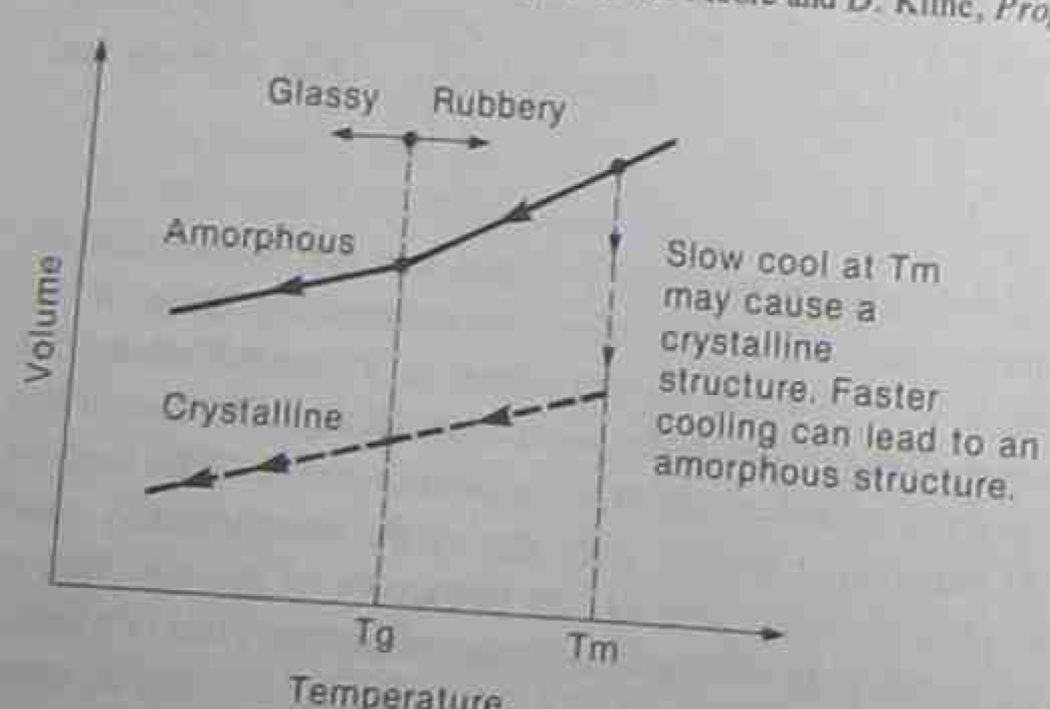


Figure 7–23 Formation of an amorphous or crystalline structure of a polymer as influenced by the cooling rate at T_m. Note that the atoms would be packed more tightly in the crystalline form.

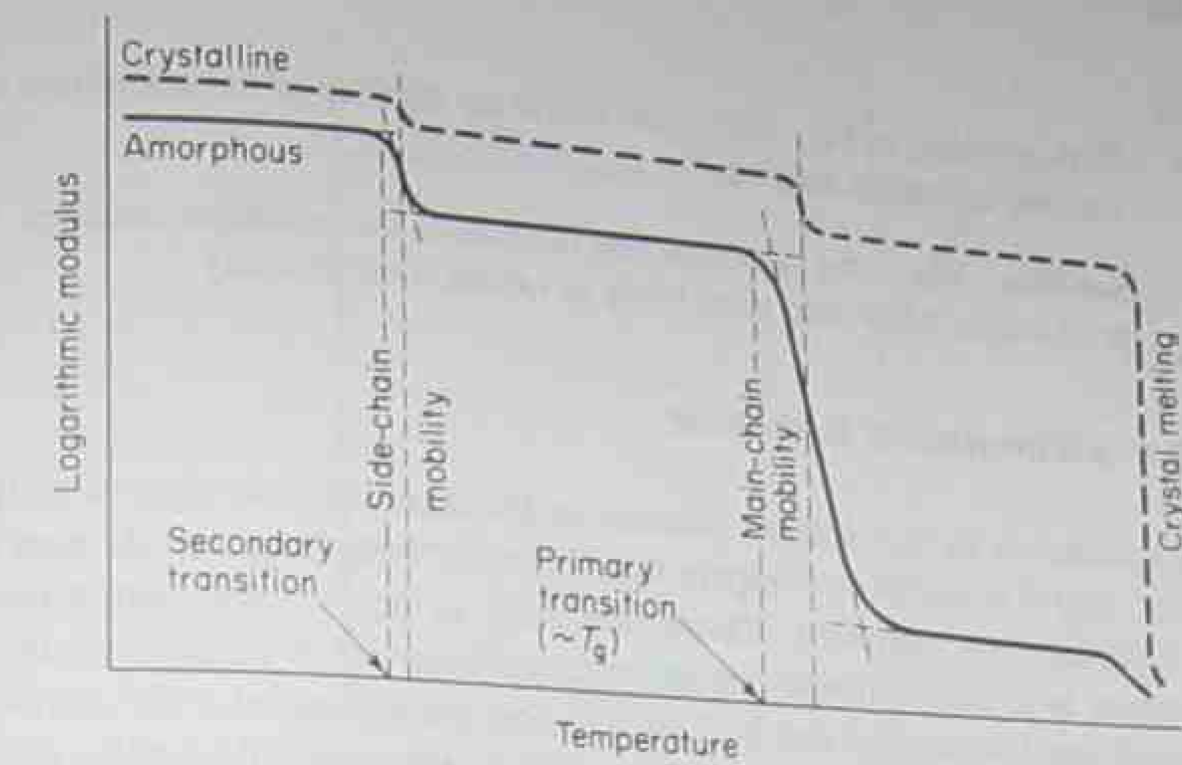


Figure 7–24 Schematic showing the influence of temperature on the modulus of a thermoplastic.

below the temperature of application, say, room temperature, would satisfy both requirements. This simple example illustrates a few reasons why T_g of polymers must be given consideration before an individual polymer is specified for an application.

Two points are worth stressing. First, T_g and its effect on certain properties is applicable, only to the amorphous portion of the structure. Second, care must be exercised in comparing the influence of the degree of crystallinity on properties. One should not conclude that polymers of high crystallinity such as HDPE are stiffer (that is, higher modulus) than polymers that are amorphous. The modulus of HDPE (highly crystalline) is decidedly lower than PC (completely amorphous). A correct conclusion results if the *same* polymer, having different percentages of crystallinity, is considered. For example, high- versus low-density PE (90 versus, say, 60 percent crystallinity) shows a higher tensile strength and elastic modulus for the HDPE.

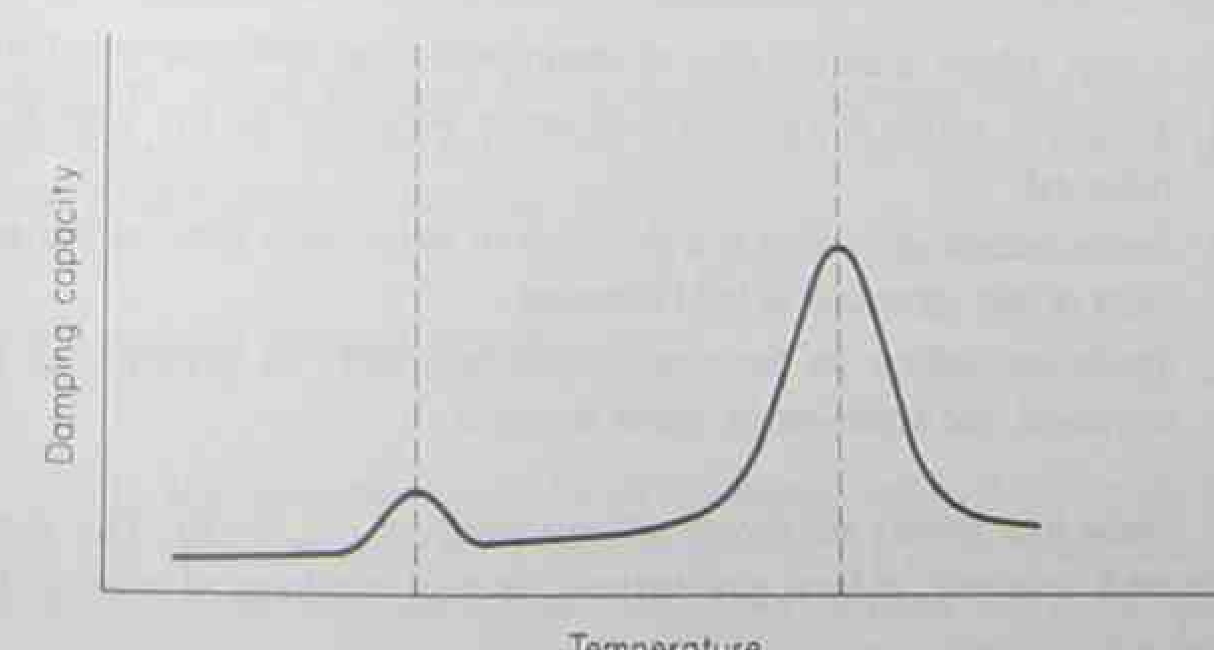


Figure 7–25 Schematic showing the influence of temperature on the damping capacity of a thermoplastic.

Example 7–5

With reference to Fig. 7–23, how would the ductility of the amorphous structure compare with the crystalline structure?

Solution Due to the tighter packing of atoms in the crystalline condition, that structure will be stronger, stiffer, and more likely to exhibit lower ductility.

7.5.4 Viscoelastic Behavior

The discussion in this section pertains to the *thermoplastic* type of polymers discussed earlier, and it is useful to compare their behavior with those metals most commonly used for load-carrying purposes. Parameters such as temperature, strain-rate, and time are important as to their influence on macroscopic behavior of metals and polymers. Although we shall not discuss why and how such factors are influential at the atomic or microscopic level, it is noted that similar behavior, at least phenomenologically, results for different reasons, since the atomic structure of these classes of solids is so different. For example, the phenomenon of *creep,* which is an increase in strain with time under constant load or stress, can occur with either type of solid. Although the observed macroscopic behavior appears similar, the fundamental *causes,* in terms of structural arguments, are quite different. To discuss them is beyond the intent of this book.

Consider a bar of steel subjected to a tensile load which causes elastic deformation at room temperature. Regardless of how much time elapses, any additional deformation would be minimal. If under similar conditions, a specimen of polyethylene had been used, we would note a continual extension as time passes; that is, creep would occur. What this really means is that under *equivalent* conditions, polymers can display a behavior not as readily exhibited by metals. This *does not* mean that metals don't creep, but to cause such an occurrence, both the temperature and magnitude of stress must be increased to much higher levels than those needed to cause such behavior with polymers.

At least four discrete types of behavior enter here. These are:

1. *Creep,* which is an *increase* of *strain* with time under constant load or stress.*
2. *Recovery,* which is a *decrease* of *strain* with time as the load is reduced or fully removed.
3. *Stress relaxation,* which is a *decrease* of stress with time as the strain or displacement of the specimen is held constant.
4. *Strain-rate effects* on the stress-strain behavior. In general, as the strain rate is increased, the stress-strain curve is raised.

These four results are shown schematically in Fig. 7–26. The simplest way to describe such behavior, at least qualitatively, is with mechanical analogs involving springs

* Creep tests are usually run under constant load for simplicity. To maintain a constant *true* stress would require a decreasing load, since the area of the test specimen decreases as elongation occurs. If nominal stress is used, then this stress is constant with constant load.

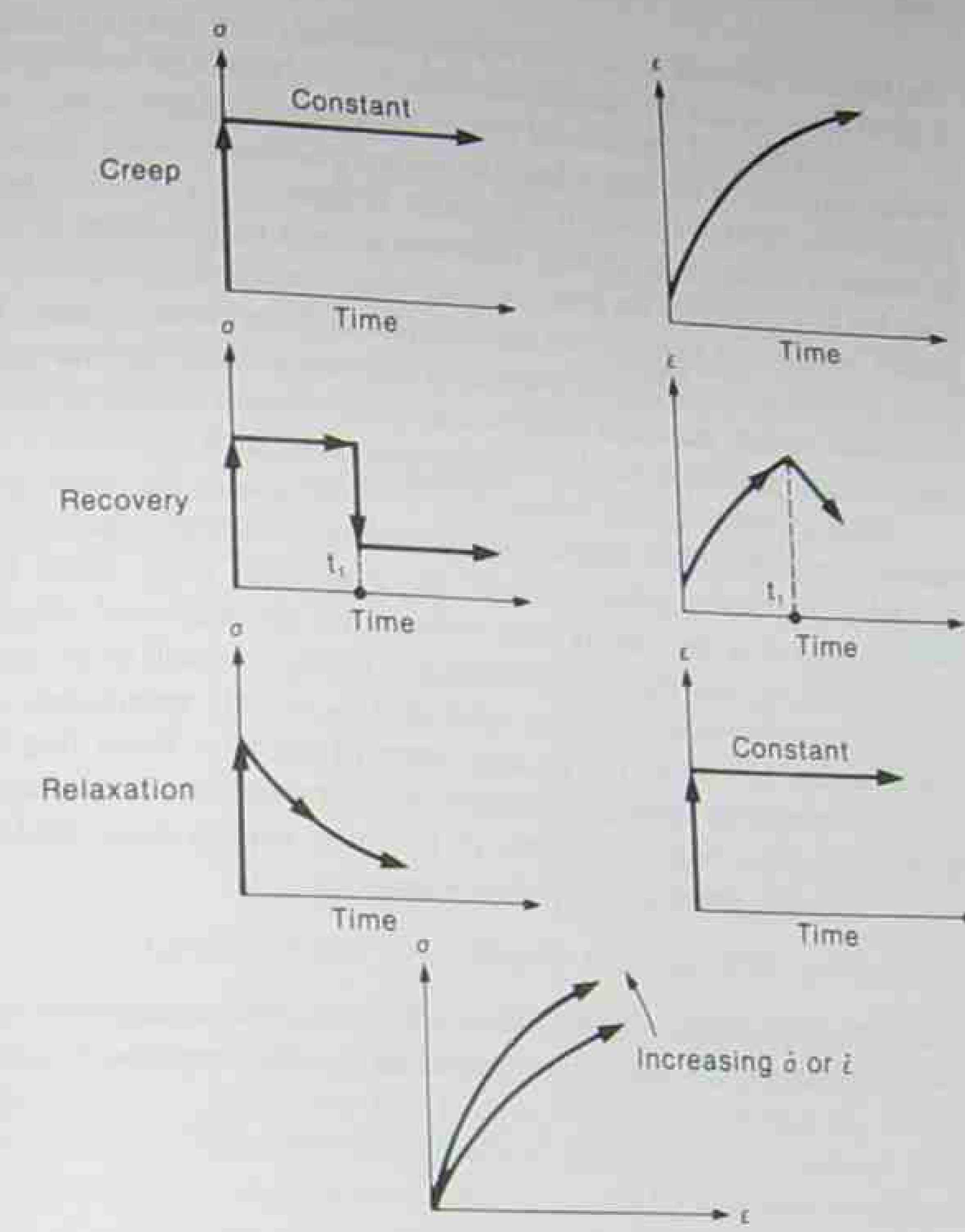


Figure 7–26 Various effects of time-dependent behavior of stress and strain.

and dashpots. These can then be expressed by equations which may then be used for predictive purposes. Shown in Fig. 7–27 are a single spring and dashpot. Under stress, the behavior of the spring is given by

$$\sigma = E\epsilon \tag{7–1}$$

while the dashpot behaves according to

$$\sigma = \eta\dot{\epsilon} = \eta\, d\epsilon/dt \tag{7–2}$$

Here both E, the stiffness or modulus of the spring, and η, the viscosity coefficient of the dashpot, are assumed constant or time independent; $\dot{\epsilon}$ is the strain rate. Three points are worth noting. First, both equations are linear in the sense that E and η are not functions of time. Second, over a broad range of temperature, a single value for E in Eq. (7–1) would not be adequately descriptive, nor would a single value of η. Finally, both E and η *are* rate sensitive with most real polymers. If one attempts to account for this, nonlinear equations result with accompanying complexities. Yet, ignoring these points still permits a good physical feel for viscoelastic behavior, and that is the main point in this discussion.

Equation (7–1) implies the type of instantaneous elastic response observed with most solids; removal of the stress leads to recovery of elastic strain. Equation (7–2) implies a linear creep behavior since, if integrated, it shows that $\epsilon = \sigma t/\eta$, but this viscous strain is nonrecoverable. Individually, these components explain some of the behavior of polymers but by themselves are quite limited. When combined in series they are called a *Maxwell* model, whereas in parallel, the name *Kelvin* or *Voigt* model is used; Fig. 7–28 illustrates these. Even these models are limited as to their prediction of the four major types of behavior mentioned earlier, but when combined as a *four-element* model, as in Fig. 7–29, their combined effects provide a much better description of such behavior.

First consider creep, where Fig. 7–30 illustrates the most general type of such behavior seen. An initial elastic response occurs immediately as the load is applied and is accounted for by the spring in series whose modulus is E_1. Next, the *increasing* strain which occurs at a *decreasing rate* occurs from the *combined* effects of the dashpot in series (that is, η_1) and the Kelvin branch (E_2 and η_2). As will be shown shortly, the Kelvin strain will tend to saturate to the value of σ/E_2 as time approaches infinity (in a mathematical sense) or essentially at long times in practice. From that time on, the single dashpot continues to strain at a constant rate. We emphasize here that *nothing* is predicted regarding *fracture* with this or any of the other models mentioned here. Although not derived here, the strain of the Kelvin branch* is

$$\epsilon = \sigma/E_2\,[1 - \exp(-tE_2/\eta_2] \tag{7–3}$$

Because Eqs. (7–1) through (7–3) are all linear, the principle of superposition can be applied; in essence the individual strains can be added together to give the total strain as

$$\epsilon_{\text{total}} = \epsilon_{\text{elastic}} + \epsilon_{\text{viscous}} + \epsilon_{\text{Kelvin}} \tag{7–4}$$

or

$$\epsilon_t = \sigma/E_1 + \sigma t/\eta_1 + \sigma/E_2[1 - \exp(-tE_2/\eta_2] \tag{7–5}$$

which in equation form basically describes Fig. 7–30 up to the tertiary region.

* See R. M. Caddell, *Deformation and Fracture of Solids*, Chapter 6 for a more detailed derivation of the various equations quoted in this section.

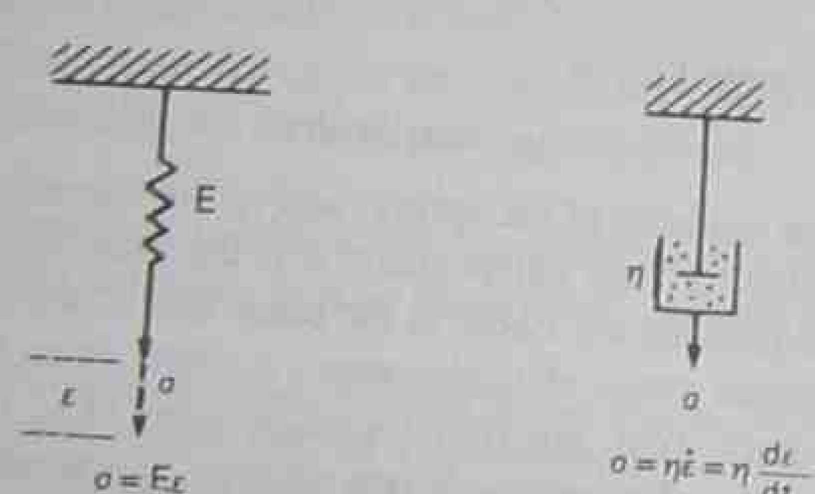


Figure 7–27 Schematic of a spring and a linear dashpot.

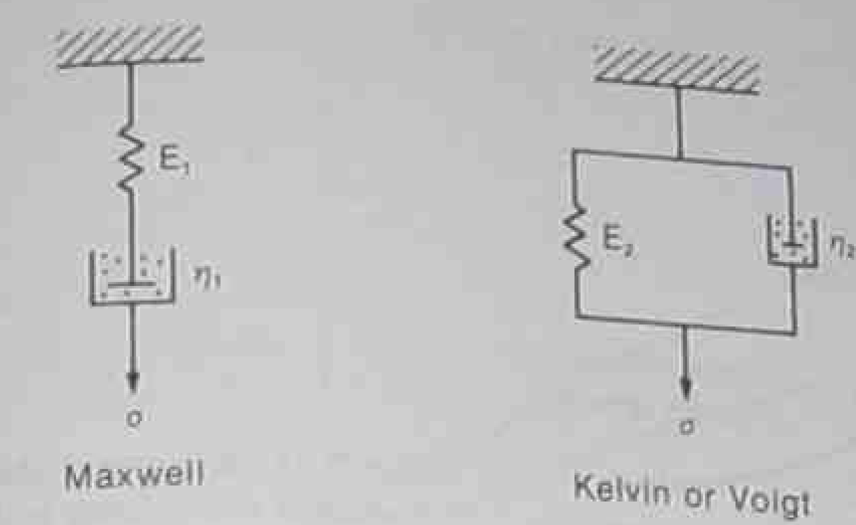


Figure 7–28 Schematic of a Maxwell (left) and a Kelvin (right) model.

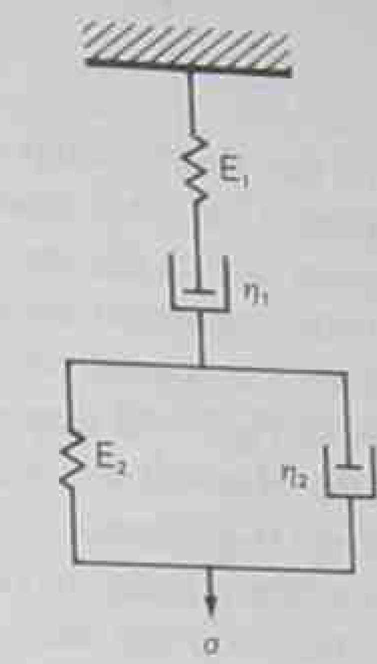


Figure 7–29 A four-element model.

Now consider recovery. Removal of the stress σ would cause an immediate elastic recovery of σ/E_1 followed by a time-dependent recovery of the Kelvin branch, which is exponential in form. Physically this occurs because the spring E_2 would tend to be restored to its initial, unstressed condition and would force the dashpot η_2 to return with it. Any strain due to the individual dashpot η_1 would be a nonrecoverable viscous strain. Figure 7–31 illustrates this behavior and is in qualitative agreement with actual behavior.

As to stress relaxation it is simplest to consider the behavior of the Maxwell and Kelvin parts individually. When the spring E_1 and dashpot η_1 are *strained* to some fixed level (which is then held *constant*), most of the strain must be accommodated by the spring, which is time-independent. As time passes, the dashpot must undergo strain (that is, $\sigma t/\eta_1$) and since the total strain is fixed, this can happen only if the spring contracts.

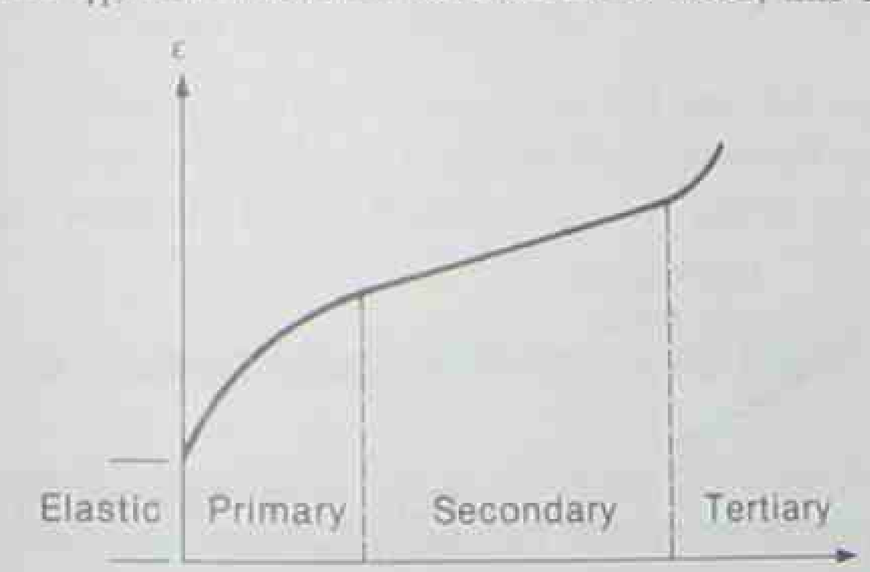


Figure 7–30 A generalized creep curve.

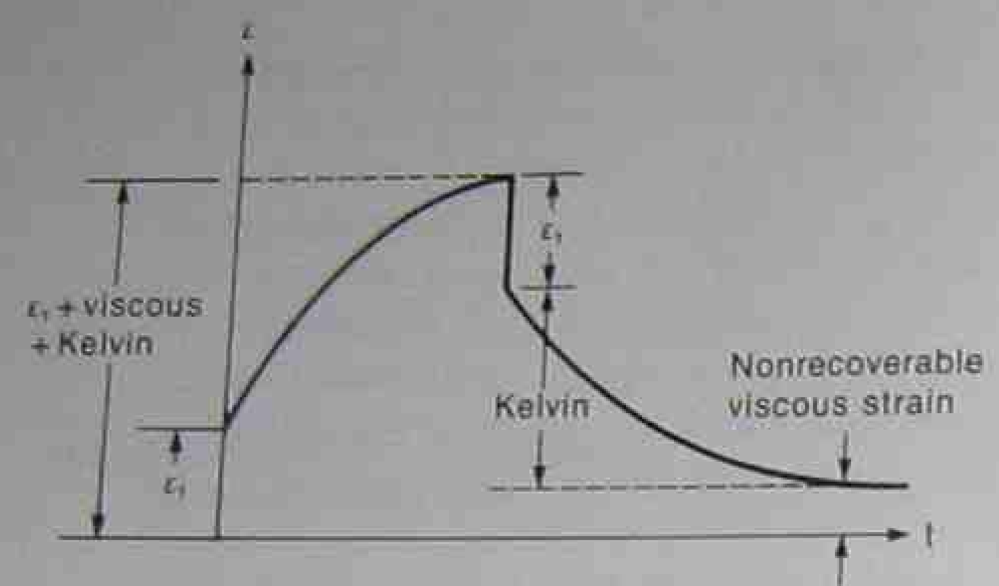


Figure 7–31 Strain-time behavior of a four-element model during loading and then unloading.

This decrease in the strain of the elastic component can result only if the stress σ decreases, hence stress relaxation. Using this same analysis of the Kelvin portion, any strain initially induced is common to both the spring E_2 and dashpot η_2 so there can be no relaxation with time, since both components are locked together. The conclusion is that any relaxation displayed by the four-component model must result because of the behavior predicted by the Maxwell portion; Fig. 7–32 displays this schematically.

As to rate effects, both the Maxwell and Kelvin models predict a raising of the stress-strain curve as the strain rate is increased, although the shapes of the curves differ. When these are combined into a four-element model, the combined effect is more realistic. Figure 7–33 illustrates this, and the interested reader can pursue the details elsewhere.*

The intent of these few pages is to provide a physical understanding of particular aspects of the behavior of certain solids and to impart the idea that for equivalent conditions of temperature, stress, strain rate, and time, polymers, as a *class* of solids, are far more susceptible to such time-dependent or viscoelastic behavior than are metals. It really comes back to the fact that the structure, as dictated by the bonding forces, is so different for these solids.

* R. M. Caddell, *Deformation and Fracture of Solids*, Chapter 6.

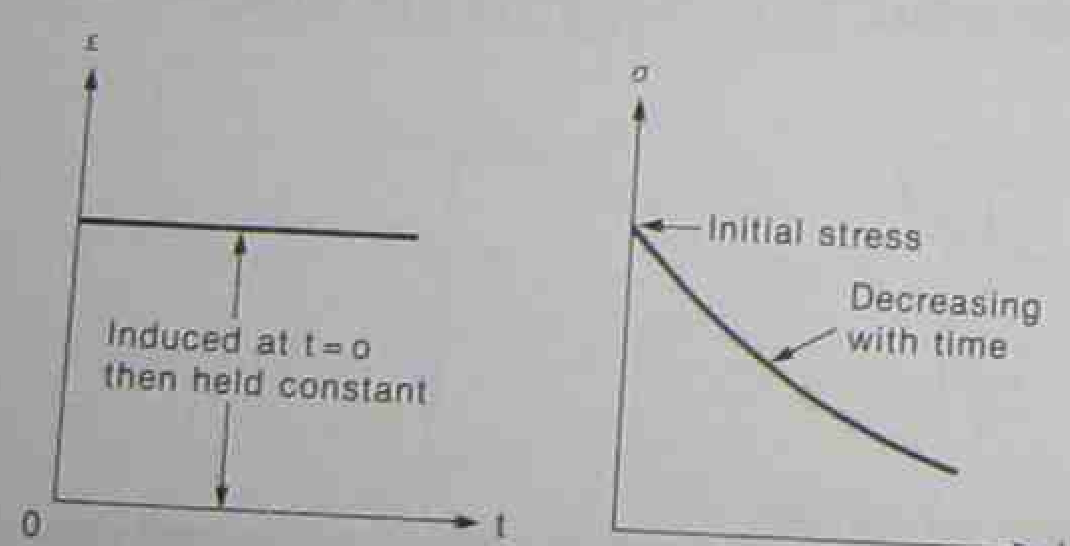
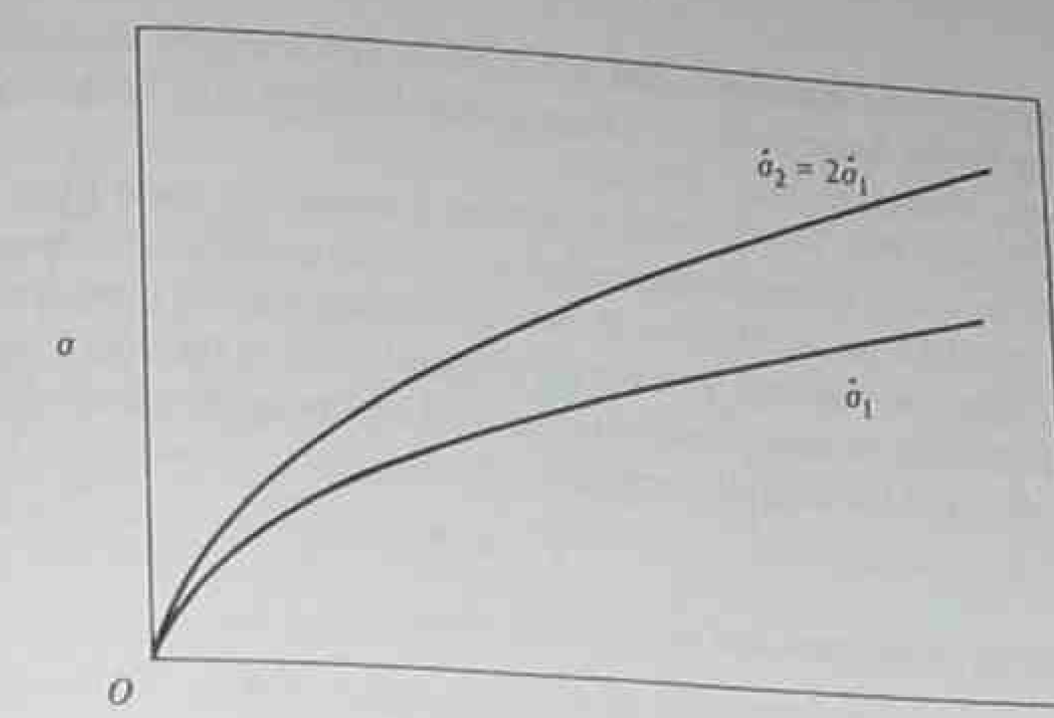


Figure 7–32 Illustration of stress relaxation.



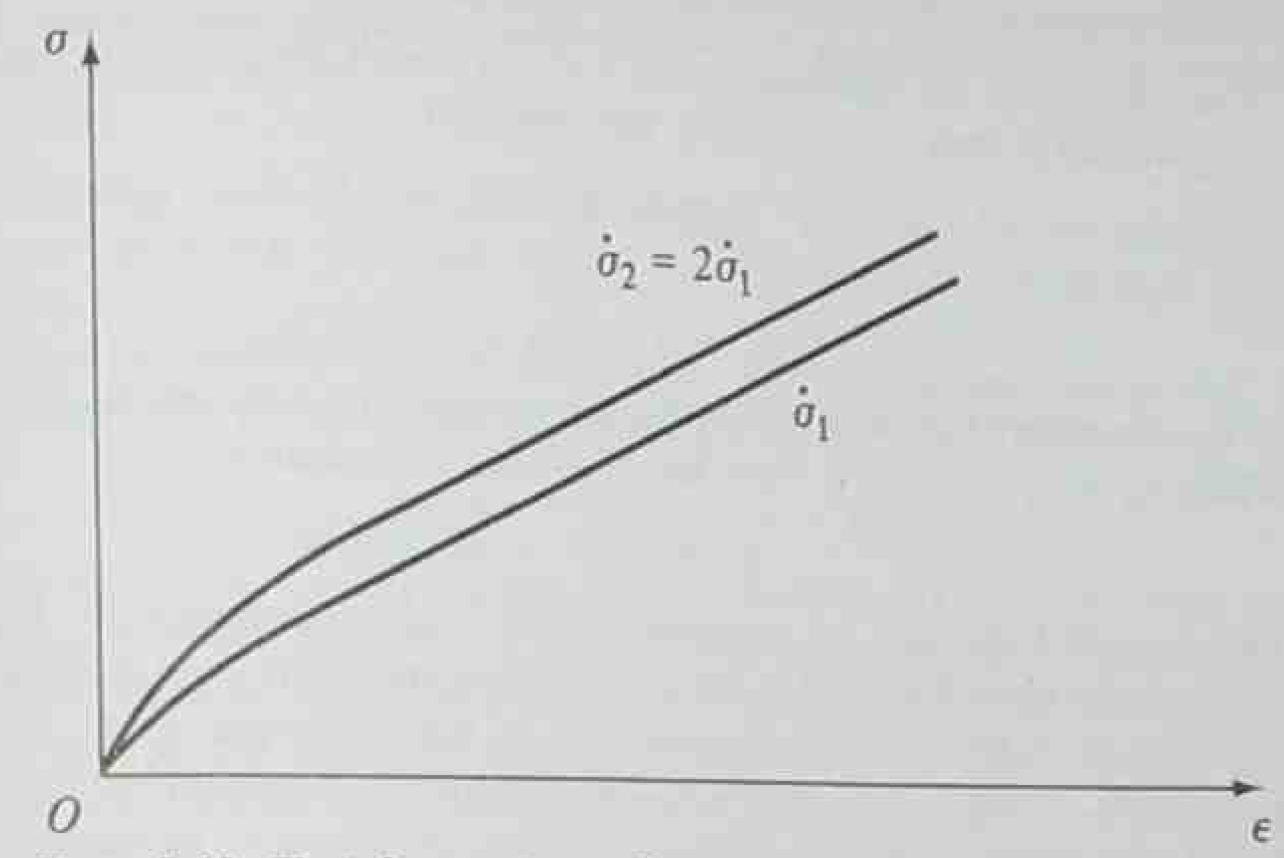


Figure 7–33 The influence of rate effects on the resulting stress-strain behavior as predicted from a Maxwell and Kelvin model.

Example 7–6

Explain why a Kelvin model by itself cannot adequately describe the long time creep behavior as shown in Fig. 7–30.

Solution Refer to Eq. (7–3) and note that as t approaches infinity (or at large values of time), the value of ϵ approaches or saturates to σ/E. Thus continued linear creep cannot be predicted by this model.

7.5.5 Composites or Reinforced Plastics

Natural composites such as wood are not considered here; rather, our attention is directed towards *man-made* composites. For our purposes, a composite is defined as a mixture of *fibers* (also called filaments) that are bonded with a *matrix* material such that a distinct interface results between the two ingredients. Matrices may be thermoset or thermoplastic polymers, metals, or even ceramics.

Consider Fig. 7–34, which illustrates a composite made from *aligned, continuous,* and *unidirectional* filaments embedded in a matrix material. It is assumed that the loading is to be parallel with the fibers as shown, and during loading there is *no slip* at the interface between fibers and matrix. This means that any strain in the fibers, matrix, and composite is equal. A force balance, involving the load supported by *all* of the fibers, P_f, all of the matrix, P_m, and the overall composite, P_c, is

$$P_c = P_m + P_f \tag{7–6}$$

and in terms of stresses and appropriate areas,

$$\sigma_c A_c = \sigma_m A_m + \sigma_f A_f \tag{7–7}$$

Since the fibers are continuous, for a given length, the areas in Eq. (7–7) must be proportional to their corresponding volumes; then

$$\sigma_c V_c = \sigma_m V_m + \sigma_f V_f \tag{7–8}$$

or

$$\sigma_c = \sigma_m (V_m/V_c) + \sigma_f (V_f/V_c) \tag{7–9}$$

If V_c is considered as unity, the *fractional* volumes of matrix and filaments are v_m and v_f, where $v_m + v_f = v_c = 1$. Then Eq. (7–9) can be written as

$$\sigma_c = \sigma_m(1 - v_f) + \sigma_f(v_f) \tag{7–10}$$

This is called the *rule of mixtures* (ROM) and is described by Fig. 7–35. This is similar to the ideas discussed in Sec. 7.4.1. In essence, if σ_m and σ_f are known, a particular value of v_f then defines the strength of the composite σ_c *in the loading direction* shown in Fig. 7–34. If such a composite were loaded at right angles to that shown, the strength of the composite would be dictated primarily by the strength

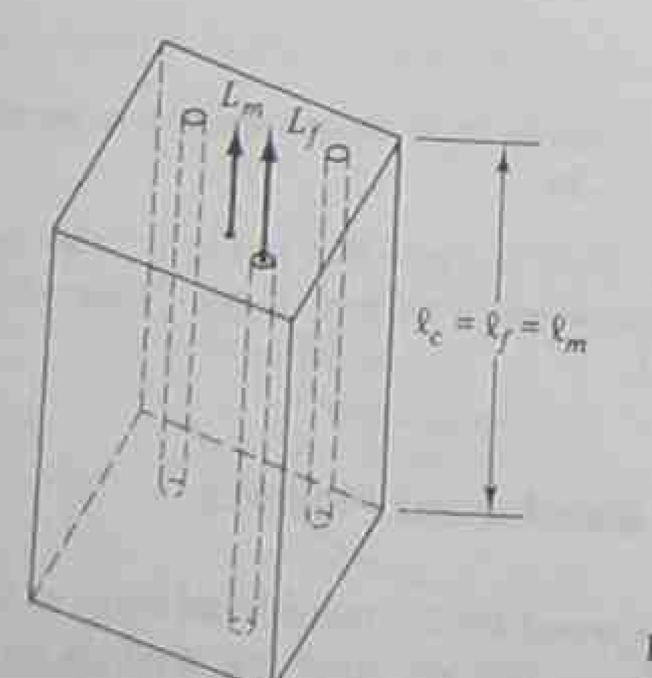


Figure 7–34 Section of a composite made with parallel and continuous fibers.

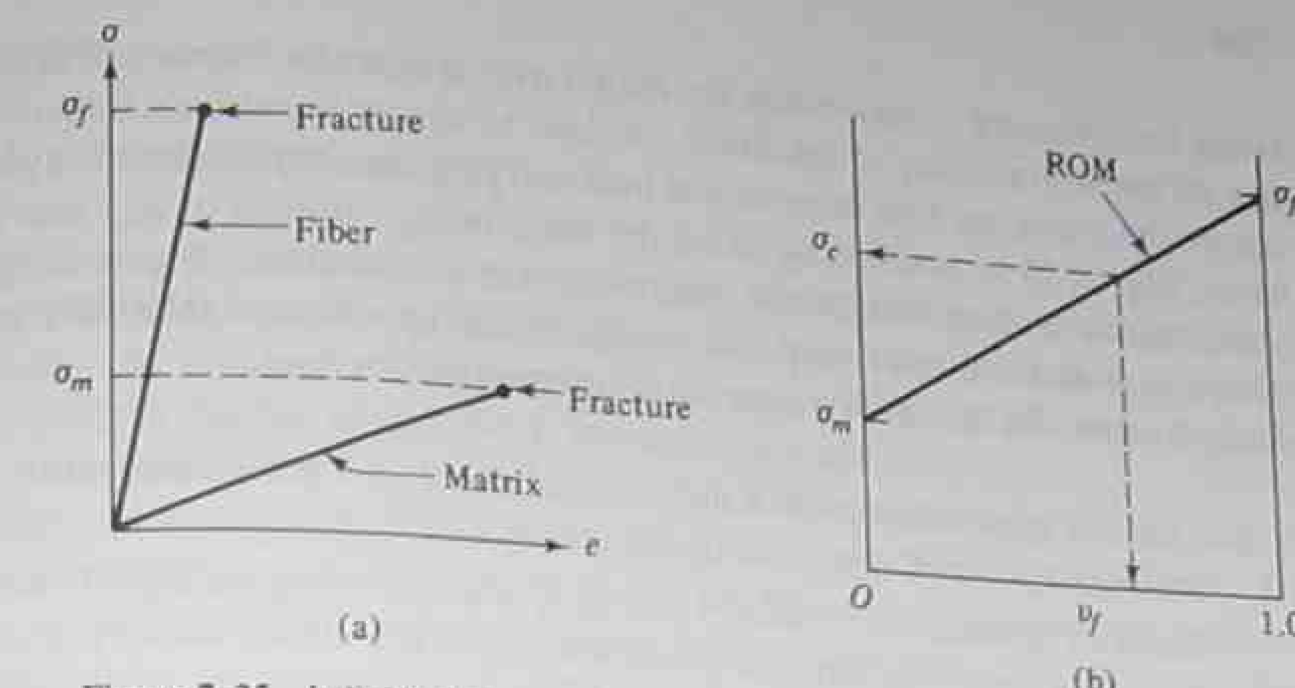


Figure 7–35 Individual stress-strain behavior of matrix and fiber materials (a) and the rule-of-mixtures approach for calculating composite strength as a function of volume fraction of fibers (b).

of the bonding *between* the two components. It would be, in general, considerably *lower* than that predicted by Eq. (7–10), indicating a decided condition of anisotropy.

Since the induced strain is common for the loading described by Fig. 7–34, the elastic *modulus* of the composite is, from Eq. (7–10),

$$E_c = E_m(1 - v_f) + E_f v_f \tag{7–11}$$

where each σ in Eq. (7–10) is given, in general terms, by $\sigma = E\epsilon$. Many experiments show that predictions from Eq. (7–10) overestimate the strength of such composites. To analyze *one* possible reason for these discrepancies, consider the *individual* stress-strain behavior of a matrix and fiber material as shown in Fig. 7–36. Because the fibers are almost always stronger and more rigid (that is, higher modulus) than the common matrix materials, the relative position of the $\sigma - \epsilon$ plots results. Under loading, the fibers will

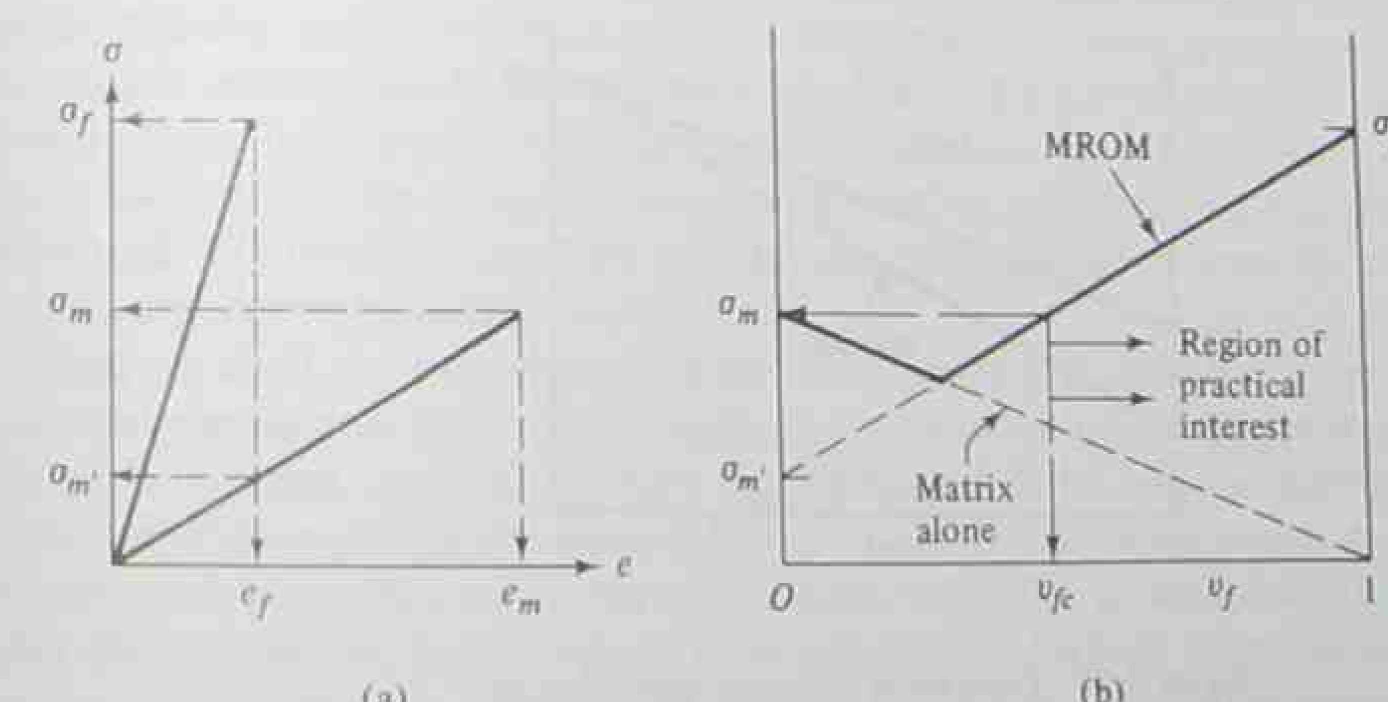


Figure 7–36 Same as Fig. 7–35 except the modified rule-of-mixtures is used.

reach their *fracture strains* before the matrix ever reaches its *fracture stress*. In many, but not all cases, fracturing of the fibers can lead to almost immediate fracture of the entire composite, since we lose some of the load-carrying capacity of the stronger component. Thus, Fig. 7–35 is modified by using the value of $\sigma_{m'}$, instead of σ_m, that is, the matrix stress which corresponds to the fracture strain of the fiber. Figure 7–37 shows this comparison and indicates why this *modified rule of mixtures* (MROM) predicts lower strengths than the ROM. We note two important points here.

1. The two approaches predict different values of σ_c for the same value of v_f, but the stiffness E_c is still governed by Eq. (7–11).
2. Unless a minimum or critical value of v_f is used (see v_{fc} on Fig. 7–36), the MROM predicts a composite strength that is *lower* than the matrix strength σ_m. In a practical vein, there is no sense in adding costly fibers unless we get a value of σ_c that reasonably exceeds σ_m so, as shown in Fig. 7–36, the practical values of v_f are indicated.

Example 7–7

Suppose you are told that the fracture strengths of a matrix and fiber material are 35 and 300 MPa, respectively, and you are to design a composite having parallel and continuous fibers acting parallel to the applied load. The strength of the composite is to be 120 MPa. What volume fraction of fibers would you specify?

Solution Since nothing is known about the fracture strains of the two components, the rule of mixtures must be used, that is, Eq. (7–10). So

$$120 = 35(1 - v_f) + 300(v_f) \quad \text{or} \quad v_f = 0.32 \text{ or about 32 percent}$$

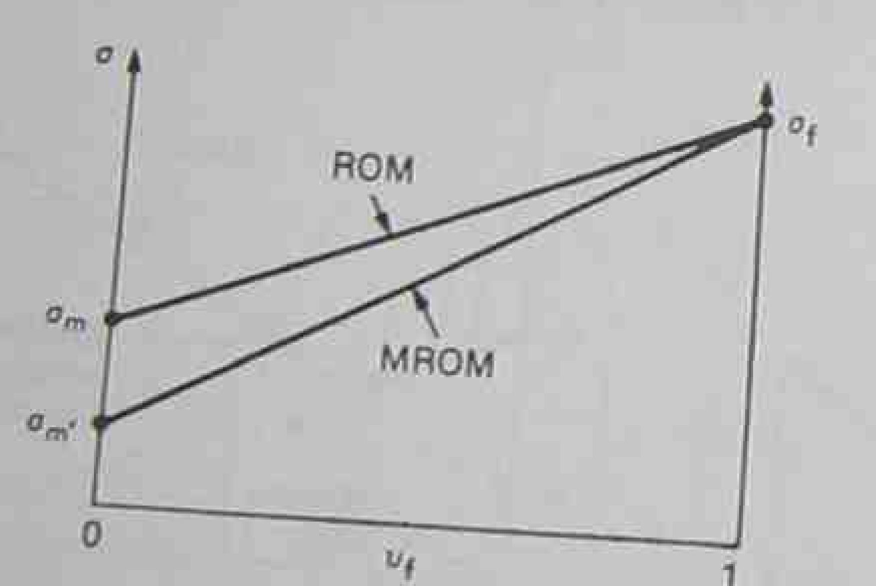


Figure 7–37 Comparison of the two approaches for predicting composite strength as a function of volume fraction of fibers.

7.6 CERAMICS

7.6.1 Introduction

In a general sense, ceramics may be defined as compounds that contain both metallic and nonmetallic elements, even though such a distinction is not as sharply defined as one might expect. However, distinguishing elements as being metallic or nonmetallic is useful, especially for the nonspecialist in materials science.

Just as we noted with polymers, there are different ways by which ceramics can be classified without going into great detail regarding structural configurations. *Traditional* ceramics would include items such as bricks, fine china, and pottery, whereas *modern* ceramics include those used as refractory solids in furnaces, catalytic converters in cars, components in jet engines, and cutting tools used in machining operations. In the discussion of the iron-carbon diagram, the phase called cementite, Fe_3C, is a ceramic even though it wasn't pointed out at that time. Other compounds such as titanium carbide and tungsten carbide are also examples of fairly simple ceramic compounds.

Regarding the primary bonding forces involved with ceramic structures, these may involve both the *ionic* and *covalent* types. As with metals, the unit cell can be viewed as the basic building block and the importance of microstructure, as it affects properties, is as crucial with ceramics as with the other materials discussed in previous sections. Although we shall not include any equilibrium diagrams here, it is noted that they are as important with ceramic materials as with metals in regard to phase transformations.

7.6.2 General Comparisons with Other Materials

As a class of materials, ceramics of certain types have been in use longer than either metals or the type of polymers that are man-made; brick and glass typify ceramics that have been used for centuries. Yet it is reasonable to state that most engineers who are not materials scientists probably possess a better understanding of metal structure and properties because ceramics are, in general, more complicated. For our purposes it is adequate to present a broad overview of comparative properties of these various classes of materials and to stress that it is the design engineer, more than the manufacturing engineer, who must contend with the selection of the most advantageous material in terms of functional as well as environmental requirements.

First, we shall compare ceramics with metals, bearing in mind that exceptions to such general comparisons can exist. If metals are considered as a class, they possess better electrical and thermal conductivity, have much greater ductility, possess greater tensile strength, and display greater resistance to fracture. Ceramics are better insulators, tend to have more stable properties at elevated temperatures, are harder, in most instances display a higher elastic modulus, but have much lower resistance to fracture. Although most bulk metals find widespread use regardless of whether the loading is tensile or compressive, ceramics possess much greater strengths in compression than in tension. Here we note one obvious exception for metals; gray cast iron, like ceramics, is much stronger in compression than it is in tension.*

Compared with polymers, ceramics are stronger and more rigid (greater elastic modulus), and retain their properties at elevated temperatures. Polymers are more ductile and can be formed to a desired shape by techniques that could not be used if ceramics were involved; in addition, polymers have lower densities, which could be important where weight considerations are of concern. Both classes are capable of varying degrees of crystallinity but *on the whole* cannot be considered to be as fully crystalline as are metals. One other concern with ceramics is that they often contain a greater degree of porosity than do either polymers or metals. Although specific processing techniques can reduce such porosity to minimal levels, the tendency for brittle fracture† is still much greater with ceramics than with either polymers or metals.

7.6.3 A Final Wrapup

In this chapter we have followed the philosophy that the manufacturing engineer is not the individual who selects or specifies a particular material for a given component. That task is left to a design engineer. It has been our experience that the great majority of design engineers have a relatively limited background in the field of materials and often select a material on the basis of strength after completing a stress analysis. But in many applications, properties related to corrosion resistance, chemical attack, temperature effects, resistance to crack propagation, and the like are of at least equal importance. Then, it seems to us, the design engineer must consult with those who possess more specialized backgrounds in metallurgy, ceramics, or polymer science for sound advice and help. It is in this same vein that the manufacturing engineer is often placed. Such a person is not expected to be a materials expert, *but* realizing and accepting one's limitations and then seeking proper assistance will usually produce a sensible and economic solution.

We also admit that the coverage of metals in this chapter has been the most extensive, whereas that for ceramics was least. Up to now the understanding of metals has been of longer standing, and their use in manufacturing has been much greater as compared with polymers and ceramics. However, this situation is rapidly changing and, due to weight considerations, for example, many applications today are seeing some type of polymeric material being substituted for metals. Again, the use of ceramics is increasing greatly and it is likely that the manufacturing engineer of the future will have to contend with a broader use of all engineering materials than did his predecessors. The brief coverage in this chapter should provide enough background in the fundamentals so that

* See Chapter 11.

† We have not discussed *fracture toughness* as a property in this text, but we mention here that this is of great importance in resisting crack propagation. Unfortunately, the broad subject of fracture is barely covered, if at all, in most undergraduate curricula today. Thus, though we are aware of this important property, we cannot in any meaningful sense cover it in this text without going into the type of detail that has been deliberately avoided on other topics. For an introduction to the subject of fracture and fracture toughness, the reader can consult D. K. Felbeck and A. G. Atkins, *Strength and Fracture of Engineering Solids*, Chapter 14, or R. M. Caddell, *Deformation and Fracture of Solids*, Chapter 8.

the reader gains a reasonable appreciation and awareness of the three classes of engineering materials.

PROBLEMS

7–1. Suppose you could control the dislocation density of a bulk metal. What advantages would result if the metal contained no dislocations? What would be disadvantageous about producing such a structure?

7–2. A piece of commercially pure copper can be made to contain a fine grain size (many small grains) or to have a much coarser grain size (fewer large grains). In which condition would the yield strength be greater? Explain why.

7–3. Two specimens of AISI 1040 steel are heated to 1600°F (871°C), homogenized, and then cooled to room temperature. One is furnace cooled, whereas the second is air cooled. Based upon the likely microstructure that results, determine the probable tensile strength of each specimen.

7–4. One-half-inch-diameter solid bars of 1040 and 4340 steel are heated to the austenite region for one hour; each is then quenched in water. Determine the maximum BHN you would expect at the surface of each bar.

7–5. Explain why a piece of steel containing a structure of as-quenched martensite might cause problems in service.

7–6. Two pieces of eutectoid, plain-carbon steel are heated to 760°C and 1050°C, respectively. Both are quenched in water. Which would display better hardenability? Explain why. Remember when γ transforms it has a choice to produce P or M.

7–7. Refer to Fig. 7–14. If a thin specimen of that metal were transferred quickly from 1400°F to 1100°F, held at that temperature for 20 seconds, and then quenched in water, what are the likely microconstituents that result after quenching?

7–8. You produce a *surface* hardness of 55 Rockwell C using a bar of 1040 steel. Unfortunately, the hardness at the *center* is about 25 R_c. If it were desired to obtain a hardness of 55 R_c throughout most of the section, what action would you recommend?

7–9. An aluminum alloy is heated to a single-phase (α) region for one hour; it is then quenched in water. The specimen is then heated to 400°C.

- **(a)** Draw a curve indicating how the hardness of the water-quenched specimen would change with time.
- **(b)** At time = 0, and for a few other time intervals, indicate how the microstructure changes by drawing schematics. On each schematic you should label the constituents to distinguish any differences in the structures.

7–10. Explain the procedure you would use to refine the grain size of a piece of pure aluminum whose current grain size is quite large.

7–11. For a certain application you want a polymer to be *glassy* and to display a *low* damping capacity. In service, the polymer will experience a temperature variation between 20°C and 40°C. Three possible choices have glass transition temperatures of −40°C, +30°C, and +100°C, respectively. Which would you select? Explain why.

7–12. Considering the overall general creep behavior displayed by many solids, where does the predicted creep behavior of a Maxwell model fit into the overall spectrum? Repeat this for a Kelvin model.

7–13. Two tensile specimens of polyethylene are individually subjected to a tensile test. With one, the strain rate induced is about 10^{-3}/s whereas the other experiences a strain rate of 10^{-1}/s. A measure of the elastic modulus shows a decidedly higher value with the higher strain rate. Three models are proposed to explain this behavior in a qualitative manner. They are

1. A spring.
2. A spring in series with a Maxwell unit.
3. A dashpot in series with a Kelvin unit.

Of the three models, which would provide best agreement with the test results? Why did you choose that particular model? No numerical calculations are required.

7–14. A glass-reinforced tape, such as that used in wrapping packages, consists of a matrix (much like Scotch tape) whose fracture strength is 35 MPa. The glass filaments possess a fracture strength of 4 GPa. By conducting a tensile test, the reinforced tape displays a strength of 420 MPa. Using the rule of mixtures, what volume fraction v_f of fibers would you expect was used?

7–15. Taking the results of Problem 7–14 one step further, the fibers are circular and have an average diameter of 0.076 mm while the *thickness* of the reinforced tape is 0.152 mm. If the width of the tape is 25 mm, and the fibers are spaced uniformly across that width, calculate the *number* of fibers involved. Note that the fibers must be continuous and parallel to the length of the tape.

7–16. A composite is to be made using a matrix that displays a fracture stress of 20 ksi and strain of 0.100. Fibers to be used show a fracture stress of 300 ksi and strain of 0.005. The σ-ϵ behavior for both of these components is linear to fracture. You want to produce a composite having continuous, parallel filaments where loading will be parallel to the filaments.

(a) Using the *modified rule of mixtures*, what volume fraction of fibers v_f should be used if the composite is to have a strength of 100 ksi?

(b) Determine the elastic modulus of the composite.

8 forming

8.1 INTRODUCTION

All readers, whether they realize it or not, have been exposed to many formed products. Such diverse items as nails, automobile fenders, casings on washing machines, metal cabinets and chairs, and plastic milk bottles have all been subjected to some type of forming operation. The principal objective in any forming process is to change the original shape of a workpiece to some final desired shape by the application of adequate forces in combination with constraints applied by specific tooling. In forming, the original material is *moved about* to obtain the end result. A few operations, such as blanking and punching, involve separation by a shearing action. These are not typical of most forming operations.

Various requirements must be met to achieve successful forming. First, forces of adequate magnitude are essential if an operation is to be effective; therefore, the manner in which forces are applied and reasonable predictions of the magnitudes of necessary forces are important in understanding forming operations, *unless* one is content with a purely descriptive coverage of hardware. Second, any individual forming process is directly dependent upon the tooling used, since that governs the final shape of the semi-finished or finished workpiece. Finally, the solid being deformed must possess the ability to move or flow as forces are applied; that is, it must have adequate ductility. These interactions are considered in much of this chapter.

The field of forming is currently dominated by the manipulation of ductile metals, and much of this chapter deals with that topic. Even so, some discussion on the forming of polymers is presented. Brittle metals (for example, gray cast iron) and other brittle materials cannot be formed successfully by usual practices, so they are excluded from further concern. Because our initial emphasis covers the *cold forming* of ductile metals, secondary effects, other than the shape change and the forces needed to cause the metal flow, must be considered. These relate to the change in the properties of the metal as it is cold formed.

There is a multitude of different types of forming machines and a variety of ways in which the tooling acts in contact with the surface of the material. For example, the power supply devices used in forging operations are sometimes classified as *path or stroke* restricted, *load or force* restricted, or *energy* restricted. Mechanical presses often employ an eccentric drive to move a loading surface up and down over a fixed length as it contacts, deforms, and then moves away from the workpiece. Here, the length of the stroke is limited to some fixed value. Hydraulic devices are designed to provide some maximum force that can be applied to the workpiece; as such, they are considered to be force restricted (of course, any such individual device will also have some maximum travel). Finally, certain presses employ a flywheel as a source of energy and, depending upon the size of the flywheel, a certain maximum energy is available for deforming the workpiece. To attempt to apply consistently such definitions to the numerous types of machines used in forming would require the kind of excessive descriptive material we seek to avoid. Instead, we shall consider forming *processes* in terms of either of two broad categories, namely *bulk forming* or *sheet forming*. This places emphasis on the shapes to be produced and leads more directly to the kind of coverage we feel is of greater importance compared with a descriptive coverage of forming machines per se.

8.2 BULK FORMING

Operations that fall under this broad category cause shape changes, in general, by the application of compressive loads to the workpiece. Thus most of the deformation results from compressive stresses and large contact between the tooling and workpiece. We now cover a sample of processes that are widely used.

Plates, sheet, and foil result from *flat rolling* in which the initial material is reduced in thickness. Here, the cylindrical rolls are driven by a motor, and the entire unit is called a *rolling mill*. When the width-to-thickness ratio of the original metal is large (at least 10 to 1), the metal exhibits little tendency to widen as it passes through the deformation zone (see Fig. 8–1), owing to the elastic constraint of material on either side of the roll gap. Thus, the decrease in thickness is accompanied by a change in length so as to maintain constant volume. This is called a *plane-strain* operation, since there is almost no movement of metal across the width, so the strain is practically zero in that direction. If the rolls are shaped (that is, not merely parallel) more complex shapes, such as I-beams, rails, and the like can be produced.

Extrusion is the process whereby constrained metal, often called a *billet*, is forced to flow by the application of pressure. If the metal is forced to flow in the same direction as the piston movement, that is called *direct* or *forward* extrusion; see Fig. 8–2. The chamber, which surrounds the billet, induces compressive stresses and as the metal is forced to flow through a converging die, the die also exerts a compressive effect on the

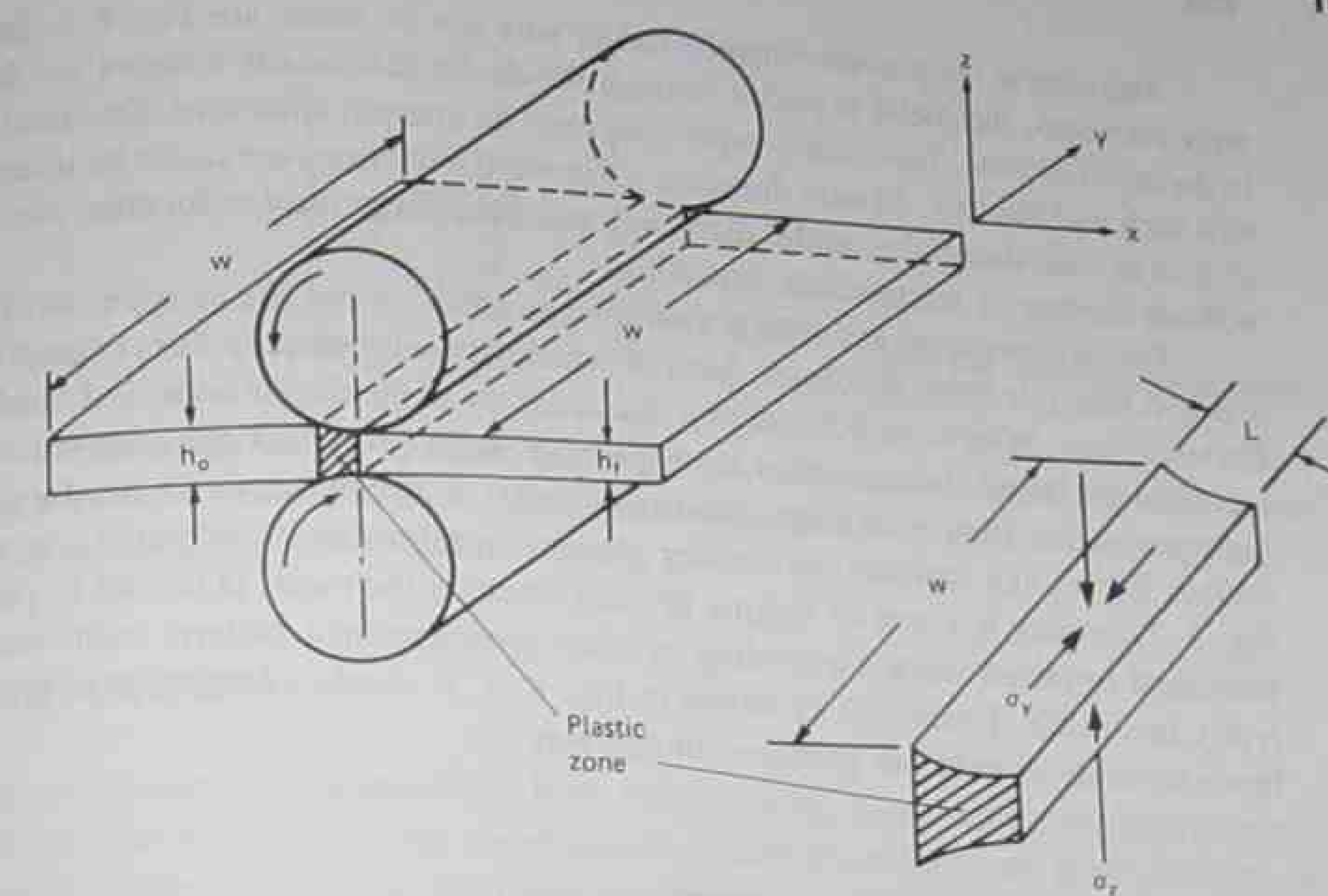


Figure 8–1 Schematic of the deformation zone in flat rolling.

metal.* Note that direct extrusion is similar to squeezing a tube of toothpaste which forces the contents through an opening. *Indirect* or *backward* extrusion is illustrated in Figure 8–3.

* For certain combinations of die angle and reduction, it is possible to develop hydrostatic tensile stresses along the centerline of the billet. (See Ref [1]). Possible consequences are discussed in Sec. 8.8.4.

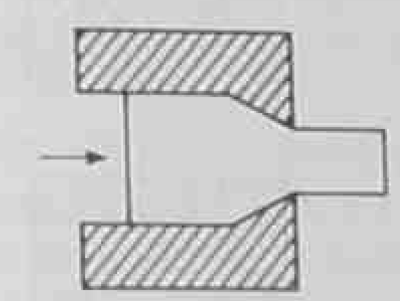


Figure 8–2 Schematic of direct or forward extrusion.

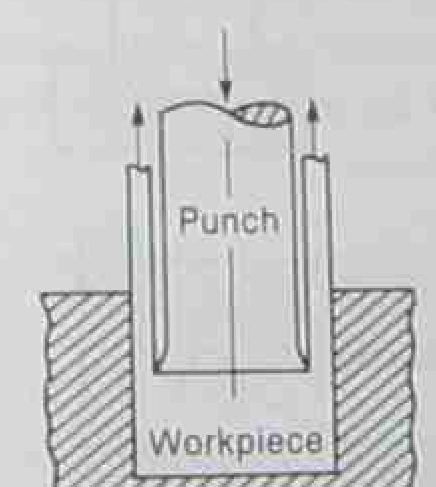


Figure 8–3 Schematic of indirect or backward extrusion.

Drawing is the process whereby rod or wire can be made; see Fig. 8–4. In contrast with extrusion, the metal is pulled through the die so that tensile stresses are developed in the outlet section. The reduction per pass must be limited; otherwise, the outlet material will neck and fracture. In wire drawing, successive reductions are made by using a series of dies of ever-decreasing outlet diameter, and quite large *total* reductions can be made without fracture or intermediate annealing.

Forging processes are usually classified as either *closed-die* or *open-die*; Figs. 8–5 and 8–6 illustrate these processes, both of which basically employ direct compression to the workpiece. With closed-die forging, the accuracy of the dies is high, and finished parts are often produced. Because dies are expensive, such processes are economical only if high production (that is, a large number of parts) is involved. Open-die forging often utilizes flat tooling surfaces, so smaller production rates can be tolerated. As shown in Fig. 8–7, *coining* is a type of forging process whereby the metal is moved to produce an embossed or raised surface according to some predetermined pattern; coins and medals typify this result. *Upsetting*, as shown in Fig. 8–8, is another forging-type process; the heads on bolts or nails are produced in this way.

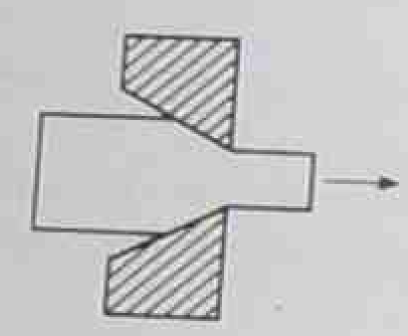


Figure 8–4 Schematic of drawing.

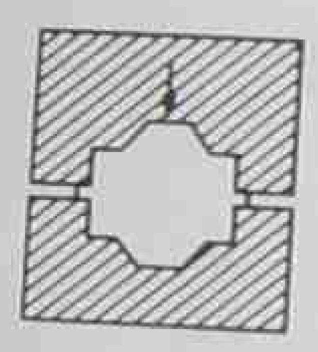


Figure 8–5 Schematic of closed-die forging.

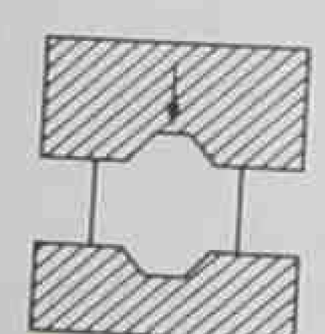


Figure 8–6 Schematic of open-die forging.

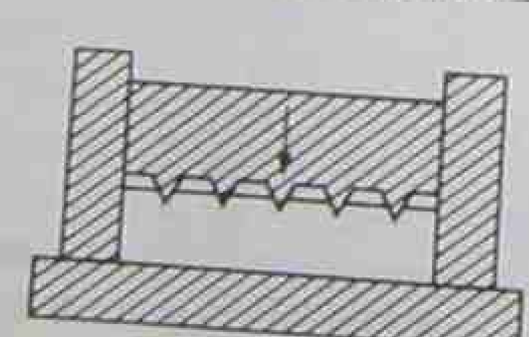


Figure 8–7 Illustration of coining.

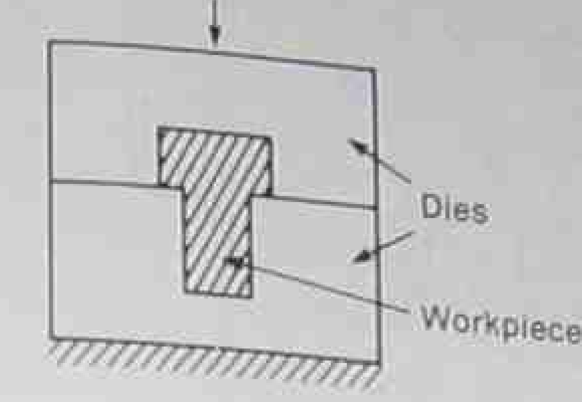


Figure 8–8 Schematic of an upsetting operation.

Other processes could be included here, such as swaging, thread rolling, and tube or pipe production, but those described above should provide an adequate introduction to the meaning of bulk forming.

8.3 SHEET FORMING

Sheet forming is distinguished from bulk forming, since the metal properties of importance, the tooling and the metal-to-tool contact conditions, are quite different. In most sheet-forming processes, the metal deforms in certain regions under tensile stresses, and the extent of deformation is often governed by the onset of necking, which is well short of actual fracture. In contrast, bulk-forming operations utilize compressive loads, and the limit of deformation (often referred to as *formability*) is the formation of cracks.

Bending, by which components such as brackets, cabinets, and the like are produced, is, perhaps, the simplest and most obvious type of sheet-forming process; it is shown in Fig. 8–9. If pure bending is used, the inner fibers experience compression while the outer experience tension. Release of the bending moment results in partial unbending or *springback* due to elastic recovery. This, by the way, is a real problem, since it is the shape of the workpiece *after* springback that is of major concern. If during bending, tensile forces are applied simultaneously, springback may be minimized, since the entire section can be subjected to net tension at both the outer and inner fibers. This is usually called *stretch-forming*.

Deep drawing (also called cupping) is used to produce items such as cartridge cases, beverage cans, flashlight cases, and the like; see Fig. 8–10. A round blank, initially clamped by means of a *hold-down* device, is forced to flow into a die by means of the punch force. The flange of the blank is subjected to *compressive* stresses, while the cup wall experiences tensile stresses. If the latter become excessive, the cup will fail, usually near the bottom, where the blank material has not been work hardened. The hold-down

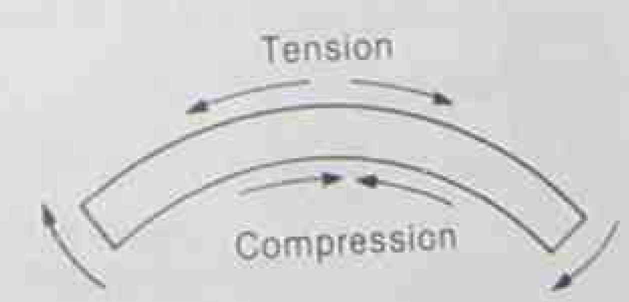


Figure 8–9 A bending operation; note the signs of the surface stresses.

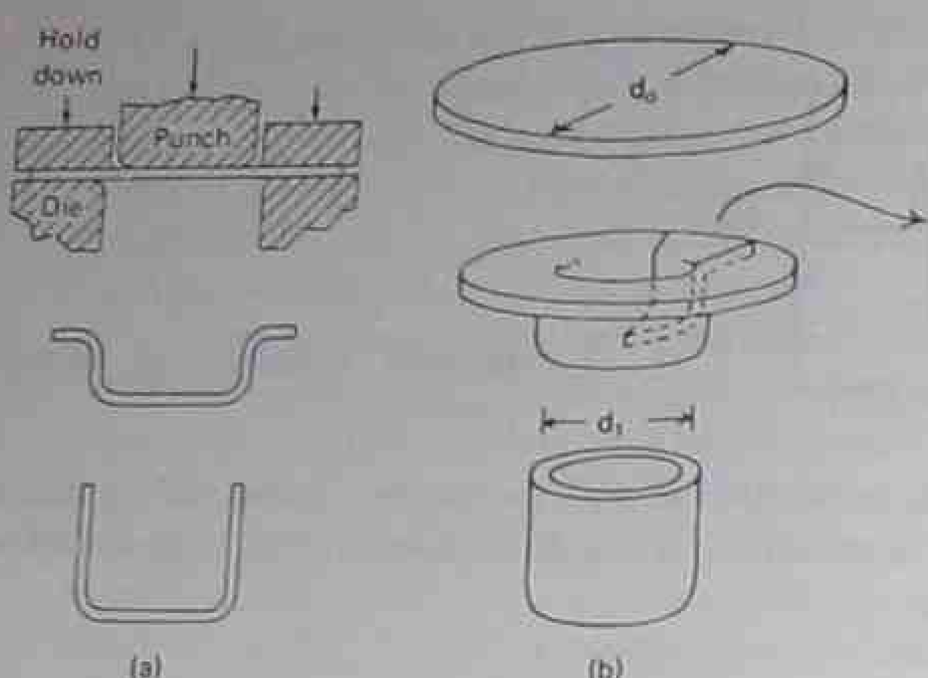


Figure 8–10 An illustration of deep drawing.

force must be of adequate magnitude, or else the compressive stresses will induce a type of buckling or *wrinkling* of the flange; see Fig. 8–11.

Often, a single drawing operation will not produce a cup of desired height, so *redrawing* is used, as shown in Fig. 8–12. As a final operation, *ironing* may be employed. Since the wall thickness from the top to bottom of the cup is not truly uniform, the cup is forced to flow through a final annular opening to produce a thinner and more uniform wall thickness; this is the ironing stage. Figure 8–13 illustrates a drawing, redrawing, ironing sequence using concentric dies and one stroke to finish the cup.

A number of sheet-forming operations is considerably more involved than those discussed above. During deformation, some regions of the sheet experience stretching

Figure 8–11 Wrinkling in a partially drawn cup due to insufficient hold-down force. From D. J. Meuleman, Ph.D. thesis, the University of Michigan (1980).

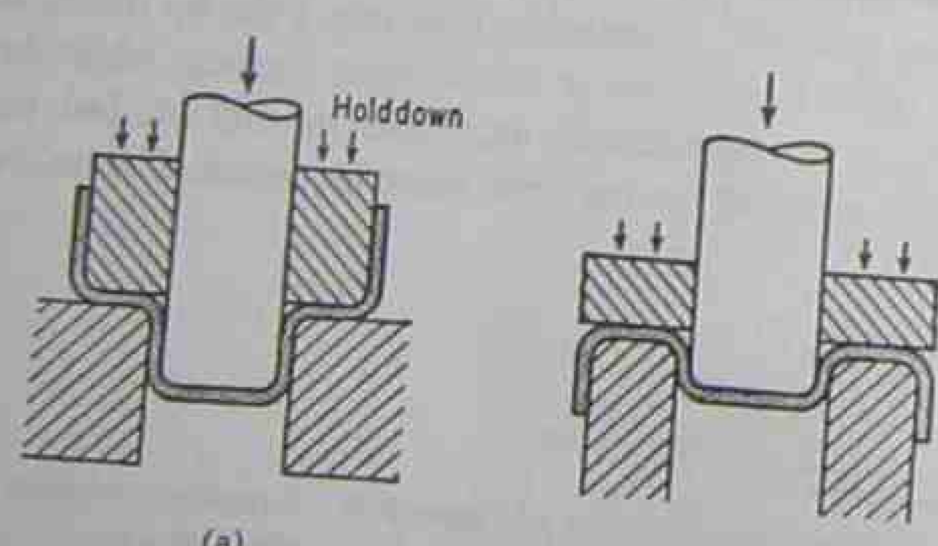


Figure 8–12 Examples of (a) direct and (b) reverse redrawing.

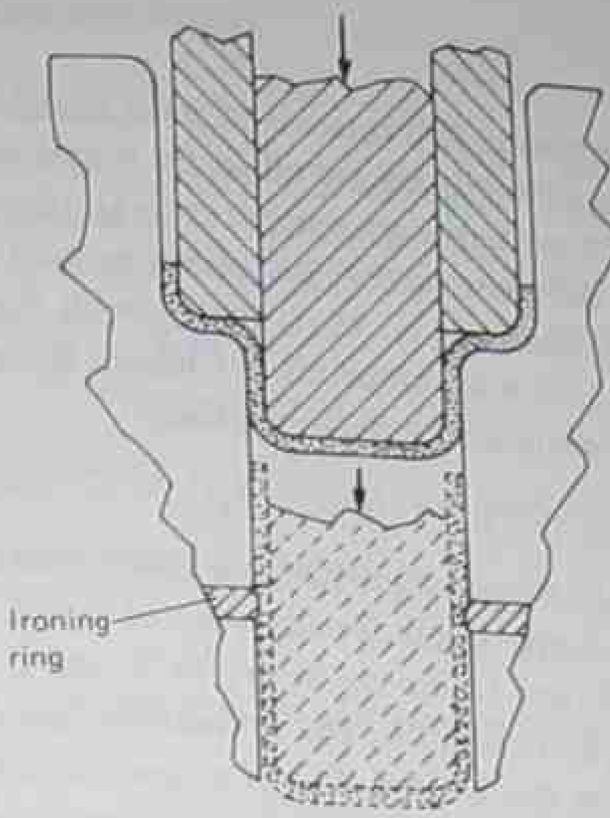


Figure 8–13 Schematic of deep drawing operation showing the final stage of ironing.

while other regions undergo bending. These can be called *complex stampings* of which car fenders, oil pans, and panels are typical. In a later section, the concept of *forming limit diagrams* will be introduced to show how limiting deformations of such complicated products are often analyzed.

8.4 YIELDING

To induce plastic deformations, adequate loads must be applied to the body so that the *applied* stresses will cause the body to yield. Note that to *continue* deformation, the stresses must be *increased* due to work hardening, as discussed in Chapter 6.

Except for the tensile test, true uniaxial loading is not encountered in forming operations, since combined states of stress are usually involved. In such cases, it is necessary to employ a *yield criterion;* this is a postulated mathematical expression involving the combination of stresses that will induce yielding. For our purposes, only one yield criterion will be used, although it is noted that others are available (see, for example, Ref. [2]). All such criteria are based upon certain assumptions, and if these are reasonably met, the more closely will predictions agree with experimental measurements. The assumptions involved are

1. The metal is isotropic and homogeneous.
2. The yield strength in tension and compression is equivalent; thus there is no Bauschinger effect.*
3. Constancy of volume prevails; thus the *plastic* equivalent of Poisson's ratio is 0.5.
4. The magnitude of the *mean normal stress,* σ_m does not influence yielding, where

$$\sigma_m = (\sigma_x + \sigma_y + \sigma_z)/3 = (\sigma_1 + \sigma_2 + \sigma_3)/3 \qquad (8\text{–}1)$$

Note that in Eq. (8–1) the stresses are added algebraically where tension is positive and

* Discussed later in this chapter.

compression is negative. Since plastic deformation of most metals occurs by *slip*, which is a shearing process, common yield criteria are all really functions of shear stresses and when particular combinations of such stresses reach a critical value, yielding is incipient.

Although referred to by different names, the criterion used here is called the von Mises (or simply Mises) yield criterion [3]. In its most useful form it is expressed as a function of *principal stresses*. This was discussed in Chapter 6, and they will *always* be designated as σ_1, σ_2, σ_3. The postulated criterion is then

$$(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 = C \quad \text{(a constant)} \tag{8–2}$$

Again, it is essential to invoke the proper signs for each stress (that is, plus for tension and minus for compression). If this is applicable for any state of stress, the constant C is most readily evaluated from a tensile test where $\sigma_2 = \sigma_3 = 0$ and σ_1 is positive. Yielding occurs when $\sigma_1 = Y$ (uniaxial yield strength), and introducing these values into Eq. (8–2) shows that $C = 2Y^2$. If a torsion test is used (this is a case of pure shear), the principal stresses are $\sigma_1 = -\sigma_3$, and $\sigma_2 = 0$.* At yielding, $\sigma_1 = k$ (shear yield strength) and using Eq. (8–2) gives $C = 6k^2$. So the Mises criterion becomes

$$(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 = 2Y^2 = 6k^2 \tag{8–3}$$

Note that Y and k are material properties *obtained from experiment*. *Any* yield criterion must be expressed as a function of such properties.

Although stated without proof, this criterion in its most general form is

$$(\sigma_x - \sigma_y)^2 + (\sigma_y - \sigma_z)^2 + (\sigma_z - \sigma_x)^2 + 6(\tau_{xy}^2 + \tau_{yz}^2 + \tau_{zx}^2) = 2Y^2 = 6k^2 \tag{8–4}$$

Since in many forming processes, the applied stresses are principal (or nearly so), Eq. (8–3) finds major use.

8.5 PLASTIC WORK

Consider a round bar of initial length ℓ_0 and area A_0 subjected to a force F that causes an extension $d\ell$. The work done is $F\,d\ell$ and on a *unit volume* basis, using the area and length under load, we obtain

$$dw = (F\,d\ell)/A\ell = \sigma\,d\epsilon \tag{8–5}$$

or

$$w = \int_{\epsilon_1}^{\epsilon_2} \sigma\,d\epsilon \tag{8–6}$$

so w is equal to the area under the true stress-true strain curve between appropriate strain limits. Using principal components, and in the most general case, we find that Eq. (8–5) becomes

$$dw = \sigma_1\,d\epsilon_1 + \sigma_2\,d\epsilon_2 + \sigma_3\,d\epsilon_3 \tag{8–7}$$

* See Chapter 6, where Mohr's circle can be used to show this result.

Example 8–1

The tensile yield strength Y of a metal is 175 MPa. A block of this metal is subjected to principal stresses $\sigma_1 = 70$ MPa and $\sigma_2 = -35$ MPa. What is the magnitude of the tensile stress in the third principal direction that would cause yielding?

Solution With Eq. (8–3),

$$(70 - [-35])^2 + (-35 - \sigma_3)^2 + (\sigma_3 - 70)^2 = 2(175)^2$$

from which

$$\sigma_3 = (35 \pm 299)/2$$

Since a tensile stress is required,

$$\sigma_3 = (334)/2 = 167 \text{ MPa.}$$

Note if a *compressive* value of σ_3 were required, the negative root would be used and $\sigma_3 = -132$ MPa.

8.6 EFFECTIVE STRESS

In plasticity problems it is useful to define an *effective* stress $\bar{\sigma}$, which is a function of the applied stresses, and which may be thought of as the equivalent single stress that has the same effect as the actual combined stresses in regard to yielding. If the *magnitude* of $\bar{\sigma}$ reaches a certain level, yielding is predicted. Thus,

$$2\bar{\sigma}^2 = (\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 \tag{8–8}$$

for the Mises criterion. Note that $\bar{\sigma}$ is *always* positive by this relationship. When $\bar{\sigma} = Y$ or $\sqrt{3}\,k$, yielding is predicted.

8.7 EFFECTIVE STRAIN

This is *defined* so as to express the incremental work *per unit volume* as

$$dw = \bar{\sigma}\,d\bar{\epsilon} = \sigma_1\,d\epsilon_1 + \sigma_2\,d\epsilon_2 + \sigma_3\,d\epsilon_3 \tag{8–9}$$

where the symbol $\bar{\epsilon}$ is taken as the total strain.* In conjunction with the effective stress defined by Eq. (8–8), the companion effective strain is expressed as

$$d\bar{\epsilon} = [2/3(d\epsilon_1^2 + d\epsilon_2^2 + d\epsilon_3^2)]^{1/2} \tag{8–10}$$

or in terms of total strains

$$\bar{\epsilon} = [2/3(\epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2)]^{1/2} \tag{8–11}$$

where the incremental strains in Eq. (8–10) can be converted to total strains by proper integration. For the derivation of Eq. (8–11), see Ref. [1]. With Eqs. (8–8) and (8–11), Eq. (6–11) can be written as $\bar{\sigma} = K\bar{\epsilon}^n$; this is the form we will use in all further work involving a strain hardening equation.

* For most of the problems of concern in this chapter, the plastic strain is nearly equivalent to the total strain, and ignoring elastic effects causes little error.

In terms of effective stress and strain, the stress-strain relations (or *flow rules*) connected with Eqs. (8–8) and (8–10) can be expressed as

$$d\epsilon_1 = (d\bar{\epsilon}/\bar{\sigma})([\sigma_1 - (1/2)(\sigma_2 + \sigma_3)]) \tag{8.12}$$

which shows certain similarities with Hooke's law in three-dimensional form, Eq. (6–29); note that $d\bar{\epsilon}/\bar{\sigma}$ is *not* a constant as is E; rather it is a positive ratio whose magnitude changes during deformation. Expressions for $d\epsilon_2$ and $d\epsilon_3$ result by interchanging subscripts.

Example 8–2

A plate is subjected to plane stress loading such that $\sigma_1 = 2\sigma_2$, $\sigma_3 = 0$, and plastic deformation results. Find the plastic work per unit volume in terms of σ_1 and $d\epsilon_1$.

Solution With volume constancy, $d\epsilon_1 + d\epsilon_2 + d\epsilon_3 = 0$, so $d\epsilon_3 = -d\epsilon_1 - d\epsilon_2$. From Eq. (8–7), $dw = \sigma_1\, d\epsilon_1 + (\sigma_1/2)(d\epsilon_2) = \sigma_1[(d\epsilon_1) + (d\epsilon_2/2)$. From Eq. (8–10), $d\bar{\epsilon} = [2/3(d\epsilon_1^2 + d\epsilon_2^2 + (-d\epsilon_1 - d\epsilon_2)^2)]^{1/2}$ or

$$d\bar{\epsilon} = [4/3(d\epsilon_1^2 + d\epsilon_1\, d\epsilon_2 + d\epsilon_2^2)]^{1/2}$$

Since $\sigma_2 = \frac{1}{2}\sigma_1$ and $\sigma_3 = 0$, from Eq. (8–12),

$$d\epsilon_2 = (d\bar{\epsilon}/\bar{\sigma})[\sigma_2 - (1/2)\sigma_1]$$

$$\text{so } d\bar{\epsilon} = (2)/\sqrt{3}\, d\epsilon_1$$

and from Eq. (8–8)

$$2\bar{\sigma}^2 = (6\sigma_1^2/4) \text{ so, } \sqrt{3}\,\sigma_1/2$$

$$\text{so } dw = \bar{\sigma}\, d\bar{\epsilon} = [(\sqrt{3}\,\sigma_1)/2](2\, d\epsilon_1 \sqrt{3}) = \sigma_1\, d\epsilon_1$$

8.8 BULK FORMING ANALYSES AND OTHER CONSIDERATIONS

8.8.1 Ideal Work Method

The simplest approach to force or pressure measurements in bulk forming operations, *where it is applicable*, involves a force or energy balance assuming 100 percent efficiency; that is why it is called *ideal*. In essence, any effects of friction or *redundant* deformation* are ignored, and what is then necessary is to envision an ideal process that could produce the same *shape change* as the actual process being considered.

As an example, consider the direct axisymmetric extrusion as illustrated in Fig. 8–14. The diameter reduction could be accomplished by pure tensile loading; it is *not necessary* that such a change could *actually* be done by such loading (that is, the induced homogeneous strain could be in excess of the necking strain in pure tension). As discussed in Sec. 8–5, the *plastic work per unit volume* for tensile loading was

* Explained in Sec. 8.8.2.

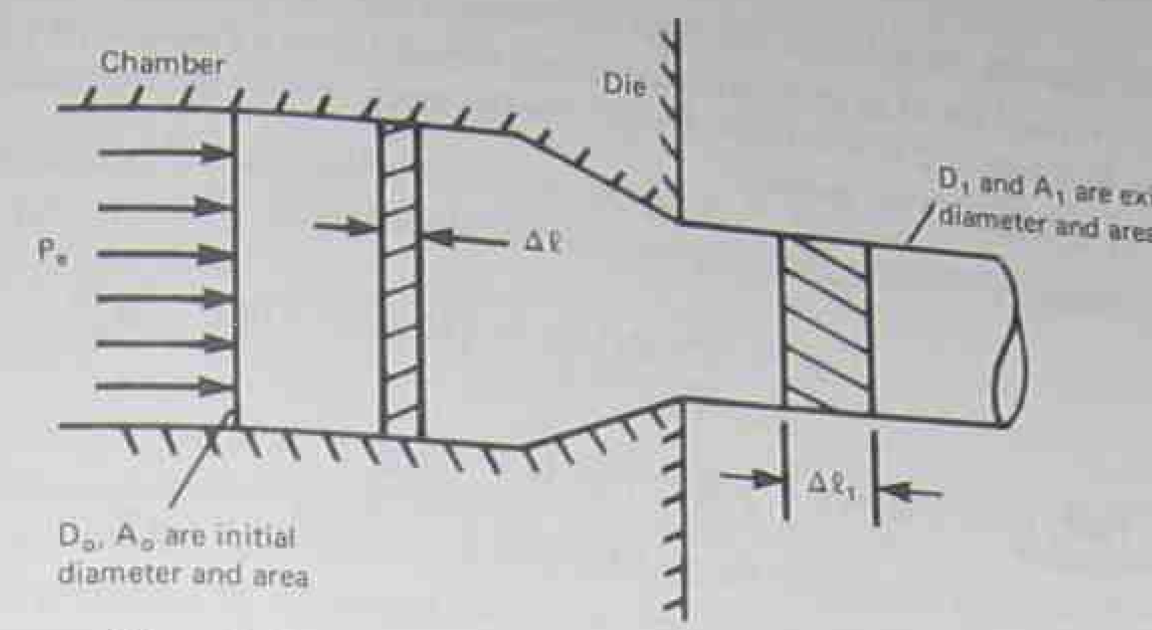


Figure 8–14 Sketch of direct extrusion if ideal deformation is assumed.

$$w_i = \int_{\bar{\epsilon}_1}^{\bar{\epsilon}_2} \bar{\sigma}\, d\bar{\epsilon} \tag{8–13}$$

the subscript i meaning *ideal*. If $\bar{\sigma} = K\bar{\epsilon}^n$ then

$$w_i = [K/(n+1)[\bar{\epsilon}^{n+1}]]_{\bar{\epsilon}_1}^{\bar{\epsilon}_2} \tag{8-14}$$

Now in reference to Fig. 8–14, and due to volume constancy, an inlet volume of $A_o\, \Delta\ell_o$ must exit as $A_1\, \Delta\ell_1$, and the *total* ideal work required is

$$W_i = F_e\, \Delta\ell_0 \tag{8–15}$$

where F_e is the *force* applied to the workpiece and the pressure P_e is simply F_e/A_0. On a per unit volume basis, $w_i = W_i/(A_0\, \Delta\ell_0)$, so

$$w_i = (F_e\, \Delta\ell_0)/A_0\, \Delta\ell_0 = P_e \tag{8–16}$$

Thus, the extrusion pressure for ideal conditions is, from Eqs. (8–13) and (8–16),

$$P_e = \int_{\bar{\epsilon}_1}^{\bar{\epsilon}_2} \bar{\sigma}\, d\bar{\epsilon} \tag{8–17}$$

where $\bar{\epsilon}_1$ refers to any strain induced *prior to extrusion* (for example, if the workpiece is initially annealed, then $\bar{\epsilon}_1 = 0$) while $\bar{\epsilon}_2$ is the homogeneous strain based upon the area change (that is, $\bar{\epsilon}_2 = \ell n(A_0/A_1)$).

Rod or wire drawing follows exactly the same arguments, so the outlet *drawing stress* is

$$\sigma_d = \int_{\bar{\epsilon}_1}^{\bar{\epsilon}_2} \bar{\sigma}\, d\bar{\epsilon} \tag{8–18}$$

since the same ideal process is involved.

Example 8–3

A round bar of an annealed metal is cold extruded from a diameter of 12.7 mm to 11.5 mm in one pass. If $\bar{\sigma} = 100{,}000\,\bar{\epsilon}^{0.2}$ psi. for this metal, what is the extrusion pressure according to the ideal work method?

Solution From Eq. (8–17),

$$P_e = \int_{\bar{\epsilon}_1}^{\bar{\epsilon}_2} K\bar{\epsilon}^n\, d\bar{\epsilon} = [K/(n+1)][\bar{\epsilon}^{n+1}]_{\bar{\epsilon}_1}^{\bar{\epsilon}_2}$$

Here, $\bar{\epsilon}_1 = 0$ and $\bar{\epsilon}_2 = 2\ell n(12.7/11.5) = 0.199$

$$\text{So } P_e = (100{,}000/1.2)(0.199)^{1.2} = 11{,}970 \text{ psi}$$

The pressure needed in forging would use frictionless compression as the ideal process, whereas rolling would utilize a plane-strain tension test. In all cases, the ideal work per unit volume is based upon Eq. (8–13), and all that is essential is to employ a process that *could* produce such a shape change and permit evaluation of the induced homogeneous strain. Note that in an operation such as hardness testing, the deformation caused by indenting would not permit a calculation of strain based upon *geometric* changes; thus, the ideal work method could not be used in such a situation. Finally, whenever equations such as Eqs. (8–17) or (8–18) are used, they *always* give a prediction that is *less than* that required for the operation; in this sense they provide a "lower bound."

8.8.2 Efficiency Factors

Besides the work or energy needed to cause homogeneous deformation, two other demands are involved. Friction at the work-tool interface consumes energy, and *redundant deformation* also occurs. The latter is simply due to internal distortion of the work in excess of that needed to produce the external shape change. In Fig. 8–14, it is easy to envisage friction at the interface, while Fig. 8–15 illustrates redundant deformation.

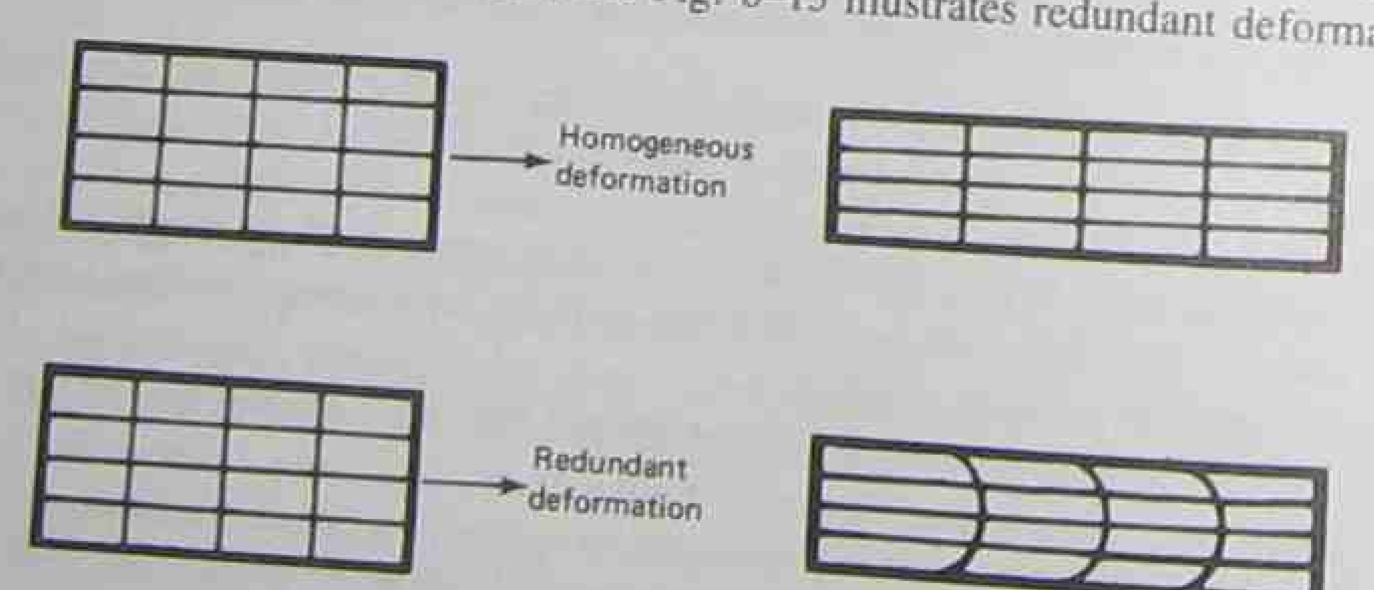


Figure 8–15 A comparison of ideal and actual deformation to show the meaning of redundant deformation.

If all of these factors are considered, then the *actual* work per unit volume is

$$w_a = w_i + w_f + w_r \tag{8–19}$$

where the subscripts are self-evident. In practice, w_f and w_r are difficult to estimate and are also not mutually independent, so a simple way to circumvent this is to define an *efficiency factor* η as

$$\eta \equiv w_i / w_a \tag{8–20}$$

thereby lumping together the non-ideal work terms. Parameters such as die angle, lubrication, and reduction will influence the value of η, but in practice a value between 0.5 and 0.65 is reasonable. If such a factor is introduced, equations such as (8–17) or (8–18) would be modified by placing a factor of $1/\eta$ in front of the integral, so that the actual values of P_e or σ_d would be greater than those based upon ideal deformation. This use of an efficiency factor, though a bit ill-defined in a fundamental sense, is an attempt to include non-ideal effects that are often difficult to assess by themselves.

Operations such as extrusion, rolling, and rod or wire drawing involve flow through a converging deformation zone. In Fig. 8–16, the three individual work contributions are indicated schematically, and one can see that as the die angle (α in Fig. 8–14) increases, w_f decreases while w_r increases. As a result, there will be an optimum angle leading to the lowest force or pressure in such operations; Fig. 8–17 shows the results found in one study noting that both α and reduction r influence the *optimum* value of α. The reason for the influence of α on w_f and w_r is easily explained: For a given reduction, a decrease in α means a greater contact length between the die and work; since the compressive stress along that interface is essentially constant, the *force* must therefore increase with the greater length. If the coefficient of friction μ does not radically change, then as α decreases, w_f must increase. With redundant deformation, as α increases there is a large degree of internal deformation due to shearing effects; as a consequence, w_r increases with α. Note that as α approaches zero, this approximates a tensile test in which $w_r = 0$.

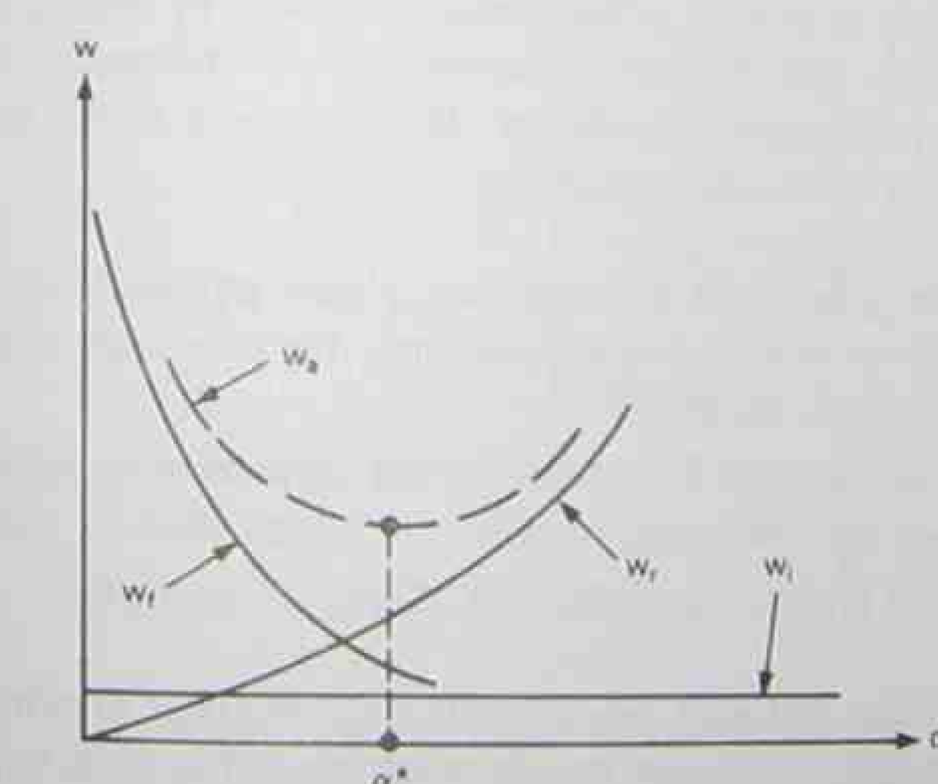


Figure 8–16 Influence of the semi-die angle α on the actual work w_a, during drawing, where the individual contributions of ideal work w_i, frictional work w_f, and redundant work w_r are shown.

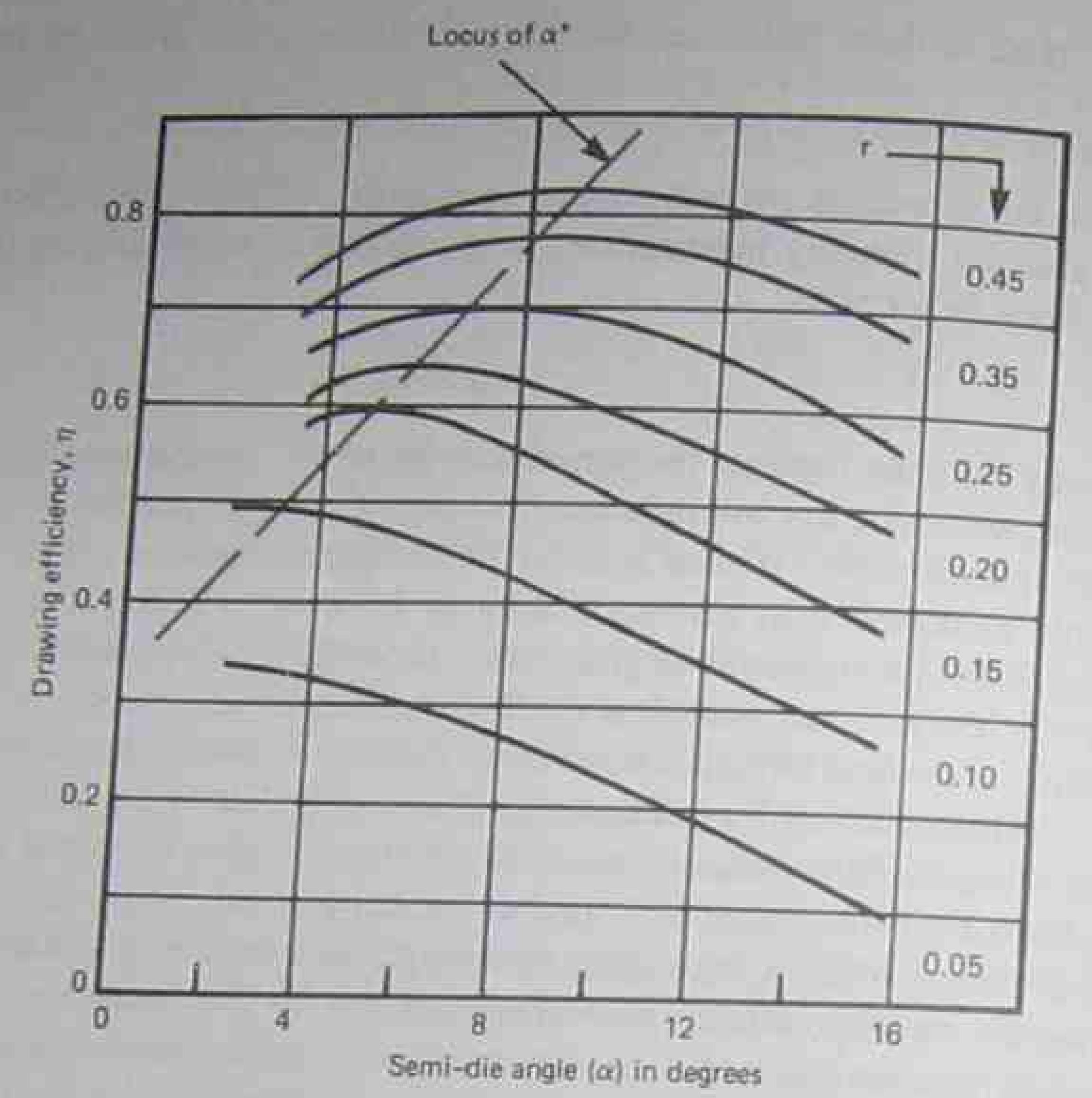


Figure 8–17 Effect of semi-die angle on drawing efficiency for various reductions; note the change in the optimum die angle α^*.

8.8.3 Force Balance or Slab Analysis

As contrasted with the ideal work method, this technique does include the effects of friction at the work-tool interface; in essence, a force balance is made by using a thin slab of metal of differential thickness. This produces a *one-dimensional* differential equation which is then solved using proper boundary conditions. The limitations of this technique arise because of the assumptions involved, these being:

1. The applied load acts in a principal direction and planes perpendicular to that direction define the other two principal directions; thus, the directions of the principal stresses are assumed from the outset.
2. Surface friction effects enter into the force balance, but these have no influence on internal distortion or upon the orientation of principal planes (that is, directions).
3. Homogeneous deformation prevails, so redundant effects are neglected.

Figures 8–18(a) and 8–18(b) illustrate plane-strain strip or sheet drawing where Coulomb friction is assumed at each interface so that the shear stress τ is taken as μP. The

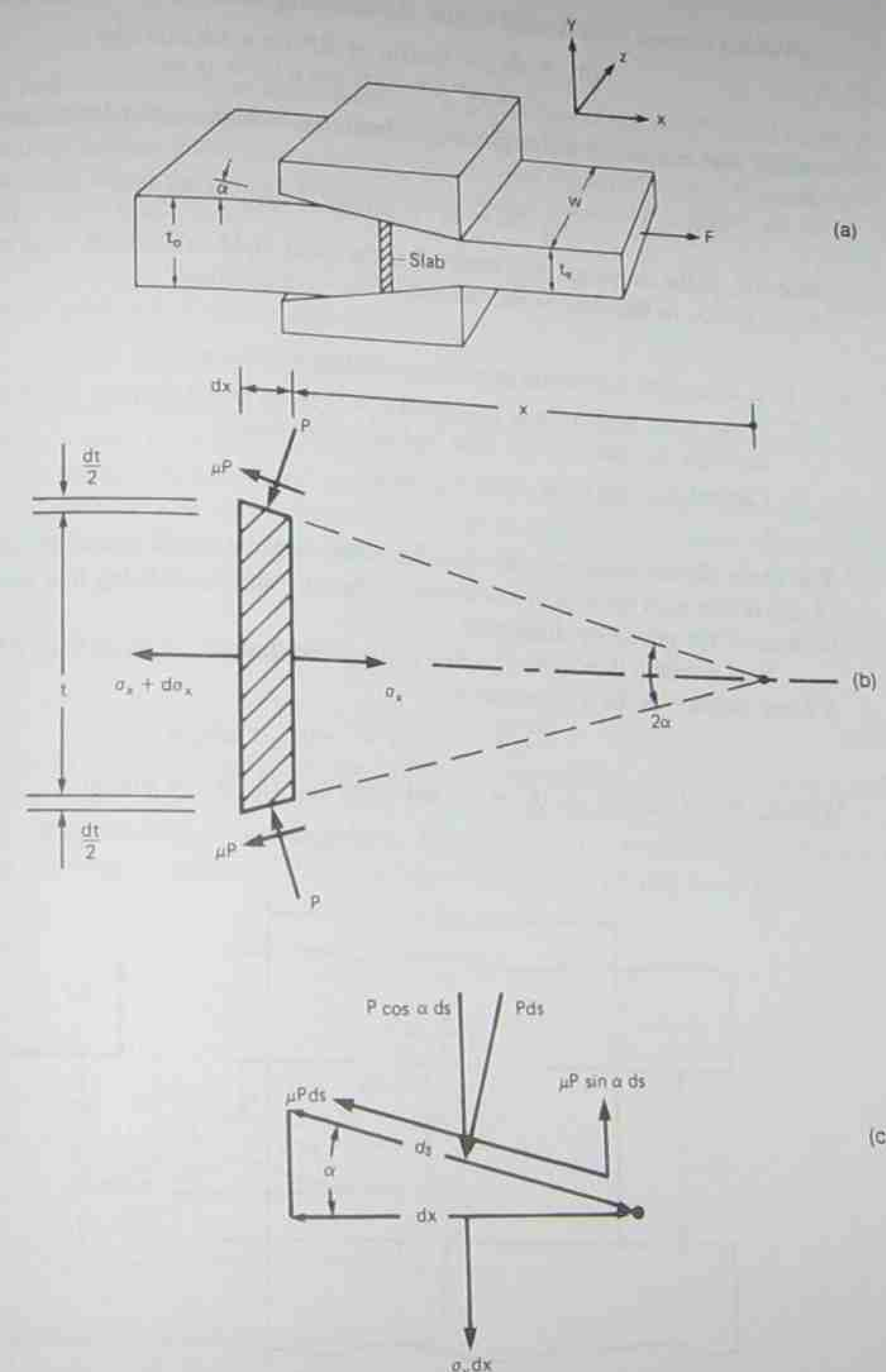


Figure 8–18 (a) Essentials of plane-strain strip or sheet drawing, (b) an enlarged view of the slab used for a force balance analysis and (c) vertical force balance.

enlarged element then is used to give the following force balance in the x-direction.

$$(\sigma_x + d\sigma_x)(t + dt)w + 2P \sin \alpha \, (dx/\cos \alpha)w + 2\,\mu P \cos \alpha (dx/\cos \alpha)w = \sigma_x wt \tag{8–21}$$

which after expanding and neglecting differentials of higher order leads upon integration to

$$\sigma_d /2k = (1 + B/B)[1 - \exp(-B\epsilon_h)] \tag{8–22}$$

where σ_d is the drawing (exit) stress, k is the shear yield strength, $B = \mu \cos \alpha$, and $\epsilon_h = \ell n(t_0/t_e)$. In this derivation, the following were assumed:

1. μ represents a constant coefficient of sliding friction.
2. Either work hardening is neglected or k is taken as the average value of shear yield between the states of the inlet and outlet material.
3. Conical dies are used, so α is constant.

The many details involved with this derivation may be found elsewhere [1]. Equation (8–22) is also used for *wire or rod drawing*, the only difference being that $\epsilon_h = \ell n\,(A_0/A_e)$ because of the geometry difference.

Next consider direct compression under plane strain ($\epsilon_z = 0$) as shown in Fig. 8–19. A force balance in the x direction is

$$-\sigma_x h + 2\tau \, dx = -(\sigma_x + d\sigma_x)h \tag{8–23}$$

With $\sigma_y = -P$ and $\sigma_x - \sigma_y = 2k$ and using Eq. (8–3), we obtain

$$dP = 2\tau \, dx/h \tag{8–24}$$

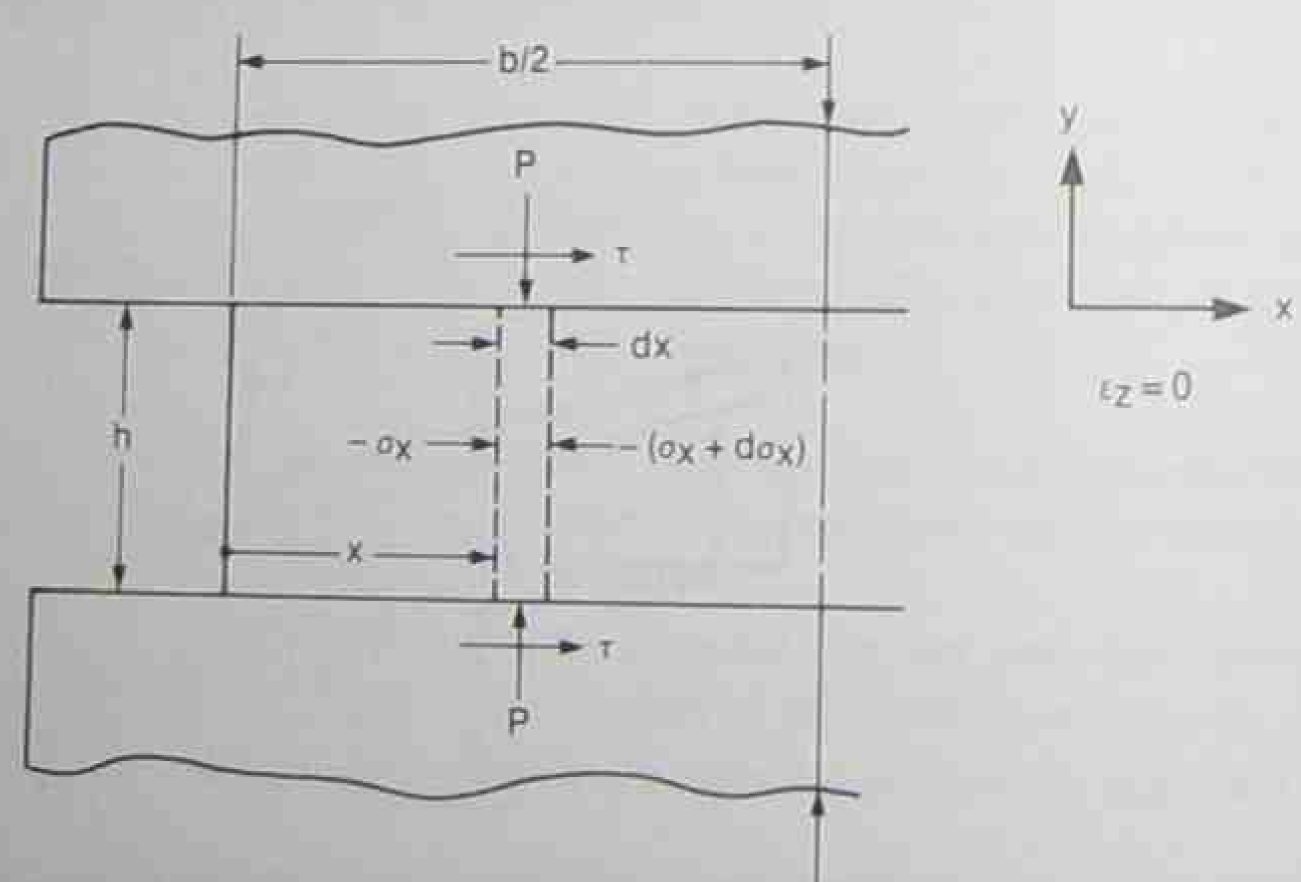


Figure 8–19 Essentials for a slab force analysis in plane-strain compression.

For *sliding* friction, $\tau = \mu P$, and since $P = 2k$ when $x = 0$,

$$P/2k = \exp(2\mu x/h) \tag{8–25}$$

noting that P is maximum at the centerline, $x = b/2$. A plot of Eq. (8–25) is shown in Fig. 8–20, and the increase of P towards the centerline is called the *friction hill*. If we assume that sliding friction does occur over the entire contact region, the *force* to cause deformation must equal the contact area times the *average pressure* P_a, where, as shown in Ref. [1], P_a is found to be

$$P_a/2k = (h/\mu b)\,[\exp(\mu b/h) - 1] \tag{8–26}$$

which for $\mu b/h << 1$ can be approximated by

$$P_a/2k \approx 1 + (\mu b/2h) \tag{8–27}$$

Note that since τ cannot *exceed* μP, and the latter cannot be $>k$ (shear strength of the workpiece), then for $P \geq 2k$, μ must be $\leq½$.

If lubrication is poor and sliding does not occur, the τ in Eq. (8–23) is replaced with k and *sticking* friction takes place (literally, shear occurs in the metal adjacent to the loading surfaces). Expressions similar to Eq. (8–25) for $P/2k$ and Eq. (8–26) for $P_a/2k$ become

$$P/2k = 1 + (x/h) \tag{8–28}$$

and

$$P_a/2k = 1 + (b/4h) \tag{8–29}$$

Figure 8–21 shows the linear fraction hill associated with Eq. (8–28); note how it differs from Fig. 8–20. It is possible to have mixed conditions where sliding prevails in one region while sticking occurs elsewhere. This will result if predictions based upon Eq. (8–25) give higher values of $P/2k$ than Eq. (8–28), since μP can never exceed k.

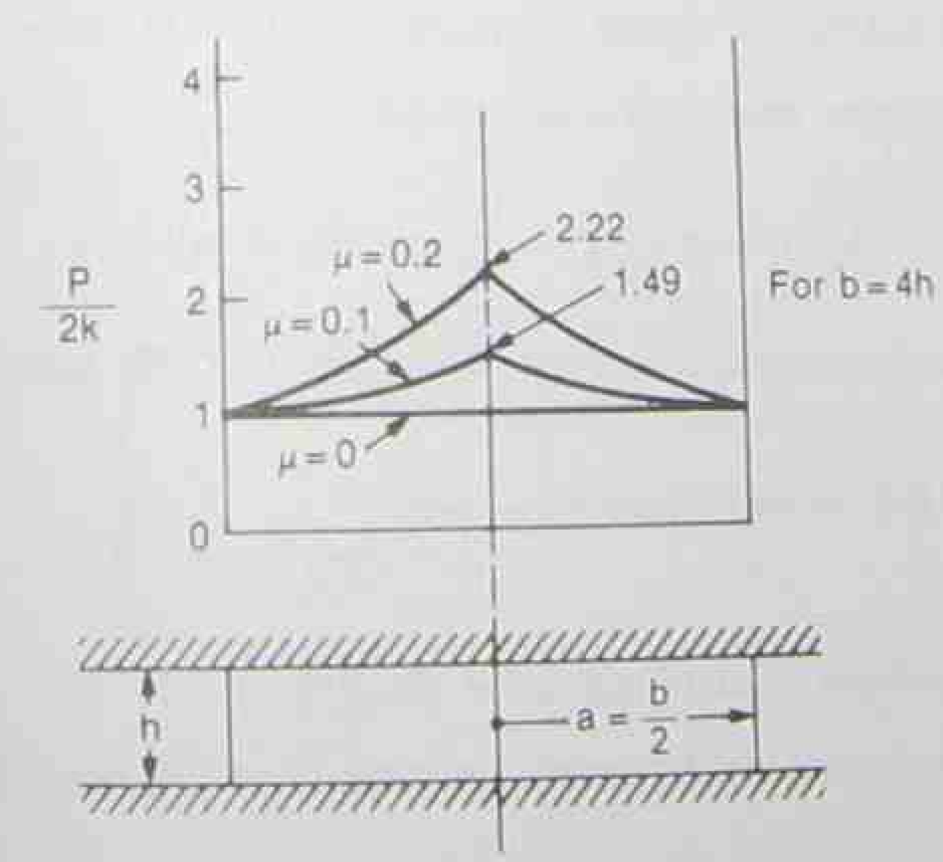


Figure 8–20 Example of the friction hill in plane-strain compression for different values of the coefficient of friction.

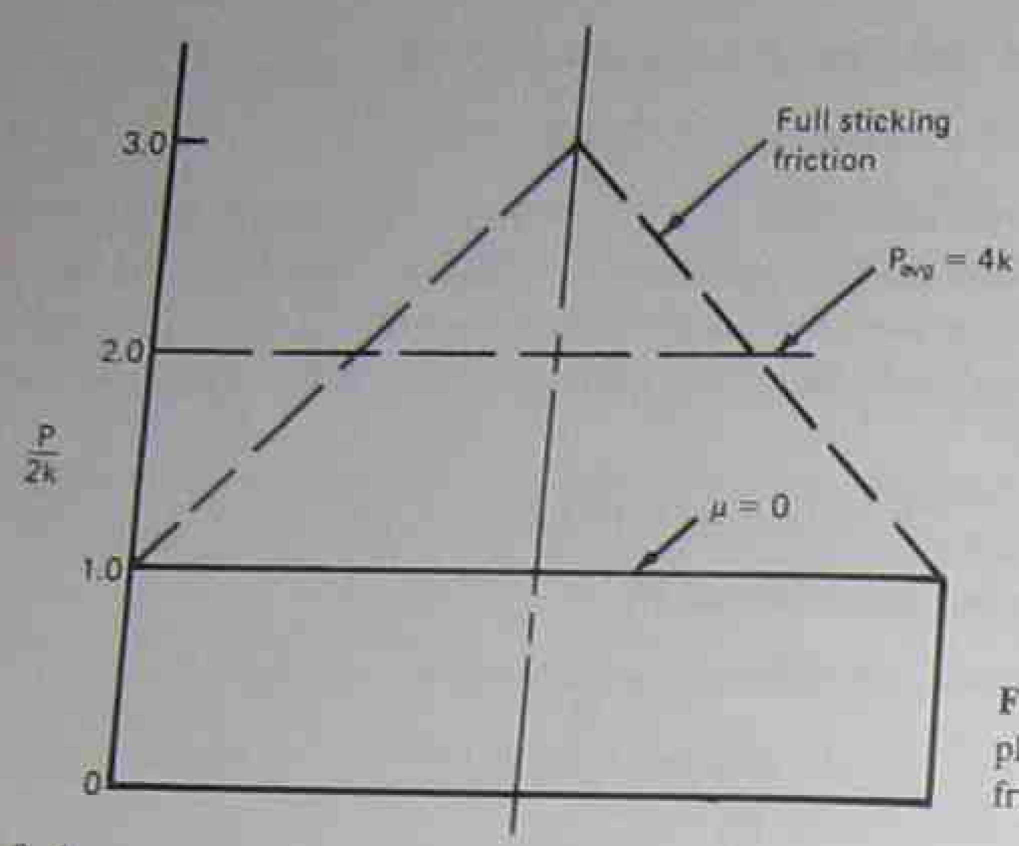


Figure 8–21 The friction hill in plane-strain compression with sticking friction.

Example 8–4

A block of metal whose yield shear strength k is 105 MPa is subjected to plane-strain compression. The block is 250 mm wide by 25 mm high. Find the average pressure at the start of plastic flow

a. If sliding friction occurs and $\mu = 0.10$.
b. If sticking friction occurs.

Solution

a. With Eq. (8–26)

$$P_a = (2kh/\mu b)[\exp(\mu b/h) - 1]$$

where $\mu b/h = 0.1(250)/25 = 1.0$

$$P_a = 2(105)/1\ [\exp(1) - 1] = 361 \text{ MPa}$$

Note, with Eq. (8–27),

$$P_a = 2k[1 + (\mu b/2h)] = 2(105)[1 + (0.1)(250)/50] = 315 \text{ MPa}$$

Here, Eq. (8–27) is not accurate, since $\mu b/h$ is *not* $<<1.0$.

b. With Eq. (8–29),

$$P_a = 2k[1 + (b/4h)] = 2(105)(1 + [250/100]) = 735 \text{ MPa}$$

A similar approach for axisymmetric compression (see Fig. 8–22 for specifics) leads to the following:

$$P/Y = \exp[(2\mu/h)(R - r)] \tag{8–30}$$

and

$$P_a/Y = (1/2)(h/\mu R)^2 [\exp(2\mu R/h) - (2\mu R/h) - 1] \tag{8–31}$$

for *sliding friction;* here Y is the *tensile* yield strength which equals $\sqrt{3}k$, according to the Mises criterion.

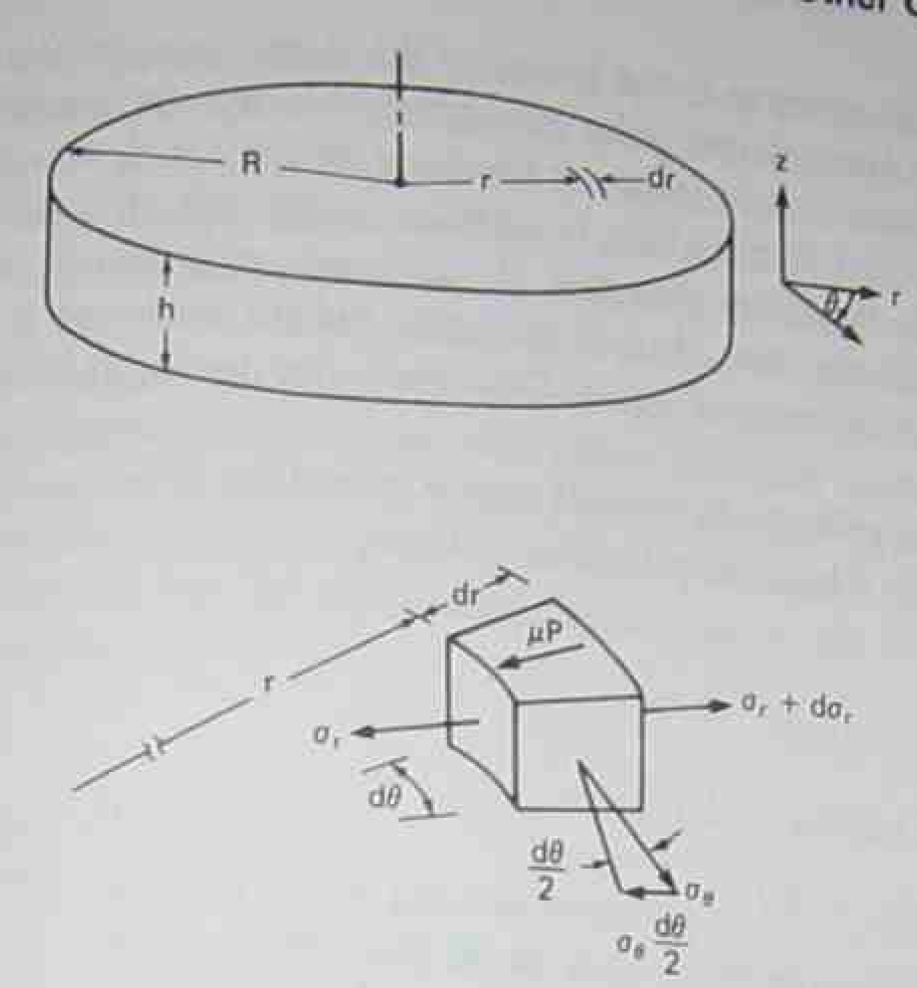


Figure 8–22 Element used in a slab analysis for axisymmetric compression.

With this geometry and *sticking friction*,

$$P = Y + (2k/h)(R - r) \tag{8–32}$$

and

$$P_a = Y + (2kR/3h) \tag{8–33}$$

The concept of a friction hill also enters into flat rolling; Fig. 8–23 is a schematic of this process. In such an operation, it is usual that $w >> h_0$, and as a consequence the inlet and outlet widths are practically equal; see Fig. 8–1. This *plane-strain* behavior results because of the constraint imposed upon the metal in the plastic zone by metal on each side of it (that is, the incoming and outlet metal). Thus, the change in thickness is

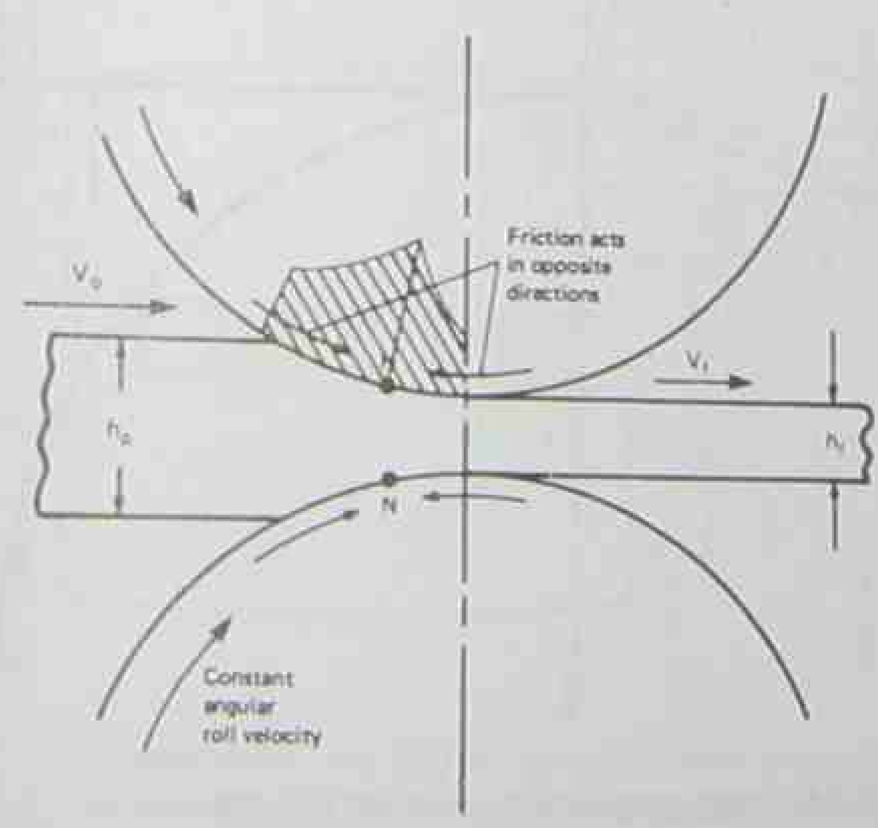


Figure 8–23 Schematic of flat rolling; note the neutral point N.

essentially accounted for by an increase in the length of the sheet, so $\epsilon_y \approx 0$ and $\epsilon_z = -\epsilon_x$. Note on Fig. 8–23 that there is one point of contact N in the deformation zone, which effectively separates the directions in which frictional effects act. It represents the point of maximum pressure of the friction hill. In addition, to the left of N (incoming), the *linear* velocity of the rolls is greater than that of the metal; once past N, the metal moves at a higher linear velocity than that of the rolls (which move at a constant *angular* velocity). N is called the *neutral point*. Of course, due to volume constancy, $V_0 h_0 = V_f h_f$.

The *average pressure* in rolling can be well approximated by

$$P_a = (h/\mu L)[\exp(\mu L/h) - 1](\sigma_p) \tag{8–34}$$

where from Fig. 8–24

$$L = (R\,\Delta h)^{0.5}$$

$$h = (1/2)(h_0 + h_f)$$

$$\Delta h = (h_0 - h_f)$$

$$\sigma_p = \text{the plane-strain flow stress (i.e., } 2k)$$

Often, to account for work hardening, σ_p is taken as the average of the flow stresses prior to and after deformation. In practical rolling operations, the strip is effectively pulled

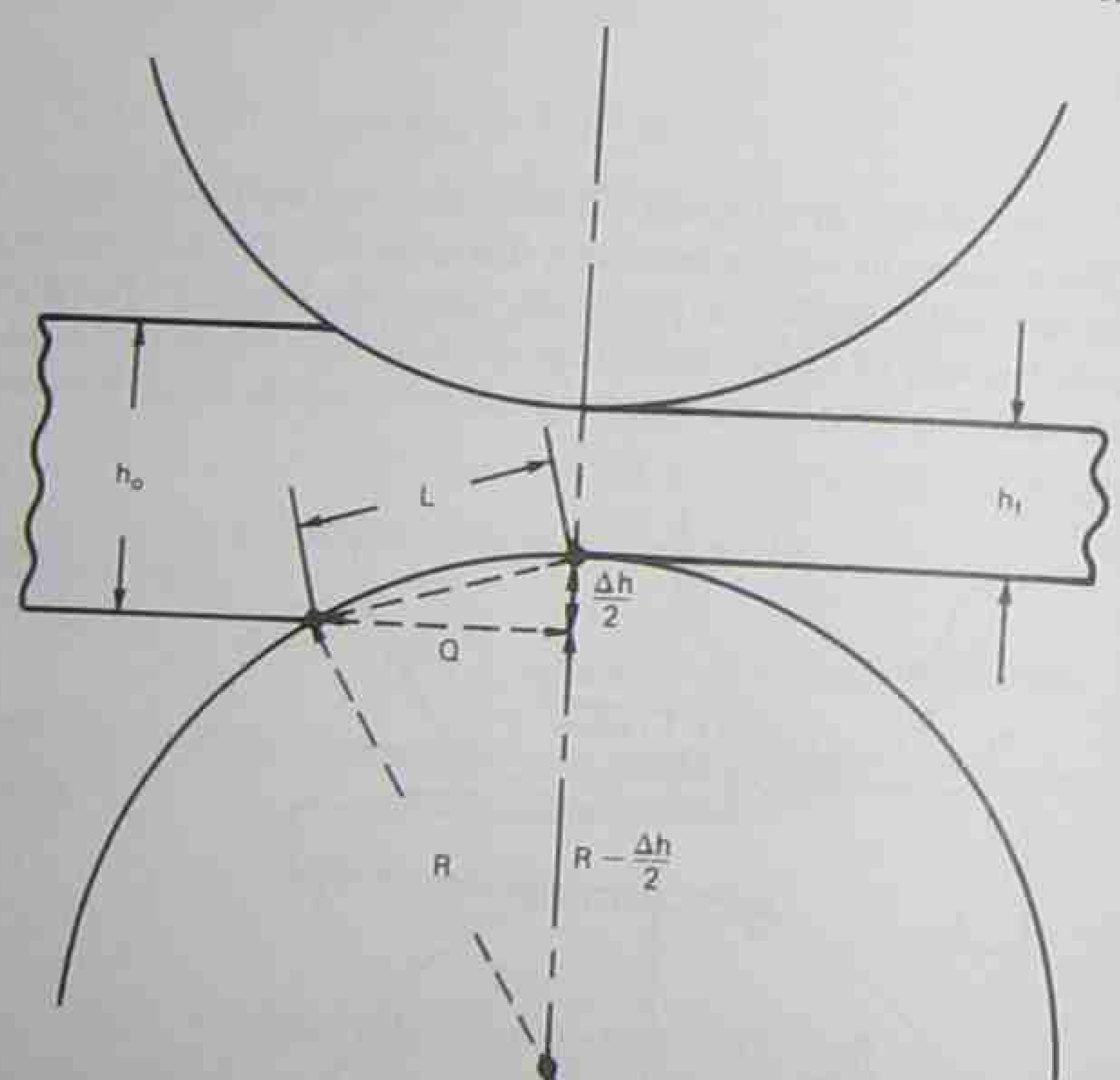


Figure 8–24 Dimensional relations in the roll gap.

through the rolls as its thickness is being decreased. Modifications to the above analysis that account for front (or also back) tension may be found in reference [1].

If P_a is multiplied by the contact area, effectively Lw, this leads to the force that tends to *separate* the rolls by elastic deformation. This force bends the rolls, causing the outlet (rolled) material to be thicker at the center than the edges. Roll *cambering*, shown in Fig. 8–25, is performed to reduce or eliminate such an effect (that is to roll sheet of more uniform cross section). Proper cambering must account for differences in material properties, so a rolling mill whose rolls are cambered correctly for aluminum will not necessarily produce a desired sheet of rolled steel. The major effects of either over- or undercambering are shown in Figs. 8–26 and 8–27. With this number of possible problems, it can be appreciated that rolling is not as simple an operation as it might first appear.

Example 8–5

A metal whose plane-strain flow stress σ_p is 20,000 psi is to be flat rolled from an inlet thickness of 0.125 in. to a thickness of 0.100 in. in a single pass using rolls of 10-in. diameter; the sheet is 18 in. wide, and spreading is negligible. If the average coefficient of friction in the roll gap is 0.08, estimate the roll separating force F_s.

Solution First determine the average pressure from Eq. (8–34), where

$$h = (0.125 + 0.100)/2 = 0.1125 \text{ in.}$$
$$\Delta h = (0.125 - 0.100) = 0.025 \text{ in.}$$
$$\text{so } L = (5 \times 0.025)^{0.5} = 0.354 \text{ in.}$$
$$P_a = [0.1125/(0.08)(0.354)][\exp[(0.08)(0.354)/0.1125] - 1]\ 20{,}000$$
$$\text{so } P_a = 22{,}740 \text{ psi}$$

The roll separating force $F_s = P_a wL$, so

$$F_s = 22{,}740(18)(0.354) = 145{,}000 \text{ lbf}$$

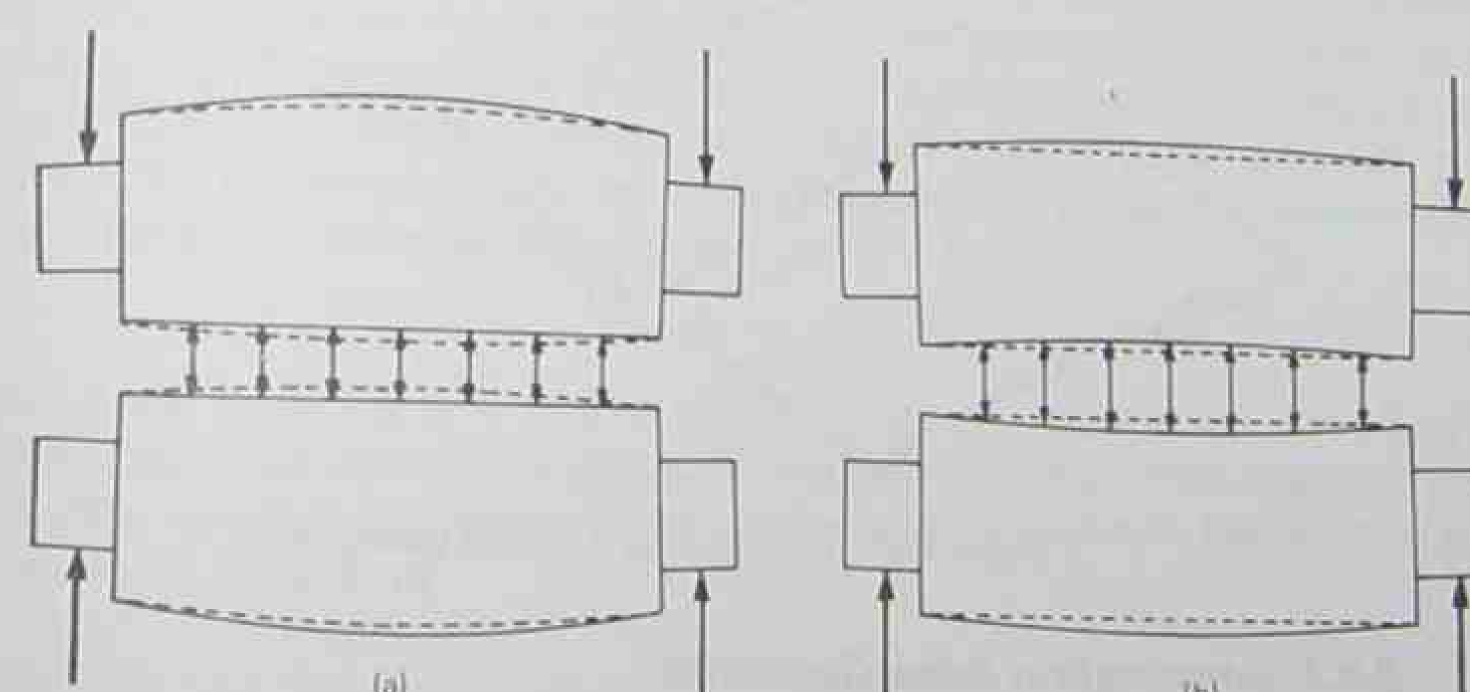


Figure 8–25 (a) Use of cambered rolls to compensate for roll bending and (b) the variation in thickness that occurs if uncambered rolls are used.

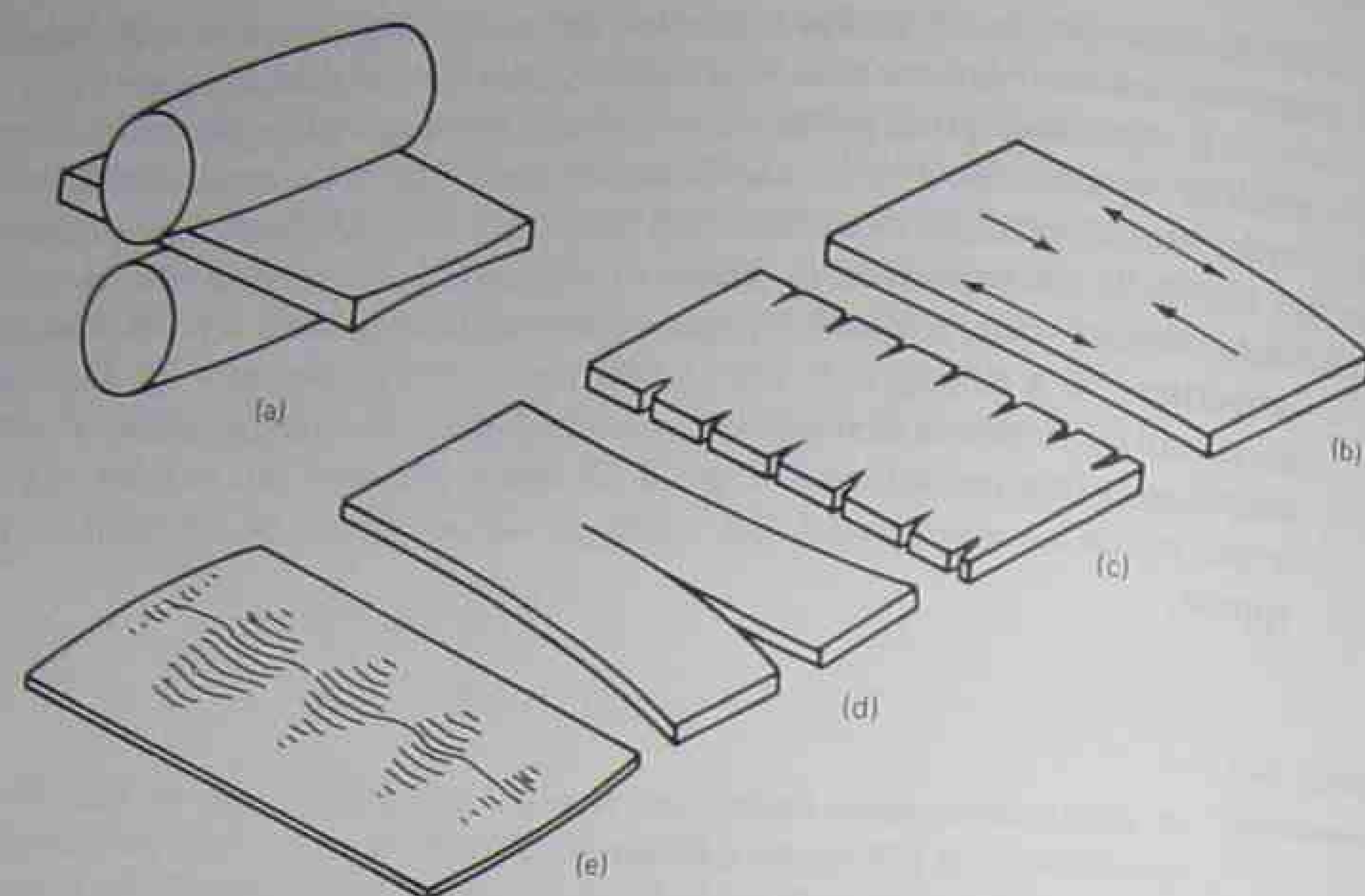


Figure 8–26 Possible effects if rolls are overcambered.

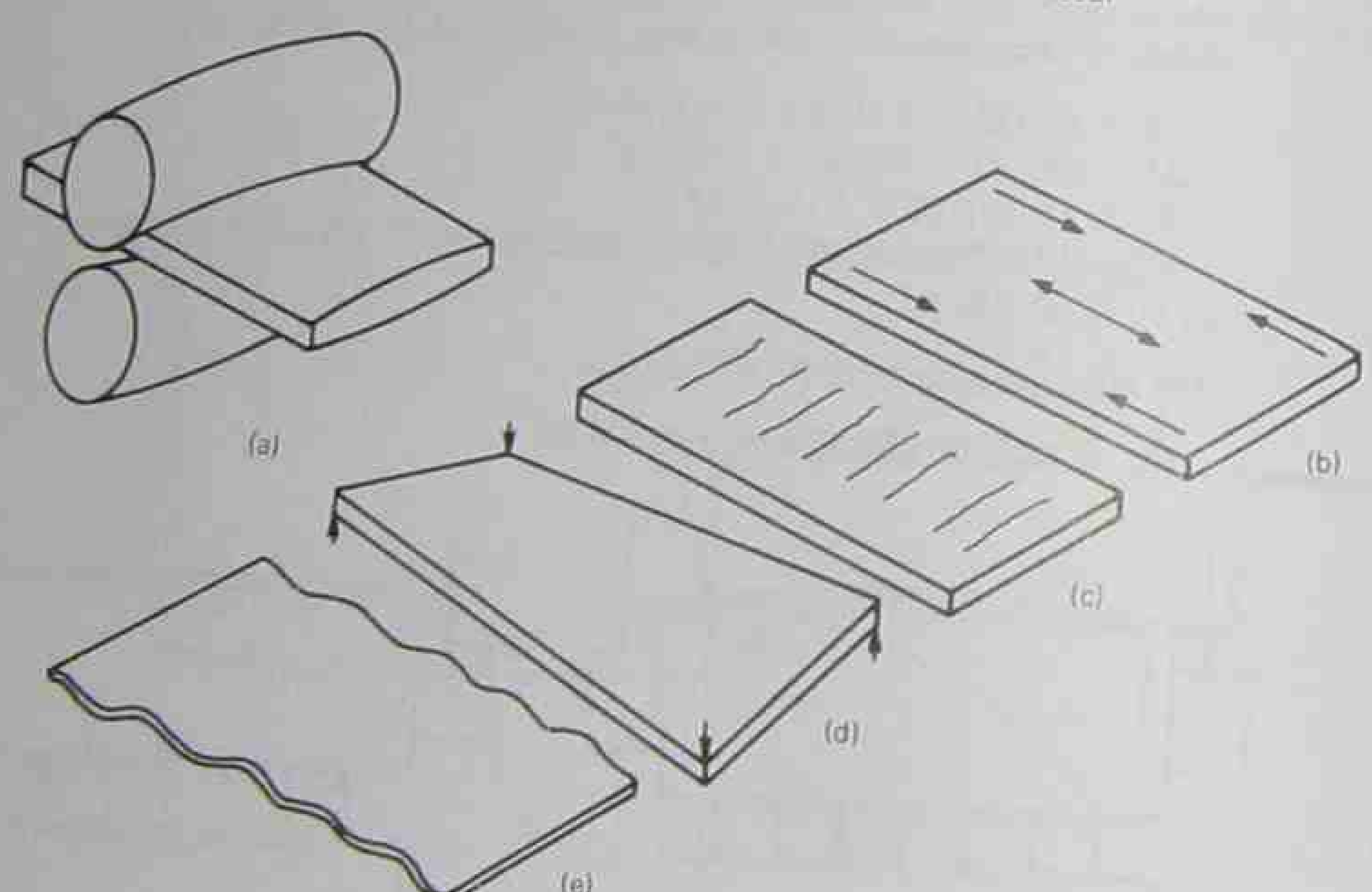


Figure 8–27 Possible effects if rolls are undercambered.

8.8.4 Deformation Zone Geometry

From previous sections, it was shown that the forces in certain bulk operations, as well as frictional and redundant effects, are all influenced by the *shape* of the deformation

zone, where in all instances the material was forced to *converge* before exiting in the desired shape. For such operations there is a simple and useful parameter which characterizes certain effects when the zone geometry is altered. It carries the symbol Δ and is defined as the ratio of the *mean* thickness *or* diameter h to the length of contact between work and tool in the zone of deformation; that is

$$\Delta \equiv h/L \tag{8–35}$$

For drawing or extrusion, $h = \frac{1}{2}(h_0 + h_f)$ and $L = (h_0 - h_f)/(2 \sin \alpha)$; thus

$$\Delta = (h_0 + h_f) \sin \alpha \,/(h_0 - h_f) \tag{8–36}$$

where α is the *semi-die angle*. If these are plane-strain operations, $r = (h_0 - h_f)/h_0$, so

$$\Delta = (2 - r) \sin \alpha/r \tag{8–37}$$

whereas for axisymmetric conditions, where h refers to *diameters*, that is, $r = (d_0^2 - d_f^2)/d_0^2$,

$$\Delta = \sin \alpha \,(1 + \sqrt{1 - r})^2/r \tag{8–38}$$

With rolling, and using the chordal length L (discussed in Sec. 8.8.3) instead of the arc of contact (this really causes little error, since the roll radius $R >> h$ in most cases), then we obtain

$$\Delta = (2 - r)(h_0/rR)^{0.5}/2 \tag{8–39}$$

where all parameters have been defined earlier. In a general sense, Δ is proportional to α/r, so small angles and large reductions tend to cause small values of Δ while the reverse holds for large α and small r. Earlier it was mentioned that redundant deformation was also influenced by α and r, so it is not surprising that such effects are tied together with Δ. Note that small values of Δ cause less redundant deformation, where unity is considered small. Perhaps the most direct method to infer redundant deformation in a quantitative way is illustrated by Fig. 8–28. There, the tensile behavior of a fully annealed specimen of stainless steel was first obtained. Then by drawing solid, round, annealed rods by different amounts, the yield strengths of the drawn rods were found; for the two cases shown, the *homogeneous* strains due to drawing were 0.090 and 0.422, respectively. Yet the strengths of the *drawn* rods implied actual strains of 0.185 and 0.500. If we consider the *redundant deformation factor* ϕ to be

$$\phi \equiv (\epsilon_r + \epsilon_h)/\epsilon_h \tag{8–40}$$

this would imply that $\phi = 0.185/0.090 = 2.5$ and $\phi = 0.500/0.422 = 1.18$ for these two cases. Using numerous combinations of α and r, and thus a wide range of Δ [via Eq. (8–38)], we obtain the results for three different metals shown on Fig. 8–29. A general expression of the form

$$\phi = C_1 + C_2\,\Delta \tag{8–41}$$

results, where the constants must be evaluated from experimentation. Note that ϕ is *never* less than unity, since that implies no redundancy, so there is a lower limit on such

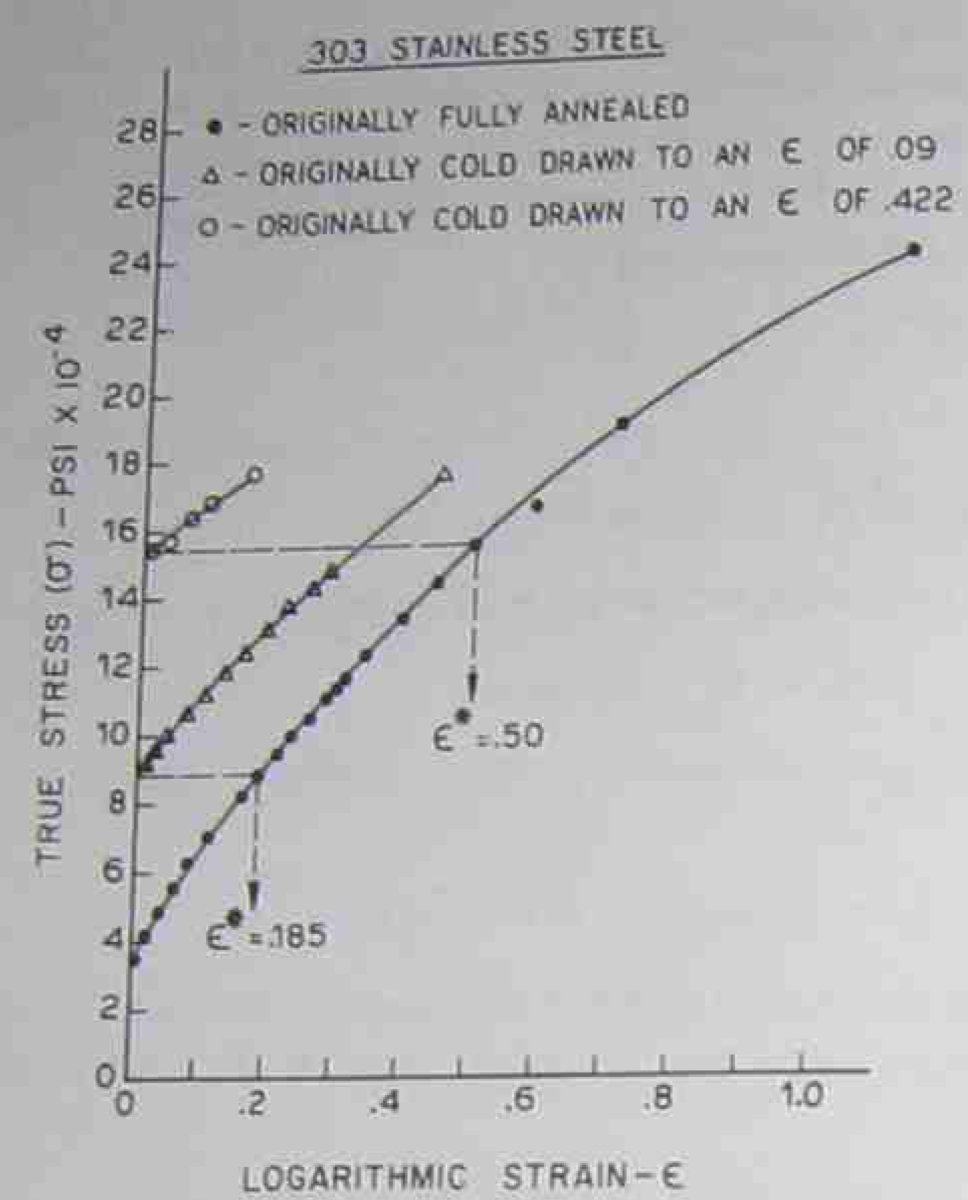


Figure 8–28 Stress-strain curves for 303 stainless steel in the annealed state and after cold-drawing.

equations. Other techniques have also been used to assess the ϕ-Δ relationship, and any user of such equations must consider how they were determined, since geometric differences and experimental techniques can lead to somewhat different forms of Eq. (8–41). More importantly, values for the empirical constants may differ. In the study related to Fig. 8–28, the empirical equations for the three metals used are

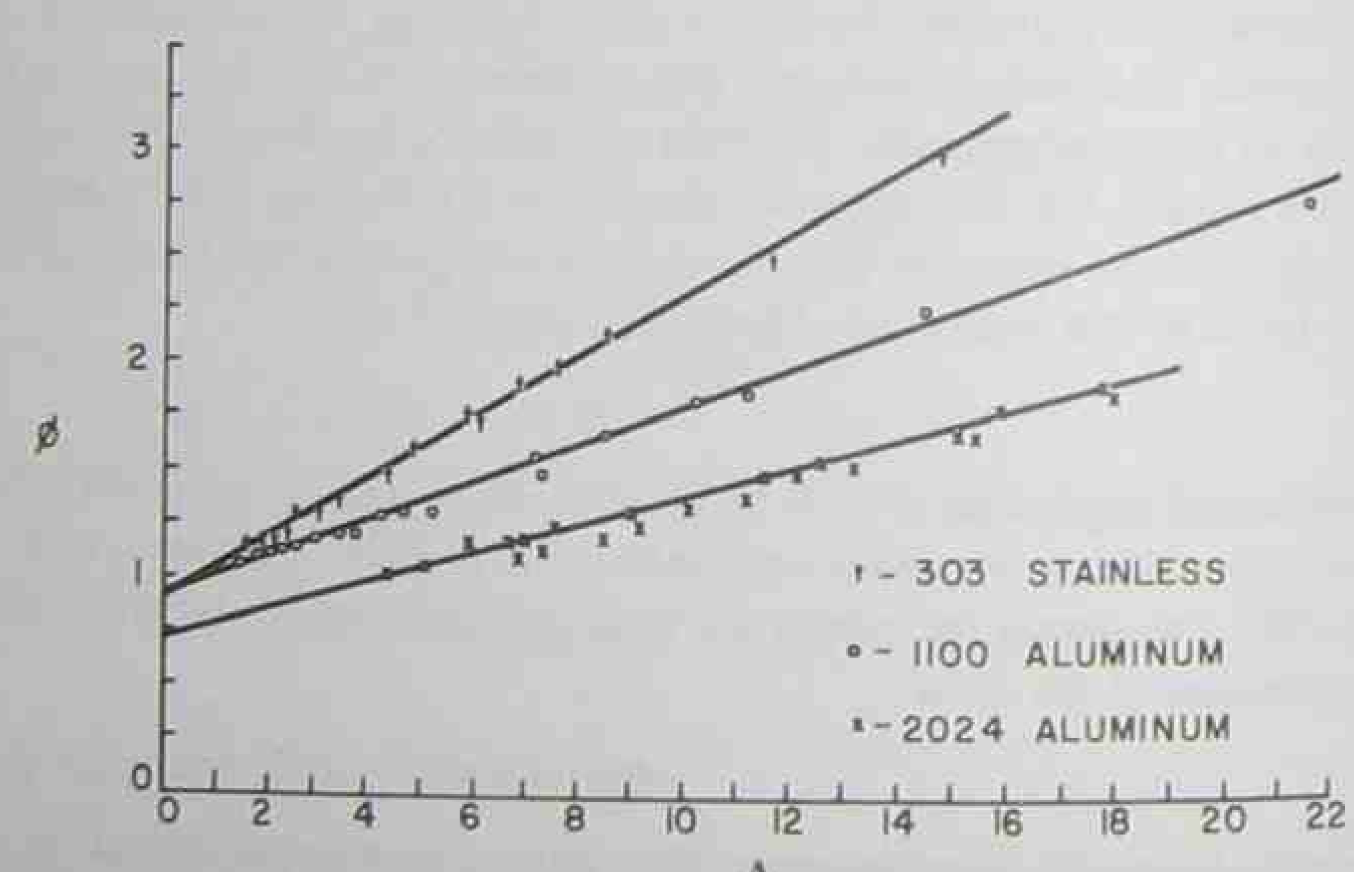


Figure 8–29 Influence of Δ on the redundant strain factor in cold rolling various metals.

303 stainless steel $C_1 = 0.87,\quad C_2 = 0.15$ (8–42)

1100 aluminum $C_1 = 0.89,\quad C_2 = 0.092$ (8–43)

2024 aluminum $C_1 = 0.72,\quad C_2 = 0.067$ (8–44)

Reference [4] gives the details.

In another study [5], using hardness measurements, the influence of Δ on ϕ was determined for both strip and wire drawing. There, Δ was found from Eqs. (8–36) through (8–38), except that α (in radians) replaces sin α. Unless α is large, these forms are quite similar. Those results gave

$$\phi = 1 + C(\Delta - 1) \quad \text{for } \Delta \geq 1 \quad \text{and} \quad \phi = 1 \text{ for } \Delta \leq 1 \tag{8–45}$$

where $C = 0.21$ for plane-strain and $C = 0.12$ for axisymmetric drawing.

Example 8–6

A one-inch round of commercially pure aluminum is extruded to a diameter of ½ inch through a die of 8 deg semi-angle. Is redundant deformation likely to be significant?

Solution First find Δ, using Eq. (8–38).

$$\Delta = \sin\alpha\,(1 + \sqrt{1 - r})^2/r$$

where $r = (1^2 - 0.5^2)/1^2 = 0.75$

$$\Delta = \sin\alpha\,(1 + \sqrt{1 - 0.75})^2/0.75 = 0.42$$

With Eqs. (8–41) and (8–43),

$$\phi = 0.89 + 0.092(0.42) = 0.93$$

Also, from Eq. (8–45) since $\Delta < 1$, $\phi = 1$; therefore, redundant deformation is negligible. Note that on physical grounds ϕ can never be less than one, so the value of 0.93 merely implies uniform straining.

The effect of redundant deformation can be seen in hardness variations, Fig. 8–30; internal damage illustrated by density variations, Fig. 8–31; residual stresses, Fig. 8–32; finally, severe internal cracking or fracture, Figs. 8–33 and 8–34. Rolling under high Δ conditions can also lead to *alligatoring*, as shown in Fig. 8–35.

8.8.5 Formability

Although force calculations are often useful, the most important concern in bulk forming is whether the desired deformation can be performed without cracking or fracturing the workpiece. Under equivalent Δ conditions, some of the consequences discussed in the previous section will occur with some metals but not others. In addition, the strains that induce failure in a single metal often depend upon the process employed. Finally, processing variables, such as geometric differences, can be important.

Results such as those shown in Fig. 8–36 indicate that in some cases, fracture strains correlate quite well with the value of ϵ_f based upon the value of A_r from a tensile test.

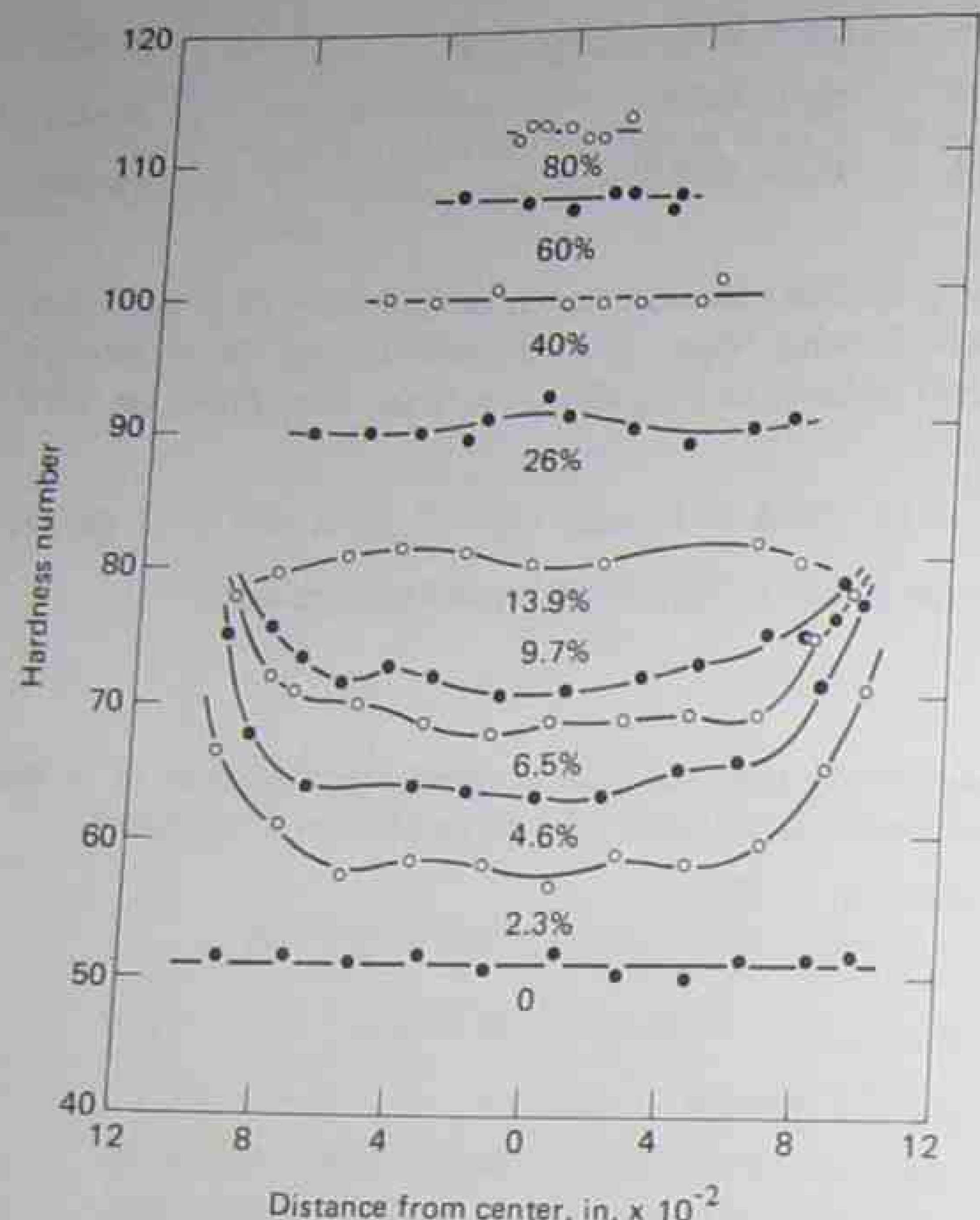


Figure 8–30 Hardness gradients in copper after cold rolling in one pass to the reductions indicated. Initial strip thickness was 0.2 in. and the roll diameter was 10 in.

However, differences in processing variables (for example, square versus round-edged strips) are also important, as that figure shows. Proper treatment of ductile fracture involves both strain and stress, as discussed in Ref. [6].

To discuss in *detail* the influence of metallurgical factors, such as the size, shape, and type of inclusions, is beyond the intent of this text; they are important, and a coverage of this can be found elsewhere [1].

The preferential alignment of grains and inclusions caused by the forming operation (this is called *texturing*) affects both the flow and fracture behavior of the formed workpiece. For example, in flat rolling, extension of the internal structure occurs in the rolling direction, and the *shape* of the grains and inclusions can change quite drastically. It is important to specify whether one is concerned with certain properties of the rolled sheet *or* fracture behavior. It is usual to expect that a cold-rolled sheet will exhibit greater *strains* to fracture if subsequent loading is parallel to the rolling direction as compared to testing *transverse* (that is, perpendicular) to the rolling direction.

The picture is not so clear in terms of mechanical properties. If tensile tests are conducted *after* rolling, one metal may display a higher yield strength in the direction of rolling than in the transverse direction, but the exact *opposite* behavior has been found in other cases.

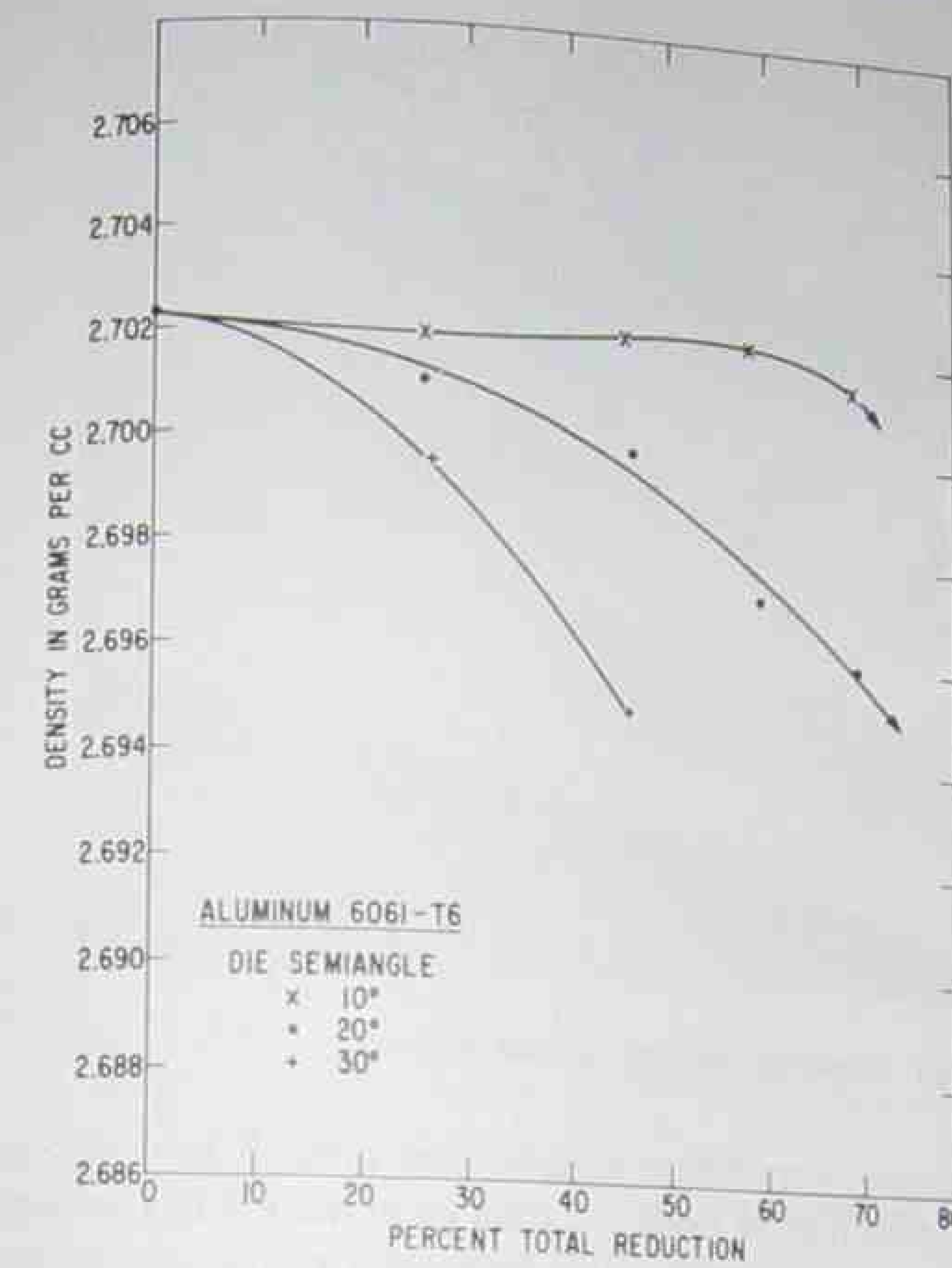


Figure 8–31 Density changes in 6061–T6 aluminum caused by drawing. Note the greater loss of density at high die angles.

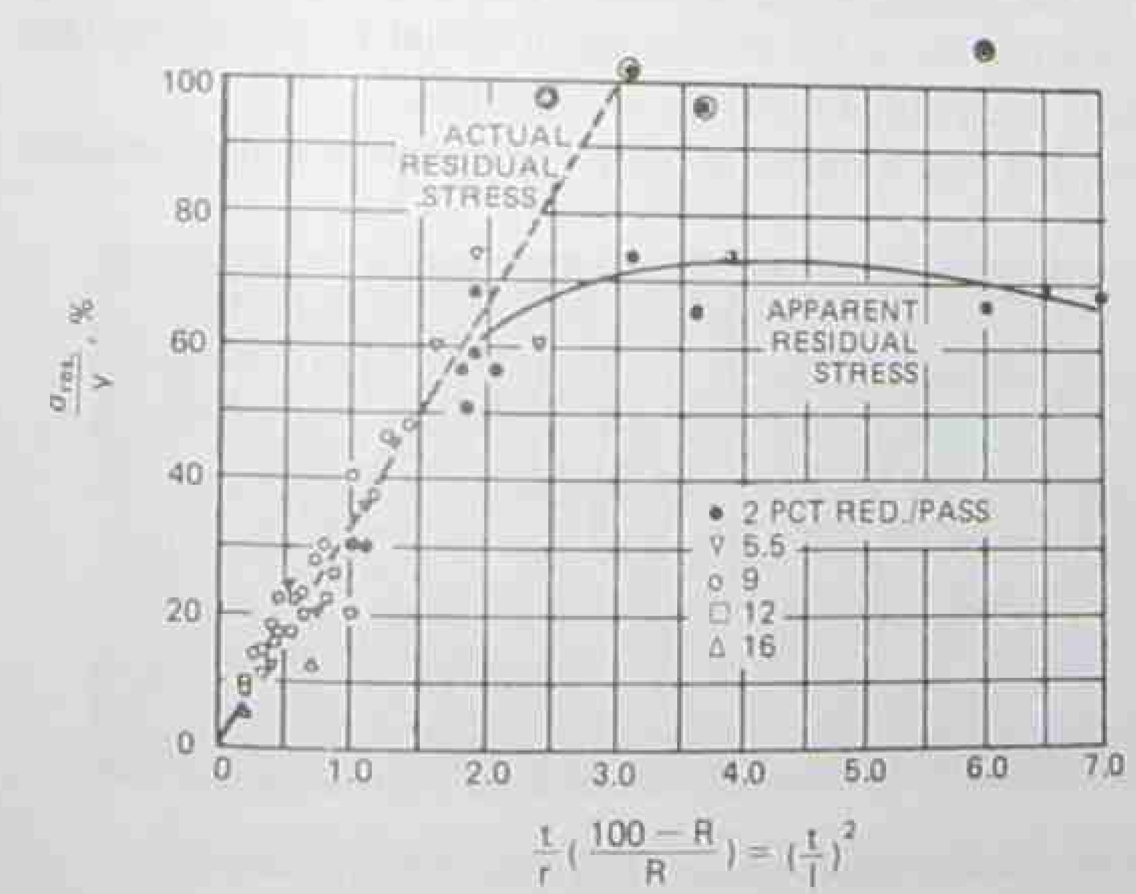


Figure 8–32 Residual stresses at the surface of rolled strip. The residual stresses are normalized by the yield strength, and the abscissa is Δ^2.

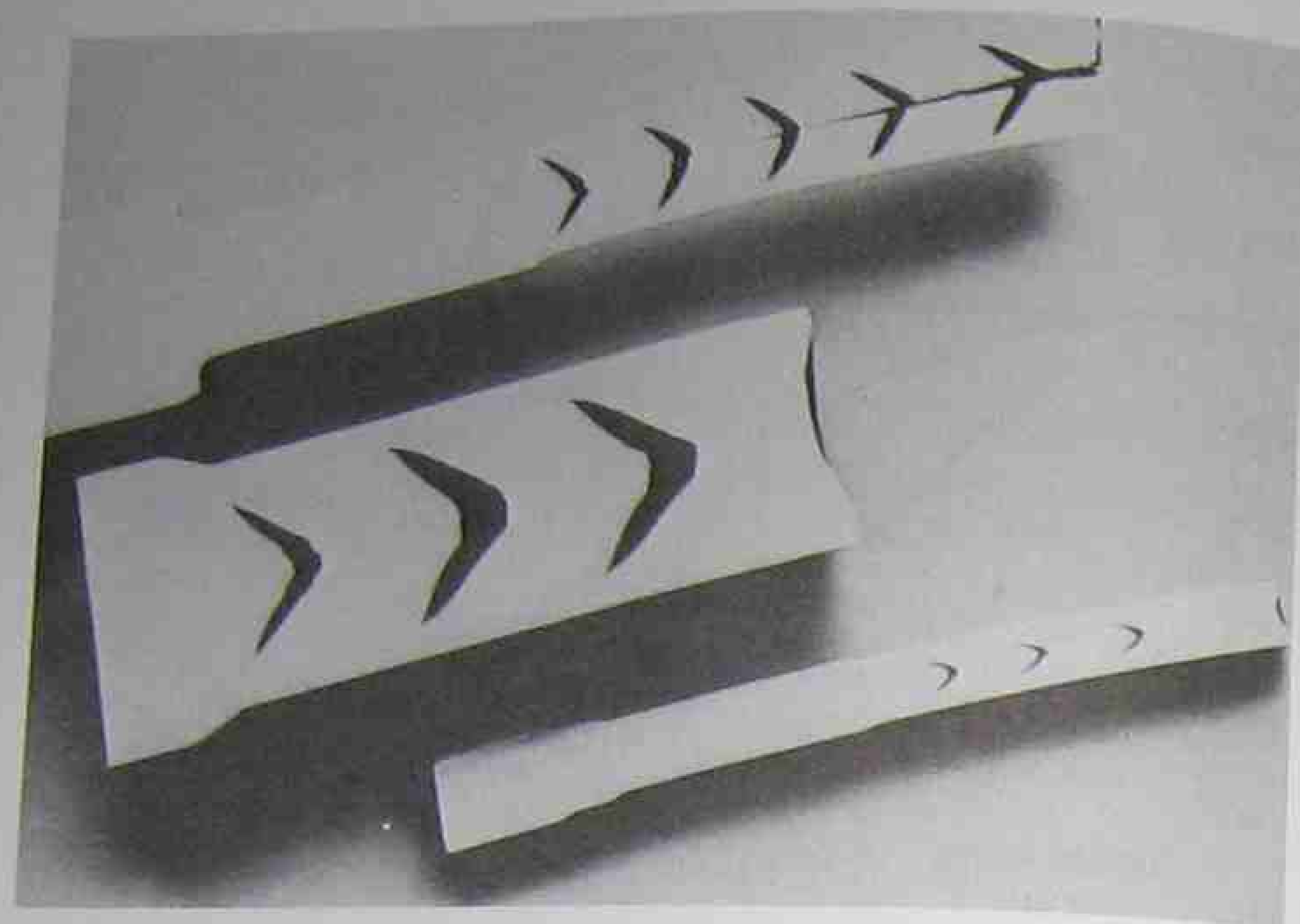


Figure 8-33 Centerline cracks in extruded steel rod.

One definite way to improve formability, though often difficult to do so, is to decrease the magnitude of the mean normal stress, σ_m (that is, to make it more *compressive*). A practical operation where this is done is called *hydrostatic extrusion*. The design of such equipment permits surrounding the workpiece with a high fluid pressure; this tends to *close up* existing voids and to suppress their growth during deformation. Figures 8–37 and 8–38 illustrate two sets of findings. In the first figure, an obvious *increase* in ϵ_f results as σ_m becomes more negative; the second figure illustrates that the workpiece *density* increases with pressure, undoubtedly due to the closing up of initial voids. Hydrostatic extrusion is especially beneficial when brittle materials are involved.

It is unfortunate, but true, that bulk formability cannot be assessed simply in terms

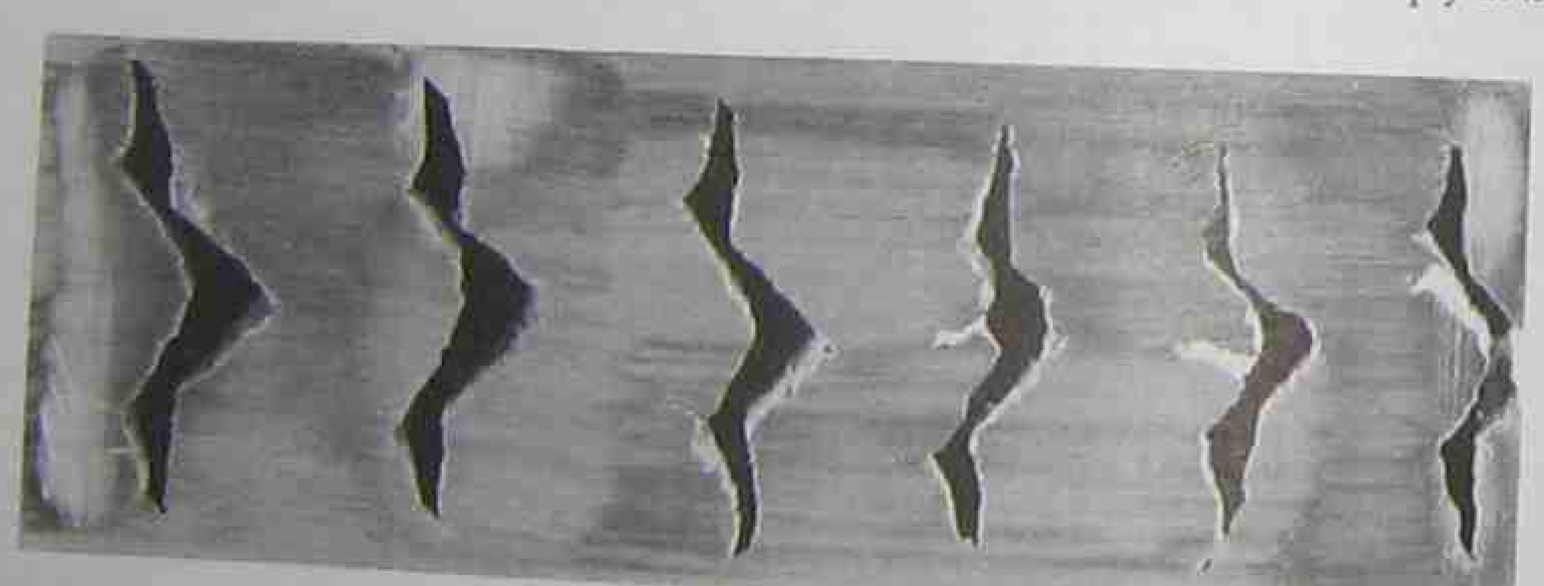


Figure 8–34 Cracks formed in a molybdenum bar cold rolled under high-Δ conditions.

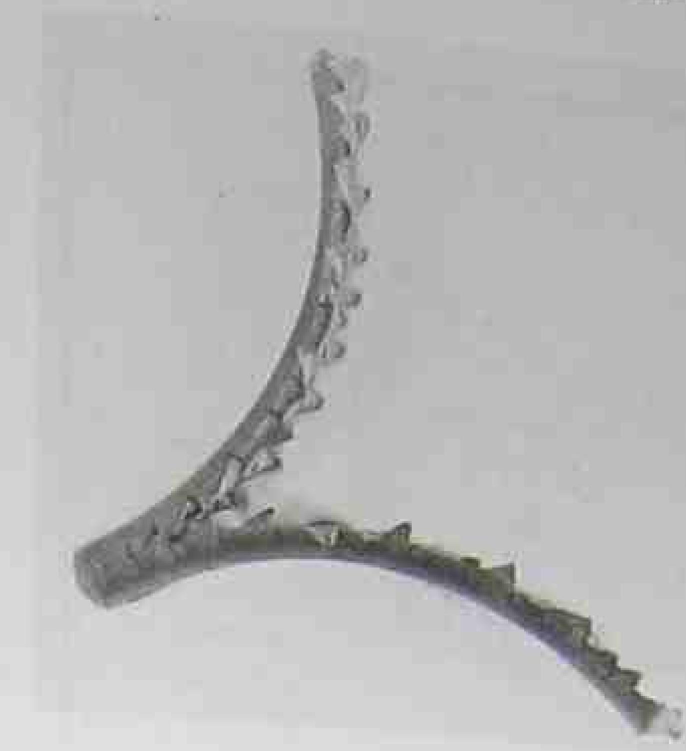


Figure 8–35 Molybdenum rod that "alligatored" from rolling under high-Δ conditions. The "teeth," caused by edge cracks, are not typical.

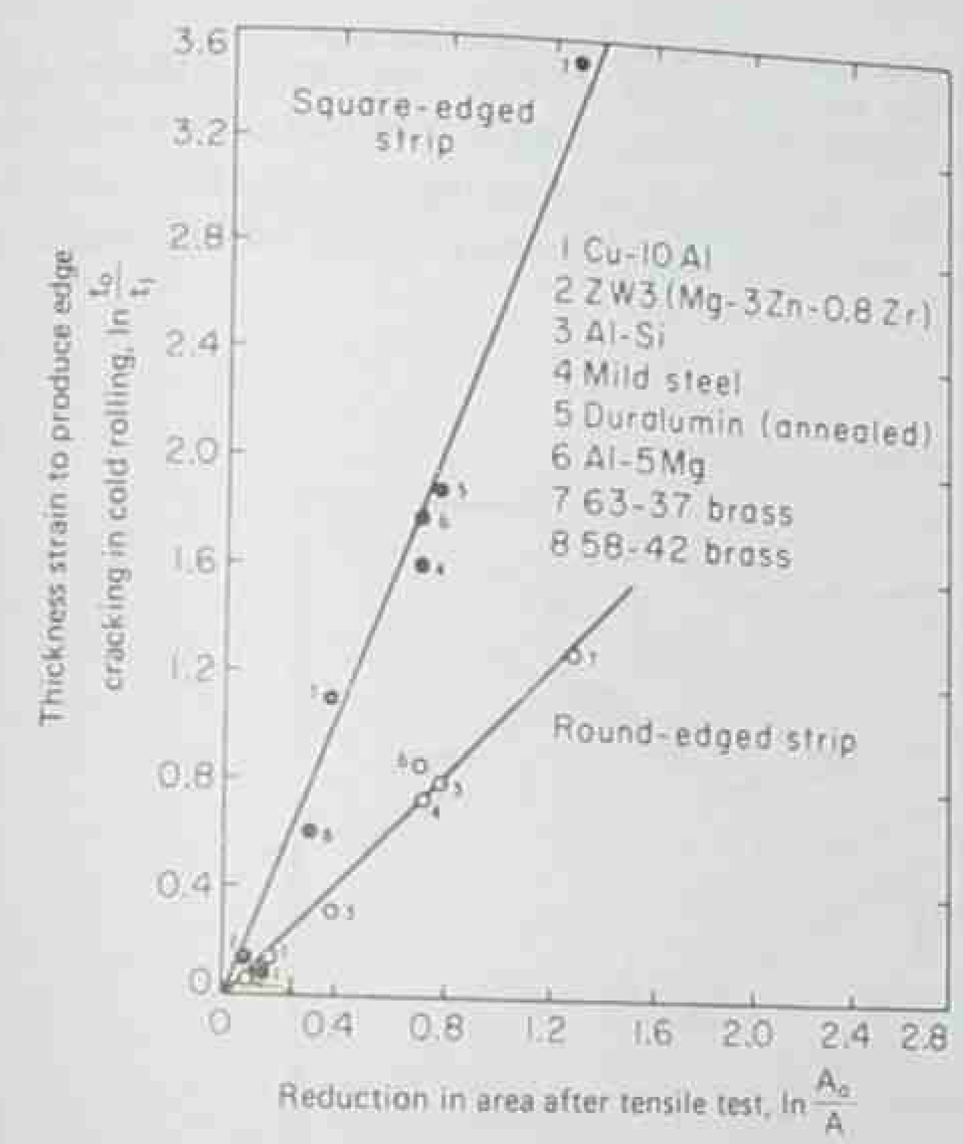


Figure 8–36 Correlation of strain at which edge cracking occurred in flat rolling with reduction of area in a tension test.

of A_r or ϵ_f found from a tension test. As one alternative, compression of solid cylinders has been used. Figures 8–39 and 8–40 illustrate the type of results from such tests (those near $\epsilon_2 = 0$ were found by bending wide specimens to fracture). In the first of those figures, the values of ϵ_1 (circumferential) and ϵ_2 (axial) were found after a barrelled specimen had fractured. Different fracture strain combinations resulted by altering the starting height-to-diameter ratio or the lubrication at the contact surfaces. The second figure illustrates how anisotropy can influence fracture.

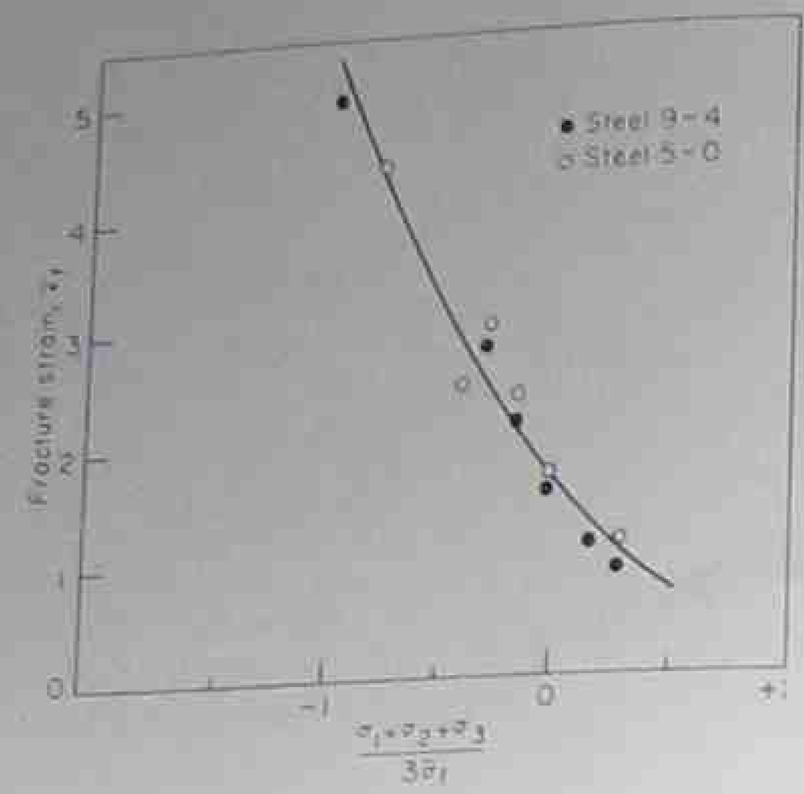


Figure 8–37 Correlation of fracture strain with ratio of the mean normal stress to effective stress.

8.9 SHEET FORMING

Products manufactured from thin metallic sheet usually experience different types of deformation in various regions during processing. That is, bending may predominate in one section, stretching elsewhere, and drawing in other regions. In some cases a combination of deformation modes may occur simultaneously during various stages of the process.

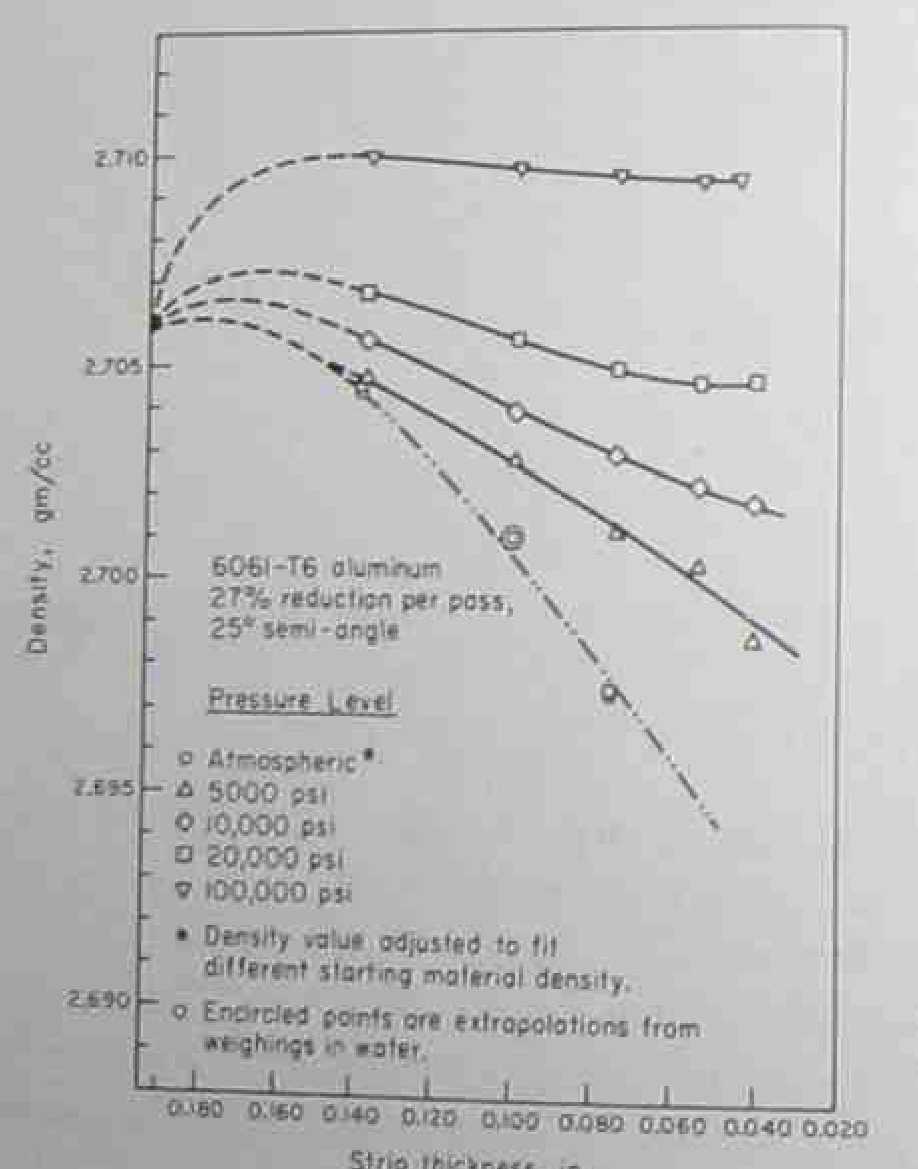


Figure 8–38 Loss of density during strip drawing and the effect of superimposed hydrostatic pressure on diminishing density loss.

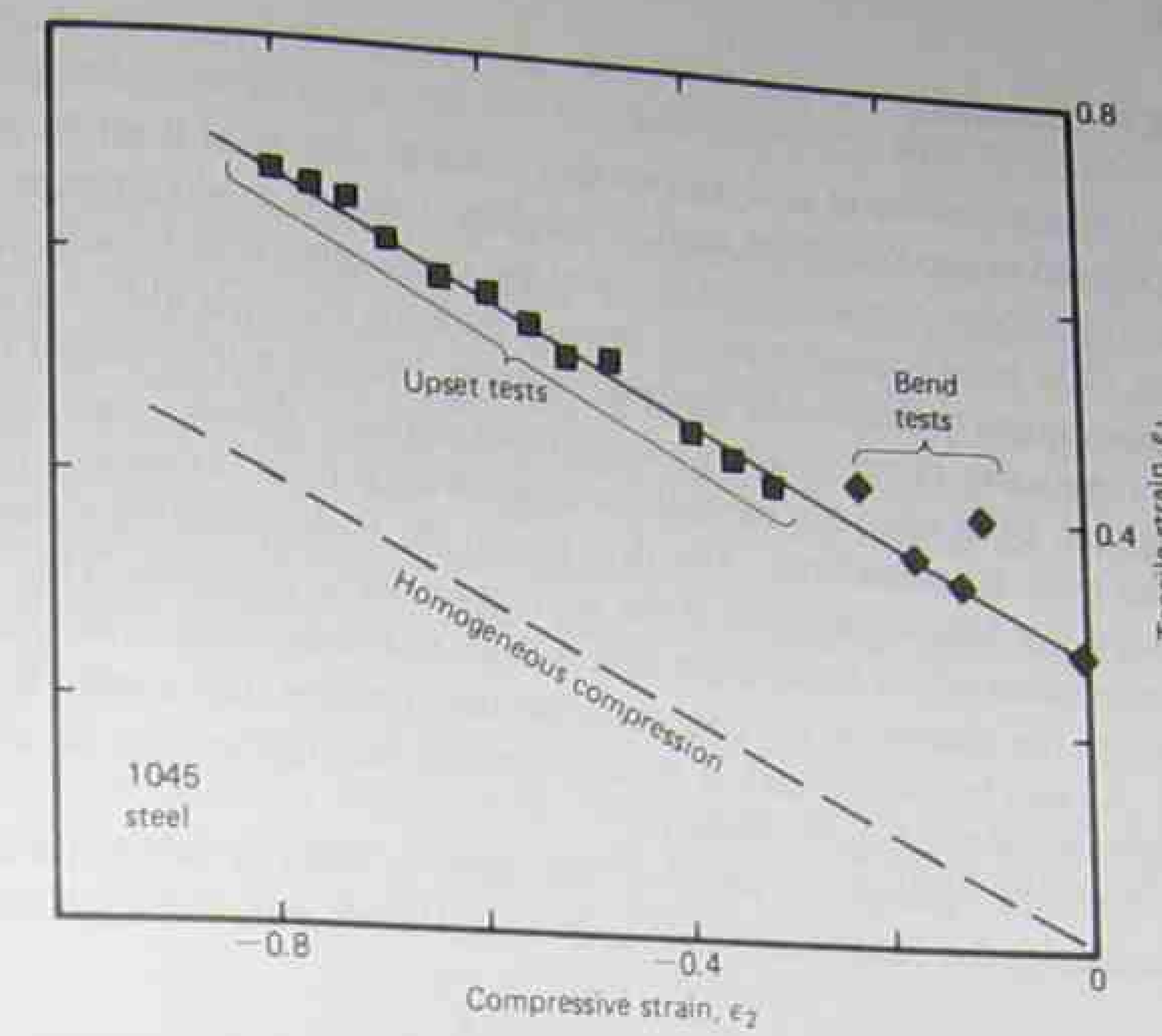


Figure 8–39 Forming limits defined by strains at fracture in upset and formability tests.

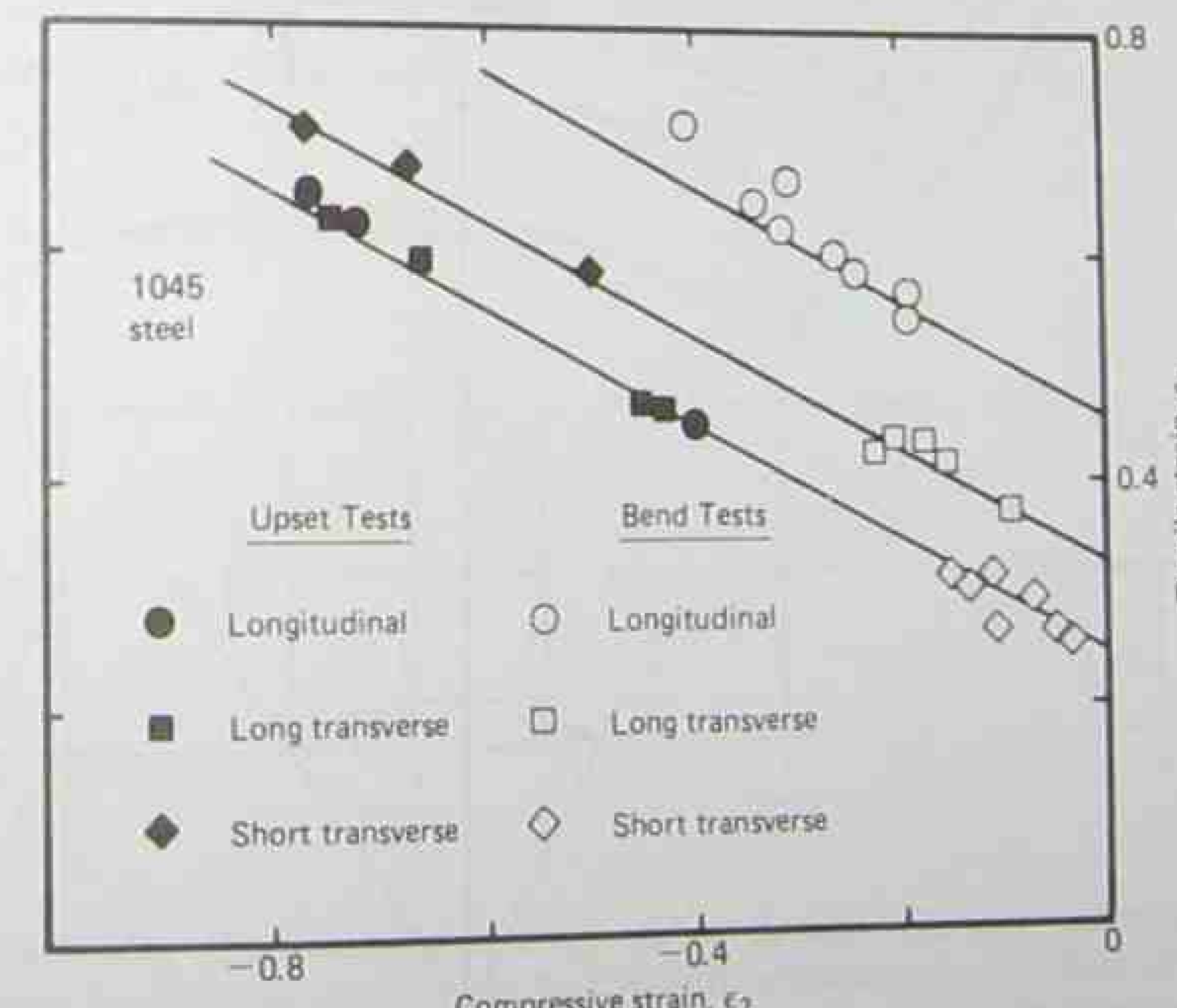


Figure 8–40 Effect of specimen orientation on forming limits.

8.9.1 Bending

The simplest process to analyze (yet this is *not* as simple as it might appear) is bending; this is used to produce items such as brackets, cabinets, and furniture components. As a thin strip is bent, tensile strains occur at the outer surface and compressive strains at the inner. If the bend is too severe, the increased tensile strains lead to *fracture;* this is usually the governing constraint. On occasion, as with bending a thin-walled tube, the compressive strains can cause buckling of the inner surface, thereby causing *failure* from that viewpoint. Even when neither of these limiting conditions is reached, two other important results must be considered. The first has to do with *springback,* which is the elastic recovery that follows removal of the applied forces or moments. This obviously leads to a final product whose bend angle differs from that under load. The second consideration relates to the change in the final cross-sectional *shape* that results in the region of the bend. As will be shown later, a section initially rectangular may be *decidedly* altered as the severity of bending is increased.

Consider springback where a thin sheet or strip is subjected to a pure bending moment using the designations in Fig. 8–41. Here any shift of the neutral axis is neglected

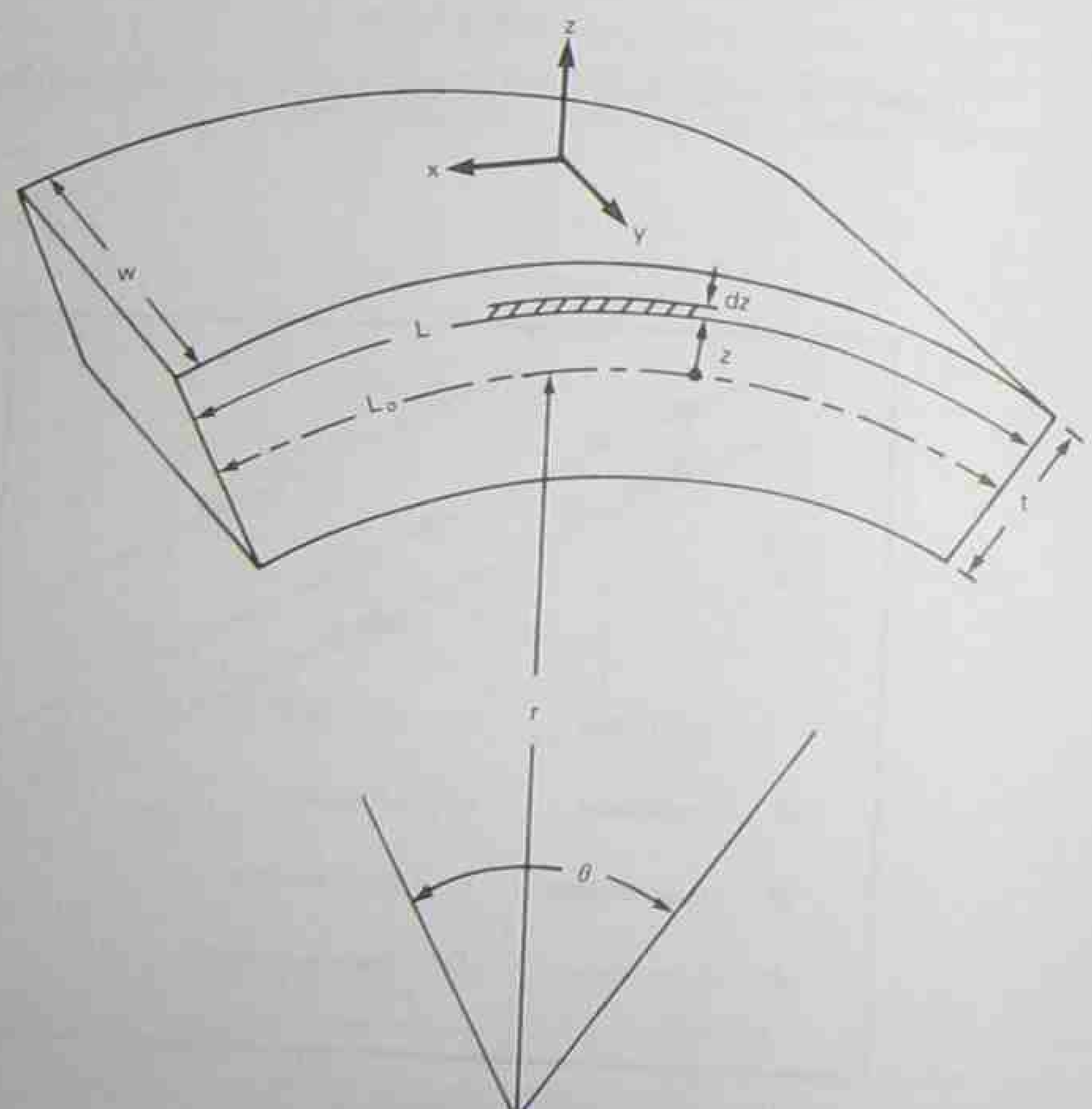


Figure 8–41 Coordinate system for analysis in bending.

and any effects of work hardening are ignored for now; in addition, if $w >> t$, $\epsilon_y = 0$, so a condition of plane strain exists.* If the bend is modest, such that $r >> t$, then

$$\epsilon_x = t/2r \tag{8–46}$$

at the outer fibers (tensile on outside and compressive on the inside) as shown in Fig. 8–42(a). For a *nonwork hardening* metal, whose stress-strain behavior is illustrated in Fig. 8–42(b), a stress distribution *under* load (plastic deformation is assumed) is given in Fig. 8–42(c). Note that the assumption of plane strain means that the flow stress σ_p equals $(\sqrt{4/3})\,Y$ if Y is the tensile yield stress [check this with Eq. (8–3)]. Thus the full section is at a stress $\sigma_x = \pm\sigma_p$, except for the elastic *core* at the center. This core becomes smaller as the bend becomes more severe.

The relation between the radius under load and after unloading, as taken from reference [1],

$$(1/r) - (1/r') = 3\sigma_p/tE' \tag{8–47}$$

where r is the radius under load, r' is the radius *after* springback, t is the thickness, and E' is the *plane-strain modulus,* which equals $E/(1 - \nu^2)$. The resulting residual stress distribution, shown in Fig. 8–42(d), comes from

$$\sigma_{x'} = \sigma_P(1 - [3z/t]) \tag{8–48}$$

noting it is compressive $(-\sigma_p/2)$ at the outside and tensile $(+\sigma_p/2)$ at the inside. If work hardening is considered and $\bar{\sigma} = K\bar{\epsilon}^n$ describes such behavior, equations similar to Eqs. (8–47) and (8–48) are given by

$$(1/r) - (1/r') = (6/[2 + n])(K'/E')(t/[2r])^n(1/t) \tag{8–49}$$

and

$$\sigma_{x'} = K'(z/r)^n[1 - (3/(2 + n))(2z/t)^{1-n}] \tag{8–50}$$

where $K' = K(4/3)^{(n+1)/2}$.

* At the free surfaces, plane stress prevails.

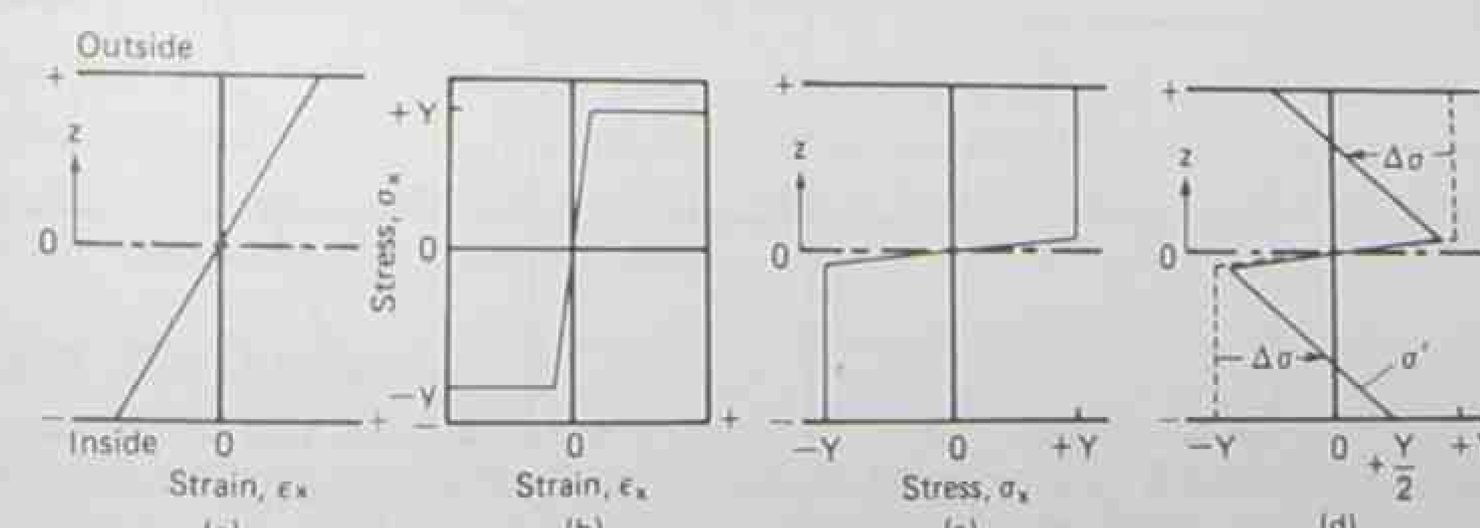


Figure 8–42 Strain and stress distribution across the sheet thickness. Bending strain varies linearly across the section (a). For a nonwork-hardening stress-strain behavior (b), the bending moment causes the stress distribution in (c). The residual stress distribution (d) results from elastic springback.

Figure 8–43 illustrates the residual stress distribution for a work hardening metal. Note that the degree of springback predicted by Eqs. (8–47) and (8–49) can be very large.

A useful method to reduce springback is to apply tensile forces simultaneously with bending; this is typical of *stretch forming*. Owing to the tensile stresses, the neutral axis moves towards the inner surface and, with sufficient tension, the neutral axis moves out of the sheet and the entire cross section yields in tension. Figures 8–44(a) through 8–44(d) illustrate various conditions when one is bending with superimposed tension using a work hardening metal. An analysis of springback can be simplified by approximating the stress-strain curve in the region of loading by a straight line of slope $d\sigma/d\epsilon$. This results in

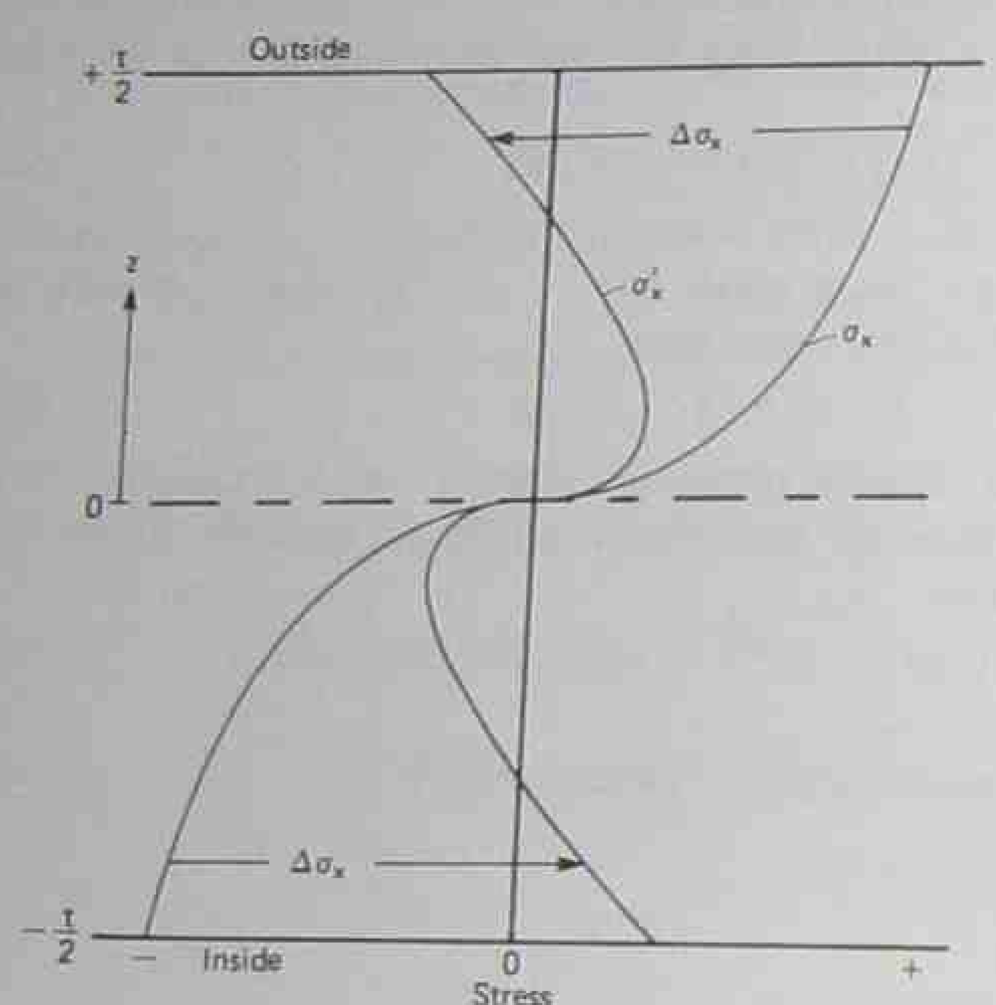


Figure 8–43 Stress distribution under bending effects and after unloading for a work-hardening material.

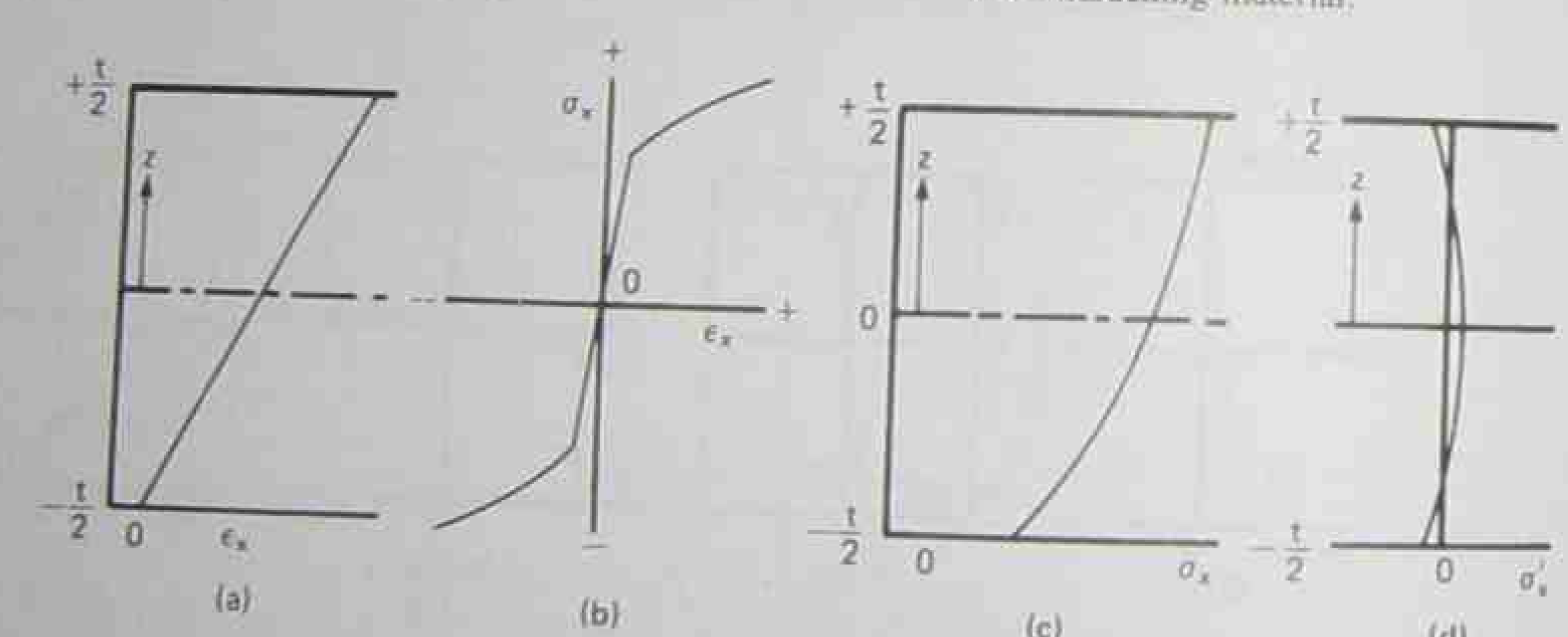


Figure 8–44 Bending with superimposed tension. With sufficient tension, the neutral axis moves out of the sheet so the strain is tensile across the full section (a). With the σ-ϵ curve in (b), the stress distribution in (c) results. Elastic unloading leaves minor residual stresses, as seen in (d).

$$r'/r = [1 - (d\bar{\sigma}/d\bar{\epsilon})/E']^{-1} \quad (8\text{–}51)$$

Because $d\bar{\sigma}/d\bar{\epsilon} << E'$, springback is greatly reduced.

Example 8–7

A piece of steel whose flow stress is 45 ksi is to be bent to a final radius r' of 10 in.; the piece is initially 0.040 in. thick. What tool radius is needed:

a. If pure bending is used?

b. If tension is applied and induces a net tensile strain of 0.025 at the centerline? Assume that $\bar{\sigma} = 100{,}000\,\bar{\epsilon}^{0.2}$ pertains.

Solution a. Using Eq. (8–47), with $r' = 10$ in., $\sigma_p = 45$ ksi, $t = 0.040$ in., and $E' = E/(1 - \upsilon^2) = 30{,}000$ ksi$/(1 - 0.3^2) = 33{,}000$ ksi, we obtain

$$\frac{1}{r} - \frac{1}{10} = \frac{3(45)}{0.04(33{,}000)} = 0.1023$$

so $1/r = 0.2023$ or $r = 4.9$ in.

b. Using Eq. (8–51), with $d\bar{\sigma}/d\bar{\epsilon} = nK\epsilon^{n-1} = 0.2(100)(0.025)^{-0.8}$ (note that K has units of ksi here), we find that $d\bar{\sigma}/d\bar{\epsilon} = 383$ so

$$\frac{10}{r} = (1 - \frac{383}{33{,}000})^{-1} = 1.013$$

therefore, $r = 9.87$ in.

This requires a tool radius that causes much less bending than in (a) and would ease the problem accordingly.

One study concerned with the minimum possible bend radius short of fracture on the outer fibers showed a reasonable correlation between sheet thickness t, percent reduction of area at fracture in tension A_r, and the radius at the inside of the bend R; this was determined by using an externally applied moment only. Results are shown in Fig. 8–45, the solid line being expressed by

$$R/t = (50/A_r) - 1 \quad (8\text{–}52)$$

Note that if $A_r > 50$ percent this analysis implies the sheet coud be bent to a zero radius without fracturing (that is, r cannot be physically less than zero). Figure 8–46 shows a ½ in. by ¼ in. bar of annealed aluminum bent to fracture; notice the final cross section, which was rectangular prior to bending.

8.9.2 The R-Value

With sheet-forming operations more complicated than bending, a material parameter that finds particular use is called the *R*-value or *strain ratio*. *R* relates to anisotropy of mechanical properties and is determined from the three orthogonal strains in a flat tensile specimen taken from the plane of the sheet, that is, from

$$\epsilon_\ell = \ell n(\ell/\ell_0), \quad \epsilon_w = \ell n(w/w_0), \quad \text{and} \quad \epsilon_t = \ell n(t/t_0) \quad (8\text{–}53)$$

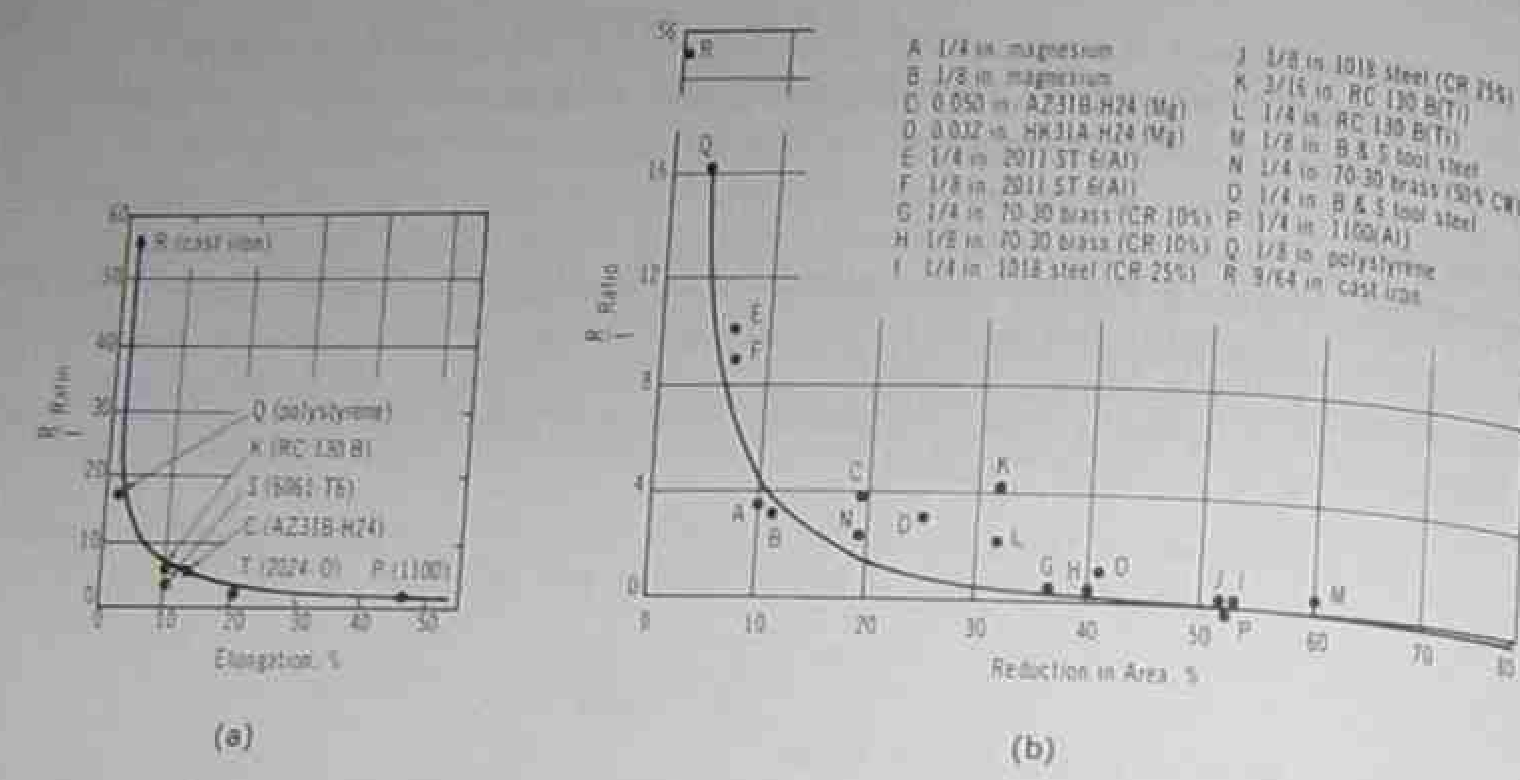


Figure 8–45 Correlation of limiting bend severity R/t, with tensile ductility.

where ℓ, w, and t refer to longitudinal, width, and thickness directions, respectively; Fig. 8–47 shows such a specimen. Now,

$$R = \epsilon_w/\epsilon_t = \epsilon_w/-(\epsilon_\ell + \epsilon_w) \tag{8–54}*$$

and since $\epsilon_\ell + \epsilon_w + \epsilon_t = 0$, R is a positive number. The physical significance is that if $R > 1$ the material is more resistant to thinning than to deformation in the plane of the sheet; the reverse is true if $R < 1$. If the sheet is truly isotropic, $R = 1$. For most ductile metals it is found that R varies with direction of testing, and the most common practice is to define an average value designated $\overline{R}$. Here,

$$\overline{R} = (R_0 + R_{90} + 2R_{45})/4 \tag{8–55}$$

where the subscript $_0$ refers to the rolling direction, $_{90}$ to the transverse, and $_{45}$ to the angle midway between the others. Observe the following:

1. If $R = 1$, the metal is isotropic.
2. If $R \neq 1$, but it is constant regardless of the angle at which the tensile specimen is made, the sheet is said to have *planar isotropy* but *normal anisotropy*.
3. If $R \neq 1$, and it varies with orientation in the plane, then the sheet exhibits both planar *and* normal anisotropy.
4. Measurements of the yield strength Y may often show little variation with orientation, and one might conclude that the sheet displays isotropy. This can be erroneous, since in these circumstances R might not be unity.

* Strictly, $R = d\epsilon_w/d\epsilon_t$ and may change with continual deformation. We assume that this ratio remains constant, so total strains are used.

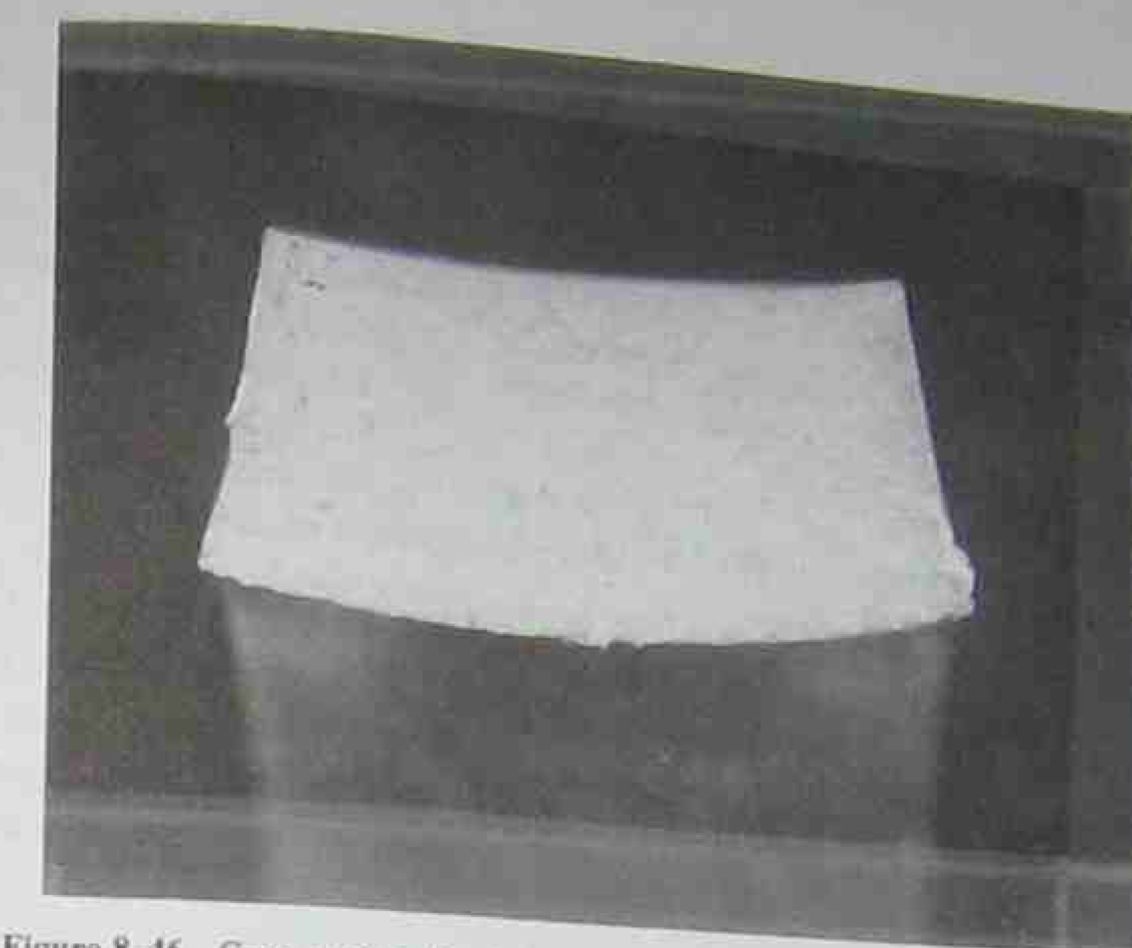


Figure 8–46 Cross section of an initially rectangular bar of aluminum bent to fracture. Photo courtesy of W. H. Durrant.

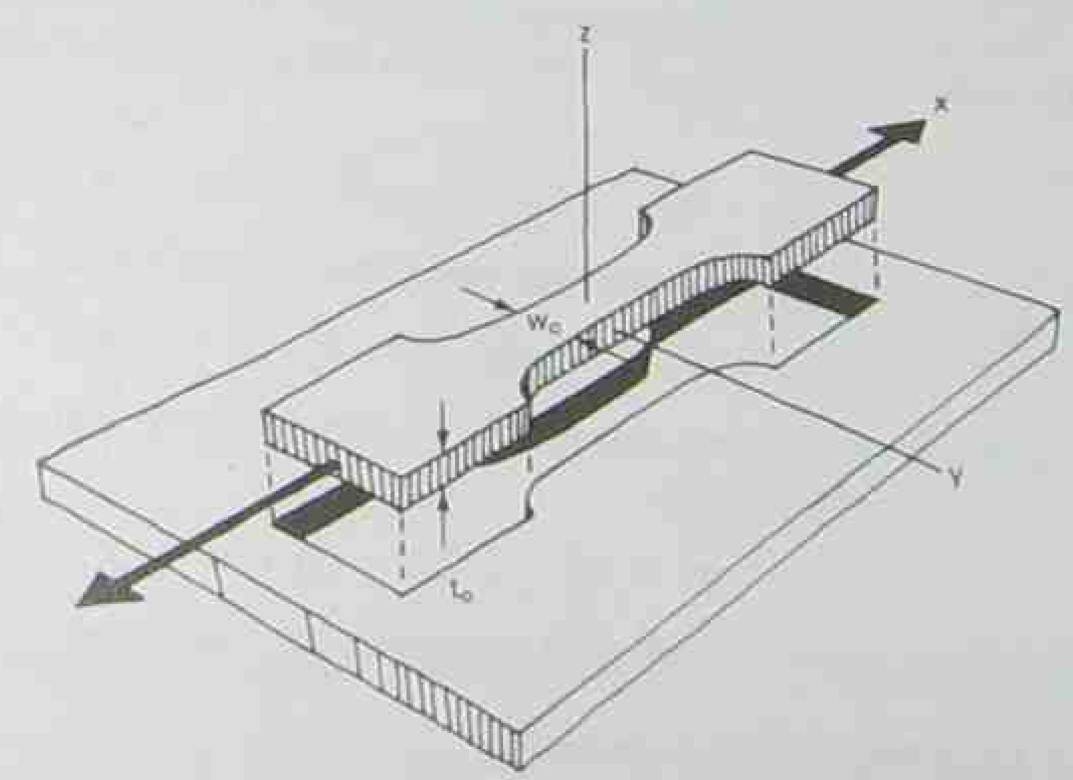


Figure 8–47 Strip tensile specimen cut from a sheet. The R-value is defined as the ratio of width to thickness strains; $\epsilon_w/\epsilon_t = \epsilon_y/\epsilon_z$.

Example 8–8

From a 0.100-in. thick sheet of metal, tensile specimens are cut at 0, 45, and 90 degrees to the rolling direction. Each specimen is then subjected to the same tensile load and plastic deformation results. Prior to loading, a rectangle of 2-in. length and 0.750-in. width was scribed on the specimen surface, where the larger dimension was parallel to the gage section. After loading, these two dimensions changed as indicated on the following page.

At 0 deg $\ell = 2.5$ in., $w = 0.65$ in.
At 45 deg, $\ell = 2.55$ in., $w = 0.63$ in.
At 90 deg, $\ell = 2.45$ in., $w = 0.67$ in.

Find the average strain ratio $\bar{R}$ for this sheet.

Solution Using Eq. (8–53) for the 0-deg direction, we have

$$\epsilon_\ell = \ell n(2.5/2) = 0.223$$
$$\epsilon_w = \ell n(0.65/0.75) = -0.143$$

and from Eq. (8–54),

$$R_0 = -(0.143)/-(0.223 - 0.143) = 1.79$$

For the 45-deg specimen, $\epsilon_\ell = 0.243$, and $\epsilon_w = -0.1744$, so $R_{45} = 2.54$. R_{90} is found to be 1.26.
With Eq. (8–55)

$$\bar{R} = (1.79 + 1.26 + 2[2.54])/4 = 2.03$$

8.9.3 Anisotropic Yield Criterion

In sheet-forming analyses it is usually essential to invoke some form of a yield criterion; otherwise, only descriptive or qualitative comments are possible.

The most widely used anisotropic yield criterion was proposed by Hill [7] in 1948; other criteria have also been proposed [8,9]. In terms of principal stresses, Hill's criterion is

$$F(\sigma_2 - \sigma_3)^2 + G(\sigma_3 - \sigma_1)^2 + H(\sigma_1 - \sigma_2)^2 = 1 \qquad (8\text{–}56)$$

where F, G, and H are parameters that characterize the anisotropy. In terms of R-values,

$$R_0 = R = H/G \quad \text{and} \quad R_{90} = P = H/F \qquad (8\text{–}57)$$

and if *planar isotropy* occurs (that is, $R = P$), then for the special case of plane stress (say $\sigma_3 = 0$), Eq. (8–56) becomes

$$\sigma_1^2 + \sigma_2^2 - (2R/[R+1])\,\sigma_1\sigma_2 = X^2 \qquad (8\text{–}58)$$

where X is the uniaxial tensile yield strength in the 1 direction.

8.9.4 Cupping or Deep-Drawing

A reasonable classification for many sheet-forming operations is to distinguish them as *deep-drawing* or *stamping*. The deep drawing of flat-bottom, cylindrical cups (also called *cupping*) is a process used to produce beverage cans, flashlight cases, and the like. There, the plane of the sheet experiences one principal strain that is tensile while the second is compressive; since any thickness change is small, that strain is usually assumed to be zero.

Stamping operations are generally complex in terms of deformation. In regions where biaxial stretching occurs, planar strains are both tensile, whereas the thinning of the

sheet causes a compressive strain through the thickness. Bending and unbending can also occur and cause redundant deformation. It should be noted that operations can include all of these factors (in addition to frictional effects), so analyses become involved.

Starting with cupping, Fig. 8–48 illustrates the coordinate system of concern while Fig. 8–49 gives details of a partially drawn cup. The crucial regions to assess are the *flange*, where major deformation occurs (the flange is forced to move in radially as the cup is drawn), and the *wall*, which must support the force needed to draw in the flange. Two of the major possibilities that must be prevented for most cupping operations are *wrinkling*, Fig. 8–11, and *fracture*, Fig. 8–50. In some applications, wrinkling is not a detriment, as noted by the flanges on aluminum foil dishes (pies, TV dinners). However, if this is to be avoided, an adequate level of hold-down force must be applied. A rule of thumb is that the hold-down-*pressure* should be about *1 percent of the yield strength* of the metal; however, parameters such as sheet thickness, cup diameter, die radius, and condition of lubrication can all influence proper hold-down pressure, so this is only a ballpark suggestion. If this pressure is excessive, greater drawing forces are needed to cause flange deformation and may cause failures, as shown in Fig. 8–50.

The parameter most commonly used to assess formability in cupping is called the

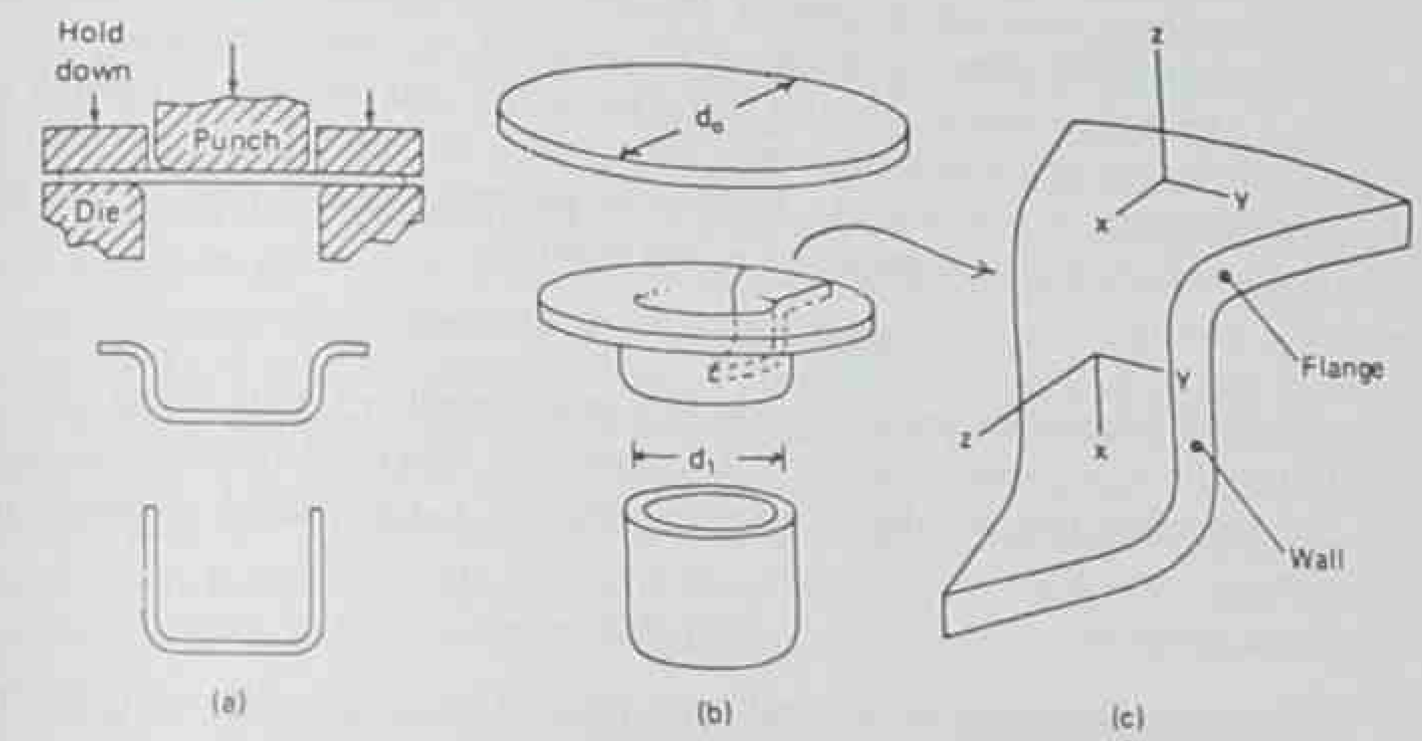


Figure 8–48 Schematic of a cupping operation showing the coordinate axes.

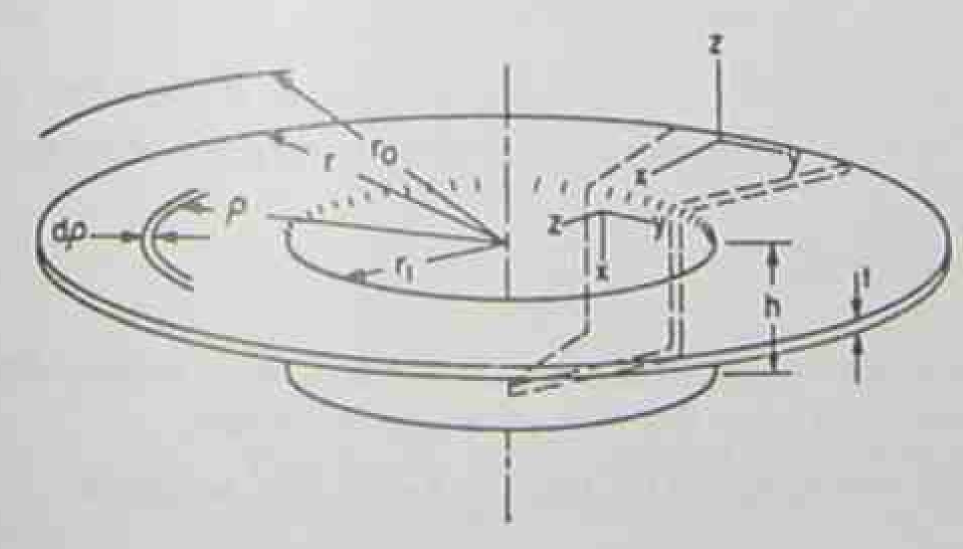


Figure 8–49 Schematic of a partially drawn cup and the coordinate system.

Figure 8–50 Drawing failures by necking at the bottom of cup wall. From D. J. Meuleman, Ph.D. thesis, the University of Michigan (1980).

limiting drawing ratio (LDR); this is the largest ratio of blank-to-cup diameters that can be drawn successfully. Too large an LDR leads to fracture. If the wall and blank thickness are assumed equal as a first approximation, the total volume of the blank can be related to surface areas by

$$\pi r_0^2 t = \pi r_1^2 t + \pi r_1 t h \tag{8–59}$$

where r_0 and r_1 are the blank and cup radii, h is the cup height, and t is the thickness as in Figs. 8–48 and 8–49. Then, in terms of diameters,

$$(h/d_1) = [(d_0/d_1)^2 - 1]/4 \tag{8–60}$$

so for a draw of $d_0/d_1 = 2$, the height-to-diameter ratio of the cup will be about 0.75. Often, deeper cups may be needed, but this could demand diameter ratios that would exceed the LDR. Other operations, to be discussed, are then necessary.

An analysis due to Whiteley [10] gives insight into the LDR. Although it involves certain simplifying assumptions (see, for example, reference [1], it provides a basis of understanding. It considers the LDR to be governed by the ratio of the flow stress *in-the-wall* σ_w to that *in-the-flange* σ_f and results in

$$\ell n(\text{LDR}) = \sigma_w/\sigma_f = \beta = \ell n(d_0/d_1) \tag{8–61}$$

where β is taken as the stress ratio. For an isotropic metal, $\sigma_w = \sigma_f$, so the LDR = 2.718. In practice it is closer to 2.1 to 2.2, because the effects of friction plus the bending and unbending over the die lip were neglected. If a *deformation efficiency* η is used, then

$$\ell n(\text{LDR}) = \eta\beta \tag{8–62}$$

and a reasonable value of η based upon practice is of the order of 0.75. To handle anisotropy, Eqs. (8–55) and (8–58) can be employed to give

$$\ell n(\text{LDR}) = \eta\sqrt{(\bar{R} + 1)/2} \tag{8–63}$$

Although many experiments show that LDR does increase with larger $\bar{R}$, the effect is not as great as indicated by Eq. (8–63). For further discussion on this point see reference [1].

Example 8–9

You wish to produce a cup having a height-to-diameter ratio of one. What limiting drawing ratio (minimum value) is needed? Assume that constant thickness prevails.

Solution Using Eq. (8–60), with d_0/d_1 = LDR (see Eq. 8–61), then we obtain

$$1.0 = [(d_0/d_1)^2 - 1]/4$$

so

$$(d_0/d_1)^2 = 5 \quad \text{or} \quad d_0/d_1 = 2.24$$

Often, a successful cup can be drawn without fracture, but *earing* occurs. Figure 8–51 shows this, the ears being high regions separated by valleys. Ear formation is undesirable, since the ends must be trimmed; this is an expense, and metal is wasted. It has been shown that earing correlates with the angular variation of R in the sheet and, although four ears are most common, other even multiples have been observed. The empirical parameter that seems to correlate best with the position of ears (with respect to orientation in the sheet) is ΔR, where

$$\Delta R = (R_0 + R_{90} - 2R_{45})/2 \tag{8–64}$$

Figures 8–52 and 8–53 relate to this discussion.

8.9.5 Redrawing and Ironing

To produce cups of depth greater than is attainable with a single drawing pass, *redrawing* and, perhaps, *ironing* are often employed. Redrawing (direct) was shown in Fig. 8–12, while a full draw, redraw, and ironing sequence using concentric punches was illustrated in Fig. 8–13. Note that ironing tends to produce a uniform wall thickness, since in typical drawing operations the wall is thicker near the top. Analyses for limiting redrawing and ironing ratios have been proposed [11] but are beyond the scope of this text. However, a few words are added here on the effects of work hardening. It has been shown [11] that in simple cupping, the value of the strain hardening exponent n has only a very minor effect on LDR over the common range of n-values observed. With regard to redrawing and ironing, however, heavily cold-worked sheets will *outperform* annealed ones. Since in this text, n is treated as a fixed value for a given metal, it is not correct to talk about one n-value for an annealed metal and a different one for that metal in some condition of cold work. Rather, it could be stated that a metal which possesses a greater degree of *uniform* tensile deformation (that is, annealed) will not be as successful in redrawing or

Figure 8–51 Earing behavior of cups made from three different copper sheets. Arrow indicates rolling direction of sheets.

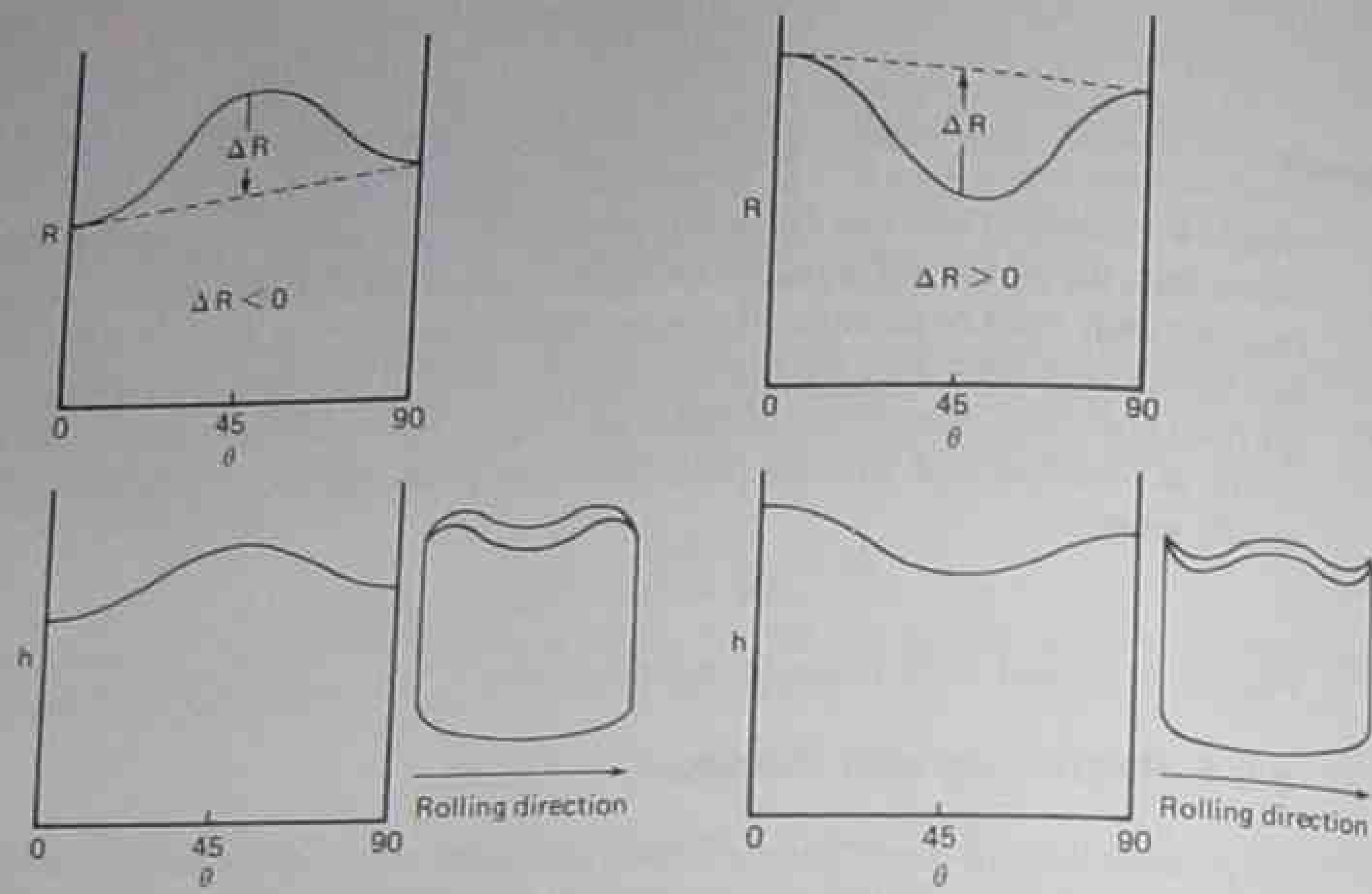


Figure 8–52 Relation of earing with angular variation of R; h is the wall height.

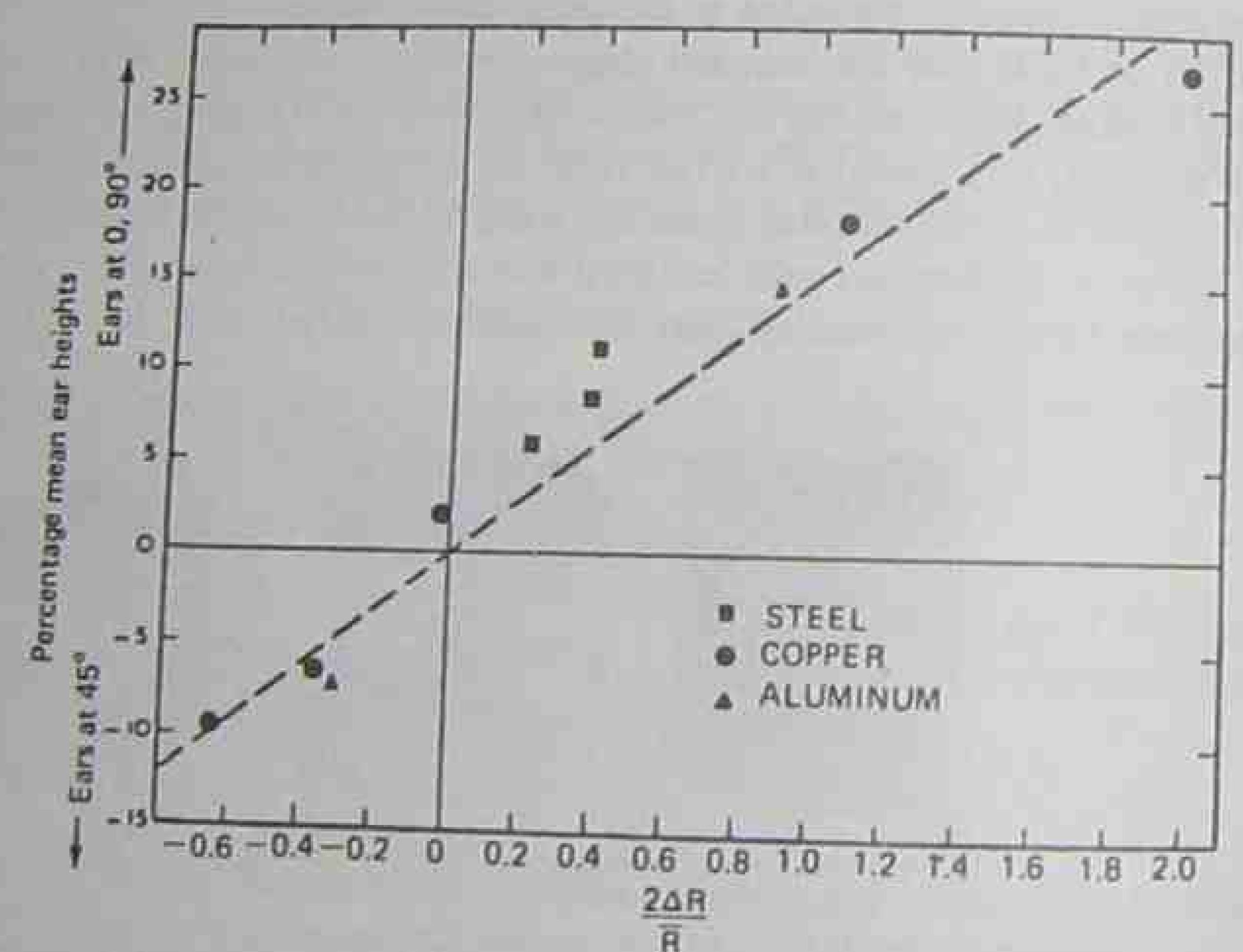


Figure 8–53 Correlation of extent of earing with ΔR.

ironing as the same sheet possessing little or no ability to strain uniformly in tension (that is, work-hardened). To support this observation, it is noted that *very heavily cold-rolled* sheet is used in the deep drawing and ironing of steel and aluminum beverage cans.

8.9.6 Forming Limit Diagrams

Many shapes formed from sheet metal are far more complicated than simple cupping, and analytical description, such as that given above, is lacking. Instead, resort is made to what is called a *forming limit diagram* (FLD). The basis behind such an approach is to print a grid of circles on the sheet. Whether the sheet is to be stamped into a prototype of an actual production part or a shape of simpler contour (usually used in research work), the procedure is identical. As the shape is formed, the circles distort into ellipses whose major and minor axes indicate the directions and magnitudes of the principal strains in the *plane* of the sheet. In such operations, the onset of any *local necking* is generally considered *failure*, although continued deformation must be induced to cause actual *fracture*. Regardless of where the process is terminated, the principal nominal strains are calculated from

$$e_1 = (d_1 - d_0)/d_0 \quad \text{and} \quad e_2 = (d_2 - d_0)/d_0 \tag{8–65}$$

where the subscripts $_1$ and $_2$ relate to the major and minor axes of the ellipse and d_0 is the original circle diameter. If true strains were used,

$$\epsilon_1 = \ell n(d_1/d_0) \quad \text{and} \quad \epsilon_2 = \ell n(d_2/d_0) \tag{8–66}$$

noting that length and *not area* changes are of concern here. Although forming limit strains are usually large, and true strains are more appropriate, many workers employ nominal strains for no apparent reason.

Figure 8–54 illustrates an example where the sheet was fractured; the plotted circles indicate strain combinations that were *safe*, those that caused *fracture*, and those very near the fracture zone that were considered borderline. In a practical sense, such tests can be used as a diagnostic tool to determine if failures are likely in production runs or if the severity of stamping is approaching a probable failure. Such tests can often save a great deal of expense, since it is of greater advantage to spot potential problems *before* a stamping line is put into actual production. Thus, the results of such tests can be used to make modifications in die design or in part requirements early in the process.

Figures 8–55 and 8–56 illustrate experimental FLDs for several commonly used metals. Factors such as sheet thickness, strain hardening exponent n, and strain-rate exponent m, all appear important. Also, the more *uniformly* strain is distributed in the plane of the sheet, the deeper may parts be stamped. In connection with these two figures, note that *lowest* values of e_1 occur when $e_2 = 0$ (that is, plane strain).

8.9.7 Concerns in Sheet Forming

As we discussed in Sec. 8.8.5, the primary concern in bulk forming operations is to avoid cracking of the workpiece. Other concerns related to residual stresses and preferred orientation of grains and inclusions are also important, *but* the avoidance of fracture must

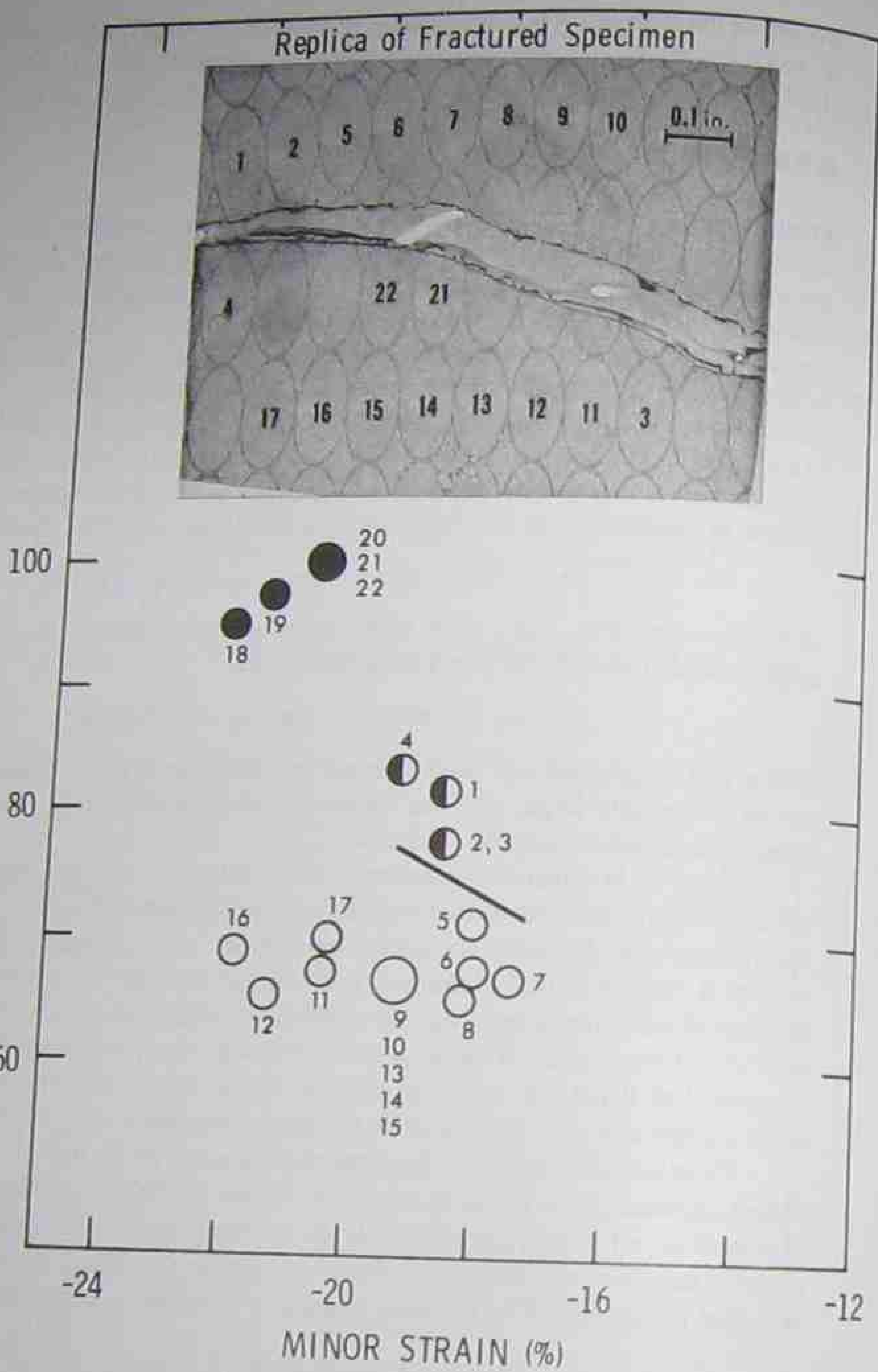


Figure 8–54 Distortion of printed circles near a localized neck (top) and a plot of the strains in the distorted circles (bottom). Solid points indicate failure; open points are away from the failure, and partially filled points are near the failure zone.

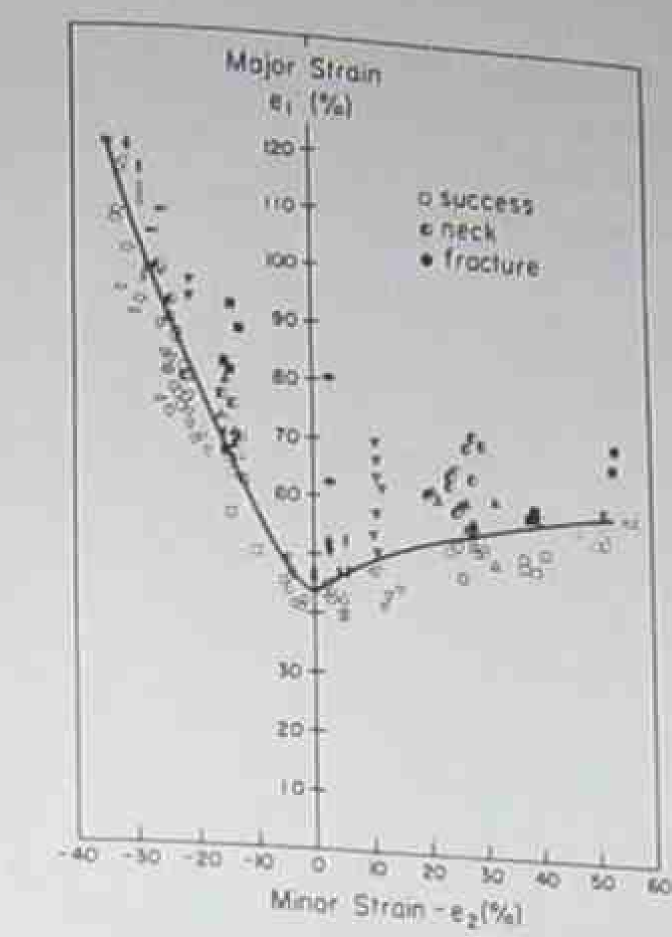


Figure 8–55 Forming limit diagrams for low-carbon steel obtained from data similar to Fig. 8–54. Strains below the curve are acceptable, while those above indicate regions affected by localized necking.

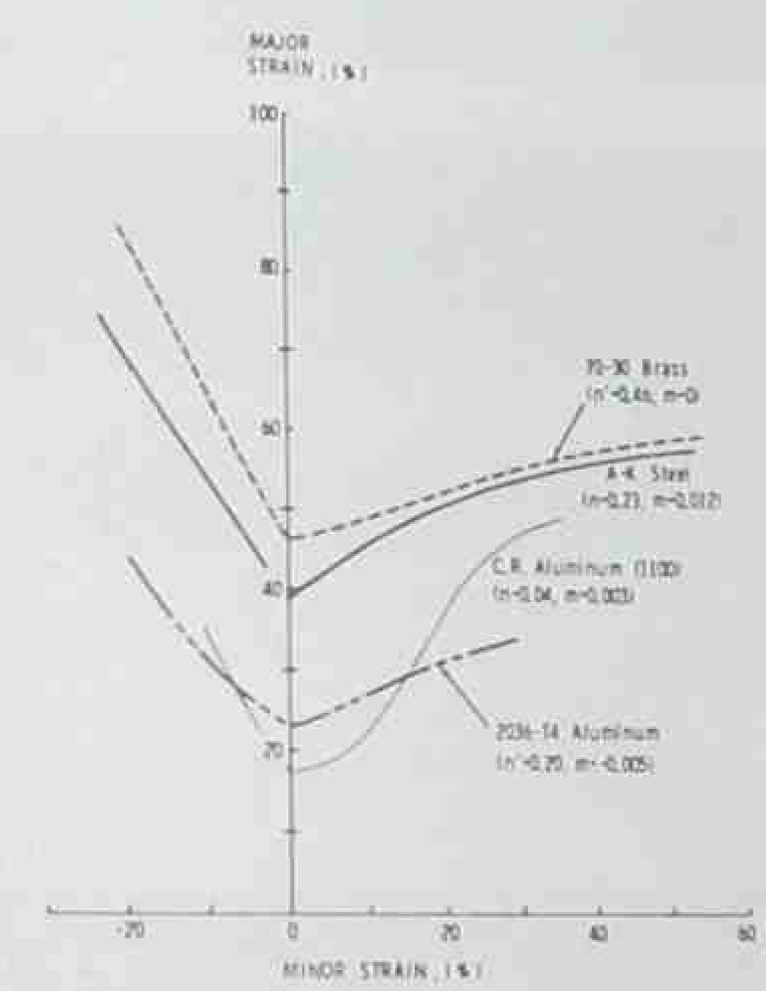


Figure 8–56 Forming limit diagrams for several metals.

come first. With sheet-forming operations the avoidance of local necking is a major concern, since that often dictates the extent of useful deformation, and the influence of $\overline{R}$, m, and n (average strain ratio, strain-rate exponent, and strain-hardening exponent) have been discussed in that regard.

Two important possibilities that influence the *appearance* of the finished surface are called the *orange-peel* effect and *stretcher strains*. The first is a surface roughness on the scale of the grain size. Due to differences in the orientation of adjacent grains on the surface, a tendency to undergo differing degrees of deformation leads to a roughened surface, somewhat like that on an orange. This result can occur only if there is a *free* surface that is not in intimate contact with tool surfaces. Using a metal with a finer grain size will reduce the overall roughening.

Stretcher strains are incomplete Lüders bands* that occur during forming. Metals such as low-carbon annealed steels display a pronounced yield point, and Lüders bands form during early straining up to strain levels of about 0.02. Because strains induced during many sheet-forming operations can vary greatly from one region to another, those regions that are strained to quite low levels are apt to display stretcher strains. One widely used method to avoid this phenomenon is to first subject the sheet to a *roller-leveling* operation. Basically, this amount of light cold working removes the pronounced yield-point affect (that is, the sheet is strained just a little). Now when the subsequent forming operation is conducted, stretcher strains do not occur. Figures 8–57 and 8–58 show these two surface defects.

* See Chapter 6.

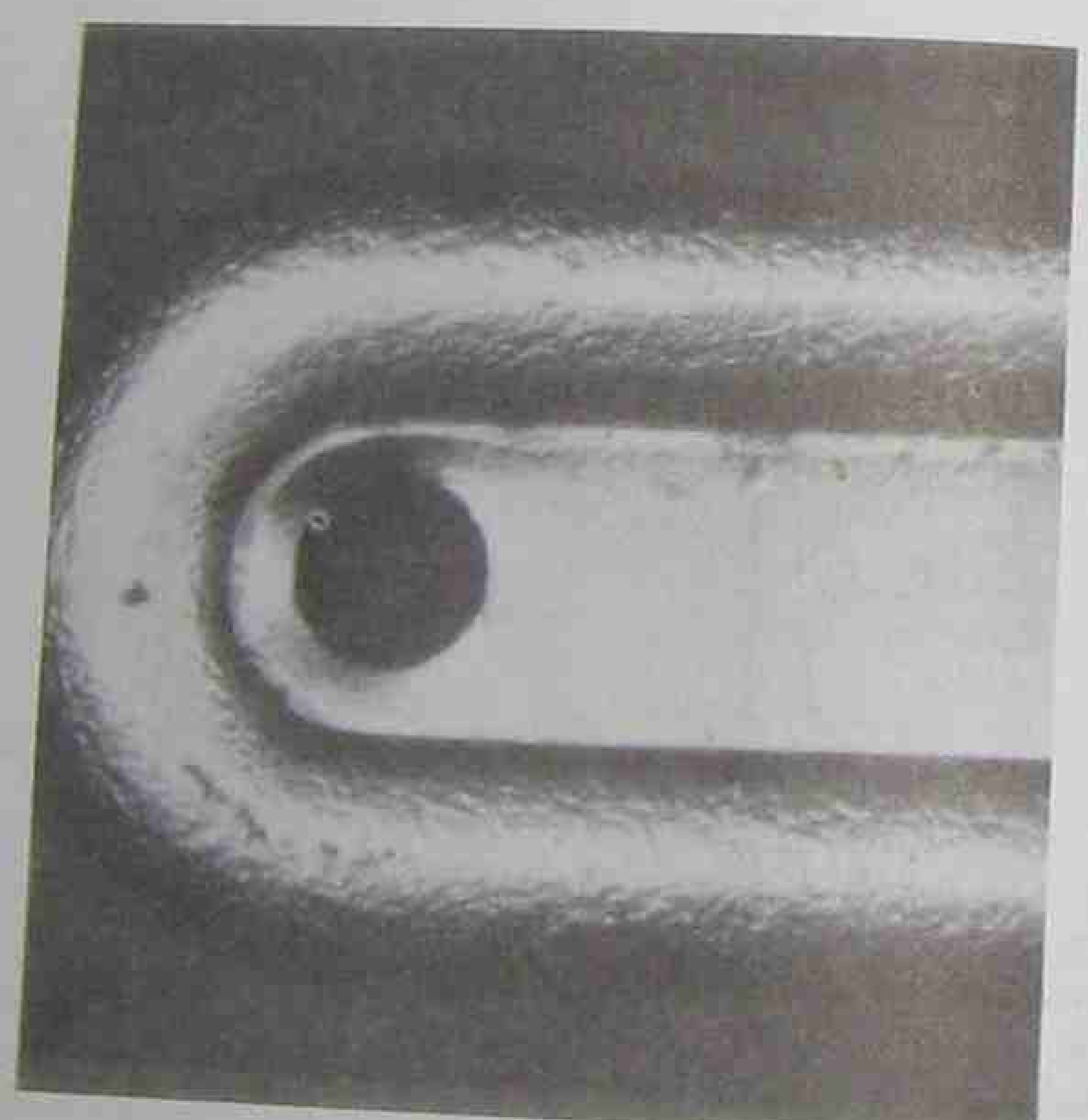


Figure 8–57 Example of orange peel.

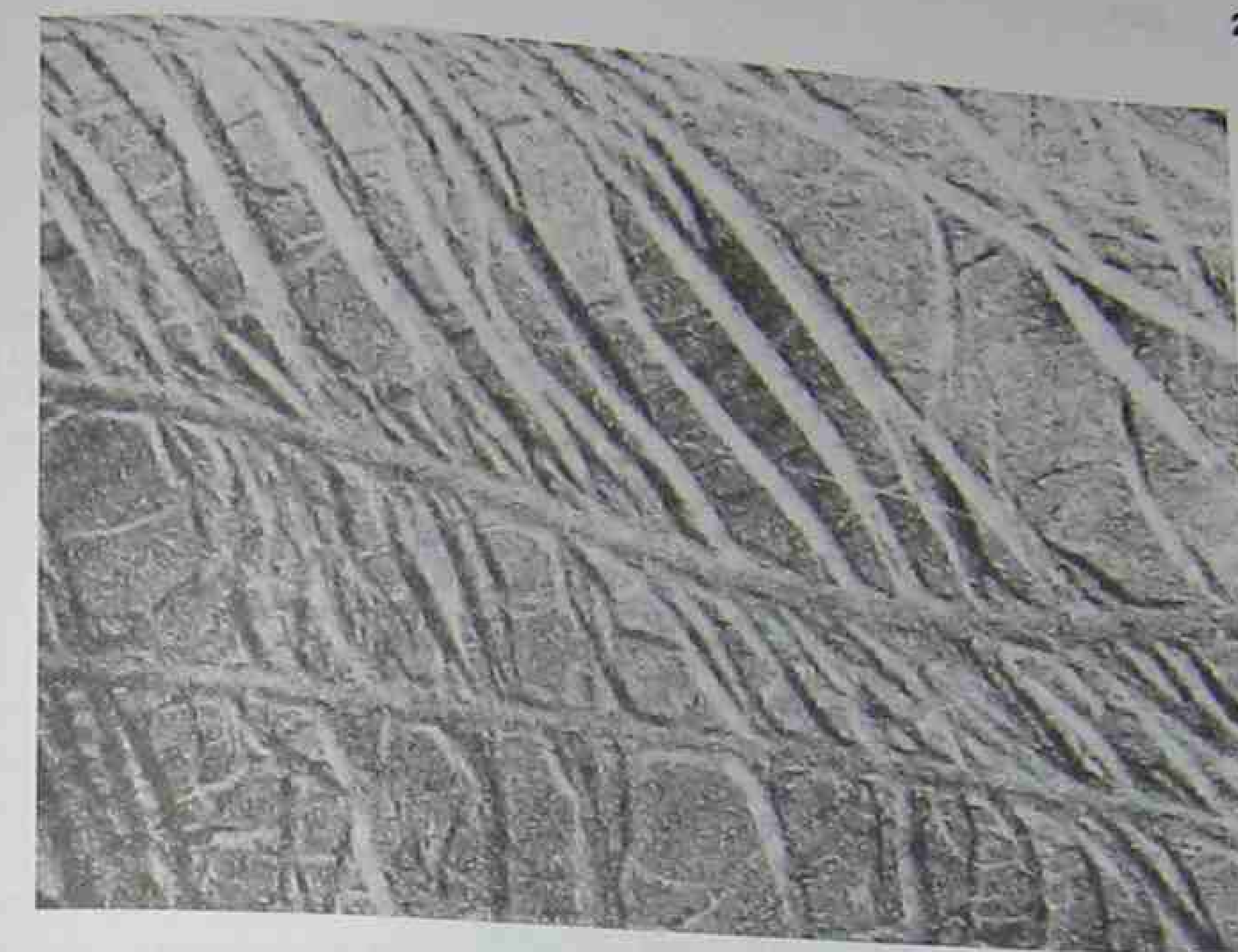


Figure 8–58 Stretcher strains on a 1008 steel sheet stretched slightly beyond the yield point.

8.10 STRAIN RATE

In many analyses, the effect of strain rate, $\dot{\epsilon} = d\epsilon/dt$, is ignored, since its affect on the responding yield or flow stress is small. However, in hot working operations and, in studies connected with forming limit diagrams, the effect of $\dot{\epsilon}$ can be important. Often, the effect of $\dot{\epsilon}$ on flow stress is presented as

$$\sigma = C\dot{\epsilon}^m \qquad (8\text{–}67)$$

where a series of tensile tests is conducted at different strain rates and the resulting stresses are chosen at some *constant* strain, as illustrated in Fig. 8–59. Implicit in Eq. (8–67) is a straight line plot on logarithmic coordinates. C is called the *strength constant*, while m is the *strain-rate* exponent. High values of m are important in *superplasticity*, where extremely large elongations result without local necking. Figure 8–60 shows a tensile specimen of a superplastic metal elongated 1950 percent. In studies devoted to forming limit diagrams, the influence of m is finding increased meaning, since it appears that metals with relatively large m values do promote large total elongation and thereby superplastic behavior.

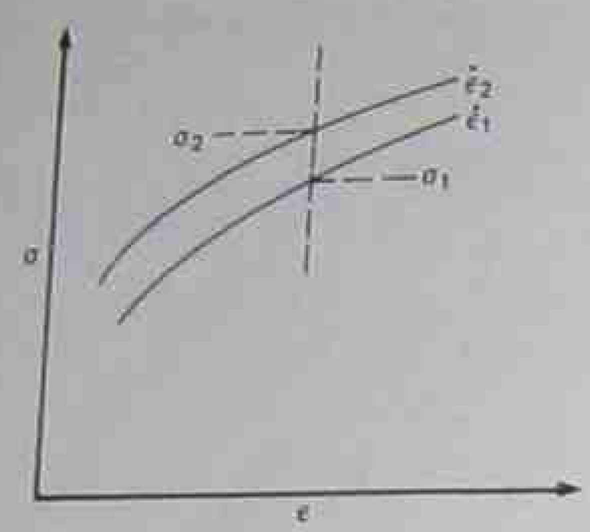


Figure 8-59 Stress-strain behavior as a function of two different strain rates, each under continuous loading. At a common strain, the strain rate exponent, $m = \ell n(\sigma_2/\sigma_1)/\ell n(\dot{\epsilon}_2/\dot{\epsilon}_1)$

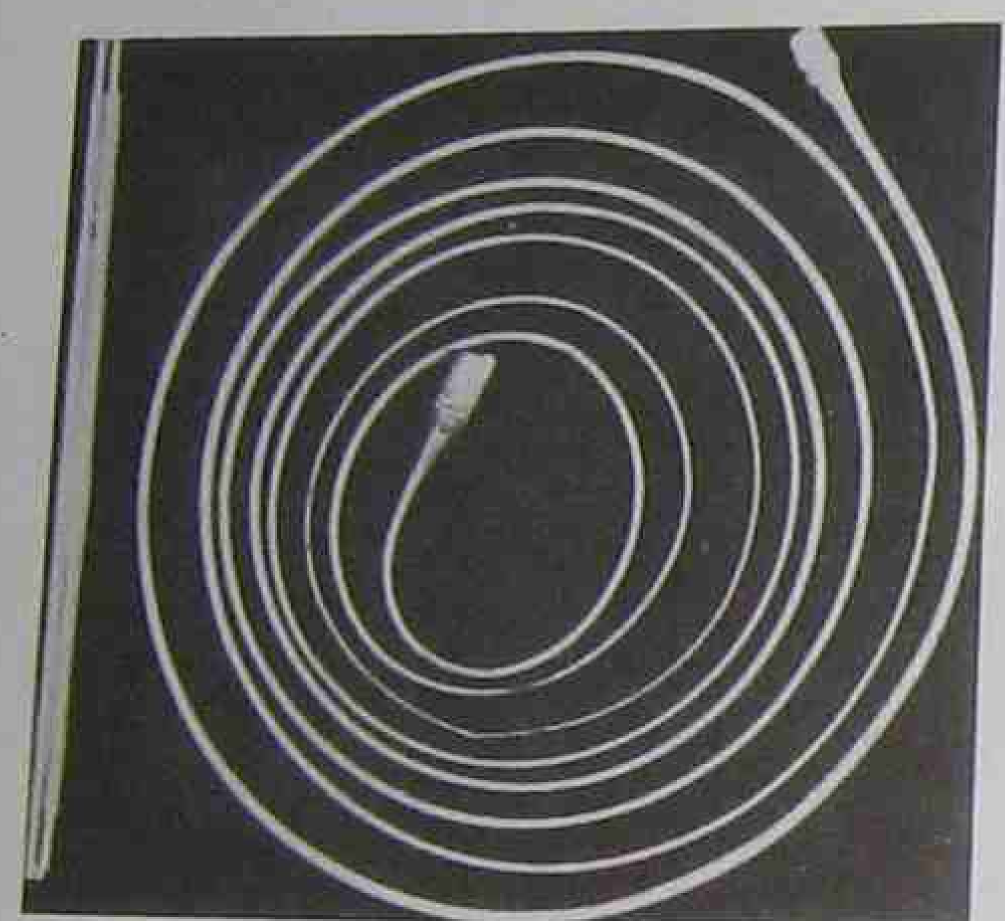


Figure 8–60 Tensile bar of Bi-Sn eutectic showing superplastic extension.

8.11 BAUSCHINGER EFFECT [12]

If a metal having an initial yield strength of Y_0 is plastically deformed, say by tensile loading, and then is unloaded, its new yield strength under subsequent tensile loading will be higher than Y_0 due to work hardening. The use of $\bar{\sigma} = K\bar{\epsilon}^n$ would permit a calculation of the *new* yield strength, which is shown as Y_1 in Fig. 8–61. Suppose, however, that the plastically strained metal is loaded in compression following the unloading from Y_1. It is often found that the compressive yield strength Y_2 will be *less than* Y_1 but greater than Y_0 (assuming initial isotropy). This is the Bauschinger effect. An explanation at the microscopic level, based upon dislocation theory, can be found elsewhere [2], while an explanation using macroscopic arguments is discussed in Ref. [13].

Much confusion exists regarding the Bauschinger effect. In these authors' experience, cold working a ductile metal by any method will usually *tend* to increase the yield strength in all directions as compared with their prestrained values, but the increase will not be necessarily uniform. It is possible that misconceptions have arisen in the translation

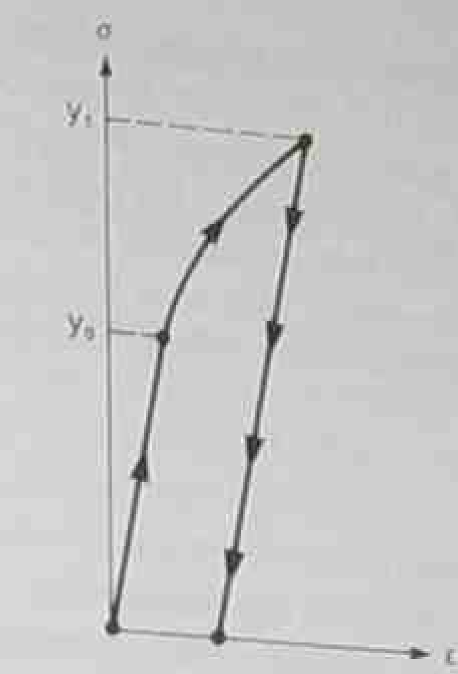


Figure 8-61 Stress-strain behavior of a bar subjected to tensile of deformation; Y_1 is the tensile yield strength of the deformed bar.

of Bauschinger's original paper (in German), and Fig. 8–62 is used to illustrate this. Because sketches in this regard are usually not drawn to scale, the unloading from Y_1 is often exaggerated. It is true that a small nonlinearity is sometimes observed, and if it is carried into the compressive region seems to imply that the *new* compressive yield strength is *less than* the original value Y_0 after cold working in tension. If this argument were carried to the extreme, where a very large degree of tensile deformation was first induced, it would be concluded that the new compressive yield strength is zero or even tensile. This is, of course, absurd. Perhaps the confusion arises from the manner in which yield strength (*not* proportional limit or elastic limit) is defined. The most sensible way to check this concept would be to take an annealed metal and determine the original Y_0 values in tension and compression by using the offset method discussed in Chapter 6. Next take another sample of the same annealed metal and subject it to plastic deformation well beyond the initial Y_0. This could be done by either tensile or compressive deformation (tensile is perhaps easier here). Now, from the *cold-worked specimen* produce a new

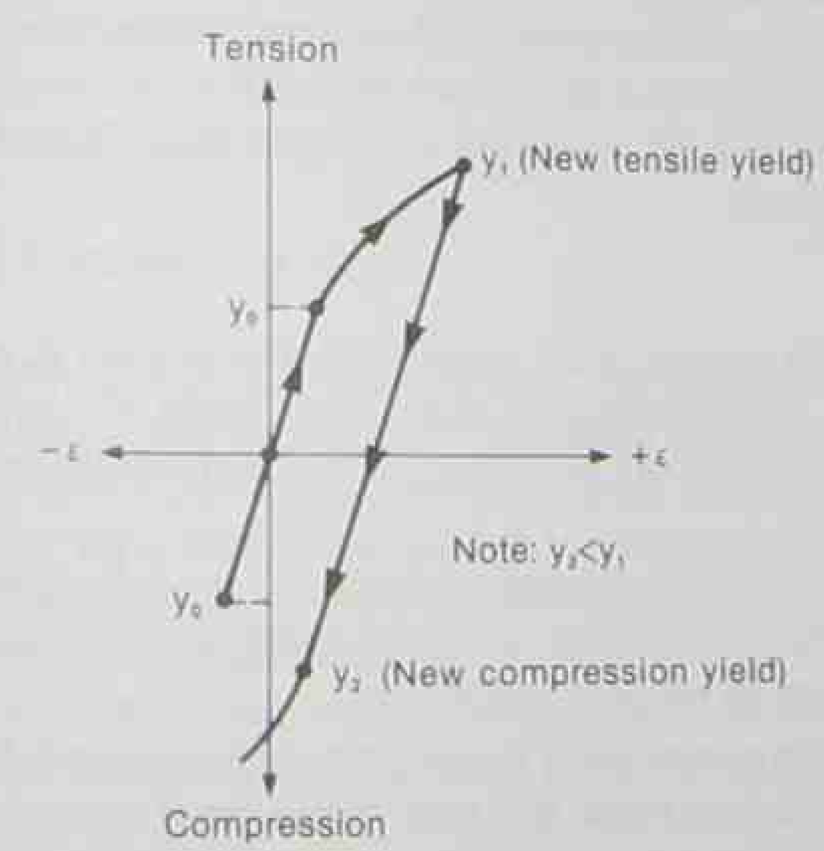


Figure 8-62 An illustration of the Bauschinger effect.

tensile specimen *and* a new compression specimen; then subject one to a tensile test and one to a direct compression test. If these results are plotted as a stress-strain curve (either nominal or true) and the new yield strengths are defined by an offset, it would be found that Y_1 (tension) is usually $>Y_2$ (compression) but *both* are $>Y_0$. Note that if prestraining is done by compression rather than tension, and the same subsequent testing is conducted, then Y_2 (compression) $> Y_1$ (tension). This is as far as the concept of the Bauschinger effect need be taken; if it is limited to this extent, confusion will be avoided.

8.12 HOT AND COLD WORKING

When metal is deformed at temperatures above that at which *recrystallization*, R_x occurs, it is called hot working; if deformation proceeds below the R_x region, it is called cold working. These are general terms; more complete coverage of R_x was presented in Chapter 7. Typical metal-forming operations induce deformation so quickly that even at elevated temperatures there is inadequate time to cause R_x *during deformation*. Instead, recrystallization occurs as the deformed metal cools down from these high temperatures. In any event, the net result is that work hardening does not occur. Much greater deformations, short of fracture, are possible with hot working, because the stresses necessary to cause flow are much lower compared with cold working; this results because the yield strength of metals is drastically lowered at typical hot working temperatures and, as mentioned above, work hardening does not occur. There are occasions, however, where failure at low strains can result during hot working; this is often called *hot shortness*. This comes about if a liquid phase is present at these elevated temperatures and results because the metal is heated into a two-phase, liquid-solid region. Understanding equilibrium diagrams makes this clear and is one reason why they were discussed in Chapter 7. One practical example of this phenomenon relates to hot working many steels. If the sulfur combines with iron to form FeS, this tends to form liquid films at grain boundaries and, at typical temperatures used, leads to hot shortness during deformation. It is common practice to add manganese to most steels, since sulfur reacts more readily with this element to form MnS *rather than* FeS. At elevated temperatures used in hot working operations, the MnS remains solid so hot shortness is avoided.

A general summary, including the advantages and disadvantages of these two broad categories of deformation, follows:

1. *Hot working*
   a. The flow stress or yield strength of the metal is lowered whether R_x occurs during or after deformation. As a consequence, much lower inputs of energy are needed to cause the desired deformation, so the size of necessary equipment and power sources is reduced.
   b. The end product is in an annealed or softened state, so any subsequent operations are begun on a structure of relatively low yield strength. Additionally, residual stresses are nonexistent for all practical purposes.
   c. Oxidation of the surface occurs, being more severe with some metals than others. This causes a rough surface finish and poor dimensional control and is detrimental in any subsequent machining operations, since it accelerates wear of the cutting tool.
   d. The use of lubricants is practically precluded, an exception being special glasses that have been developed for this purpose (they are relatively expensive).
2. *Cold Working*
   a. The initial yield strength and its increase due to deformation are greater during cold working than during hot working; thus, for similar deformations, cold working demands larger power inputs.
   b. Both the strength and ductility of the end product are different from the initial material.
   c. Surface finish and dimensional control are vastly superior to their counterparts of hot-worked products.
   d. Due to work hardening, the yield strength of the end product is greater than the initial material.

A few other comments related to this section are

1. Certain reactive metals, such as titanium and zirconium, are sometimes hot worked under a protective atmosphere to eliminate oxidation; this does introduce an added cost that would be prohibitive in most cases.
2. Large ingots of steel are initially hot worked to the nearly desired size, *pickled* in acid to remove the oxide scale, and then cold worked to final size.

8.13 FORMING OF POLYMERS AND CERAMICS

As mentioned at the start of this chapter, and reflected in the coverage so far, the forming of metals, certainly on a tonnage basis, predominates this field of processing. However, the same types of processes are employed in the production of parts made from polymers and ceramics; a brief coverage involving applications to these classes of solids will conclude this chapter.

8.13.1 Polymers

Many thermoplastics (see Chapter 7) are produced in various cross-sectional shapes by *extrusion*. There it is usual to feed small pellets of the polymer into a heating chamber after which a force is applied that causes the pliable polymer to exit from a die whose shape governs that of the extruded workpiece. Although different *methods* are used to apply the necessary force, the fundamentals of extrusion are really identical to those covered in earlier sections.

In *compression molding*, a quantity of the polymer is placed into a cavity, heated, and then subjected to compressive forces; in essence, this is a forging process, as shown in Fig. 8–63.

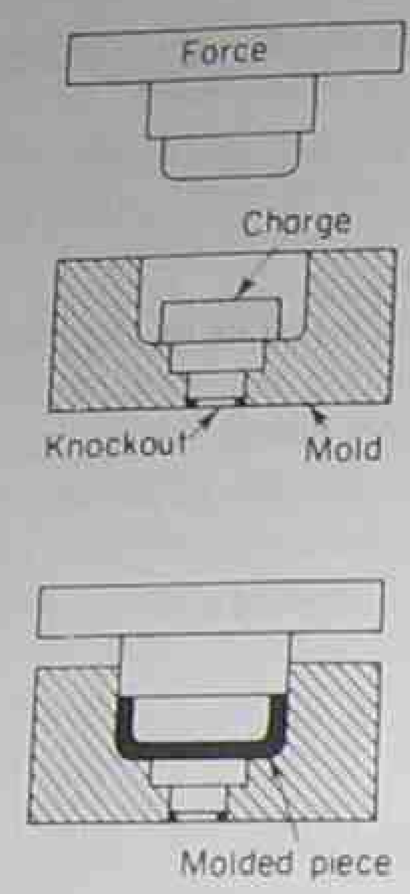


Figure 8–63 Schematic of compression molding.

Blow molding, shown in Fig. 8–64, is used to make items such as plastic bottles; it involves the use of a tubelike shape which is first heated and then placed into a mold. The application of internal pressure causes the tube to bulge until it touches the walls of the mold. Since *recovery** after bulging is to be minimal, the polymer must be strained to a level that induces permanent deformation. This may require stable neck propagation, as discussed in Chapter 6.

Thermoforming involves the use of sheet material which is first heated. With the application of air pressure, the sheet is forced to fit over a shaped mold; essentially this is a stretch-forming operation. Often, instead of air pressure, the opening between the sheet and mold surface is evacuated, and since many polymers used with this process have low strength, atmospheric pressure on the top of the sheet is sufficient to force the workpiece to make contact with the mold surface. Figure 8–65 illustrates the details.

Polymeric sheets are often made by *calendering*. There, a heated mass of bulk polymer is fed into and through a series of rolls to produce some final thickness; this is a rolling process.

8.13.2 Ceramics

The basics behind powder processing, usually involving *sintering*, are covered in Chapter 10. Here we will discuss only those aspects involving forming procedures. Prior to the final operations that cause ceramic products to become extremely hard and brittle (at that point they *cannot* be subjected to further plastic deformation), the ceramic powder is mixed with various additives to bind the particles into a coherent but somewhat porous mass. This mass is usually placed into a mold cavity and then subjected to compressive

* See Chapter 7.

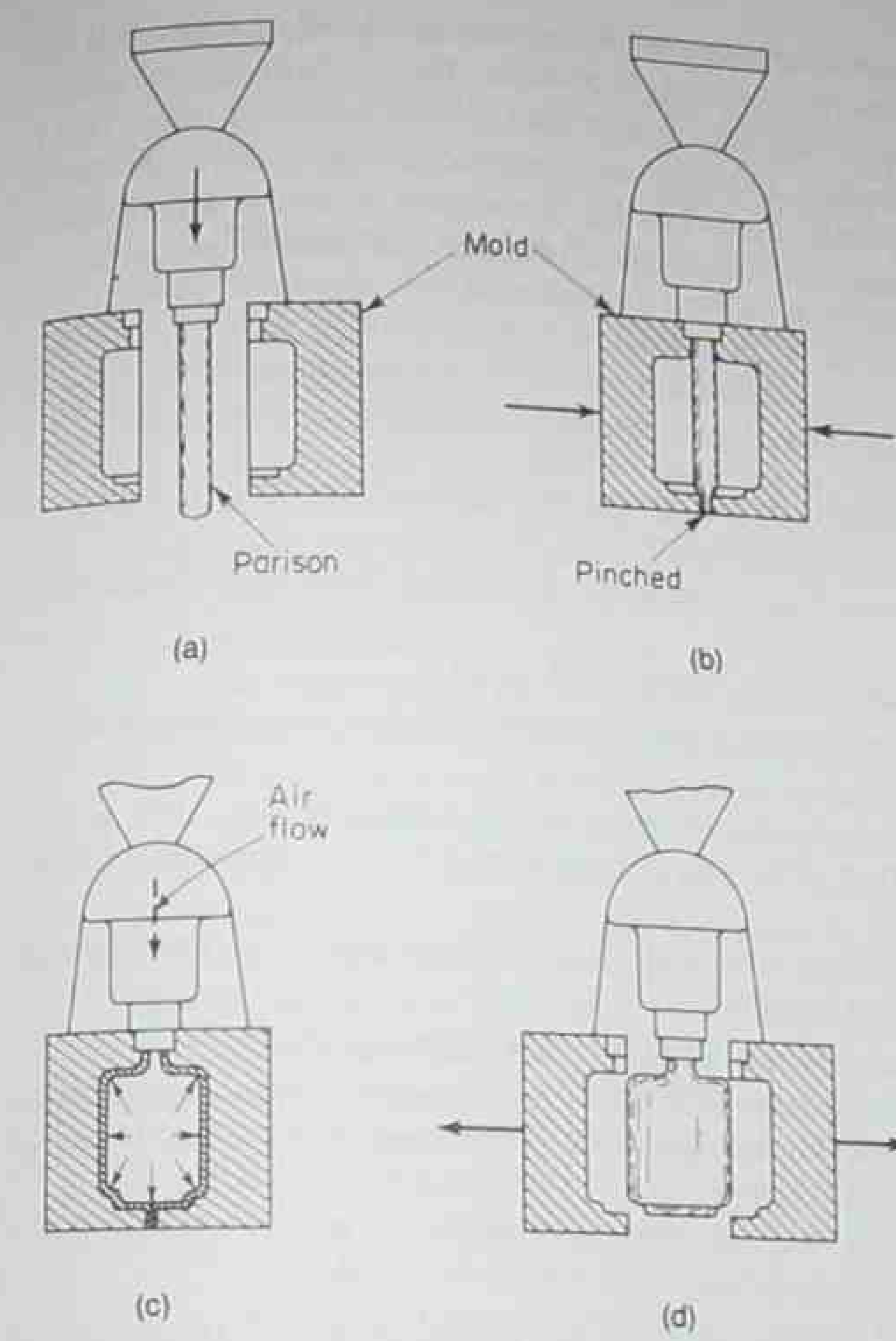


Figure 8–64 Schematic of blow molding.

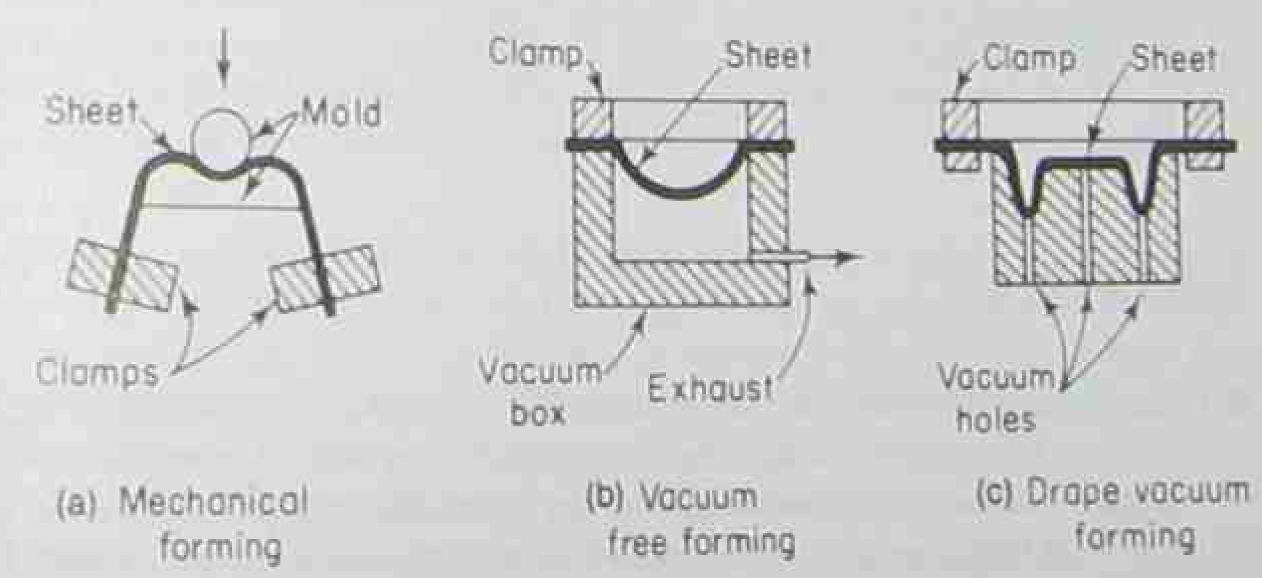


Figure 8–65 Examples of thermoforming.

loads which compact the material. A reasonable increase in density results that can approach the bulk density of the raw material; in addition, this *preforms* the material to some desired shape prior to producing the compact into its final hardened form. This is really a closed-die forging operation. Extrusion of a ceramic powder-additive mixture is also done. Regardless of the process used, forming of ceramics is *always* accomplished prior to the latter operations that lead to the typical brittleness of ceramic products.

REFERENCES

1. W. F. Hosford and R. M. Caddell, *Metal Forming: Mechanics and Metallurgy*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1983, Chapters 9 and 14.
2. R. M. Caddell, *Deformation and Fracture of Solids*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1980, Chapter 4.
3. R. von Mises, Gott. Nach; math.—phys., Kasse, 1913, p. 582.
4. R. M. Caddell and A. G. Atkins, *Journal of Engineering for Industry, Transactions of ASME*, Series B, **90** (1968), p. 411.
5. W. A. Backofen, *Deformation Processing*. Reading, Mass.: Addison Wesley, 1972, p. 137. Also, J. Burke, ScD. thesis, Dept. of Metallurgy and Materials Science, Massachusetts Institute of Technology, 1968.
6. A. G. Atkins and Y. W. Mai, *Elastic and Plastic Fracture*. Chichester, U.K.: Ellis Horwood Ltd.: New York: John Wiley and Sons, 1985.
7. R. Hill, *Proceedings of the Royal Society*, London, **193 A** (1948), p. 281.
8. W. F. Hosford, *7th North American Metal Working Res. Conference Proceedings*, SME, Dearborn, Mich., (1979), p. 191; also R. Logan and W. F. Hosford, *International Journal of Mechanical Science*, **22** (1980), p. 419.
9. R. Hill, Math. Proc. Cambr. Phil. Soc., **85** (1979), p. 179.
10. R. Whiteley, *Trans. ASM*, **52** (1960), p. 154.
11. W. F. Hosford, "Formability; Analysis, Modeling and Experimentation," Metall. Soc. of AIME, N.Y. (1978), p. 78.
12. J. Bauschinger, *Civilingenieur*, 1881, p. 289.
13. N. H. Polakowski and E. J. Ripling, *Strength and Structure of Engineering Materials*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., (1966) pp. 144–46.

PROBLEMS

8–1. Each face of a metal cube has an area of 3 in.^{2}, and the tensile yield strength of this metal is 40 ksi. For normal loading, frictionless conditions may be assumed and an *x-y-z* coordinate system indicates directions normal to the faces. If *compressive* loads $F_x = -20{,}000$ and $F_y = -25{,}000$ pounds force are applied to appropriate faces, what *tensile* force F_z is required to cause yielding according to the Mises criterion?

8–2. A thin-walled tube having closed ends is made from a metal whose yield strength in pure shear is $k = 20$ ksi. The diameter of the tube is 4 in. and the wall thickness t is 0.050 in. For such

a geometry under internal pressure P the stress state can be approximated quite accurately by

$$\sigma_\theta = (Pr/t) \text{ (hoop stress) where } r = \text{radius}$$
$$\sigma_\ell = (Pr/2t) \text{ (axial or longitudinal stress)}$$
$$\sigma_r \approx 0 \text{ (radial direction)}$$

Determine the value of P to cause yielding.

8–3. The tensile yield strength Y of a metal is 140 MPa. In service it will experience known stresses of $\sigma_x = 50$ MPa, $\sigma_y = -80$ MPa, $\tau_{xy} = 30$ MPa and $\tau_{yz} = \tau_{zx} = 0$. What magnitude of a tensile stress σ_z will cause yielding?

8–4. Prior to the loading of a testpiece, a square grid having dimensions of 6 mm per side is marked on one region at the surface. Under loading, the grid changes into a rectangle of 7 mm by 5 mm. Determine the effective strain in the region of the grid under this plastic deformation.

8–5. Determine the plastic work per unit volume for the situation where $\sigma_1 = 2\sigma_2$ and $\sigma_3 = 0$; express the answer in terms of σ_1 and $d\epsilon_1$.

8–6. For the stress state $\sigma_x = 10$, $\sigma_y = 5$, $\tau_{xy} = 3$ (all in MPa), $\sigma_z = \tau_{yz} = \tau_{zx} = 0$, determine the magnitudes of the principal stresses from a plot of Mohr's circle.

8–7. A solid shaft of 2 in. diameter is subjected to the simultaneous loading of an axial tensile force of 80 kN and a torque that causes a shear stress at the surface of 20 MPa. With a plot of Mohr's circle find the magnitudes of the principal stresses and the largest shear stress.

8–8. Aluminum is extruded from a 4 in. to a 1 in. diameter in a single stroke. If the average yield stress is 10 ksi and the efficiency factor η is 0.50, what extrusion pressure is required according to the ideal work method?

8–9. A sheet of metal 24 in. wide and 0.150 in. thick is cold rolled to a thickness of 0.100 in. in a single pass; the 24-in. dimension remains essentially constant. The inlet speed of the sheet is 300 ft/min and $\bar{\sigma} = 25{,}000\,\bar{\epsilon}^{0.2}$ is the strain hardening relation. Here, $\eta = 0.65$ and the von Mises criterion is applicable. Ignoring any effects of strain rate, find the horsepower needed to perform this operation, using the ideal work method.

8–10. You are to estimate the force needed to coin an American quarter; the process is done cold. The action involved may be considered as axial compression, and the entire workpiece flows plastically. The outer diameter is about 0.95 in., the mean thickness after forming is 0.060 in., the average flow stress is 25 ksi, and sticking friction is reasonable.

8–11. A metal has a constant flow stress of 35 MPa. It is to be drawn from a diameter of 100 mm to 50 mm in a single pass using a die of 30 deg semi-angle. Using the slab method, compute the drawing stress if the frictional condition at the interface is
(a) frictionless
(b) $\mu = 0.20$

8–12. Hot forging is to be done on a slab of metal whose initial cross section is one inch by one inch, the length being 10 in.; after forging, the length remains at 10 in. while the cross section is 0.500 in. high by 2 in. wide. A flat-faced *drop hammer* supplies the necessary force, and sticking friction at the interface may be assumed. The flow stress of this hot billet is 2000 psi.
(a) Determine the *force* needed to perform this operation.
(b) Determine the *work* needed.
Use the slab method of analysis.

8–13. Wire drawing is to be performed by using a die of semi-angle (α) of 6 deg and a reduction $r = 0.20$.

(a) Calculate the ratio of the contact area between the tool and workpiece to the average cross-sectional area of the deformation zone.

(b) Determine Δ.

8–14. The parameter Δ has been defined as the ratio of the length of an arc across the midsection of the deformation zone to the contact length, the arc being centered on the apex of a cone or wedge formed by extrapolating the die walls.

(a) For wire drawing, show that this definition leads to

$$\Delta = \alpha(1 + \sqrt{1 - r})^2/r$$

(b) Using this definition, compare the values of Δ with that given by Eq. (8–38), where

(1) α = 10 deg, r = 0.25
(2) α = 45 deg, r = 0.50

8–15. As sheet is cold-rolled it is wrapped around a coiler (physically much like a roll of scotch tape). The coil diameter must be large enough so that coiling involves only elastic bending. If the sheet is 4 feet wide, 0.035 in. thick, and has a yield strength of 30 ksi, what should be the minimum diameter of the coiler?

8–16. A strip tensile specimen has a surface grid imprinted on its face where ℓ_o = 2 in. and w_o = 0.75 in. When it is loaded into the plastic region at a given instant, it is observed that ℓ = 2.35 in. and w = 0.710 in. What is the R-value based upon these observations?

8–17. Calculate the height-to-diameter ratios for drawn cups with drawing ratios of 1.8 and 2.5. Assume constant thickness.

8–18. A typical beer can (aluminum) is 5.25 in. high by 2.438 in. in diameter; the wall thickness is about 0.005 in. while the thickness on the bottom is 0.016 in. What starting blank diameter would be needed to make this can?

9 machining or cutting

9.1 INTRODUCTION

Machining can be envisaged by considering an example that should be known to the reader. Starting with a block of wood and a sharp knife, the process of whittling involves the cutting or carving of small pieces (or chips) of wood from the initial block until some desired shape is produced. In the nomenclature of machining, the initial block is the *workpiece*, while the knife is the *cutting tool*. In this process, one hand holds the block and is considered as the *workholder*, and the second hand holds the knife and provides the necessary force *and* relative motion between the work and tool; this hand is viewed as the *toolholder*. All of the essential components involved in any machining setup regarding the workpiece, cutting tool, their relative positions, and motion to produce some desired surface shape are embodied in this simple illustration.

Using this example, we may define *machining* as that process which produces a part of a desired shape and size by *removing* material (called chips) from a parent workpiece through the use of sharp-edged cutting tools. In effect, an oversized workpiece is carved to its desired shape. Since the vast number of industrial materials that are machined are metals, and most experimental or empirical information is available for these solids, the expressions "metal cutting" and "machining" are often used interchangeably. Near the end of this chapter some comments will be made about the machining of plastics, since interest in these materials has been on the increase.

The three basic components of any common machining operation are the machine tool, the cutting tool, and the workpiece. The *machine tool* consists of a group of subassemblies whose purpose is to maintain the relative positions of the workpiece and the cutting tool, and to provide the necessary power input and relative motions between them in order to produce the shape and size desired. The *cutting tool* may have one or more edges, the geometry of which may be accurately and deliberately controlled or which may have a random configuration.

Finally, it is crucial to realize that whatever shape is produced from the initial workpiece, it is *always* the result of the relative motions of the workpiece and cutting tool. From certain *basic* machine tools, modifications have been made to produce differently named tools, yet if one considers the relative motions involved, it will be apparent that these more advanced machine tools are really off-shoots of the basic ones we shall emphasize in this text.

9.2 BASIC MACHINE TOOLS AND INVOLVED MOTIONS

9.2.1 Engine Lathe

Figure 9–1 is a sketch of a typical engine lathe where the major components are indicated. There is no need to be too concerned about the various component names (there are several texts—for example, Refs. [1 through 4]—which do this in great detail, and these can be consulted if the reader wishes to pursue this in depth); what is more important is

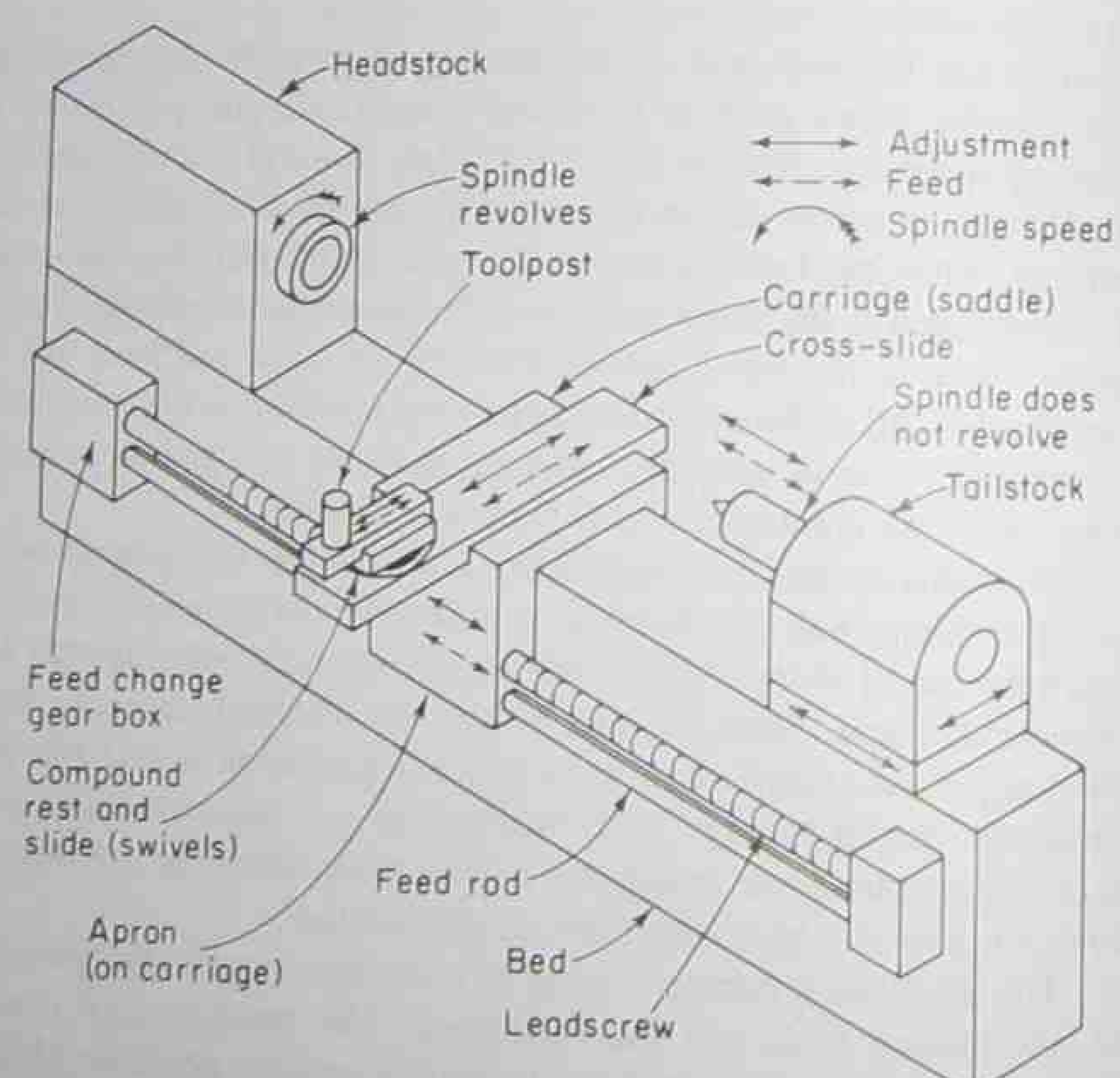


Figure 9–1 Schematic showing major components and movements of an engine lathe.

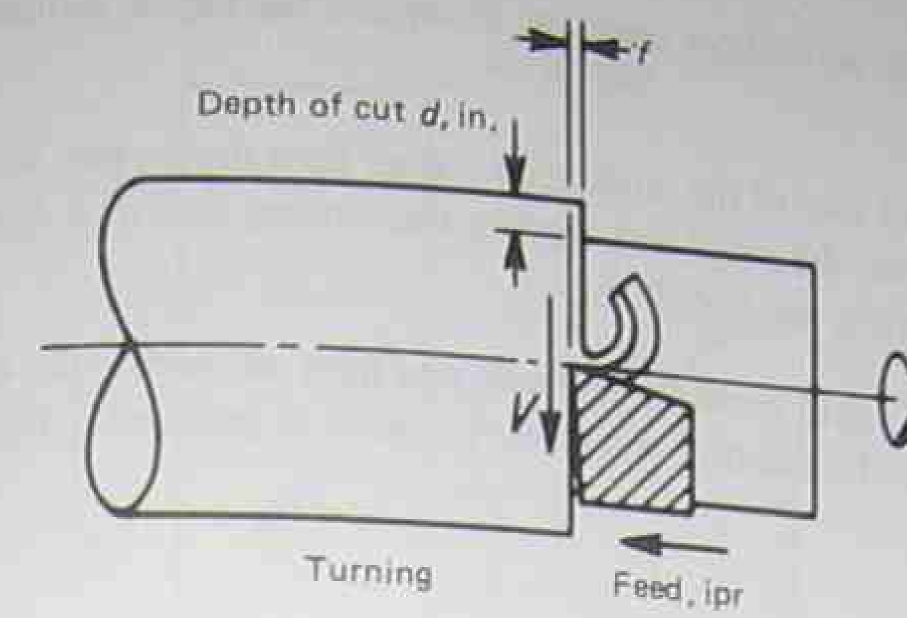


Figure 9–2 Illustration of turning, showing tool and workpiece motions.

to note that machine tools which are categorized as a lathe, or an offshoot of a lathe, possess one thing in common. The workpiece is *rotated* and the cutting tool is *almost* always fed or moved in a straight line.* Any surface that results from this combination of motions can, at least theoretically, be produced on a lathe. Note that this may not always be the most efficient way to machine such a surface, since other machine tools, by using their relative motions, may provide a more economical way to do this. Of the more common *operations* performed with lathes are

1. Turning. Here the work is rotated and the cutting tool moves in a straight line parallel to the axis of rotation of the work. The net result, regardless of what cross-sectional shape the initial workpiece had, is a cylindrical surface.† In most cases, the workpiece would be initially cast, rolled, extruded, and so on to be approximately circular prior to machining. Figure 9–2 is a schematic of a turning operation.

2. Tapering. Here the work rotates and the tool moves in a straight line at some angle with the axis of rotation of the workpiece. The manner by which the tool is made to move at such an angle can be done by offsetting the workpiece through the tailstock or

* One obvious exception is where curved surfaces are produced by guiding the tool to move in two directions in a plane simultaneously so as to produce a curve-like surface.

† Note that if the workpiece had a square cross section, the cutting tool must project far enough into the section if the resulting shape is to be cylindrical throughout. See Fig. 9–3.

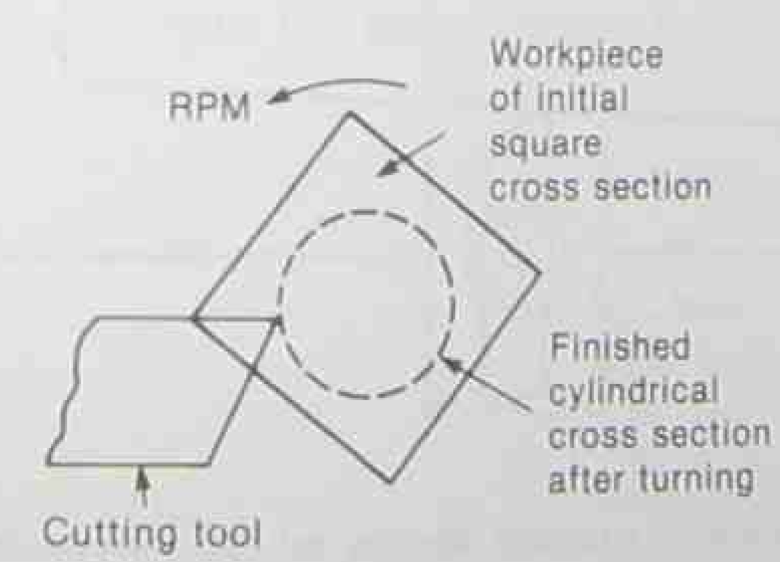


Figure 9–3 Turning an initially square workpiece to produce a cylindrical surface.

by using a taper attachment. Again we will not stress the hardware, but Fig. 9–4 illustrates this process schematically.

3. Facing. Often, the end face of the workpiece, rather than the length, must be machined. This is done by rotating the work and moving the cutting tool in a direction perpendicular to the axis of rotation; see Fig. 9–5.

4. Drilling. By adapting a drill bit to the tailstock and then advancing the drill in a straight line into the center of the rotating workpiece, a hole is produced; Fig. 9–6

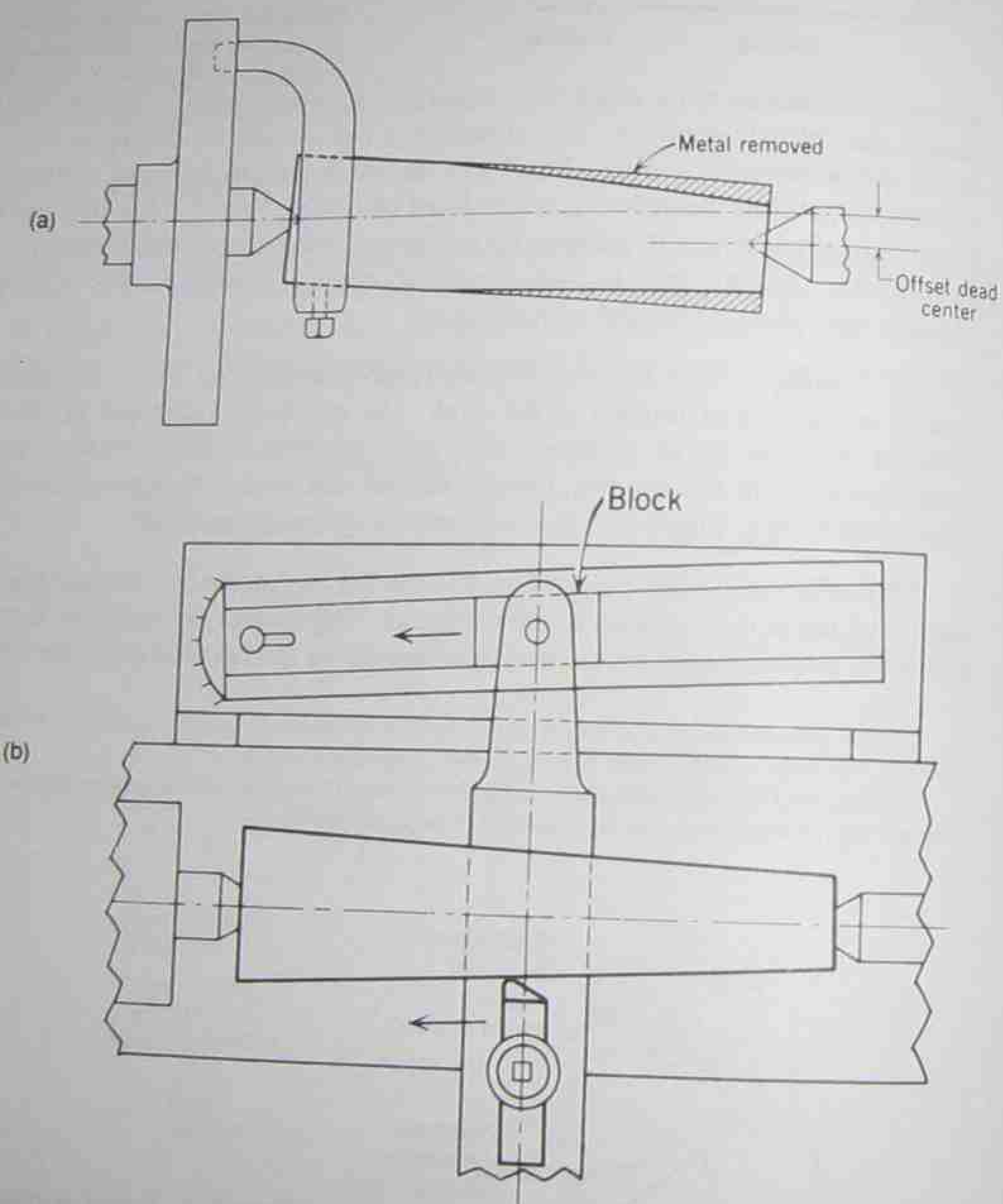


Figure 9–4 Taper turning using a tailstock offset (a) and a taper attachment (b).

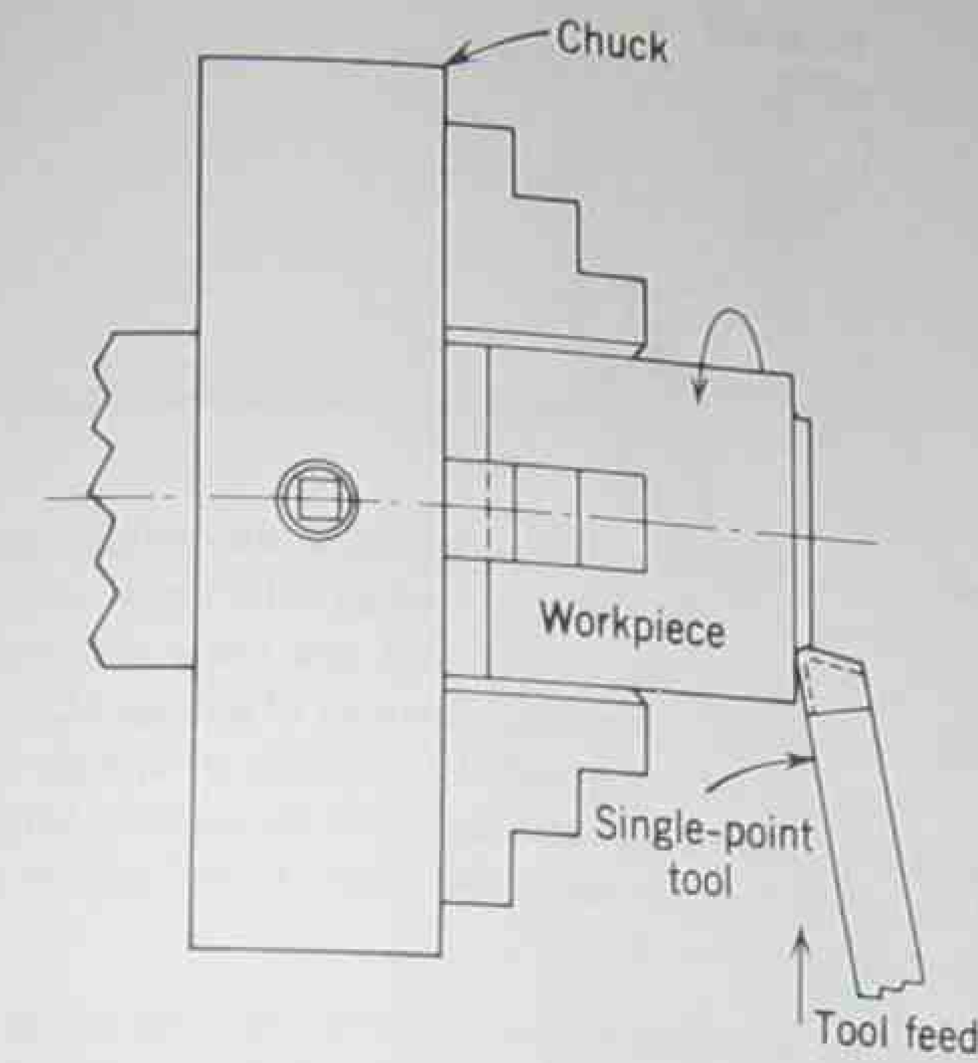


Figure 9–5 An illustration of facing.

illustrates drilling. Drill presses, to be discussed shortly, are used to machine most *initial* holes, but in cases where a hole is to be made in reference to other surfaces machined on a lathe, it is most sensible to drill the hole with lathe motions. In many instances, a drilled hole possesses adequate dimensional accuracy for a desired function; for example, it may serve simply as clearance for a bolt to fit through. Other requirements may demand a hole of closer size variation and roundness that exceeds the capability of drilling, so *finishing* operations follow drilling. To employ such finishing operations *always* requires a hole to start with; these subsequent operations simply enlarge the initial hole to some final desired size and surface finish. Two such operations are discussed next.

5. Boring. This is really *internal* turning. A workpiece containing a drilled hole is rotated while the cutting tool moves in a straight line parallel to the axis of rotation. Rather than a cylindrical surface on the outer diameter (turning), an internal cylindrical hole results. Figure 9–7 shows a boring operation.

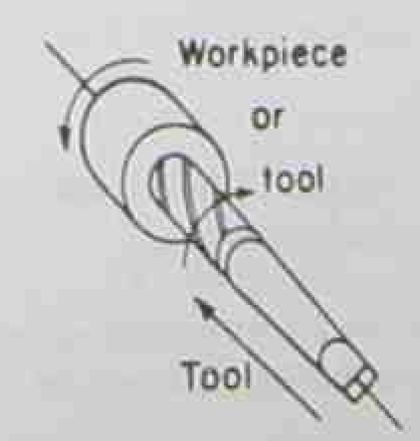


Figure 9–6 The essentials of drilling, showing tool and workpiece motions.

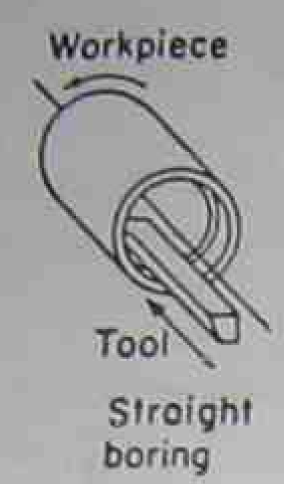


Figure 9–7 An illustration of boring.

6. Reaming. On occasion, functional needs may require a hole whose desired size variation and surface finish exceed those produced by boring.* In single point operations such as boring, elastic deflection of the cutting tool can cause dimensional variations on the cut surface. In reaming, a tool containing a number of cutting edges (see Fig. 9–8) is adapted to the tailstock and moved in a straight line into the previously bored hole. Each cutting edge removes a chip of small thickness and since the cutter is balanced in a radial sense regarding the forces involved in cutting the chips, a final hole of quite uniform diameter and *smoothness*† results.

7. Threading. In this operation, the cutting tool is ground as indicated in Fig. 9–9 and, via the lead screw and gearing in the headstock (see Fig. 9–1), is moved at a predetermined rate so as to provide a helical type of cutting action, thereby producing a thread of a constant pitch. The full thread is not cut in one pass of the cutting tool; instead, after each pass, it is moved back to the beginning of the cut and advanced radially inward, so that each successive pass for the full length of thread increases its height until the final depth is achieved. Note that there are other methods for producing threads.‡

8. Other comments. It is important to highlight certain aspects involved in the discussion up to this point. First, the motions involved have two general characteristics. The relative *linear* velocity, called the *cutting speed*, between the cutting tool and workpiece is *high*. Although this word is ill-defined at this point, we shall return to it later in this chapter. In addition, the velocity at which the cutting tool moves into the workpiece (called the *feeding* motion or velocity) is always *much lower* than the cutting velocity (see Fig. 9–11). These motions have tremendous significance in machining operations regardless of the type of machine tool employed.

Additionally, two broad types of cuts can be defined. In one, except for the surface finish that results, the major shape of the final surface is caused by a *generating* type of cut. There, the shape of the cutting tool has a negligible effect on the resultant surface

* In certain current practices, boring at very high speeds using diamond cutting tools produces a finished surface of excellent finish and dimensional consistency.

† Surface finish was discussed in Chapter 3.

‡ *Taps* for internal threads and *dies* for external threads are also used (see Fig. 9–10). With taps or dies, the full thread depth is cut in one pass. Rolling of threads can also be accomplished, but this is not pursued further.

Figure 9–8 Several types of reamers. Photo courtesy of W. H. Durrant.

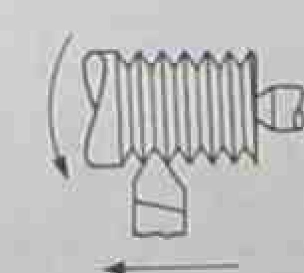
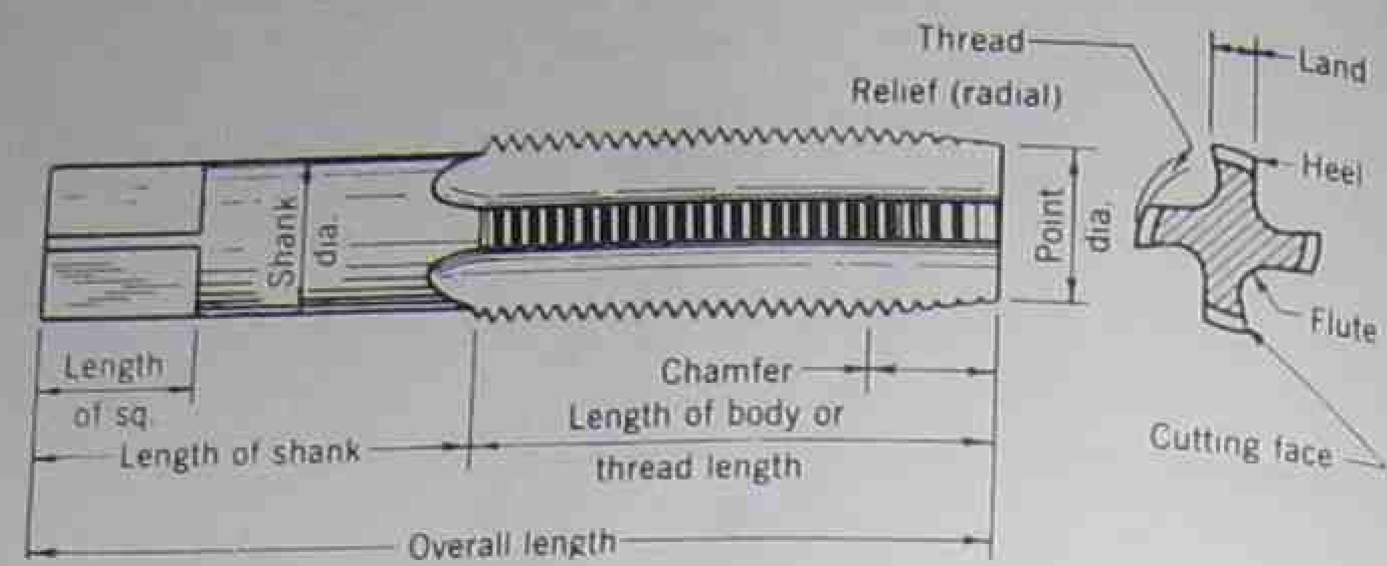


Figure 9–9 A schematic of threading using a single tool.

(a)



(b)

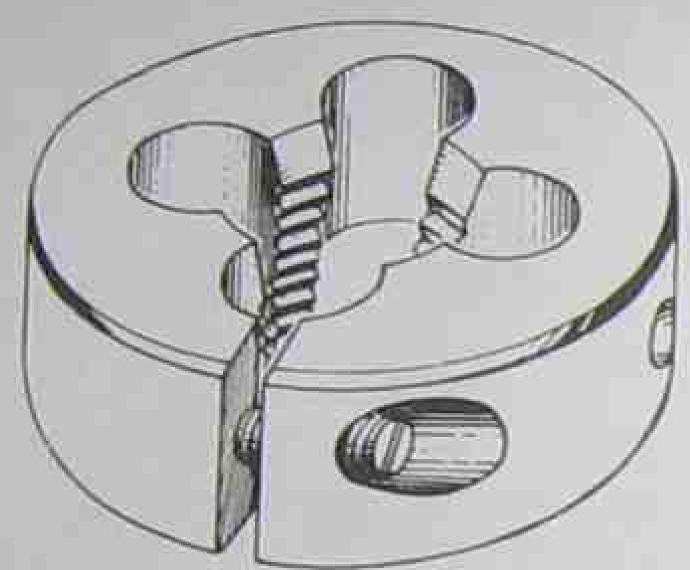


Figure 9–10 (a) Components of a tap for cutting internal threads and (b) a die for cutting external threads.

shape. Figure 9–12 illustrates this point. For some operations, the ground shape of the cutting tool produces the final shape of the workpiece. These are called *forming* cuts, and Fig. 9–13 illustrates this point.

What now follows is a similar coverage of other basic machine tools and operations; again the emphasis is not upon detailed descriptions of machine tool components; rather the relative workpiece-tool motions are stressed.

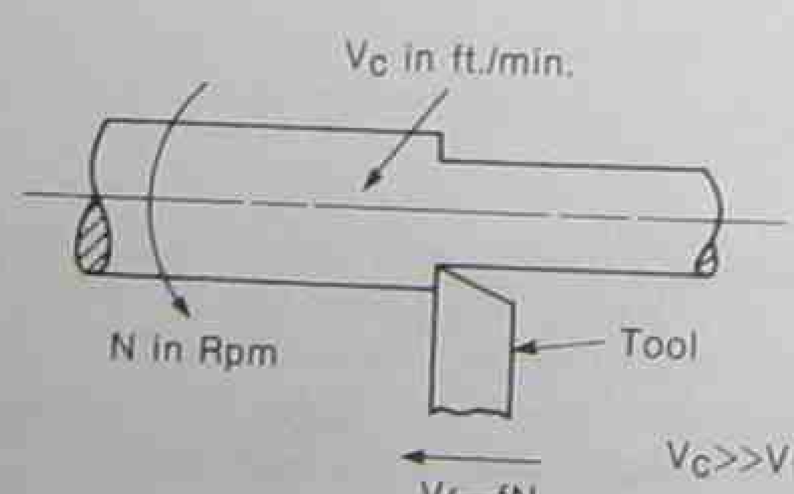


Figure 9–11 Schematic to illustrate the cutting and feeding velocities in turning.

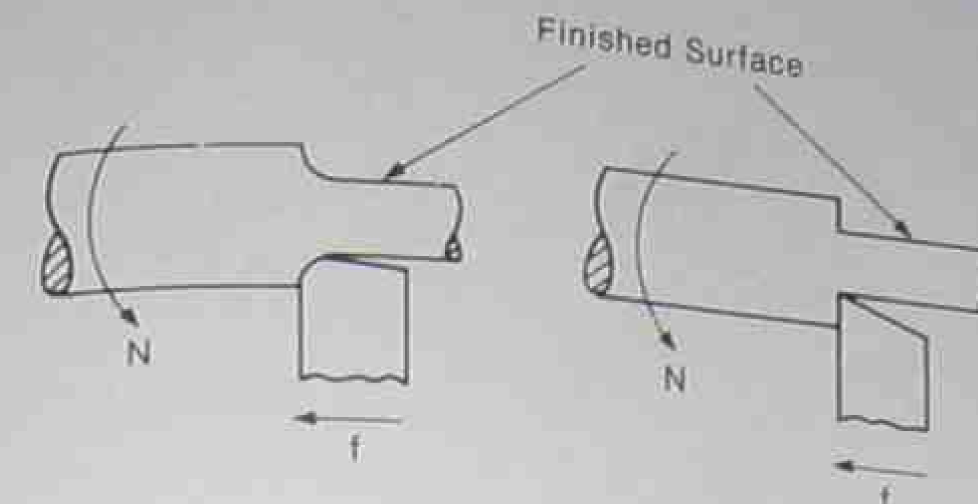


Figure 9–12 Illustrations of generating cuts.

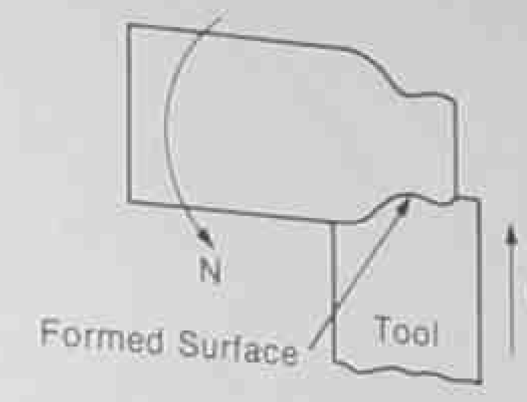


Figure 9–13 Illustration of a forming cut.

9.2.2 Milling Machines

By far the most common motions involved here utilize a cutting tool that *rotates* and a workpiece that moves in a *straight line*. Figure 9–14 illustrates a number of the more common combinations. In most instances, flat surfaces are generated, although formed surfaces can be produced. The cutting tool may rotate about a vertical axis (vertical milling) or a horizontal axis (horizontal milling). In fact, if a drill bit is adapted to the rotating head, holes can also be produced; this is done on occasion, but is usually not the primary manner for cutting holes. One of the principal differences between lathe and milling operations is that the latter utilizes a cutting tool composed of a *number* of cutting edges rather than a single cutting edge. Later, when rates of metal removal are discussed, the impact of this difference will be more meaningful.

9.2.3 Drill Presses

The most common method used to produce a hole employs a type of drill press. Here, the cutting tool called a *drill* (see Fig. 9–15) is fitted into the spindle of a drill press shown in Fig. 9–16, most often by the mating of two tapered surfaces (that is, the drill and the spindle of the press). In the majority of cases the workpiece is clamped to a supporting surface called the *table*, and the drill itself is then rotated about a fixed axis and moved in a straight line motion so as to sever chips and produce a hole. Although numerous names are given to many types of drill presses, the fundamental motions just discussed are similar to all.

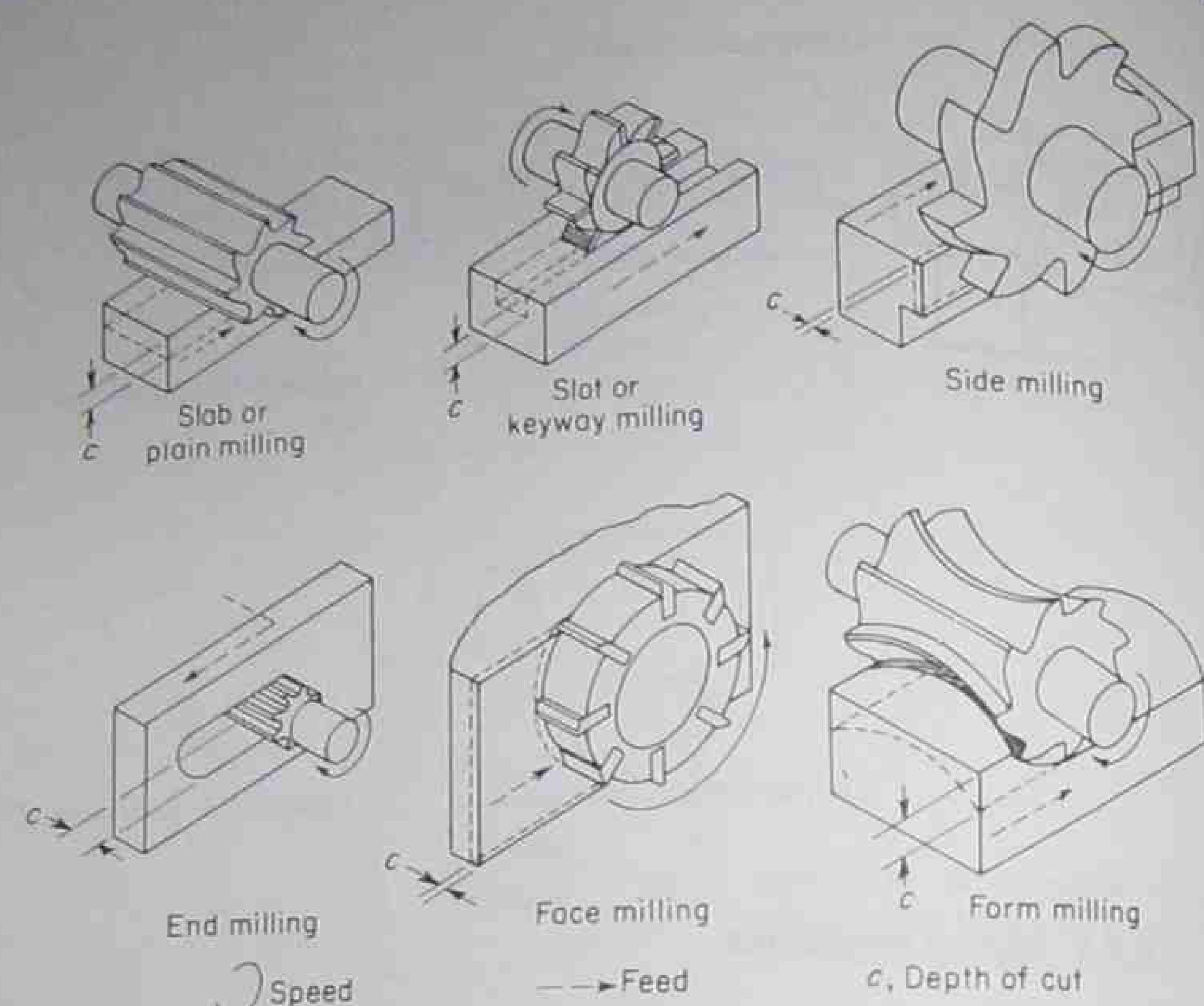


Figure 9–14 Various types of milling operations to produce different surface shapes.

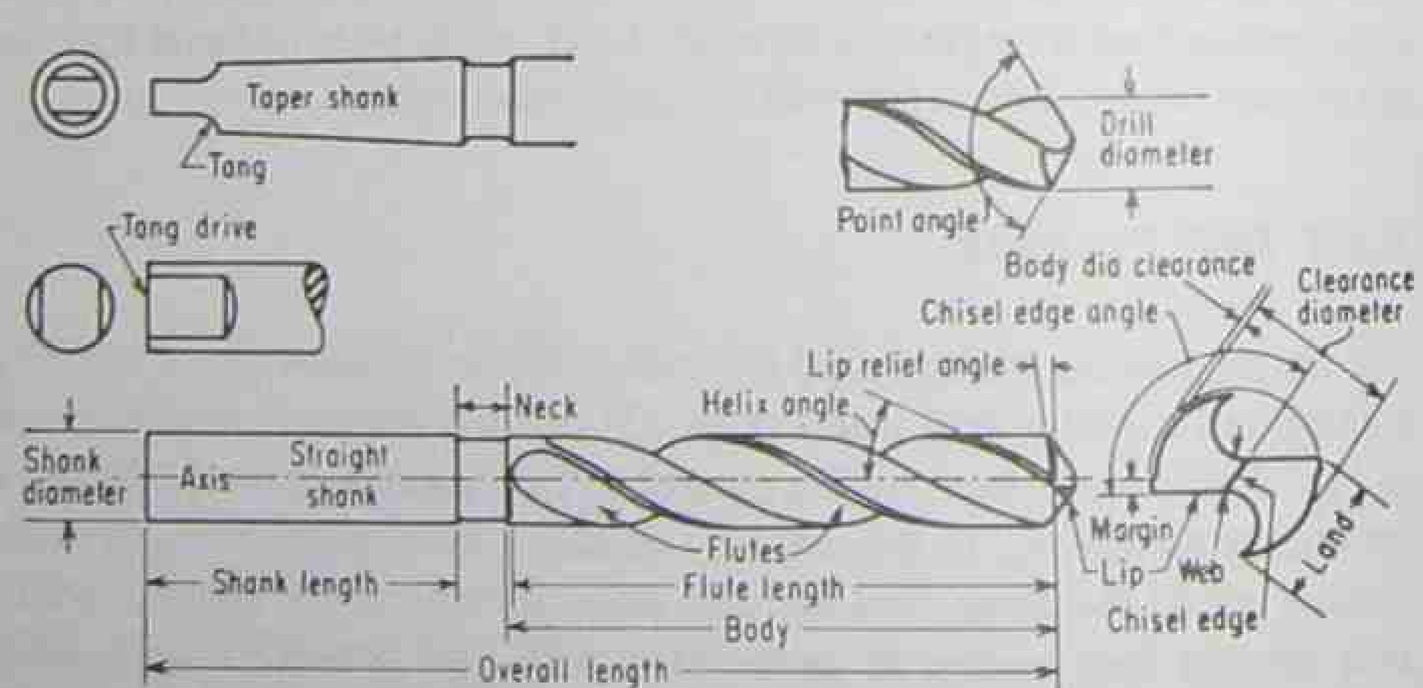


Figure 9–15 Twist drill nomenclature.

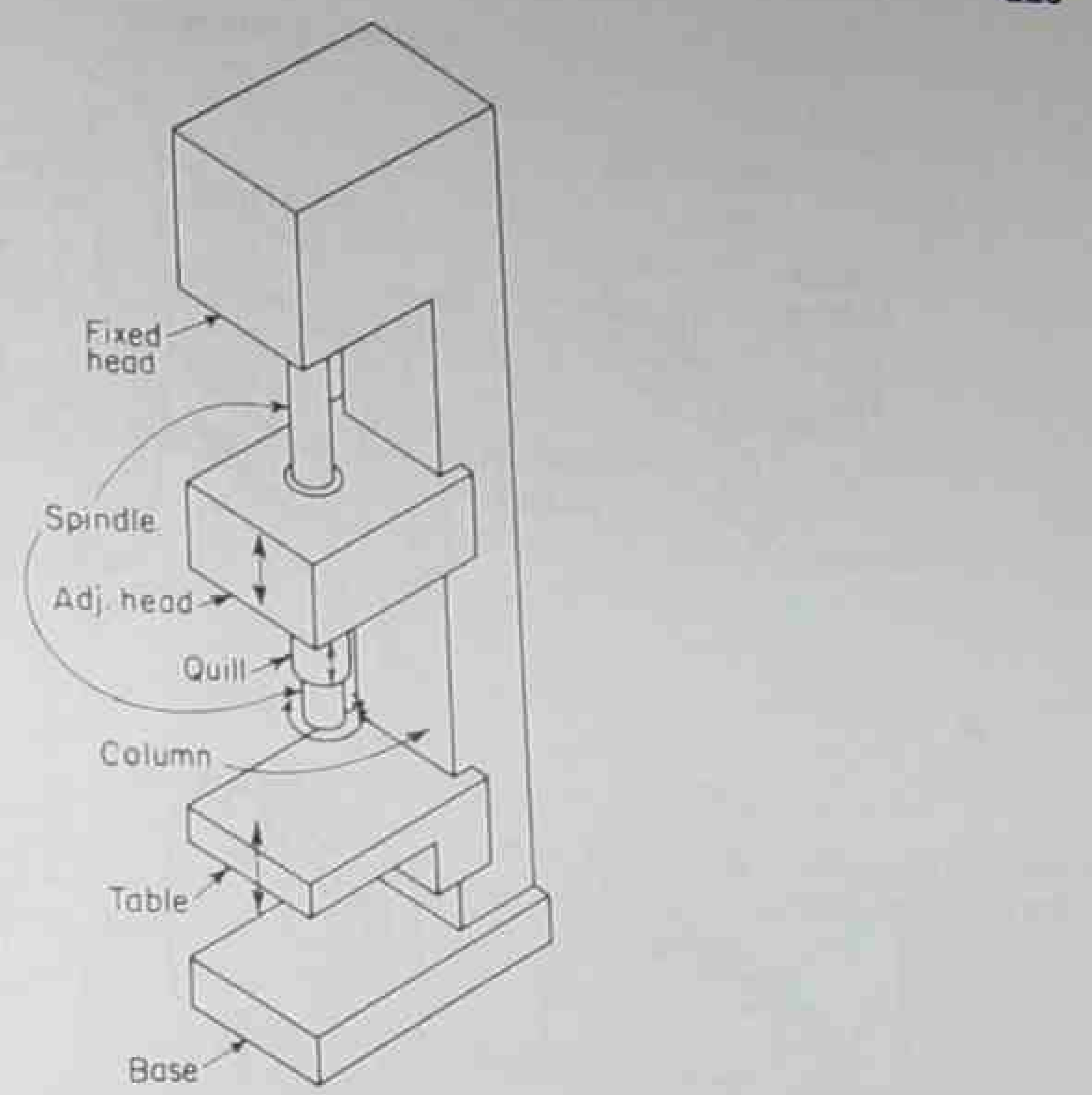
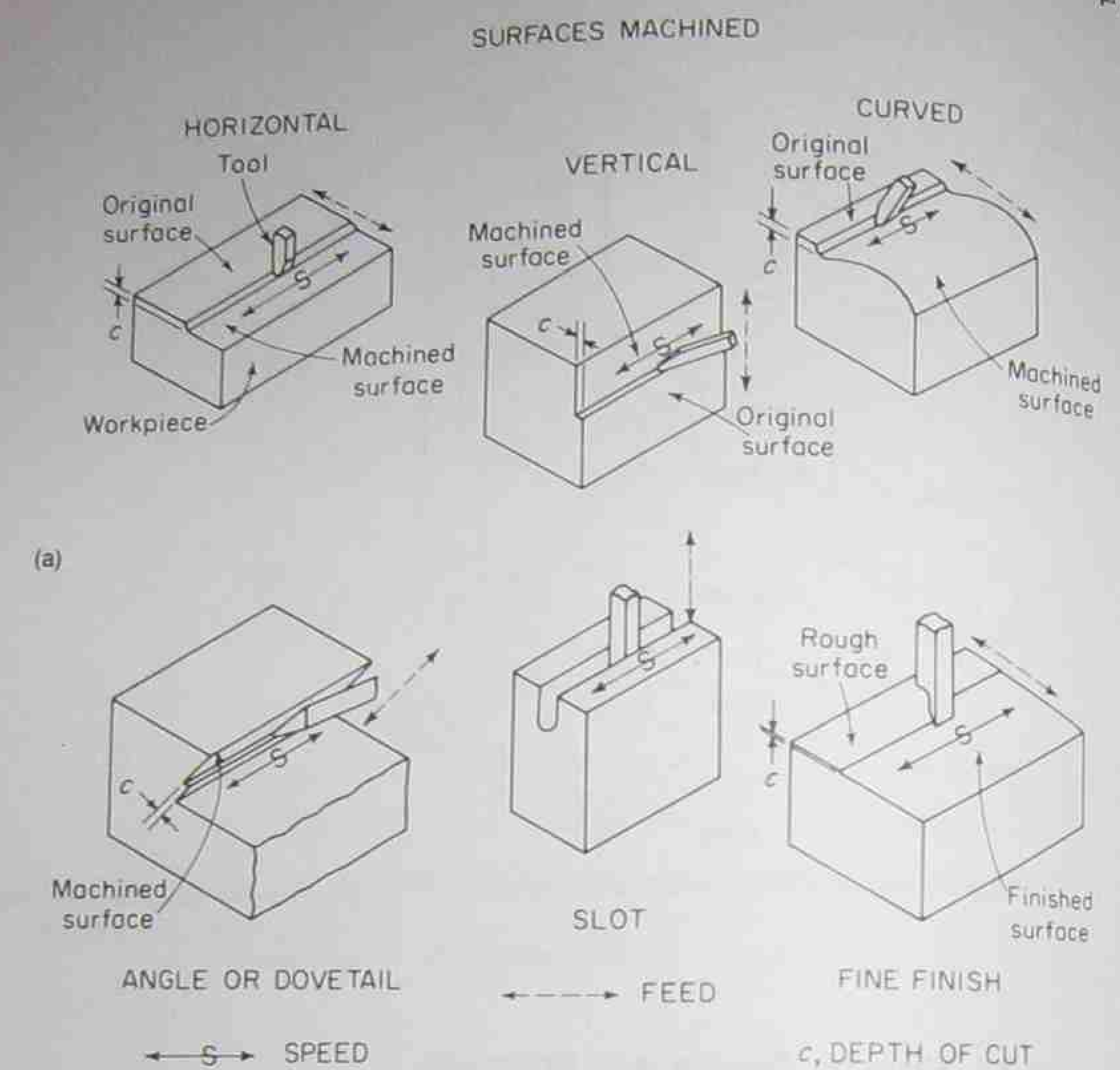


Figure 9–16 Schematic of a drill press showing major components and movements.

9.2.4 Shapers, Planers, and Broaches

Shaping is illustrated in Fig. 9–17. There, a single point cutting tool (that is, one cutting edge) reciprocates back and forth on each *stroke* of the tool. During the return portion of the stroke, when *no* cutting takes place, the workpiece, which is securely clamped to the table, moves or *feeds* at right angles to the tool motion. Flat, vertical, or angled surfaces can be produced by shaping, as indicated in Fig. 9–17. Because the cutting action is not continuous, as in turning, and since only one cutting edge removes chips during a *portion* of each back and forth stroke, the rate at which metal is removed by shaping is rather low. A *planer*, although usually larger in size, is very similar to a shaper; the major difference is that while the work reciprocates back and forth, the tool *moves* during each return stroke in a direction perpendicular to the cutting direction. Note that while the tool is moving prior to each cut, the workpiece is clamped to a table that simply moves back and forth along a fixed path (see Fig. 9–18). Of all the machine tools discussed to this point, the *rate* at which metal is removed is lowest for a shaping operation. By making one major alteration, this same operation produces a great increase in metal removal. This operation, called *broaching*, is illustrated in Fig. 9–19. In essence, a number of cutting edges are



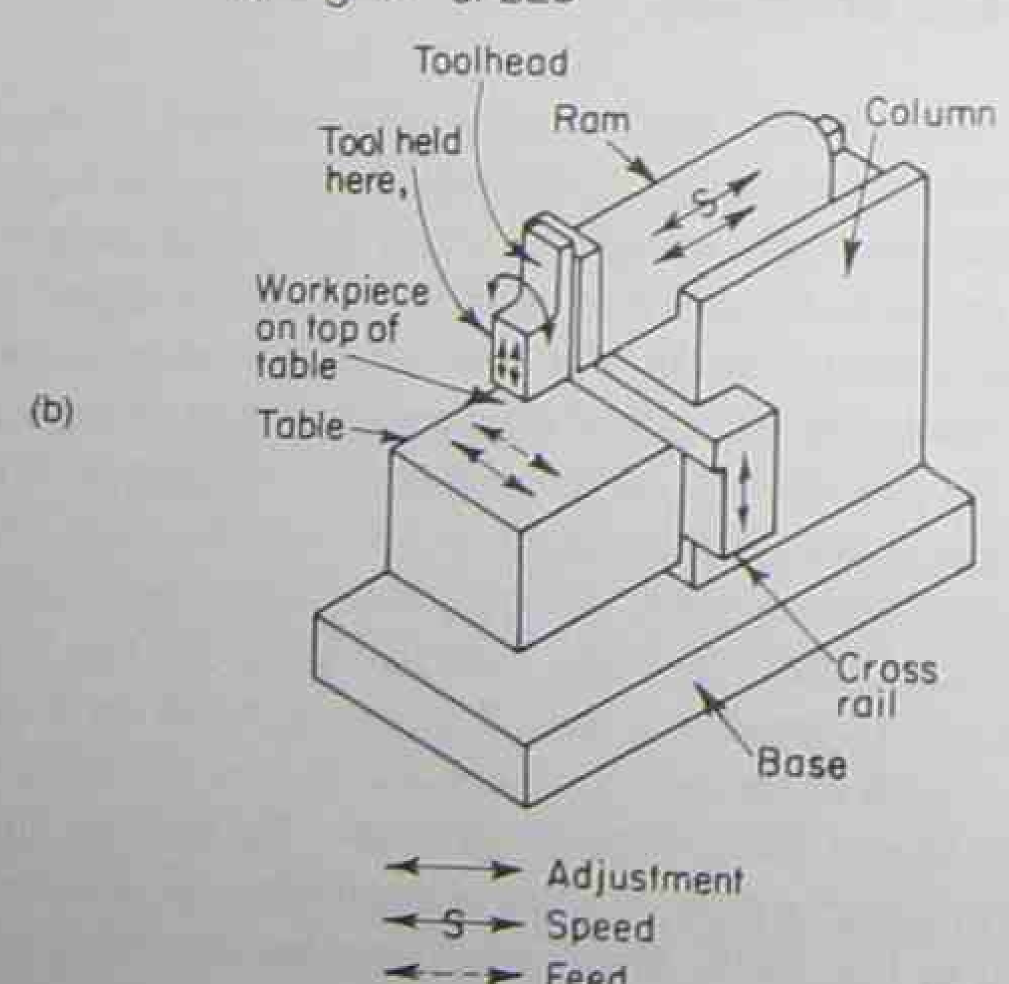


Figure 9–17 (a) Typical surfaces produced by shaping and (b) essentials of a horizontal shaper.

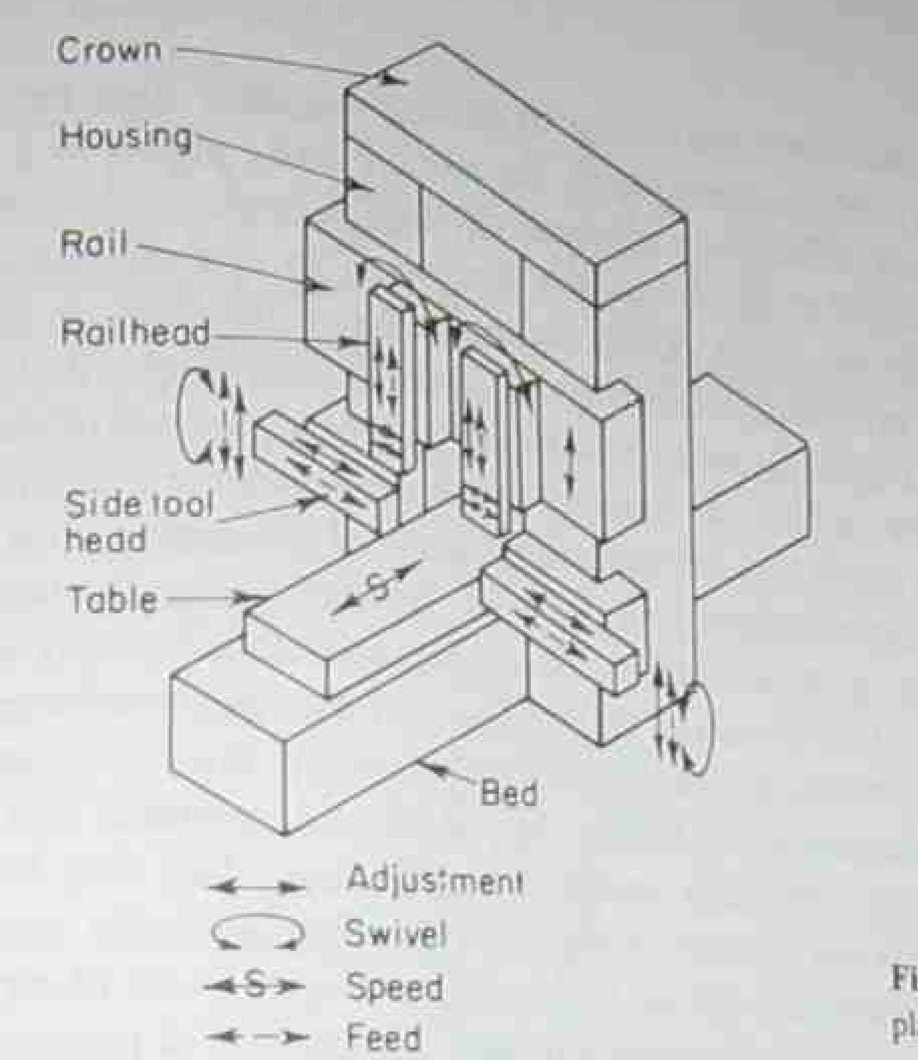
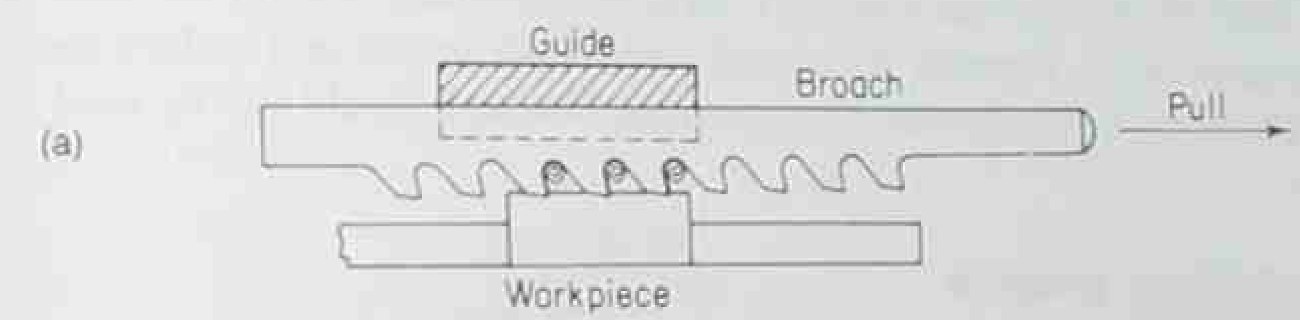


Figure 9–18 Schematic of a horizontal planer showing major components and movements.



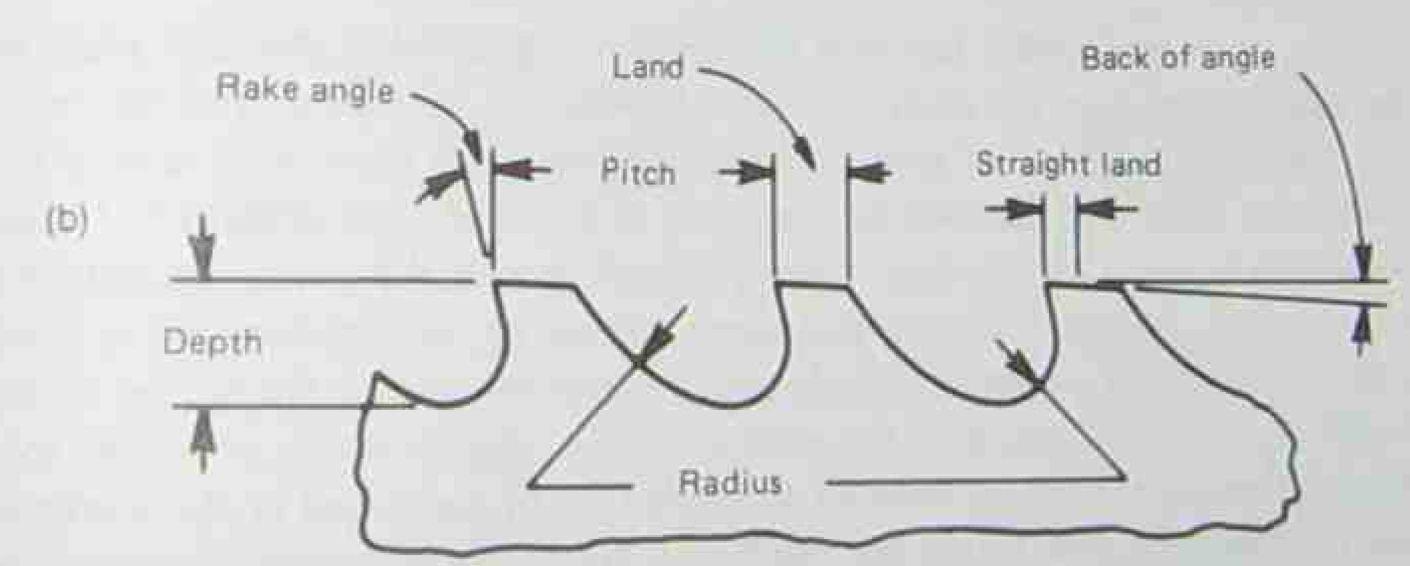


Figure 9–19 (a) Schematic of a broaching operation and (b) nomenclature of one type of cutter.

adapted to a single toolholder and moved into and through the workpiece much like a shaping operation. Each cutting edge removes a chip successively lower than the previous edge. Since a number of edges are employed, the rate of metal removal increases dramatically per stroke as compared with shaping, yet the fundamental work-tool motions are identical.

9.2.5 Grinders

In one respect, *grinding machines* are similar to milling machines since the cutting tool, here called the *grinding wheel*, rotates while the work usually moves in a straight-line motion. However, there are at least two fundamental differences between grinding and all the basic operations discussed to this point. The first has to do with the cutting tool itself. In all previous operations, the edges of the cutting tool (whether one, as in turning, or multiple, as in milling) are ground carefully to some desired shape; that is, they are geometrically controlled edges. Grinding wheels are usually composed of small particles of aluminum oxide, silicon carbide, or diamond which are *bonded* in a matrix to form the grinding wheel. Here the small bits are random in terms of orientation; thus a geometrically controlled series of cutting edges does not exist. Besides this randomness of cutting-edge orientation, grinding, as with other abrasive operations called *honing* and *lapping*, always involves the removal of *very small* chips compared with the other basic operations. In terms of rate of metal removal, these abrasive operations involve exceedingly low levels, but this is the fundamental reason that they produce exceptional size control and surface finish as compared with other machining operations. This will become more evident when cutting forces are covered later. Also, grinding is usually essential where extremely hard metals or ceramics are to be machined. Figure 9–20 illustrates several grinding operations. Any engineer involved with the manufacture of products should be aware that machining, and *especially* abrasive operations, should be used only when functional necessity demands the type of dimensional control and surface finish that can be produced by such methods. They are expensive and should not be used indiscriminately.

9.2.6 A Short Wrapup

To this point, machining has been defined and certain *basic* machine tools have been discussed in terms of the relative motions between the workpiece and the tool.

As an illustration of how the basic machine tools have been altered to increase production rates, consider two modifications of the basic lathe. Figure 9–21 shows what is called a *turret* or capstan lathe. There, depending upon the cutting operations required, a number of different tools are positioned and set on various faces of a movable tool holder—a turret. Suppose an initially round workpiece is to be finished as shown in Fig. 9–22(a). As with a basic lathe the work is held in position and is driven by the spindle. Now, however, there is no tailstock to support the opposite end of the workpiece. Note that this limits the free length that can be machined in this manner, since the workpiece is analogous to a cantilever beam; deflections during cutting can be excessive if the "beam" is too long. Instead, the turret replaces the tailstock but provides no end support. Figure 9–22(b) illustrates the various tools mounted in position on the turret. Beginning with the starting drill and proceeding in sequence through the tapping of the thread, all internal operations are accomplished as shown in Figure 9–22(c). Turning of the outer diameter and then cutting off the part to the desired length are accomplished with tools adapted to the *square* turret. Note from Figure 9–21 that this square toolholder, which is

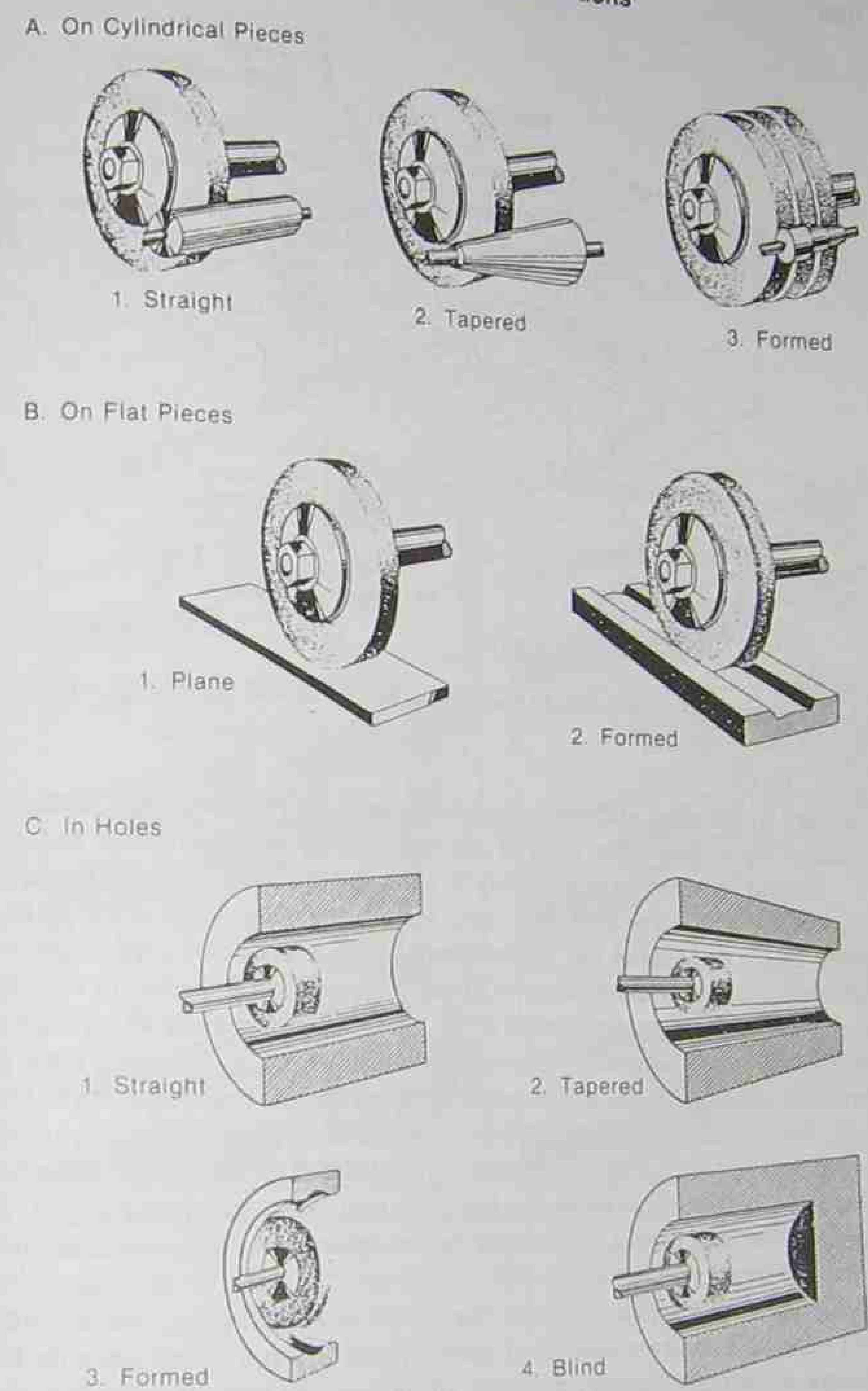


Figure 9–20 Various types of grinding wheels and operations.

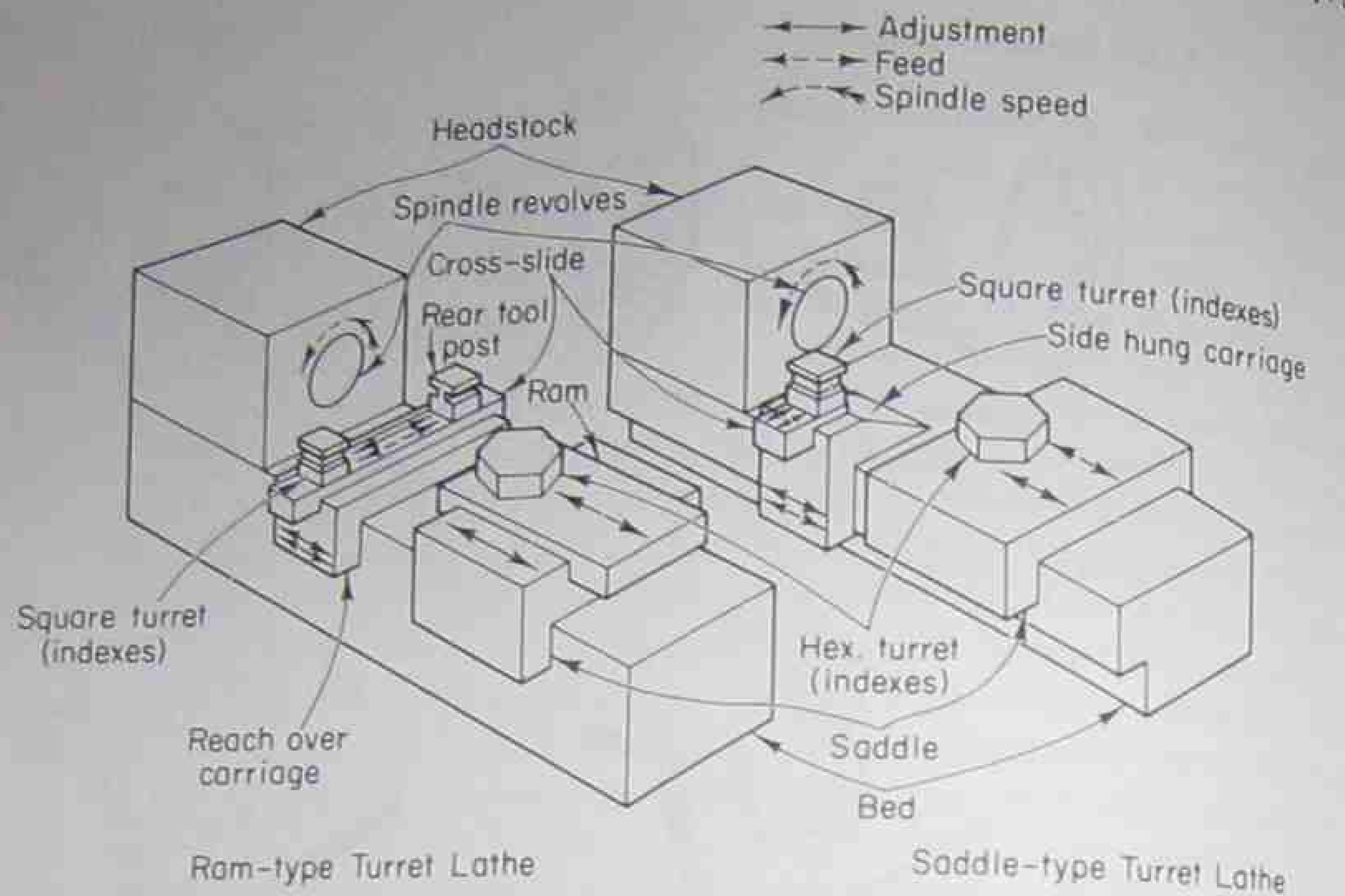


Figure 9–21 Schematic of a horizontal turret lathe showing major components and movements.

connected to the cross-slide, can be indexed in 90-degree increments, thereby providing the use of up to four other tools in addition to those mounted on the hexagonal turret.

Several comments are necessary to complete this discussion. First, each individual tool, whether adapted to the turret or the square toolholder, must be preset to govern the desired length and size of cut. As an example, the turning tool must project beneath the work surface a specified amount to produce the desired diameter; in addition, it must travel a particular distance parallel to the axis of the workpiece to produce the desired length of the turned surface. Various types of *stops* are used for the latter purpose; in essence, these stops cause a disengagement from the drive source at the desired point of travel. Second, after any cutting action is completed by a tool on the hexagonal turret, the turret is returned to its original position (by the hand of the operator or often by automatic means) at which point it rotates to bring the next preset tool into position. During this return, no *cutting* takes place as far as the hexagonal turret tools are concerned. Finally, if this part were to be produced with a basic lathe, it might be necessary to set one tool in position, complete that operation, then break down that setup, and so forth. With the turret lathe, this constant revision of setup for each tool is avoided, since the initial setup, although sometimes lengthy, positions all tools at one time. It is for this reason that a turret lathe will outproduce a basic lathe in terms of pieces per hour; however, this advantage of a turret lathe diminishes if only a few pieces are to be made. As an extreme example, if only one part were needed, a skilled machinist could produce this on a basic lathe probably before the tooling could be completely set on the turret lathe. Therefore, the *volume* of parts may often dictate which machine tool is most efficient.

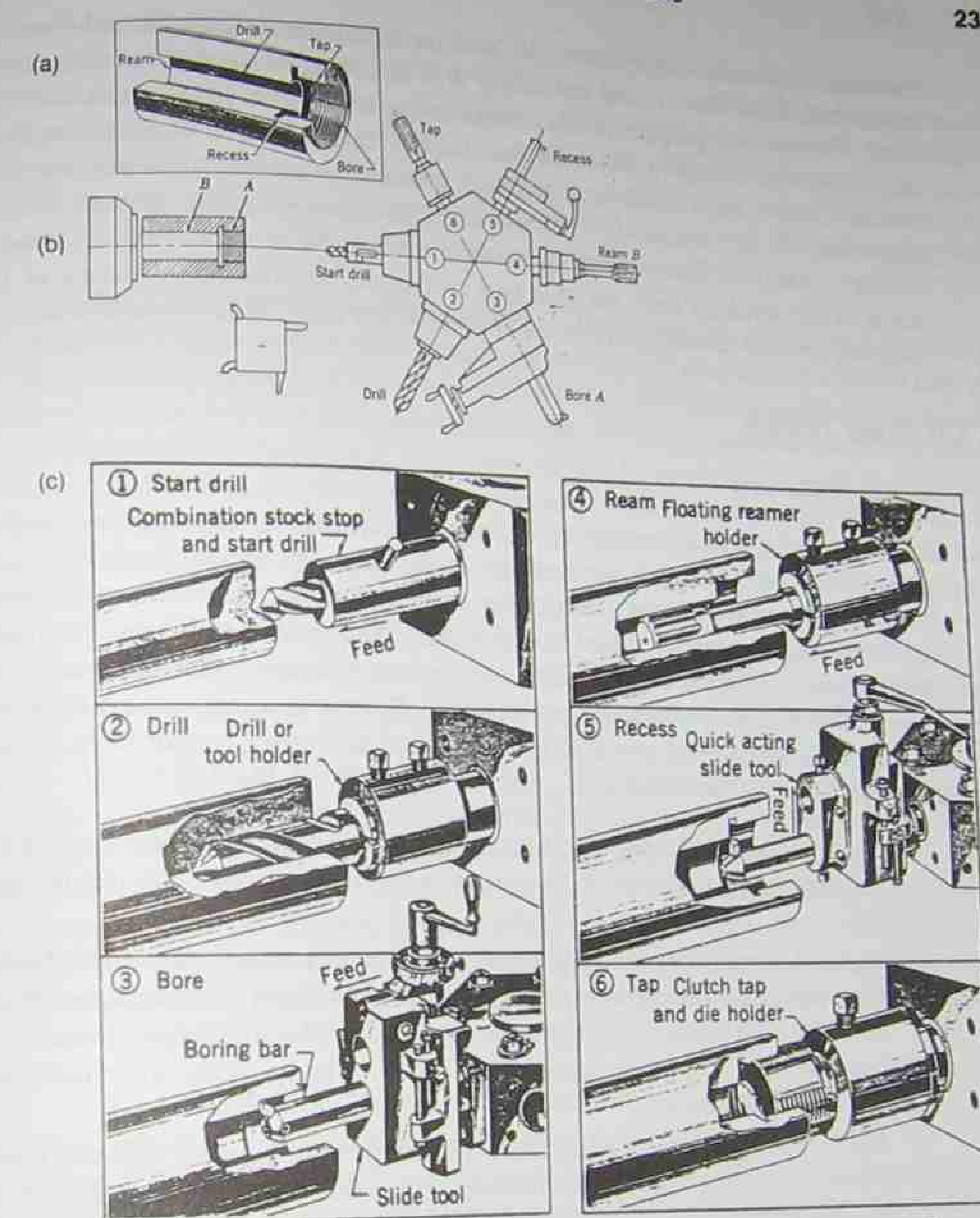


Figure 9–22 (a) Part to be produced on a turrent lathe, (b) setup illustrating sequence of operations related to turrent tolls, and (c) details related to internal cutting operations.

As a final step in evolving from a basic lathe to a similar machine tool capable of extremely high production rates (that is, pieces per hour), the *automatic turning* (*or screw*) *machine* was developed. Rather than employing a turret, all tools are prepositioned on various independent toolholders whose feed rate and length of travel are governed by various devices such as cams and screws. Unlike the turret lathe, where the turret must be retracted to position the next tool, all tools on an automatic screw machine are already positioned. Since there is less *noncutting* time involved, production rates are higher.

The principal purpose in this comparison is to illustrate how a basic machine tool may be modified to give increased production, *yet* the relative work-tool motions are identical. Other basic machines have been modified in an analogous manner, and it is hoped that this one illustration will prove adequate in presenting the concepts behind such changes. Note also that with modern control devices, NC or CNC may be added to the name of the machine tool, yet it is best to think of the machine in terms of relative motions, thereby considering it as an offshoot of the more basic machine tools.

9.3 CUTTING TOOLS

9.3.1 Materials

In all machining operations, the removal of chips involves an input of work from some power source. If we restrict the emphasis to metals, there is usually a large amount of plastic deformation involved in producing and severing the chip from the workpiece; the work or energy input is converted to heat in the cutting zone. Added heat is developed as the chip rubs over the tool. (Chip formation is discussed later.) This somewhat complex combination of high temperatures and frictional effects at the tool-work interface requires an acceptable cutting tool produced from a material possessing certain basic characteristics. These include the following:

1. The tool *must* be *harder* and *more wear resistant* than the workpiece. If we try to cut steel with a tool made from commercially pure aluminum, we will not get very far.
2. The tool should *not* be susceptible to undesirable changes in microstructure at the temperature prevailing during cutting. As an example, a tool of plain-carbon steel may possess a martensitic structure initially, but at elevated temperatures this structure may change due to tempering,* and the initial hardness and wear resistance will decrease considerably.
3. The tool should possess adequate *toughness* to avoid fracture, especially with intermittent cutting (such as in milling), since impact loading can induce severe stresses in the tool.
4. If, in addition to the above properties, the tool possesses high thermal conductivity, heat may be more rapidly transferred from the cutting zone into other members of the machining setup. This would aid in prolonging the life of the tool.

Based upon these requirements, the various tool materials are as follows:†

9.3.1.1 Plain High-Carbon Steels. These contain about one percent carbon and are initially heat-treated to produce a microstructure of martensite and cementite (iron

* Discussed in Chapter 7.
† For typical industrial use, some of these tool materials find much greater application than do others.

carbide). Their initial hardness is of the order of 60 to 64 Rockwell C. However, at temperatures in the range of 200°C to 300°C and for modest time at such temperatures, these materials become softer and less wear resistant. This change is due to tempering of the martensite and since higher cutting speeds at the work-tool interface lead to higher temperatures, this tool material must be used at low speeds when cutting metals (thus they find limited use in most industrial applications). With low-strength materials such as wood and plastics, they can be used successfully at higher speeds.

9.3.1.2 High-Speed Steels (HSS). In addition to containing carbon of the order of 0.7 to 0.8 percent, these materials utilize various alloying elements that may total as much as 25 percent of the mass. Although their proper heat treatment is a bit more involved than that used with plain carbon steels, the basic maximum hardness is similar. Perhaps the key advantage in such a comparison is the ability of high-speed steels to maintain their initial hardness and wear resistance at temperatures up to 500°C. Thus, they can perform satisfactorily at higher cutting speeds than can the plain-carbon type; that is why they are called high-speed steels. There are literally hundreds of commercial grades of these tool materials, the major types being designated T, where tungsten is the major alloying element, or M, where molybdenum predominates. More specific and detailed information about compositions and recommended uses can be best obtained from the producers of these tool steels.

9.3.1.3 Cast Nonferrous Alloys. These contain the elements cobalt, chromium, and tungsten, which comprise about 95 percent of the total structure. A hardness of about 60 Rockwell C is typical and is obtained without heat treatment; in essence they are cast and then ground to the desired shape. They can be satisfactorily used at temperatures of 600°C but are more brittle than high-speed steel tools. Their introduction somewhat overlapped that of the next tool material, sintered carbides, and since these carbides were, in general, superior cutting tools, the cast nonferrous alloys never caught on to the extent anticipated.

9.3.1.4 Sintered (Cemented)* Carbides. Elements such as tungsten, titanium, tantalum, and chromium are combined with carbon to form hard, wear-resistant carbides. They are *sintered*† to shape and, in the presence of a binder (usually cobalt), display a hardness and wear resistance in excess of all previous materials. (They also have an elastic modulus about three times that of steel, which leads to smaller deflections under a given load.) Perhaps their major drawback is their relatively low fracture toughness, which makes them susceptible to fracture or chipping under impact loading. As with high-speed steel tools, there are so many types of carbides that recommendations as to specific use should be sought from the producers.

9.3.1.5 Ceramics. Aluminum oxide is the most widely used material of this

* *Coated* carbides usually consist of a base of tungsten carbide that is coated (using chemical vapor deposition) with a thin layer of titanium carbide, titanium nitride, or aluminum oxide.
† See Chapter 10.

type. Not only is it used in grinding wheels (as are silicon carbide and diamonds) but is sometimes used as single point cutting tools. Ceramics are extremely hard and wear resistant and can operate best in continuous cutting (for example, turning) rather than interrupted cutting.w

9.3.1.6 Cubic Boron Nitride. A more recent development in cutting tool materials involves the bonding of a thin layer of cubic boron nitride (CBN) to a carbide base material. Shock resistance is provided by the carbide substrate, while the CBN (which next to diamonds is the *hardest* material currently available) provides extremely high wear resistance. It is also chemically inert to elements such as iron and nickel and displays excellent oxidation resistance at elevated temperatures.

9.3.1.7 Diamonds. Man-made diamonds as well as low-grade natural diamonds are the hardest and most wear resistant of all tool materials; however, they are also very brittle. This usually restricts their usage to grinding wheels or in final finishing operations, such as precision boring, where they operate at high speeds but at a small size of cut. In such situations they can produce parts of close tolerances and excellent surface finish.

In closing this section it is important to note that no single tool material is best in all situations. Depending upon the operation, conditions of cutting speed and size of chip, intermittent or continuous cutting, and so forth, any one of the materials mentioned may prove most economical in a particular situation. For example, diamonds find little success in machining ferrous metals such as steels, while carbides often chip if the cutting speed and feed rate are too low. No attempt is made here to provide a detailed list of the pros and cons of each type of tool material regarding the most appropriate application.

Figure 9–23 illustrates the influence of temperature on the hardness of some of the common tool materials. Since temperature increases with cutting speed, it is obvious why, for example, sintered carbide tools can operate satisfactorily at much higher speeds than can plain-carbon tool steels. Ceramics, CBN, and diamonds are used at even higher speeds.

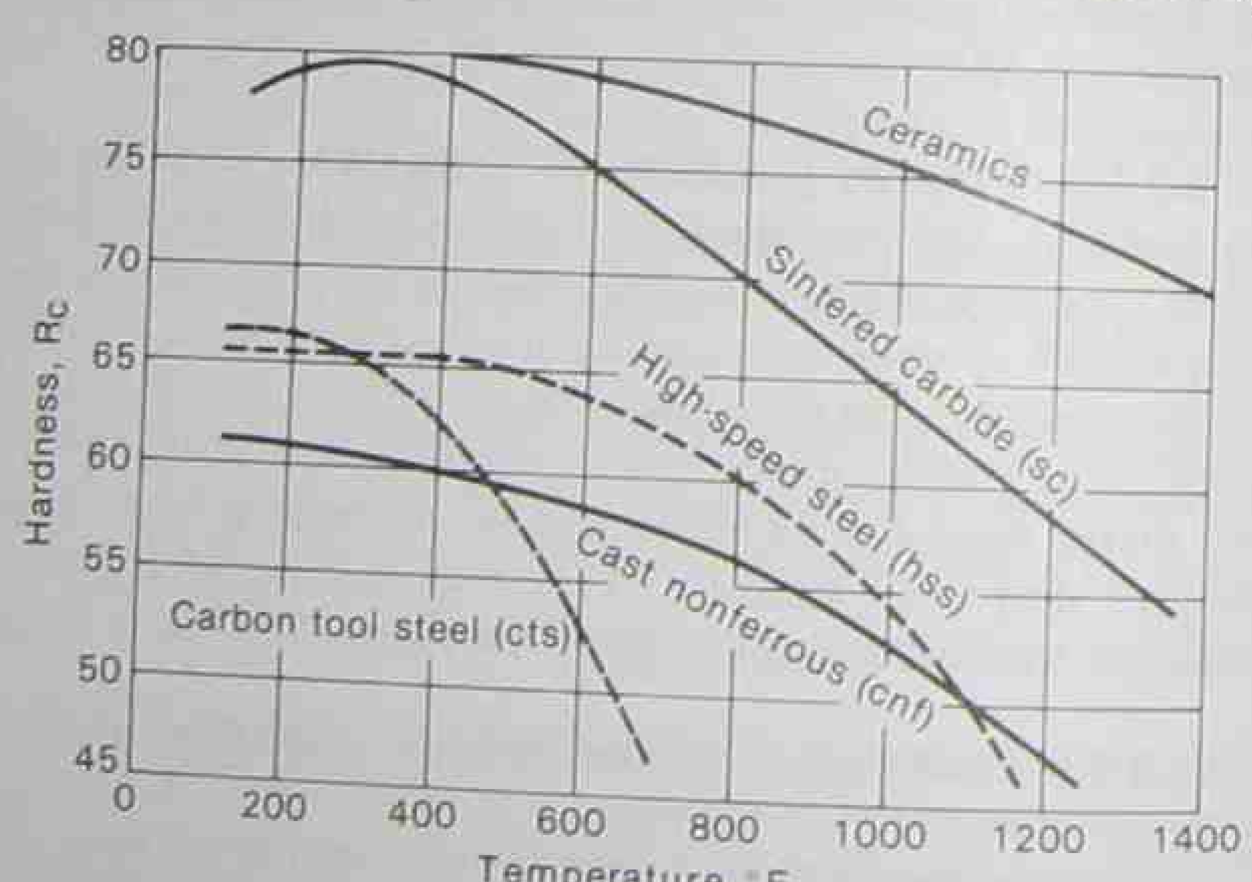


Figure 9-23 Effect of temperature on hardness of various cutting tool materials.

9.3.2 Tool Shape (Signature)

On several past occasions, mention was made of the shape of the cutting tool and whether it was geometrically controlled or of a random nature. Let's dispense with the *random* type first. In abrasive operations* such as grinding, particles such as aluminum oxide are, in fact, the cutting edges. (*Note:* they do remove *chips* just as in turning, but the chip size is so small it is hard to see by eye.) These particles are bonded together, perhaps in the form of a circular disc or *grinding wheel*. As the wheel rotates and is fed into the workpiece, all edges at the wheel periphery enter and then exit from the cutting zone, each removing a small chip. The key point to emphasize is that the *orientation* of each particle as it proceeds to cut is not controlled; that is, angular relationships between the cutting particle and the workpiece do vary.

With controlled cutting edges, we can more logically discuss tool geometry. Included here in some detail are single point tools, drills, and milling cutters. Again, there is no attempt made to cover *all* types of such cutting tools. Certain handbooks and texts of a descriptive nature can be consulted to fill in missing gaps. Here, we are attempting to show that although different names and nomenclature are involved, there is still a degree in similarity of *purpose* for producing controlled angles on cutting tools.

9.3.2.1 Single Point Tools Although there are different systems† used to describe the *nomenclature* of single point tools, our preference is the one discussed below since, for an introduction to this subject, it is more direct and simpler to envisage than other systems.

Figure 9–24(a) shows three views of a single point tool, while Fig. 9–24(b) provides an overall perspective; also shown are the tool angles given in a sequence called the *tool signature*. Each of the three faces constitutes a plane whose relation to the initial tool blank is developed by a pair of angles. (The faces are usually produced by grinding away material from the initial blank.) For example, the back and side rake angles produce a surface where both angles are measured from the top of the blank, whereas the end cutting edge and end relief angles are referred to the original front of the tool blank. If a nose radius other than zero is specified, this refers to the radius of an arc that is ground to blend in smoothly with the end and side cutting edges. The combined influence of the *rake* angles is important in controlling the direction of chip flow; in addition, these angles also influence the tool life and magnitude of the forces needed for chip removal.‡ The basic function of the *relief* angles

* *Honing*, most often used to finish holes, and *lapping*, used to finish flat surfaces, are, in essence, fine-scale grinding.

† See for example, G. Boothroyd, *Fundamentals of Metal Machining and Machine Tools*. New York: McGraw-Hill Book Co. (1976), pp. 167–84.

‡ It is more correct to consider an *effective rake angle*, which results from the values of the back and side rake angles. For our introductory purposes this refinement need not be considered. We note too that as tool hardness increases but toughness decreases, smaller rake angles are used. This is why carbides, ceramics, and diamonds often possess *negative* rake angles to avoid early chipping of the cutting edge.

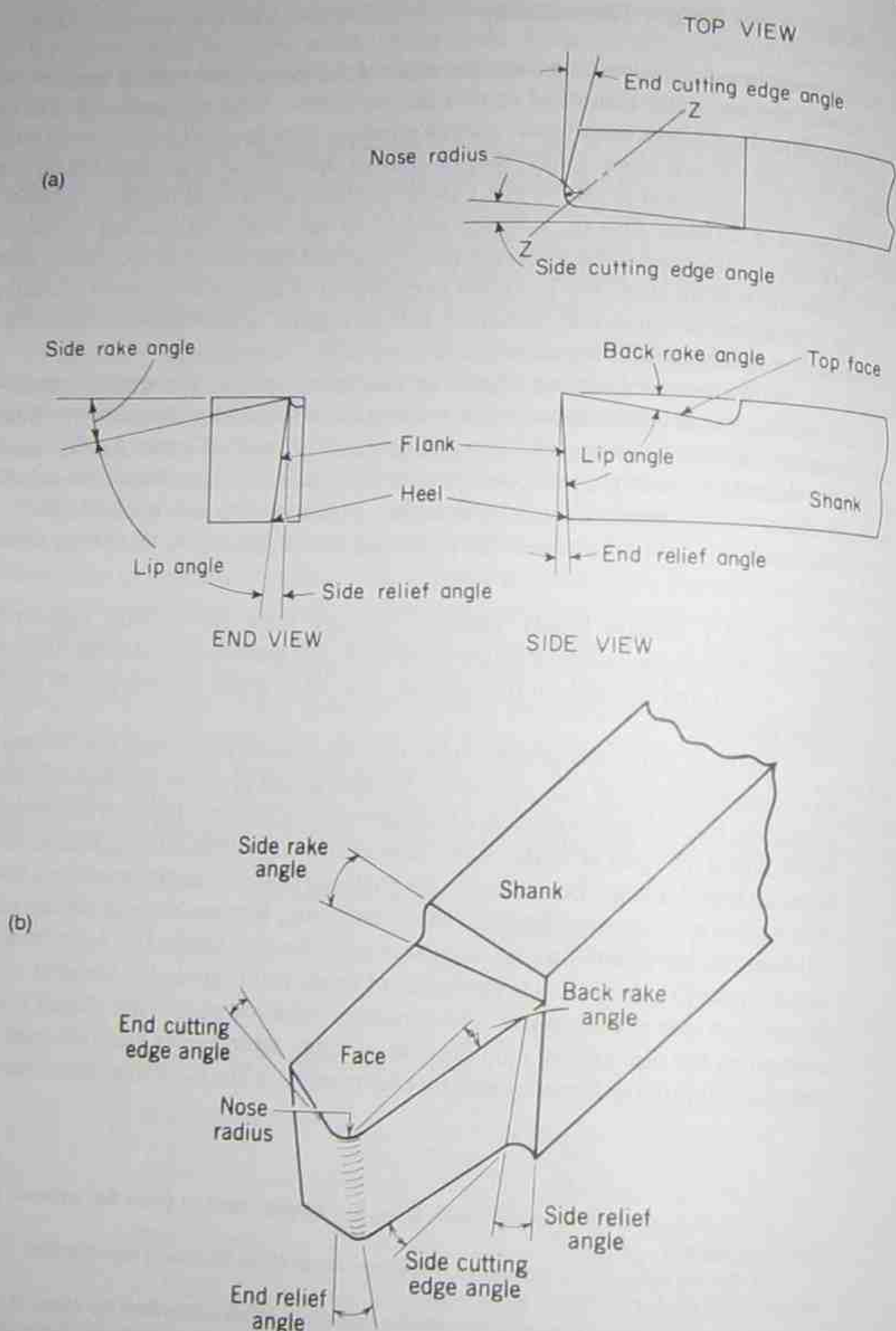


Figure 9–24 (a) Three views of a single-point cutting tool showing the tool signature and (b) overall view of a finish-ground tool.

is to avoid rubbing between the tool and workpiece; the *end* cutting edge angle is ground for the same reason. For cuts such as turning, the *side* cutting edge angle provides the major cutting edge, and this angle also influences the direction of chip flow and contact area between the chip and tool. In addition, as discussed later, tool life and surface finish are both affected by the side cutting edge angle. Finally, the primary effects of the nose radius are connected with tool life and surface finish. Although most of these effects will be covered later in more detail, the reason for introducing them at this time is to indicate that there are important consequences connected with the manner in which a single point tool is ground to a certain signature.

9.3.2.2 Drill Geometry. Figure 9–15 shows the geometry and nomenclature related to a two-fluted twist drill. Although it appears very different from a single point tool, there are definite analogies involved. Each cutting edge angle (two here) provides the same function as the side cutting edge angle of the single point tool, while the helix angle is a "combination" of the end and side rake angles. The flutes, which provide the means for the chip to be removed from the cutting zone, have no direct counterpart with single point tools unless one considers them analogous to the top face of such tools. There is, however, no correspodence between the nose radius and the chisel edge. Note that drills do not have a conical "point"; we shall discuss why later.

9.3.2.3 Milling Cutters. Since there are a variety of milling operations used in machining (see Fig. 9–14), various cutting tool descriptions are encountered. For illustrative purposes, consider a face milling cutter, as in Fig. 9–25, where the pertinent angles are indicated. Note that for the cutter shown, a number of single point tools would be inserted and clamped into a solid body. Although this is a multiedged tool, just as is the two-fluted twist drill, the entire unit is called a *milling cutter*. Table 9–1 provides a useful comparison of the approximate equivalence of the angles mentioned to this point.

With operations such as shaping and planing, the tool nomenclature is really identical to the single-point signature discussed earlier, whereas broaching and reaming include certain modifications. Yet with all of these operations, the basic angles provide similar functions.

9.3.2.4 Grinding Wheels. We have noted earlier that a grinding wheel consists of many particles which are held in position by some type of bond. Abrasive particles most commonly used are aluminum oxide, silicon carbide, and diamond; cubic boron nitride is also used on occasion. Of the important characteristics of any grinding wheel (besides the *type* of abrasive particles) are the following:

1. The particle size called the *grit number*. In a relative sense this ranges from coarse, number 10, to very fine, number 500. Going to finer grit numbers within that scale improves the capability of getting better surface finish which, in turn, requires smaller sizes of cut (that is, feed rates and depth of cut).
2. The type of *bond*. There are a number of materials used to bond the abrasive grains

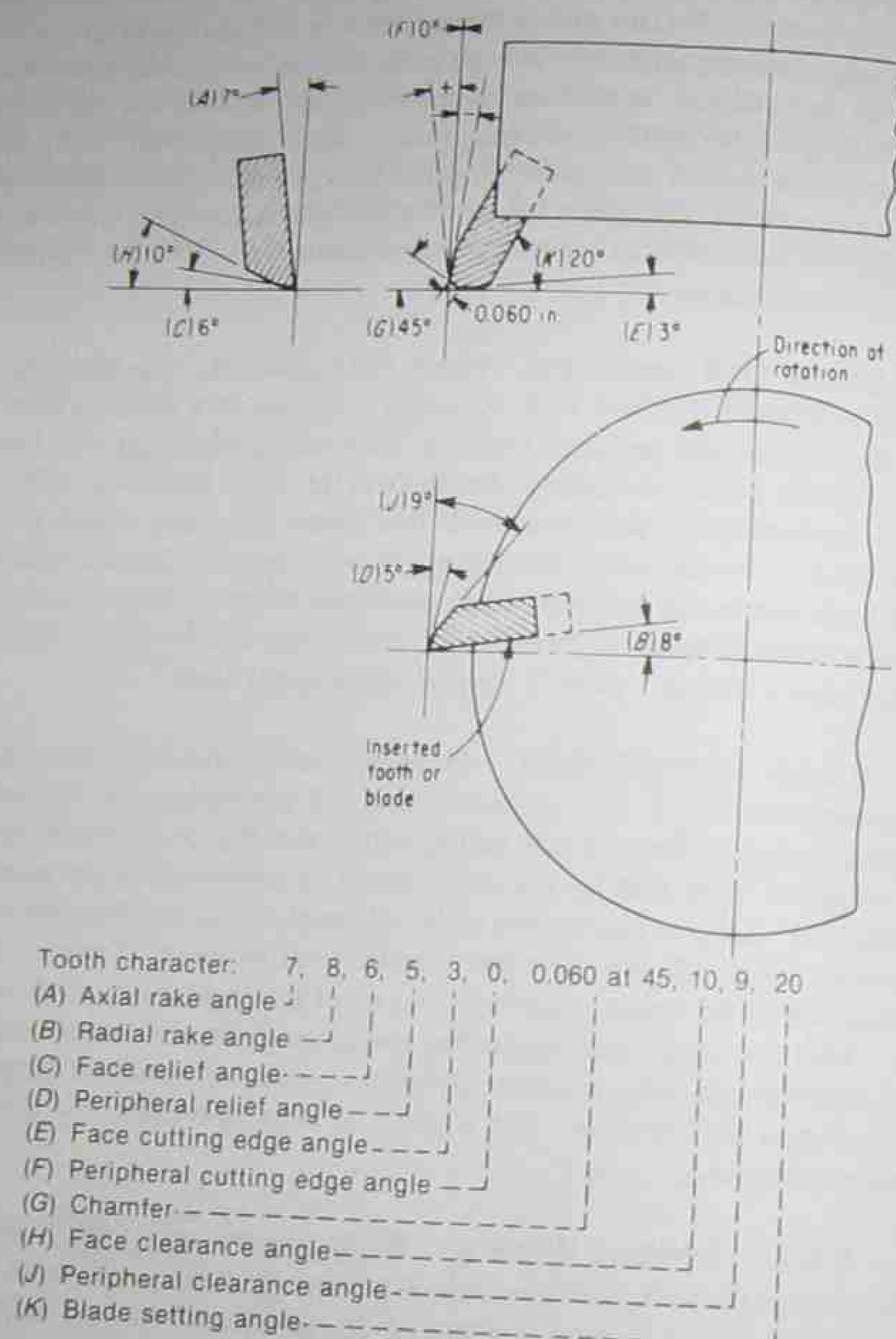


Figure 9–25 Tool signature of a face milling cutter.

together so as to form a solid mass that becomes a grinding wheel. Among the more common types of bonds are the *vitrified, resinoid,* and *rubber;* other materials are used, so the ones stated serve as illustrations. Vitrified wheels are quite strong but also brittle, since the bond is basically a glass. The resinoid type employs a thermosetting resin (see Chapter 7) as the primary bond material. Because the elastic modulus is lower than the vitrified type, these wheels have greater flexibility

TABLE 9–1* COMPARISON OF APPROXIMATELY EQUIVALENT TOOL ANGLES FOR VARIOUS TYPES OF CUTTING TOOLS

Single Point Tool	Drill	Milling Cutter
Back rake	Helix	Axial rake
Side rake	Helix	Radial rake
End relief	Lip relief	End relief
Side relief	Lip relief	Peripheral relief
End cutting edge	Semi-point	End cutting edge
Side cutting edge	Semi-point	Corner angle

*For greater detail, see *Machining Data Handbook*. Metcut Research Associates Inc., Cincinnati, Ohio, 1966, Section 2, p. 411.

and somewhat lower strength. With the rubber-bonded type, even lower stiffness and greater flexibility results; however, they can be made quite thin and these are often used as *cutoff* wheels (in essence, they perform as a type of circular saw).

3. The *grade* of a wheel. This is a measure of the strength of the bond or tenacity with which the bond resists a tearing out of particles from the wheel. Grade is also related to the so-called *hardness* of a wheel, where harder wheels imply higher bond strength.
4. The *structure* of a wheel. This designates the relative porosity, which has several implications. A relatively *dense* structure indicates a greater number of abrasive particles per unit volume, but a lesser degree of surface area into which particles of the work material (called *swarf*) can fit as they are removed. The reverse comments apply as porosity increases; this implies a more *open* structure.

We again make no attempt to document the numerous combinations of such designations used; many types of wheels are commercially available, and proper advice from producers of grinding wheels must often be sought. However, there is one overall concept to keep in mind. Ideally, one would like to use a grinding wheel until the major number of abrasive particles that contact the workpiece become worn (or are dulled) to the point that effective grinding diminishes. At that time, the worn particles would literally be pulled away from the bond so as to reveal new, sharp particles for subsequent grinding. A proper combination of grade and structure thus becomes important in such a self-sharpening wheel. Often the wheel has an overall combination of characteristics such that a self-sharpening action does not occur. Then it is necessary to *dress* the wheel by some additional act that severs the worn grains in order to expose fresh ones. Dressing can be accomplished by moving a supported and shaped diamond (for example, pointed) across the periphery of the wheel so as to remove a thin layer from the outer diameter. This also *trues* the wheel; that is, it tends to return the wheel to a circular configuration. If wheels are not *self-sharpening* or *dressed,* small pieces of the material being ground will *load* the regions between the abrasive particles; at the same time, the particles themselves become worn. Several consequences can result. First, the temperature of the workpiece in the area or region being ground can increase to undesirable levels. With steel, for example, the temperature can reach a level that will transform the original structure to austenite; then as the wheel passes that region, the rate of heat transfer into the body of the workpiece can be fast enough to convert the austenite to martensite. This can induce undesired residual surface stresses or even cracking of the workpiece. In general, by maintaining sharp cutting edges, flooding the cutting region with a copious flow of coolant, and reducing the size of cut will preclude such undesirable results.

9.4 CUTTING FLUIDS

A truly large number of cutting fluids is available commercially; to document all would be a prohibitive and, for our purposes, an unnecessary task. In many industrial situations, users will most often consult with manufacturers of fluids to seek advice as to which is most suitable for a particular machining operation.

In a most general sense, cutting fluids are used for one of two purposes. For operations where low friction is of greatest concern, thereby leading to excellent dimensional control and surface finish, so-called *lubricants* find major use. Many such operations are carried out at relatively low speeds. At the other extreme are those fluids whose primary purpose is to conduct heat away from the cutting zone, thereby cooling the workpiece; they are often called *coolants*.*

Lubricants are often composed of mineral oil, sulfur, chlorine, and fatty oils and, in comparison with coolants, have lower thermal conductivity. Coolants often are composed of a mixture of water and soluble oils; in many cases the composition involves water, emulsifiers, fatty oils, and rust inhibitors. Note that water itself is an excellent coolant but, because it would produce rust on the machine tool, cannot in any practical sense, be used by itself. Perhaps the most important point to remember is that coolants are used where tool life is of primary concern, whereas lubricants find major use where surface finish and dimensional control are most important. One can certainly take exception to such a division, since instances can be cited wherein the use of coolants at cutting conditions of high speeds and a small size of chip can produce excellent surface finish. Yet to provide the most *general* insight regarding cutting fluids, the distinction proposed is certainly reasonable.

9.5 CHIP FORMATION AND TOOL WEAR

From previous discussions, one can now envisage just what is used to conduct a machining operation. Depending upon the desired shape of the part to be produced, a machine tool, in conjunction with a cutting tool made from a particular tool material and having a certain signature, plus the use of a pertinent cutting fluid, are selected. Now we are ready to start cutting and a good question to ask is "just what happens when material is cut?" First, two (but not all) answers are now given.

* Cutting fluids also flush chips away from the cutting zone.

9.5.1 Chip Formation

To cause a chip, enough power must be supplied by the machine tool, since a certain amount of energy is needed to sever or shear the chip from the workpiece; in addition, the cutting tool must possess certain desirable properties, as we discussed earlier. Depending to a large extent on the properties of the workpiece, two broad classifications of chips can result. In machining *brittle* metals such as gray cast iron (which has limited ductility), the chips tend to come off in small pieces and are called *discontinuous* chips; this is illustrated in Fig. 9–26(a). In contrast, the machining of ductile metals often causes a continuous ribbonlike chip to form*; thus, the word *continuous* is used in this case.

It is sensible to further subdivide continuous chips into those which do *or* do not form with a *built-up edge* (BUE). Figures 9–26(b) and 9–26(c) illustrate this point. BUE formation is a phenomenon of decided significance, since its presence can have a pronounced effect on both tool life and the surface finish of the machined part. Most probably, a built-up edge results when metal from the underside of the chip adheres to the tool due to the effects of friction at the tool-chip interface and the extremely high pressure that can result at that interface. Although this problem has been studied for many years, it is still found as a subject in learned publications. It does appear, however, that the BUE reaches some critical size; then some of it breaks loose and is deposited on the finished work surface as well as the underside of the chip; see Fig. 9–27. This overall process repeats itself as cutting continues. The results of much research indicate that to diminish or to preclude BUE formation, one should:

1. Increase the cutting speed.
2. Decrease the thickness of the chip (that is, use smaller feed rates).
3. Use a lubricant.

It is perhaps ironic that in view of certain detrimental aspects of the BUE, it actually covers the cutting edge during machining and forces the moving chip to make contact on the top face of the tool; this may actually prolong tool life. The manner by which *tool wear* occurs, as discussed next, is also affected by the presence of a BUE.

9.5.2 Tool Wear

Due to the combined effects of elevated temperature, forces of rather large magnitude, and frictional effects between a chip and cutting tool, it should not be surprising that cutting tools wear as the operation proceeds. Two general types are usually specified, these being *flank* and *crater* wear. Figure 9–28 illustrates this. Flank wear (for example, with a single point turning tool) results on the side face in a direction downward from the cutting edge. Literally, tool material is worn away by the rubbing (frictional) action at the flank-tool

* Such a result is detrimental in practice, since chips must be removed from the machine tool periodically. Continuous chips, being like tangled wire, are an expensive nuisance. To avoid this, *chip breakers* are often used. Regardless of how a chip breaker is introduced, the major effect is to force the chip to bend so severely that it periodically fractures into a number of discontinuities, thereby avoiding a long, continuous chip.

(a)

(b)

(c)

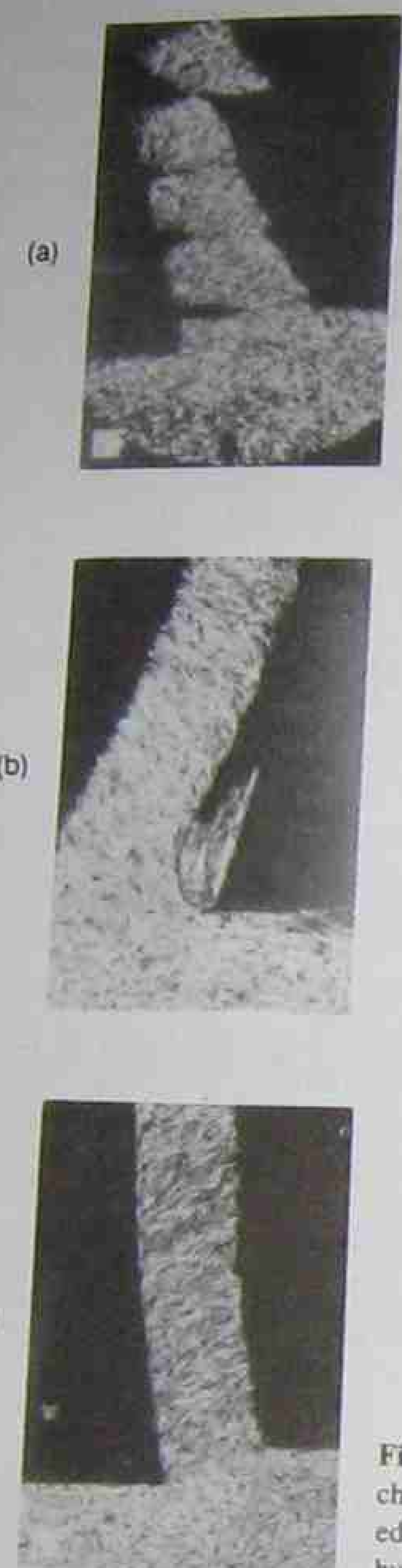


Figure 9–26 Examples of a discontinuous chip (a), a continous chip with a built-up edge (b), and a continuous chip without a built-up edge (c).

interface. Crater wear, also illustrated in Fig. 9–28, most often results when a BUE forms, thereby causing the major frictional effect between the chip and tool to occur on the top face of the tool (that is, away from the cutting edge). Of course, either type of wear is undesirable, and either may dictate the useful time a tool operates before it must be removed and reground, or replaced, to produce a new or fresh signature. Perhaps crater wear can be considered more serious, since if it becomes excessive, the remaining section of the tool can no longer support the forces involved and catastrophic fracture of a section

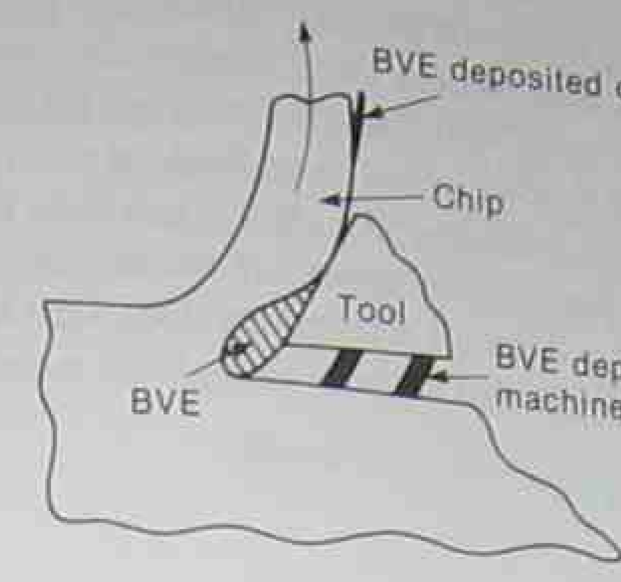


Figure 9–27 Sketch showing built-up edge deposits on underside of chip and on the finished work surface.

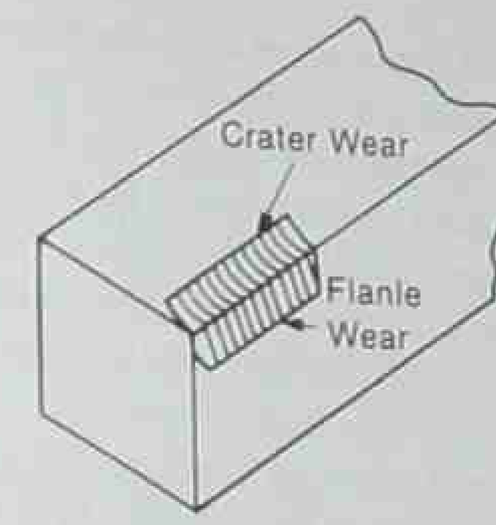


Figure 9–28 Sketch showing the difference between flank and crater wear.

of the tool results; this requires excessive regrinding, which leads to greater expense. Although machining economics is covered in Chapter 13, it is prudent at this point to state a sensible general philosophy. One should *never* use tools to catastrophic failure* in *industrial* practice, since machining costs will increase.

9.6 BASIC MACHINING PARAMETERS

The word *basic* as used here may seem confusing, since one could argue that inputs such as the tool material and signature, type of cutting fluid, and work material are all important in analyzing any machining process. This is certainly true. However, *once* the machine tool and workpiece are chosen (this comes about usually from design specifications including the *shape* of the desired part and the mechanical properties needed), and once a sensible tool material and shape are specified, along with an acceptable cutting fluid, there then remain only *three* parameters that can be adjusted to influence such things as tool life, power requirements, and surface finish among other concerns. These three parameters are the *feed rate*, *depth of cut*, and *cutting velocity* (or cutting speed), and they are the only factors that can be manipulated or controlled by alteration, once all other parameters have been defined (that is, work material, tool material, and the like). In fact,

* In tool life testing, this is often done (see Sec. 9.9).

the combined effects of these basic quantities are the final *inputs* that govern just what goes on at the cutting edge. Much has been written about numerically controlled (NC) machines or computer-aided manufacturing (CAM),* and one may get the impression that somewhat exotic control devices are all that is needed to understand machining. (Many experts in controls or computers also seem to believe this.) Yet if one does not understand the importance of the *basic machining parameters,* which are the factors to be properly controlled, all else can become secondary. After all, a computer does what it is told to do, so if an improper combination of feed, depth, and velocity are fed into a computer, it will be controlling poor inputs.

9.6.1 Feed or Feed Rate

In its most general sense, the *feed* is analogous to the thickness of a chip and is most often defined as the distance that the cutting tool moves into or along the workpiece surface per some relative unit time. Several examples should help to clarify this. Turning cuts involve a feed rate that is set on the machine tool in terms of *inches per revolution* (ipr).† During each revolution, the tool advances parallel to the axis of rotation a certain distance at a constant rate; this means that the thickness of the chip is uniform for whatever feed rate is set. With the spindle speed set for a particular rpm (call this N), then in one minute, the tool would have moved fN inches; Fig. 9–29 shows this. With drilling operations the concept is identical; if we set a feed rate of so many ipr and a spindle speed of N, the drill advances fN inches into the workpiece in one minute. Figure 9–30 illustrates this.

With milling, the situation is not so simple, for several reasons. Here, as the cutter rotates (see Fig. 9–31), and the workpiece, which is clamped to the table, is fed into and past the cutter, the chip thickness (that is, feed) is not constant; rather, it varies during the cutting action. It is most common to select a desired *average* thickness of chip (usually specified as *inches per tooth*), then to multiply this by the number of teeth in the cutter to give inches/rev., and finally to multiply by N (rpm of the cutter as set by the operator) leading to a feed in *inches/min.* It is this feed that is *set* on the machine which then, by this backward calculation, leads to the desired average chip thickness, so

$$F = fnN \tag{9-1}$$

where F = feed of table *set* in in./min
N = spindle speed in rpm as *set*
n = number of teeth in the cutter
f = resultant *average* chip thickness in inches per tooth

With these three widely used operations as a guide, the feed rates for other operations should be understood if it is kept in mind that the feed always comprises one of the dimensions related to the cross-sectional *area* of the chip as it is removed from the

* See Chapter 5.

† Many machine tools used in the United States are still calibrated in English (not SI) units; thus we employ such units here. Whether inches or mm are used does not alter these basic concepts. Conversion to SI units is a simple matter.

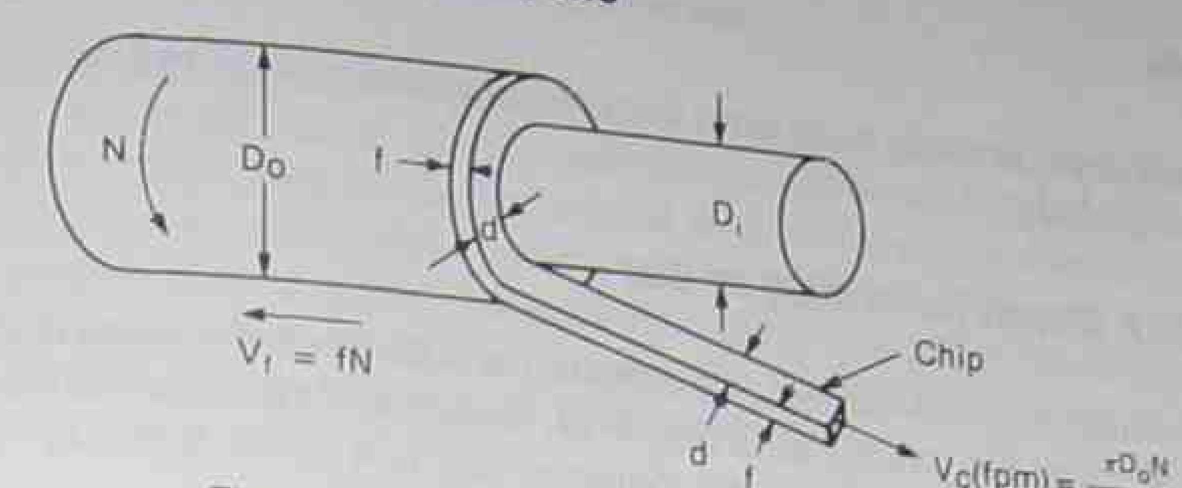


Figure 9–29 Machining parameters related to a turning cut.

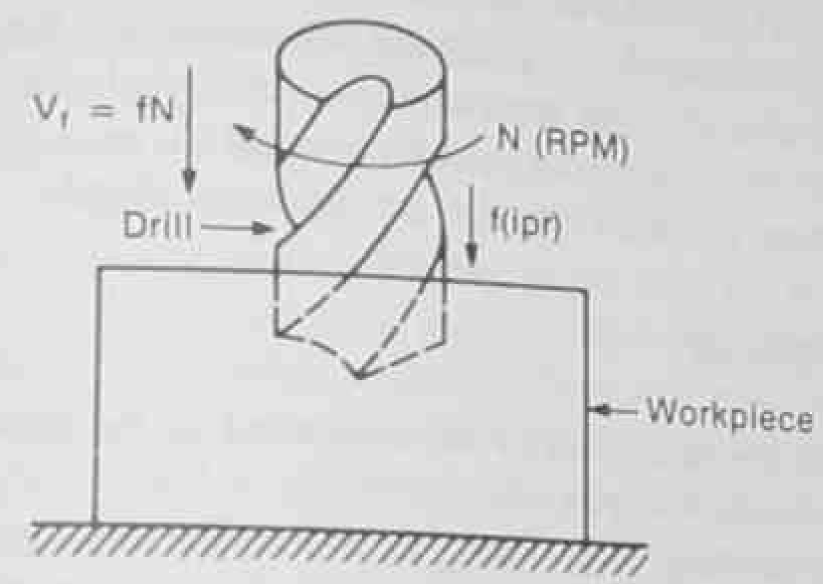


Figure 9–30 Machining parameters related to a drilling cut.

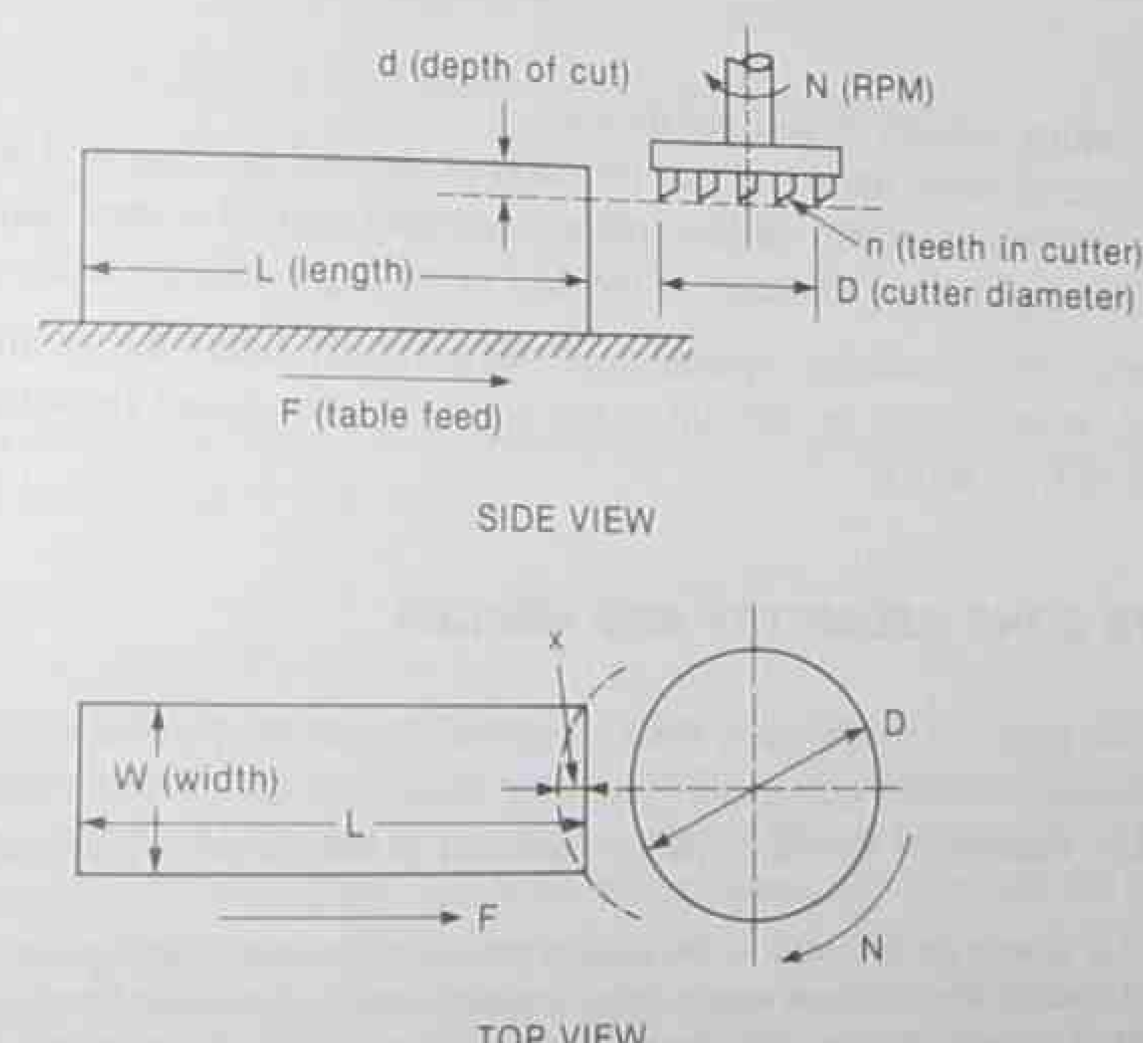


Figure 9–31 Machining parameters related to a face-milling cut.

workpiece. In nearly every case, the feed is the smaller of the two dimensions and carries the symbol f.

9.6.2 Depth (or Width) of Cut

In conjunction with the feed, the depth of cut is the other dimension that defines the area of the chip. For turning, as in Fig. 9–29, the depth of cut is the *radial* distance that the tool point is set from the outer diameter; that is it is one-half $(D_0 - D_i)$; it will be designated by d.

For drilling, the depth of cut is literally the radius of the drill, whereas in milling, as in Fig. 9–31, the depth of cut is the distance the tools project beneath the original work surface. A *note of caution* about milling. From geometric considerations, the *area of chip* removed by each *tooth* per revolution is the product of fw, where w is the *width* of the workpiece. This illustrates one instance where definitions must be used with care; this should become clearer when *rates of metal removal* are discussed shortly.

9.6.3 Cutting Velocity (Speed)

The cutting velocity V will, in this text, *always* be considered as the *maximum linear* speed existing between the tool and workpiece.

Continuing with turning, drilling, and milling as above, the cutting velocity V is given by

$$V = \pi dDN/12 \qquad (9\text{–}2)^*$$

where

V = cutting velocity in *feet* per minute
N = rpm of either the workpiece (turning) *or* the tool (drilling and milling)
D = diameter in *inches* of the component being rotated (outer diameter of work for turning, but the diameter of the drill or milling cutter for those operations)

For any other machining operation, if one carefully notes the relative geometries involved, it should not be difficult to use some common sense in order to determine the values of f, d, and V.

9.7 CUTTING ZONE GEOMETRY AND FORCES

Now that topics of tool angles, wear, and basic machining parameters have been covered, we continue with the question of just what goes on at the cutting edge. In Fig. 9–32(a), the basic manner by which chips are removed is illustrated.† The cutting tool, having a

* If SI units are desired, V in feet/min can be converted to m/s by multiplying by 0.0051.

† Strictly, this illustrates what is called orthogonal or *two-dimensional* cutting. Turning the full wall thickness of a tube or shaping, where the width of the cutting edge is greater than the width of the workpiece, both constitute orthogonal cutting. Although most operations are three-dimensional, the concepts presented here still portray a realistic picture of the basic mechanism of chip removal.

rake angle α and *relief angle* θ, projects below the original work surface at some distance t (strictly, this is the feed or chip thickness as set), and moves through the workpiece at a velocity V. Chips form by causing material ahead of the tool to shear along the *shear plane* which lies at a *shear angle* Φ that is oriented as shown. The actual thickness of the chip t_c depends upon α and Φ. Often, the ratio of t to t_c is called the *cutting ratio* r and in actual practice r is always less than one; that is, $t < t_c$. Reference to Fig. 9–32(b) where the length of the shear plane is taken as unity, leads to

$$r = t/t_c = \sin\Phi/\cos(\Phi - \alpha), \qquad (9\text{–}3a)$$

so

$$\tan\Phi = (r\cos\alpha)/(1 - r\sin\alpha) \qquad (9\text{–}3b)$$

Since α is ground to a desired angle and t can be set accurately, a reasonable measure of t_c allows a prediction of the shear angle Φ with Eq. (9–3b). Also, only when $\alpha = \Phi$ is t_c equal to the length of the shear plane.

Figure 9–32(c) is a velocity diagram, or *hodograph*, which is used to determine velocities of concern. Relative to the tool, the workpiece moves at a velocity V taken in the horizontal direction. As the chip forms, the metal is suddenly forced to move at a velocity V_c parallel to the tool face; this requires a jump discontinuity along the shear plane V_s at an angle Φ to the horizontal. By setting V to some scaled value and using the angles Φ and α as shown, the magnitudes of V_c and V_s can be measured from the holograph as functions of the value of V. An alternative approach is to note that continuity of mass, if we assume no change in the width of the chip, requires that

$$Vt = V_c t_c \quad \text{or} \quad V_c = rV \qquad (9\text{–}3c)$$

Noting that $V_s = A + B = V\cos\Phi + V_c\sin(\Phi - \alpha)$ and

$$V_c = V\sin\Phi / \cos(\Phi - \alpha)$$

leads to

$$V_s = V\cos\alpha / \cos(\Phi - \alpha) \qquad (9\text{–}3d)$$

Although not derived here,* an expression for the *shear strain* involved in machining can be found in terms of Φ and α. It is

$$\gamma = \cot\Phi + \tan(\Phi - \alpha) \qquad (9\text{–}3e)$$

For example, if Φ is 30 deg and α is 10 deg, then γ is 2.1. What this implies physically is that the shear strain in machining is *extremely high* relative to other deformation processes; values as high as five are not uncommon. This would be roughly equivalent to a reduction of area exceeding 90 percent in a tensile test, so the amount of strain hardening induced in a chip is very large. Hardness measurements of chips and the region surrounding the shear plane support this observation.

Figure 9–32(d) illustrates force relationships pertinent in orthogonal cutting. The

* See S. Kalpakjian, *Mechanical Processing of Materials* Van Nostrand Co., Inc., 1967), pp. 245–50.

total force F_t acting between the chip and tool can be resolved into different sets of components. First consider the forces that are horizontal and vertical; these are the cutting force F_c and vertical or thrust force F_v. Next, the force F_t could be resolved into components parallel and normal to the tool face, these being F_f and F_n, respectively. Finally, F_t can be resolved into components normal and parallel to the shear plane, these being N and F_s. Depending upon which pair of components is of greatest interest, various relationships as functions of F_t, Φ, α, and the *friction angle* β can be developed where the coefficient of friction between the chip and tool is taken as $\tan \beta$. One illustration of such relationships has been the extensive research work, conducted for many years, concerned with predicting the relationship or interconnection of Φ, α, and β. Using the model shown

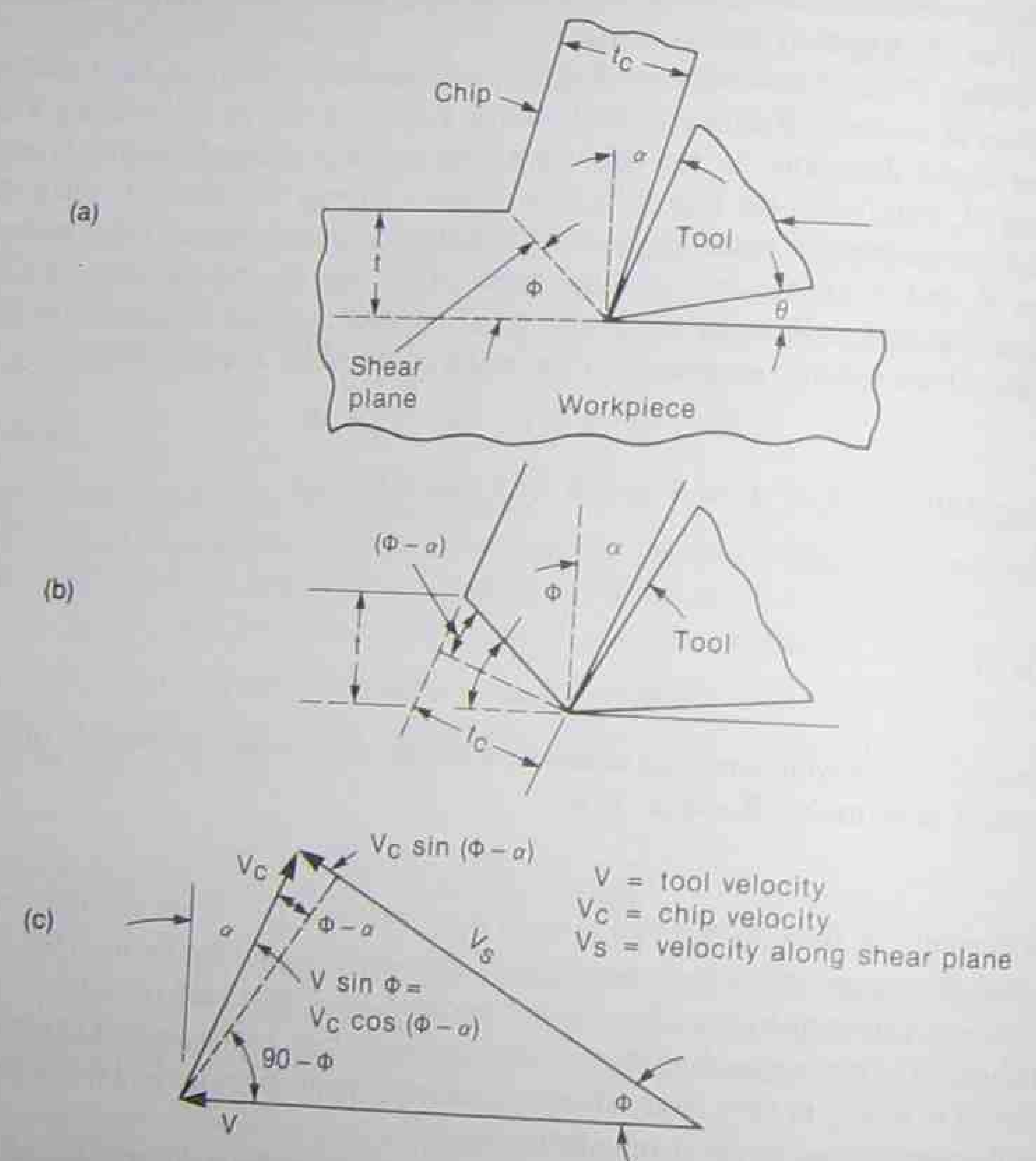


Figure 9–32 (a) Orthogonal machining indicating machining parameters and angles of concern, (b) geometry relating chip thickness t_c, to thickness as set t, (c) velocity vector diagram or hodograph, and (d) force relationships in plane and reduced to a force circle.

in Fig. 9–32(d), assuming that β is independent of Φ, and considering that the shear angle results so that the maximum shear stress occurs on the shear plane (or alternately that the cutting force F_c is a minimum) leads to

$$\Phi = 45 + \alpha/2 - \beta/2 \qquad (9\text{-}3f)^*$$

Many other expressions have also been developed where other models and assumptions were employed. Unfortunately, none is in acceptable agreement with experimental results using a *wide range* of cutting conditions. Yet, in a qualitative sense, Eq. (9–3f) provides some correlation with reality. If α is fixed for a given tool, then decreasing β, perhaps by using an effective lubricant, should cause Φ to increase, with a corresponding decrease in the chip thickness t_c (t being constant here). Thinner chips, leading to less deformation and lower forces, should, if anything, improve surface finish. Such a result is observed in practice.

* This is often called Merchant's analysis after M. E. Merchant, *Journal of Applied Physics* 16.5 (1945) p. 267.

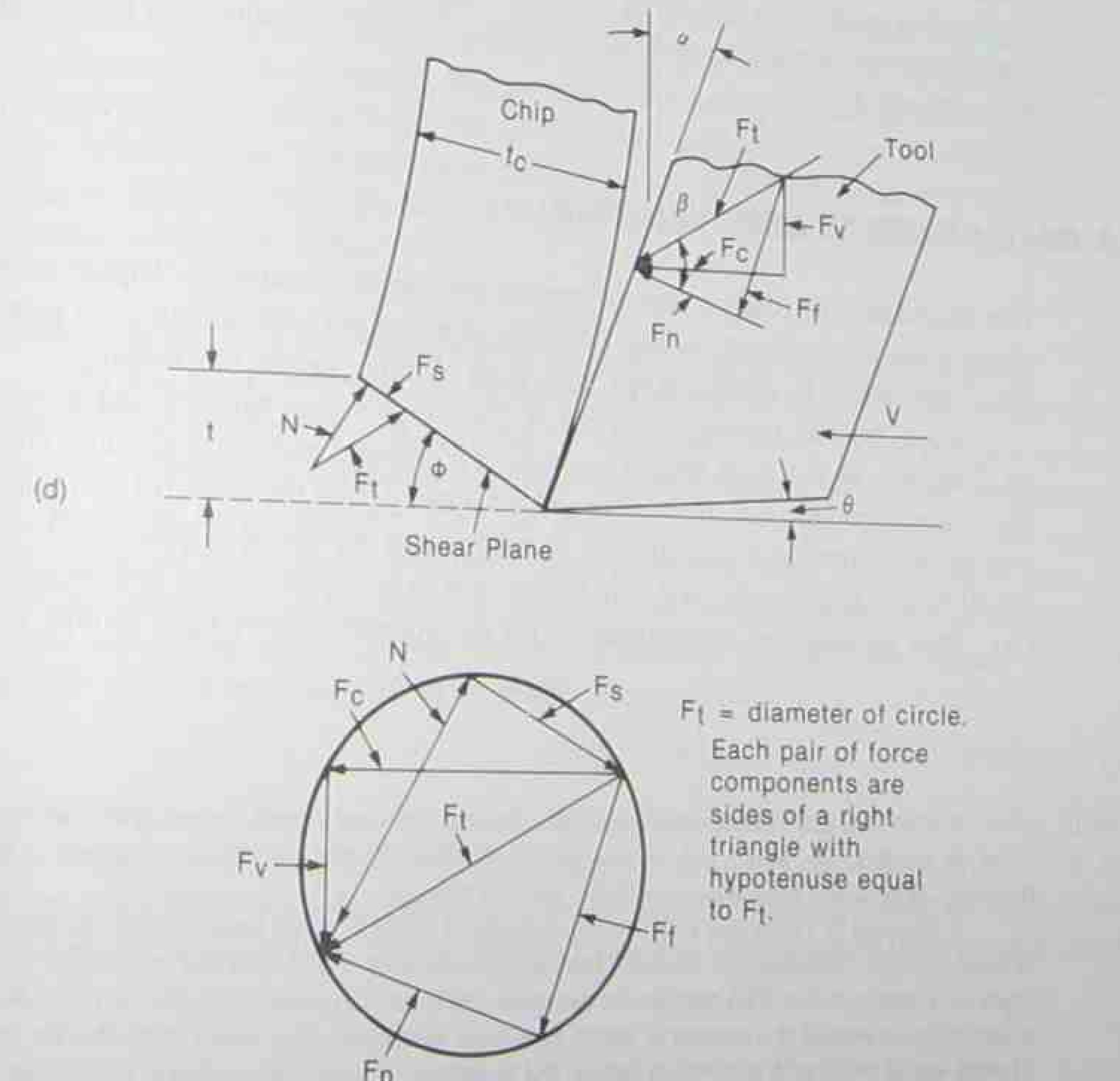


Figure 9–32 (Continued)

In closing this brief section, it is appropriate to discuss why analyses and models such as those used above do not always provide the degree of agreement with experimental results one would desire.

1. Assuming that shear occurs on a single plane is not verified in practice; rather, a shear *zone* is observed. Thus considering that there is a distinct shear angle Φ is incorrect.
2. Plastic deformation ahead of the shear zone as well as in a small substrate beneath the finished work surface is not taken into account.
3. Built-up edge formation is not considered.
4. Tools are not perfectly sharp, regardless of the care exercised in grinding the various angles. In addition, tool wear begins quickly as cutting commences.

In spite of these shortcomings, such efforts are still beneficial in terms of improving our understanding of some of the basic mechanisms of machining, and for that reason alone they have been worthwhile. Perhaps more realistic models and assumptions will one day lead to predictions that agree more closely with experimental data. We note here that considerations of cracking and the inclusion of fracture toughness in such studies is one such attempt.*

9.8 RATE OF METAL REMOVAL (RMR)

The importance of RMR is that it governs the rate of production. Higher rates of metal removal lead to producing a larger number of pieces per unit of time, say parts per hour. Usual practice is to employ RMR with units of cubic inches per minute.

First consider turning, as in Fig. 9–29, where values for feed and depth have been set. Imagine, as the work begins to rotate at N rev/min and the tool feeds into the work, that you grabbed the end of the chip and led it away from the work for one minute. (*Don't* ever try to do this because chips can be razor sharp and are also quite hot.) What results, for all practical purposes, is a chip whose area is fd square inches and whose length is V feet. Thus an excellent approximation for the volume removed per minute is

$$\text{RMR} = 12Vfd \tag{9-4}$$ †

where

* A. G. Atkins, "Fracture Toughness and Cutting," *Internnational Journal of Production Research* 12, 2 (1974), p. 263. See also A. G. Atkins and Y. W. Mai, *Elastic and Plastic Fracture* (Chichester: Ellis Horwood, New York: John Wiley and Sons, 1985), Chap. 10.

† The data in Tables 9–2 to 9–5 were originally obtained by using English units. To avoid unduly small or large numbers, velocities, for example, have traditionally been specified in feet per minute, whereas feed and depth are given in inches. This requires the use of the coefficient 12 to give a velocity in inches per minute. Such a conversion is needed in a number of places in sections that follow. One could argue that the exclusive use of SI units would avoid such conversion factors, but to convert all such original data is unnecessary. If an answer is desired in SI units, then one can find the answer in English units and simply make the proper conversion at that time.

V = cutting speed in *feet*/min
f = chip thickness (feed) in inches
d = depth of cut in inches
RMR = in.^{3}/min

To preempt a question that almost *always* arises at this point, let's compare the approximate value for RMR above with an exact calculation, using reasonable values for the various parameters.

Suppose a four-inch-diameter bar of a solid is to be turned with a feed rate of 0.015 ipr, depth of cut of 0.075 in. and an rpm of 150. From Eq. (9–2) the value of V based upon the outer diameter is 157 fpm. Then, using Eq. (9–4), we obtain

$$\text{RMR} = 12(157)(0.015)(0.075) = 2.12 \text{ in.}^3\text{/min}$$

Now consider a more exact calculation. Since the annular area of metal removed is $(\pi/4)(4^2 - 3.85^2)$ and the tool moves fN inches in one minute, then the exact value is

$$\text{RMR} = (\pi/4)(4^2 - 3.85^2)(0.015)(150) = 2.08 \text{ in.}^3\text{/min}$$

This leads to an error of less than 2 percent and is one of the reasons that the simplicity of Eq. (9–4) finds wide use. Of course, as the relative ratio of the depth d and diameter D increases, Eq. (9–4) becomes less accurate.* For most practical situations and certainly for our purposes, we will use Eq. (9–4) when computing values of RMR in *turning*.

With drilling, and starting at the time when the full drill diameter is cutting, the value of RMR is simply the area of the hole times the linear velocity of drill movement into the work piece. Thus,

$$\text{RMR} = (\pi D^2/4)(fN) \tag{9-5}$$

where D is the drill diameter in inches and f and N carry their usual meaning.

At this point, we stress that memorizing a number of such equations is not essential and can lead to errors if *used in situations where they do not apply*. It is best to look at the given operation and sensibly determine which dimensions and speeds are directly tied to the rate of metal removal. With this suggestion, let's return to the milling operation shown in Fig. 9–31. As the workpiece enters the cutting zone, it is moving at some speed of F in./min as set on the machine. Once the full width of workpiece is being cut (see dotted line on Fig. 9–31), in each succeeding minute a volume of metal equal to the cross section of dw times the value of F will be removed; thus,

$$\text{RMR} = Fdw \tag{9-6}$$

where d and w are the linear dimensions shown. Two points are instructive here. First, no single equation satisfies RMR for all operations, so a little logic must be used for operations other than those described. The second point is that regardless of the operation, RMR is a direct function of feed, depth of cut, and cutting speed (that is, rpm here). It

* This is because the velocity based upon D deviates more from the *mean* velocity based upon the mean diameter.

is hoped that these few illustrations will provide adequate background to this important concept.

We close this section with a few comments related to *machining time*, that is, how long does it take to actually make the desired cut? In all cases it is simply a matter of determining the length of workpiece to be machined and then dividing by the velocity *parallel* to or *in the direction* of that length. For example, if a part that is L inches long is to be turned, then the machining or cutting time t_m is simply

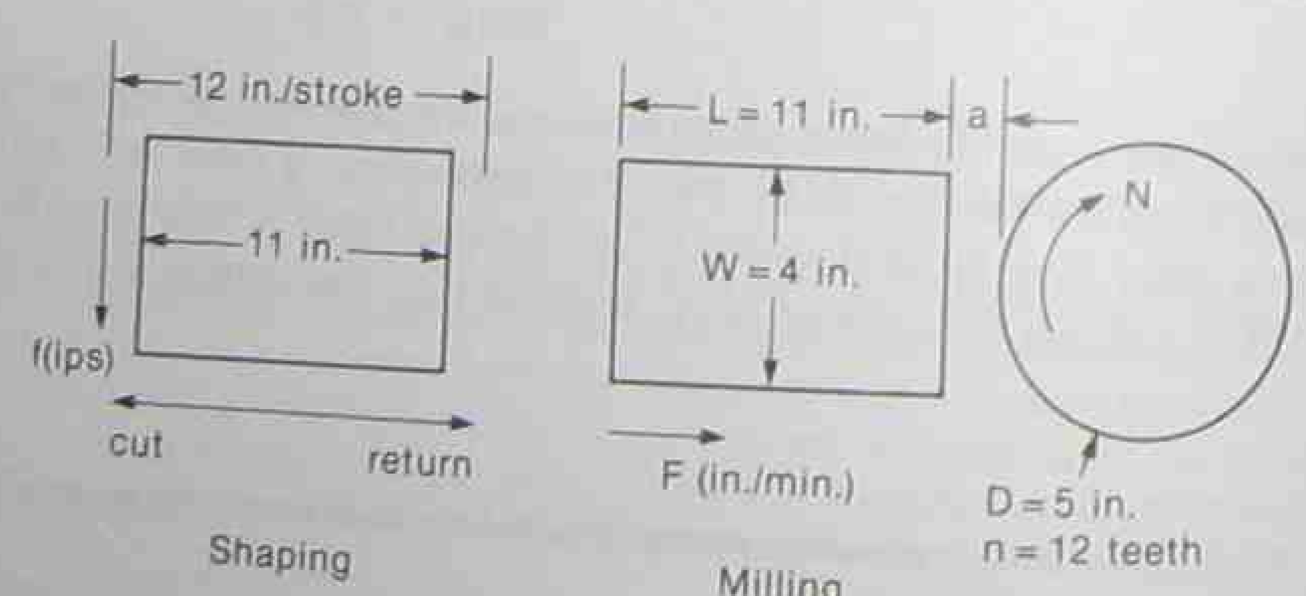
$$t_m = L/fn \tag{9-7}$$

Similarly, if a hole of depth L is to be drilled, $t_m = L/fN$ also. This does neglect the small distance the drill point must move before the full drill diameter comes into play. This can be determined with a little trigonometry and added to L if greater accuracy is desired. Milling as in Fig. 9–31 requires a bit more explanation. When the cutter first touches the work, it must travel a distance x before the full width is being machined. (This is similar to the distance the drill point must move as just mentioned, although x is usually much larger.) Again, x can be found because pertinent dimensions are known. Now if the cutter travels the distance L, then $t_m = (L + x)/F$, F being the table feed. In many cases, however, the cutter is moved the entire length of the work to produce a somewhat more uniform surface. In that case, $t_m = (L + D)/F$, where D is cutter diameter.

It is of interest here to note that if certain other parameters are altered, that alone will *not* change the rate of metal removal *or* the machining time. For example, using carbide tools will, in general, permit the use of higher cutting speeds as compared with high-speed steel tools. But *unless* the cutting velocity itself is increased, the use of the same values of feed, depth of cut, and cutting speed will still produce the same RMR and machining time. This is one reason that we have considered these parameters as being basic.

Example 9–1

A block of metal, 4 in. wide by 11 in. long by 3 in. high, is to be reduced to a height of 2.8 in. Either a shaper or milling machine could be used and the sketches indicate the manner by which this might be done. In shaping, the single point tool must have a small overtravel at each end of the stroke (use 0.5 in. here); thus the length of either the cutting or return stroke is 12 in. in this case. In milling, consider that the approach distance a is also 0.5 in. and that



the milling cutter of diameter D will travel the full 11-in. length *plus* the cutter diameter. If the feed rate is to be 0.006 (inch per *stroke* on shaper and inch per *tooth* on the mill) and the cutting velocity is to be 60 feet per minute, find:

1. The time to machine the piece by each method.
2. The rate of metal removal when full cutting applies.

Solution 1. First consider the shaper. The tool is set to reciprocate back and forth in terms of strokes per minute (spm), where the linear velocity goes from zero to a maximum and then back to zero at each end of the stoke. If we base strokes per minute on an *average* cutting velocity of 60 fpm (that is, only the cutting portion of the stroke), then

$$V \text{ (fpm)} = S \text{ (spm) } L \text{ (length of stroke in feet)}$$
$$V = 60 \text{ fpm} = S(11 + 1)/(12) \quad \text{so} \quad S = 60 \text{ spm}$$

The time to complete the 4-in. width (if it is assumed that the first stroke makes a full cut) is

$$t = W/fS = 4/(0.006)(60) = 11.1 \text{ min}$$

With the milling operation, the rpm comes from

$$V = \pi DN/12 \quad \text{or} \quad N = 12(60)/5\pi = 46 \text{ rpm}$$
$$F \text{ (ipm)} = fnN = 0.006\ (12)46 = 3.3 \text{ ipm}$$

The total distance to travel is $0.5 + 11 + 5 = 16.5$ in.

$$t = 16.5/3.3 = 5 \text{ min}$$

2. RMR for the shaper is the volume per stroke times the strokes per minute, or

$$\text{RMR}_s = 11(0.006)(3 - 2.8)(60) = 0.79 \text{ in.}^3/\text{min}$$

For the milling machine,

$$\text{RMR}_m = (\text{width})(\text{depth})F = 4(0.2)(3.3) = 2.64 \text{ in.}^3/\text{min}$$

This example illustrates that milling is faster than shaping, because the multitoothed cutter provides a higher RMR and thus lower machining times.

9.9 TOOL LIFE

It can be argued that the single most critical aspect of machining is the time* that the cutting tool operates in an acceptable manner; this is called the *tool life*. Various criteria have been used to define tool life, some of which are

1. Complete destruction of the cutting edge.
2. A specified amount of flank wear.
3. A loss of dimensional accuracy.
4. Degradation of surface finish to an unacceptable level.
5. An increase in the forces associated with cutting to an undesirable level.

* Number of pieces produced or the production rate, say pieces per hour, are basically equivalent.

In this text, the first criterion will be used. All that is involved is to measure the accumulated time that the tool actually cuts under a combined set of cutting parameters until failure occurs. Not only is this simple to do and easy to understand, but it negates the need for special sensing or measuring devices which are inevitably associated with most of the other definitions. Finally, more quantitative information, based upon *experiment*, is available in the literature where this definition has been used.

9.9.1 Tool Life (Taylor) Lines

In 1906, F. W. Taylor* showed that for a fixed set of cutting conditions (that is, workpiece, feed, depth of cut, tool material and signature, and cutting fluid), tool life and cutting velocity were directly related. The plot or line describing such a relation is called a *tool life line*.

Since far more empirical information has been published on this subject, where turning served as the cutting operation, we shall restrict our coverage to that one process. However, the ideas and concepts developed are applicable for other operations such as drilling, milling, and the like. The word *machinability* has been around for many years; we note this here because in assessing the ease with which various metals are machined, a relative ranking has most often been made in terms of some measure of tool life.

Now suppose that we intend to turn a material, say an annealed low carbon steel, using a coolant as a cutting fluid and high-speed steel tools ground to a recommended signature. For discussion purposes, a particular feed rate and depth of cut have been selected. With all of the above conditions kept *constant*, we are ready to obtain a tool life line. Using a freshly ground tool, a cut is started at some initially arbitrary cutting velocity, and the total cutting time is measured until tool failure.† Once the failure zone is cleaned up,‡ a new tool is inserted, the cutting velocity is altered (either higher or lower than the first), and a second test is conducted. Repeating this procedure leads to a number of combinations of cutting velocity-tool life values. With few exceptions, these data points, when plotted on *logarithmic* coordinates, can be reasonably described by a straight line whose equation has the form

$$VT^n = C \tag{9-8}$$

In Fig. 9–33, for example, the data give $VT^{0.1} = 100$. Here, V is the velocity in ft/min, T is the tool life in minutes, n is the tool life exponent, and the value of C accounts for the affects of all of the other parameters which were *fixed* for *this series of tests*.

The numerical value of C is simply the magnitude of V for a tool life of one minute (that is, the intercept of the line where $T = 1$) while n is the *absolute* value of the slope

* F. W. Taylor, "On the Art of Cutting Metals," *Transactions of American Society of Mechanical Engineers* (ASME), **28** (1906), p. 31.

† Tool life is sometimes expressed in cubic inches of metal removed to failure. For reasons that will be shown, we prefer to keep the machining parameters as discrete quantities rather than lump them in a volume rate.

‡ Just prior to tool failure, the temperature in the cutting zone is high enough to form a structure of austentite (for steels), and stopping the cutting action can be followed by rapid cooling of this zone. It is not only possible but probable that a thin layer of martensite results. This must be removed prior to the next test, since this is not the structure being studied in this case.

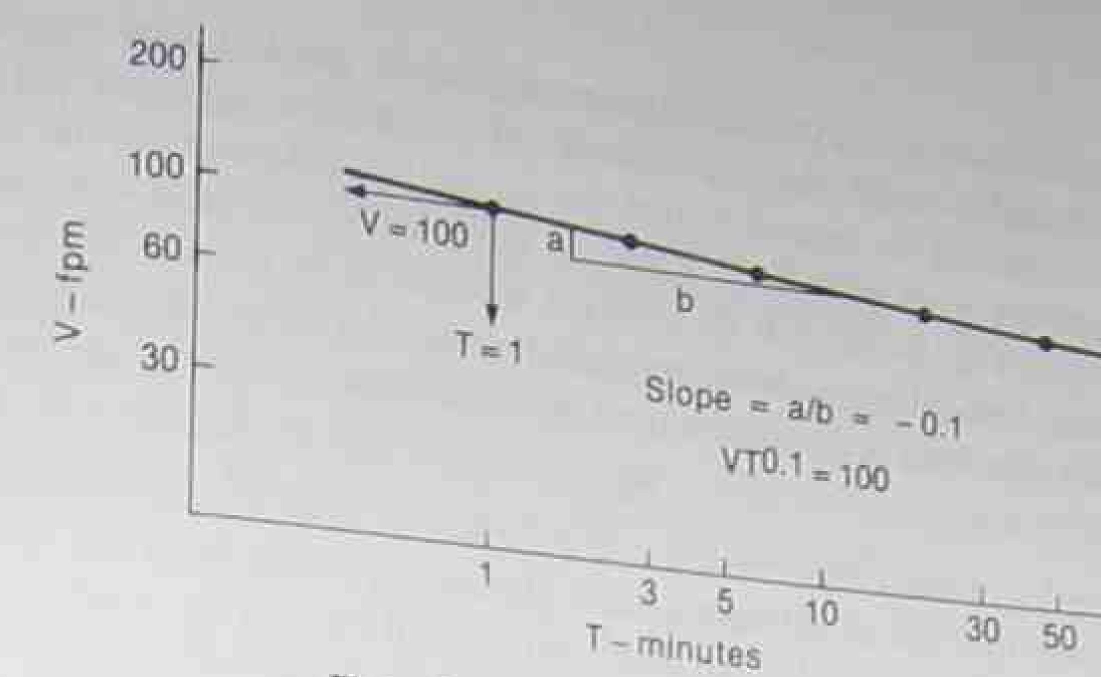


Figure 9–33 A typical tool life line.

of the line. (Strictly, $V = CT^m$, where m is negative as shown. Recasting to $VT^n = C$ means that n is taken as positive.) Thus, for the set of cutting conditions used, once the line or equation is obtained, the cutting velocity needed to produce a desired tool life can be readily found. Certainly, C must have units to make this equation dimensionally correct, but they are rather messy to include. The reader need not be concerned about this if it is realized that the product of V (in feet per minute) and the tool life T (in minutes) raised to the power n will give a numerical value equal to C. *This same physical concept will pertain for using many other empirical expressions that follow in this chapter.* At this point, some practical comments are offered. As with any empirical information, regardless of the field of engineering, the equation of the tool life line describes an observed relationship between two parameters. It is not *exact;* that is, there will always be modest variations in predicted versus observed values, yet such predictions are both useful and reasonable. This often bothers students, since much of their earlier background involved the derivation of equations based upon first principles. Yet one must realize that much of engineering entails the use of empirical work leading to results such as Eq. (9–8). Note also that since the magnitude of n is always less than unity, modest changes in V cause rather *severe* changes in T.

9.9.2 Generalized Tool Life Equations

The development of Eq. (9–8), being based upon a number of constant parameters (that is, f, d, and the like), poses restrictions regarding the use of that equation. If, for example, a constant feed rate of 0.010 ipr had been involved during those tests, and it was desired to use a feed rate of 0.020 ipr, then Eq. (9–8) could not be expected to provide a reliable prediction of T for a certain value of V, if it is assumed that *all* other parameters were identical to the test conditions. Intuition should indicate that cutting a thicker chip would require greater input energy and lead to higher cutting temperatures, thereby giving a shorter tool life for the same cutting speed. This is exactly what does occur. The tool life line developed with the higher feed rate would be displaced below. However, it is usually parallel to the first. In essence, n would be the same, but the magnitude of C would differ and reflect the difference in the combination of cutting parameters. If a constant but different feed rate of, say, 0.005 ipr is used, still a third line results, being displaced *above* the other two (since thinner chips would be cut and less heat would be generated) but again parallel to the others. This *family* of tool life lines is shown on Fig. 9–34(a). Again, it is important to remember that except for the difference in the feed rates, *all* other parameters were held fixed when the three individual lines were determined. Now, as shown in Fig. 9–34(b), if an *arbitrary* tool life of 60 minutes is chosen, and if the three discrete velocities involved are plotted versus the three feed rates, another straight line usually results; for the values used in this discussion this has the equation of

$$V_{60} = K_{60} f^a = 2.6 f^{-0.7} \tag{9–9}$$

where V_{60} means the velocity for a 60-minute tool life. If instead we had chosen an arbitrary tool life of 20 minutes and plotted V_{20} versus feed, the resulting line would be parallel to the one for V_{60} but displaced above it (that is, the constant would be larger than 2.6, but the exponent would remain at -0.7). What has been accomplished is that the feed rate is no longer fixed, instead, for each equation such as Eq. (9–9), the *tool life* is fixed. One could rightly ask, just what have we accomplished? To answer this question, let's pursue a similar family of tool life lines where a different depth of cut had been used throughout, but *all* other parameters are still fixed. Arbitrarily we assume that the depth of cut during the first set of experiments was 0.125 inch while the second set used a depth

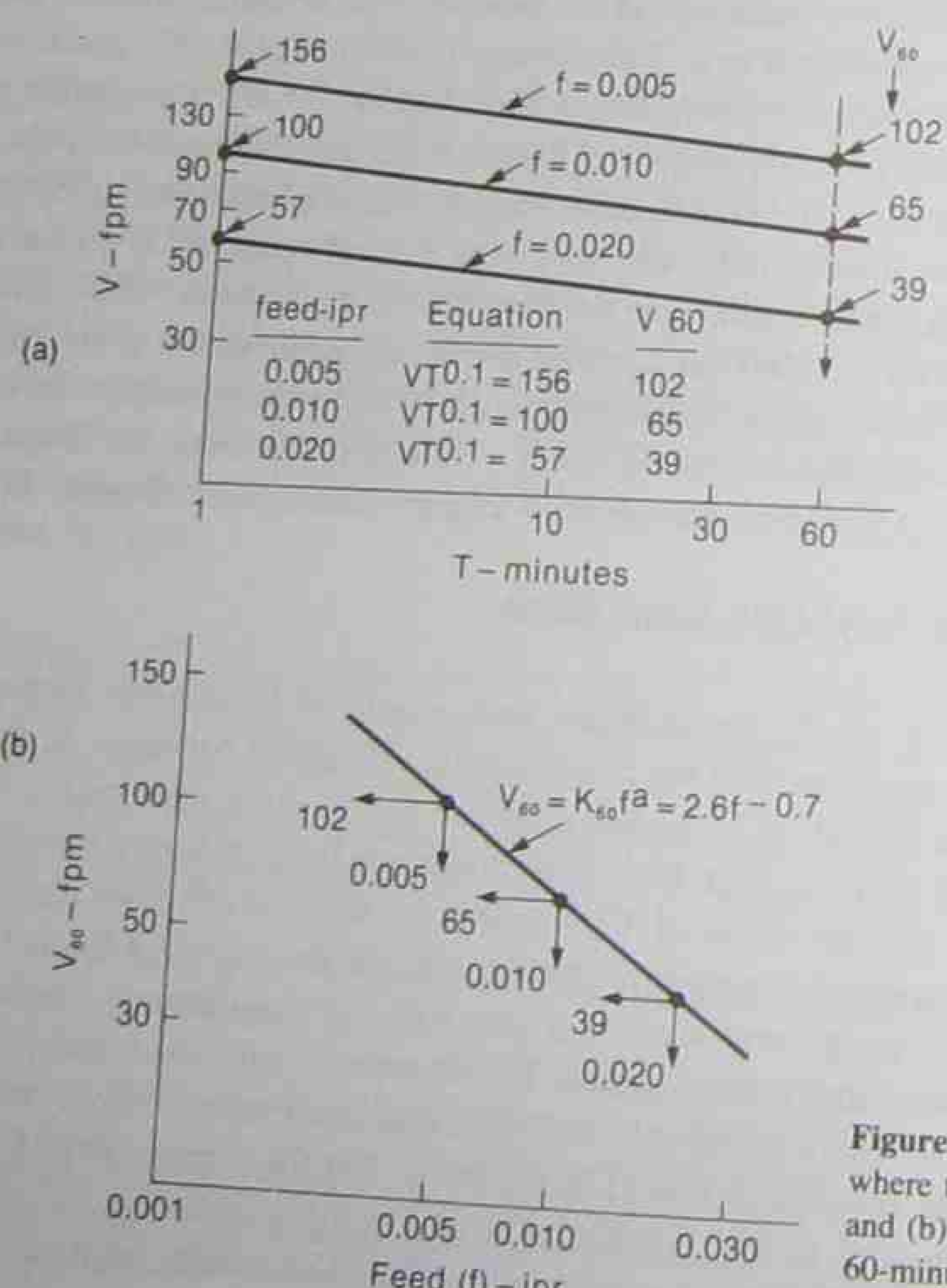


Figure 9–34 (a) Family of tool life lines where the feed rate is the only variable, and (b) a plot of the velocity for a 60-minute tool life versus feed rate for the data shown in (a).

of 0.250 inch. Figure 9–35(a) includes these findings plus those from Fig. 9–34(a). Following the procedures that led to Eq. (9–9), we see in Fig. 9–35(b) that

$$V_{60} = 2.6 f^{-0.7} (0.125 \text{ in. depth}) \tag{9–10a}$$

$$V_{60} = 2.1 f^{-0.7} (0.250 \text{ in. depth}) \tag{9–10b}$$

The coefficient K_{60} (that is, 2.6 and 2.1) varies with the depth of cut and plots of these points are shown in Fig. 9–35(c). *Note:* If other values of K_{60} were determined for other values of d, they would also fall on the line shown; only two values are used here to avoid repetitive discussion. The line on Fig. 9–35(c) fits the equation

$$K_{60} = C_{60} d^b = 1.4 d^{-0.3} \tag{9–11}$$

Combining Eqs. (9–9) and (9–11) in both symbolic form and using the numerical values deduced from the plots, we obtain

$$V_{60} = K_{60} f^a = C_{60} f^a d^b = 1.4 f^{-0.7} d^{-0.3} \tag{9–12}$$

Recalling that V_{60} denotes a 60-minute tool life and with $VT^{0.1} = C$ for any single tool life line, we obtain

$$V_{60}(60)^{0.1} = C \quad \text{or} \quad V_{60} = C/60^{0.1} \tag{9–13}$$

Combining Eqs. (9–12) and (9–13) gives

$$V_{60} = C/60^{0.1} = 1.4 f^{-0.7} d^{-0.3} \tag{9–14}$$

or

$$C = (60)^{0.1}(1.4) f^{-0.7} d^{-0.3} \tag{9–15}$$

so

$$C = 2.11 f^{-0.7} d^{-0.3} \tag{9–16}$$

But, since $VT^{0.1} = C$, substitution into Eq. (9–16) and rearranging give

$$VT^{0.1} f^{0.7} d^{0.3} = 2.11 \tag{9–17}$$

Thus, we have extended the more restricted, individual tool life lines into a more *generalized* form. However, such an empirical expression is accurate only if the other fixed parameters are maintained (that is, tool signature, work material, and the like). Note that the individual parameters whose numerical values would be introduced into Eq. (9–17) carry the units mentioned in the comments leading up to this equation. The constant of 2.11 does have units which are indeed cumbersome and serve no useful purpose; this need not be carried further.

Example 9–2

The generalized tool life equation for a particular turning operation is $VT^{0.1} f^{0.75} d^{0.3} = K$. For the same work material, tool material and signature, and cutting fluid combination, it is known that $VT^{0.1} = 120$ when the feed rate is 0.010 ipr and the depth of cut is 0.150 in. Using this information, determine the cutting velocity that should be used to produce a 30-minute tool life, using a feed of 0.005 ipr and depth of cut of 0.200 in.

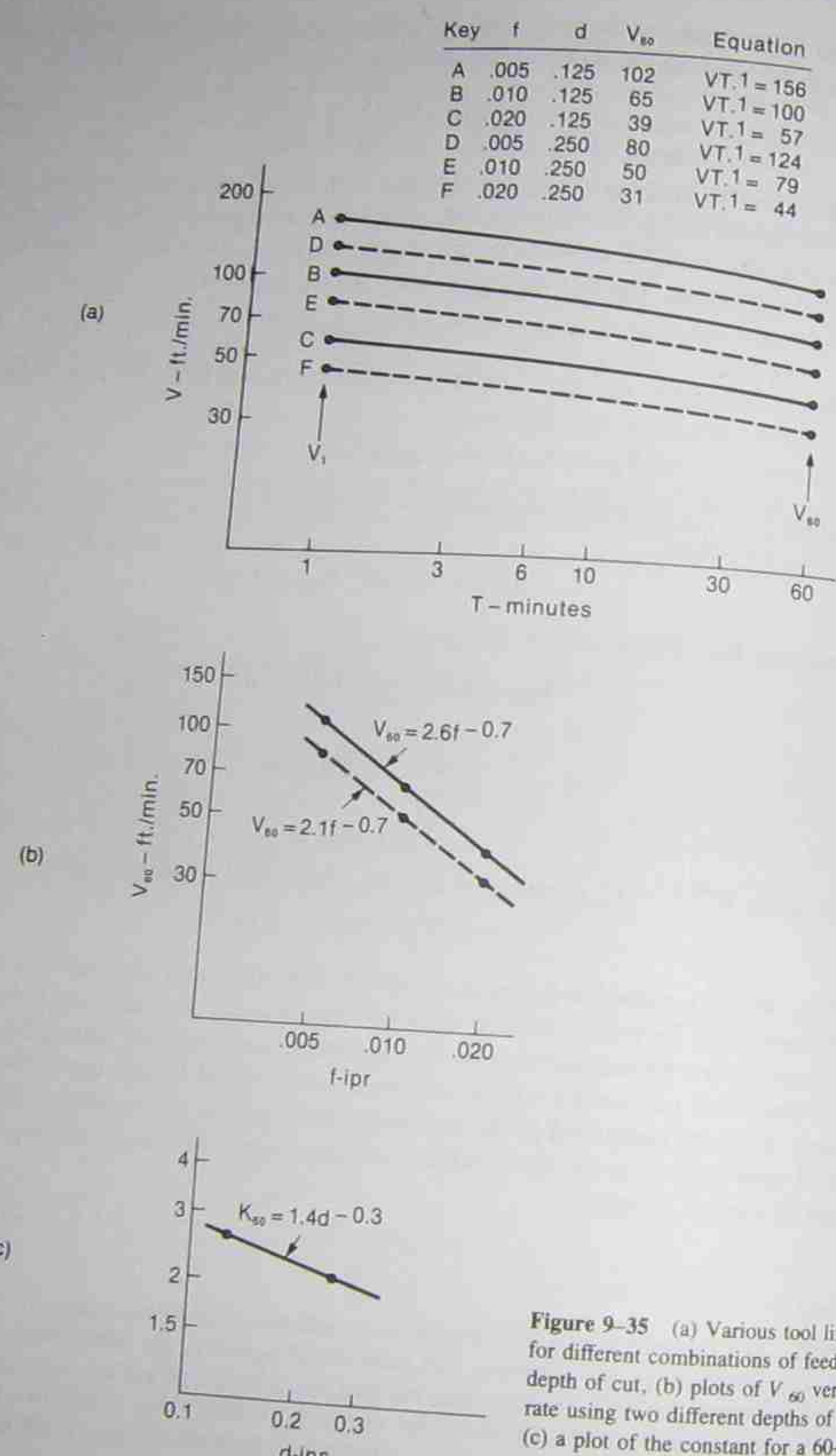


Figure 9–35 (a) Various tool life lines for different combinations of feed rate and depth of cut, (b) plots of V_{60} versus feed rate using two different depths of cut, and (c) a plot of the constant for a 60-minute tool life versus depth of cut.

Solution Although this is not the quickest way to solve this problem, let's find K first. Since the results from an individual tool life line are known,

$$VT^{0.1}f^{0.75}d^{0.3} = K = 120(0.010)^{0.75}(0.150)^{0.3}$$

so $K = 2.15$.

Then

so

$$V(30)^{0.1}(0.005)^{0.75}(0.200)^{0.3} = 2.15$$

$$V = 2.15/(1.405)(0.019)(0.617) = 131 \text{ fpm}$$

Note also,

so,

$$120(0.010)^{0.75}(0.150)^{0.3} = V(30)^{0.1}(0.005)^{0.75}(0.200)^{0.3}$$

$$V = 120/(30^{0.1})(2)^{0.75}(0.75)^{0.3} = 131 \text{ fpm}$$

Since vast amounts of information can be found in the literature, we include only typical tabular or graphical data here to give some idea of how an equation such as Eq. (9–17) might be modified to reflect the influence of parameters such as work material, signature, and cutting fluids. These are shown in Table 9–2. It is unfortunate that such parameters cannot be introduced directly into Eq. (9–17), but since their effects cannot be adequately described by a power-law expression, such action is precluded.

Even with such limitations there are important insights to be gained from Eq. (9–17). Recall that for turning, the rate of metal removal was given by

$$\text{RMR} = 12Vfd \tag{9–18}$$

and, from above,

$$VT^{0.1}f^{0.7}d^{0.3} = 2.11 \tag{9–19}$$

If any of the three parameters V, f, or d is altered, RMR changes in a direct ratio. For example, doubling f would double RMR if V and d are fixed. Note that in Eq. (9–18) all *exponents* are unity. However, because the *exponents* in Eq. (9–19) are different, their influence is not equivalent. To illustrate this point, suppose that a certain tool life is desired and that the turning cut involves a particular depth of cut to produce a desired diameter. Then, since T and d are fixed, Eq. (9–19) reduces to

$$Vf^{0.7} = 2.11/(T)^{0.1}(d)^{0.3} = \text{constant} \tag{9–20}$$

Once we select a feed *or* a cutting velocity, the other parameter is automatically defined. As an illustration, suppose that a feed rate of 0.010 ipr and velocity of 100 fpm provide the desired tool life T. What results if we double f and still want the same T, d being constant? Using Eq. (9–20), we obtain

$$100(0.01)^{0.7} = \text{constant} = V(0.02)^{0.7} \tag{9–21}$$

so the *new* velocity is found to be 62 fpm.

TABLE 9–2* SOME TYPICAL VALUES FOR C AND n IN $VT^n = C$ FOR VARIOUS METALS, TOOL SIGNATURES, AND SIZES OF CUT

Tool		Work Material	Size of Cut		C	n
Material	Signature		f-ipr	d-in.		
High C steel	8,14,6,6,6,15,3/64	Yellow brass	0.013	0.100	300	0.10
HSS	"	Gray cast iron	0.026	0.050	172	0.10
HSS	8,14,6,6,6,0,0	SAE 1035 steel	0.013	0.050	130	0.11
HSS	8,14,6,6,6,15,3/64	SAE 1045 steel	0.013	0.100	192	0.11
↓	↓	SAE 3140 steel	0.013	0.100	178	0.16
		SAE 4350 steel	0.013	0.100	78	0.11
		SAE 4350 steel	0.026	0.100	46	0.11
HSS	8,22,6,6,6,15,3/64	Monel metal	0.013	0.100	170	0.08
		Monel metal	0.026	0.150	127	0.07
T64 carbide	6,12,5,5,10,45,0	SAE 1040 steel	0.025	0.062	800	0.16
↓	↓	SAE 1060 steel	0.025	0.125	660	0.17
		SAE 1060 steel	0.025	0.250	560	0.17
		SAE 1060 steel	0.042	0.062	510	0.16
		SAE 1060 steel	0.062	0.062	400	0.16
High C steel	8,14,6,6,6,15,3/64	Bronze	0.013	0.100	232	0.11
HSS	8,14,6,6,6,0,0	SAE B1112 steel	0.013	0.050	225	0.11
Stellite (Cast nonferrous)	0,0,6,6,6,0,3/32	SAE 3240 steel	0.031	0.187	215	0.19

*Adapted from O. W. Boston, *Metal Processing* (New York: John Wiley and Sons, Inc., 1951), p. 150. The high-speed steel tools (HSS) were all of the 18-4-1 type, and some of the quoted values for f, d, C, and n have been rounded off.

Now compare the rates of metal removal for these two cases, using Eq. (9–18). For the first case,

$$(\text{RMR})_1 = 12(100)(0.01)d = 12d \tag{9–22}$$

whereas

$$(\text{RMR})_2 = 12(62)(0.02)d = 14.9d \tag{9–23}$$

Thus, the increase in RMR is

$$\% \text{ increase} = 100(14.9 - 12)/12 = 24\% \tag{9–24}$$

while the tool life remains the same.

Consider the alternative of doubling the velocity. Then

$$(100)(0.1)^{0.7} = (200)f^{0.7} \tag{9–25a}$$

so the computed feed rate is about 0.0037 ipr. Then

$$\text{RMR}_2 = 12(200)(0.0037)d = 8.9d \tag{9–25b}$$

Thus, there is a *decrease* in RMR of

$$\% \text{ decrease} = 100(12 - 8.9)/12 = 25.8\% \tag{9–26}$$

The lesson to be drawn relates to the most logical *setting* of the machining parameters to produce high rates of metal removal for a desired tool life. It is the *magnitudes of the exponents* in Eq. (9–19) that are significant. Since, in the usual case, the depth of cut has the smallest exponent, feed the intermediate exponent, and velocity the largest exponent, it is *always most sensible* to use the largest depth of cut permissible, and then to use as large a feed rate as acceptable, and, finally, to adjust the velocity to give the desired tool life. The above example shows why *roughing* cuts, where dimensional control and surface finish are not of primary concern, are performed with low speeds, producing chips of relatively large cross-sectional area (that, fd). For a desired tool life, the rate of metal removal is of prime importance. So-called *finishing* cuts, designed to satisfy both dimensional control and surface finish specifications, are carried out at high speeds, producing small chips. Operations such as grinding epitomize such a set of conditions; this indicates why abrasive-type operations are often the last employed, since they do produce parts having small dimensional variations and excellent finish.

9.10 FORCE AND POWER REQUIREMENTS

There are a number of practical reasons why it is important to make reliable estimates of the magnitude of forces necessary to cut materials. Some of the most important reasons are

1. To estimate the motor size for new machine tools.
2. To determine the limiting rate of metal removal with existing machine tools.
3. To assist in the design of clamping and workholding devices.

Variations in cutting forces may also be used in determining the effectiveness of cutting fluids, in machinability evaluations of different materials, and as one sensing source in the area of adaptive control of machine tools. Other uses could be stated; however, in this presentation the major emphasis will be placed upon those listed above as 1 and 2.

9.10.1 Single Point Turning

Because it is the simplest operation to understand and because more numerous studies have been reported in the literature in comparison with other cutting operations, the basic understanding of forces is presented first for a single point turning operation. Figure 9–36(a) shows a schematic of the tool-workpiece setup with the feed and rotational motions indicated; the round bar is being reduced from diameter D_1 to D_2 in one pass. The total force acting on the tool is resolved into three components *for convenience*. These are designated as the radial force F_r, which acts at right angles to the axis of rotation, the longitudinal force F_ℓ, which acts parallel to the axis of rotation and in the same horizontal plane as F_r, and, finally, the tangential force F_t, which acts tangent to the rotating workpiece and is perpendicular to the plane containing the other two components.* Figure 9–36(b) illustrates these three components and the total cutting force as viewed head on with the single point tool.

There is no known method to *accurately* predict the magnitude of these force components without resorting to experimentation. This should not be too surprising when one realizes that, to varying degrees, the following list of common variables enters into force requirements:

* F_t is often referred to as F_c, the cutting force, while F_ℓ is sometimes noted as F_f, the feeding force; see Fig. 9–36(b).

(a)

(b)

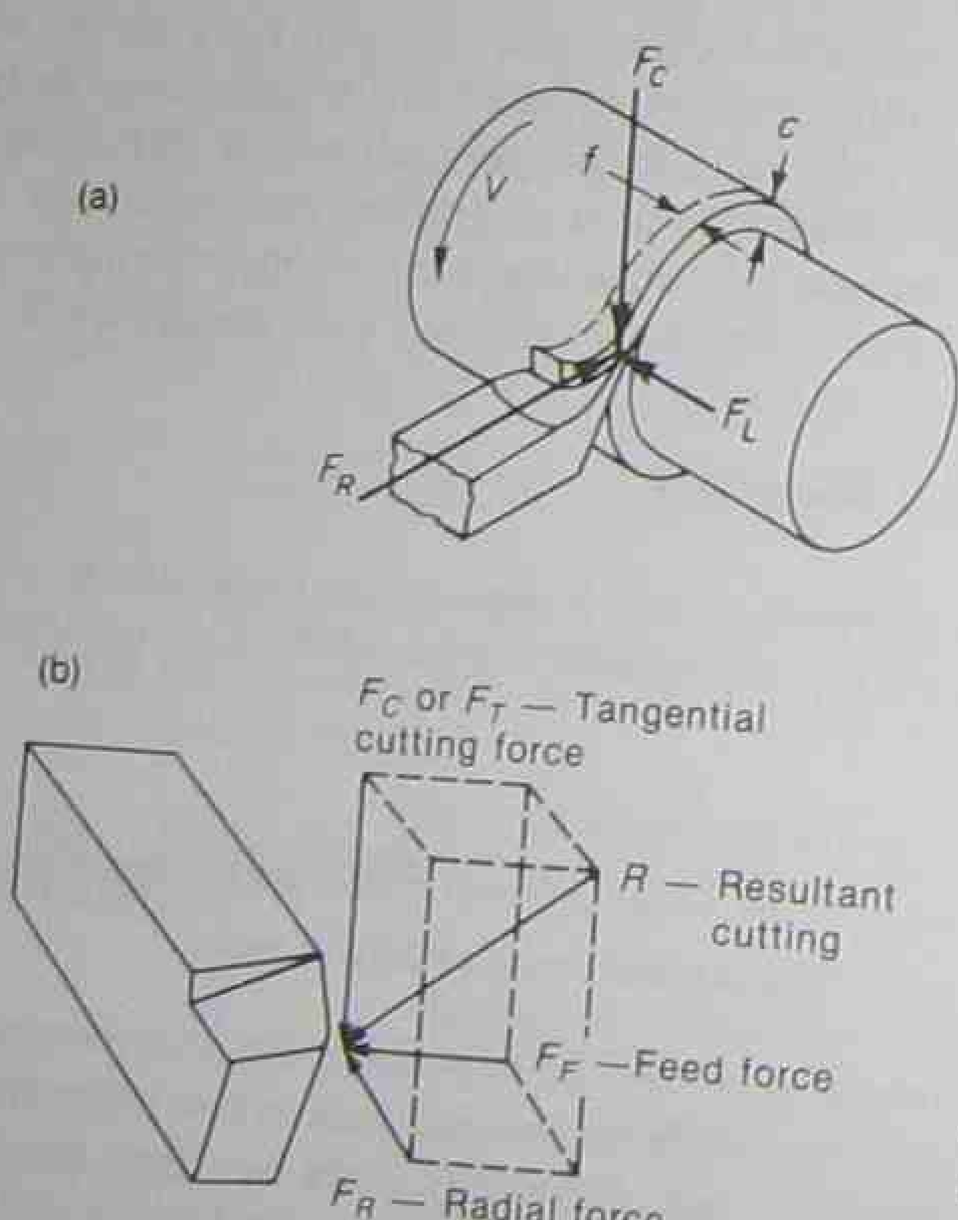


Figure 9–36 (a) Schematic of tool and workpiece to display the three force components, and (b) resolution of the three components into a total force acting on a single-point tool.

1. Feed rate or chip thickness
2. Depth or width of cut
3. Cutting velocity
4. Combination of tool angles (that is, tool signature)
5. Condition of lubrication
6. Work material and/or type of microstructure
7. Sharpness (or dullness) of tool edge

A common practice used in force studies is to employ a *force dynamometer* designed to measure the three components simultaneously for a particular combination of cutting conditions. With a number of such tests on the same workpiece, a range of each important cutting parameter may be investigated to indicate the manner in which the force components behave. Here we shall concentrate on the effects of items 1, 2, and 3 in the above list and, eventually, present some comments on the other parameters.

For a thorough explanation of the manner in which the force components are studied and used, it is necessary to concentrate only on any one of them, since an identical approach would follow for the other two. Because it is the predominant component in power requirements, we shall temporarily concentrate on F_t and then bring in F_r and F_ℓ when the discussion about F_t is completed. Here, the procedure is similar to that used in the discussion on tool life, where a number of experiments are to be conducted using a particular work material, tool material and signature, and condition of lubrication; these parameters are to be fixed for the discussion that follows. Let us further assume that tool dullness, as caused by wear, is not considered for now. What remains as potential variables are the three parameters that constitute the machine tool variables, that is, the feed rate, depth of cut, and cutting velocity.

As an opening investigation, consider that the depth of cut and cutting velocity are fixed, since it is desired to determine the manner in which F_t varies with the feed rate. Once all parameters are fixed, the tool is fed into the workpiece at a selected feed rate, say 0.005 ipr, and when the force output from the dynamometer indicates a quasi-steady-state condition, the tool is disengaged. A number of such tests are conducted over a range of feed rates of interest and what is usually observed, within a very small scatter band, is that the influence of feed rate on F_t can be described quite adequately by a power law expression of the form

$$F_t = K_1 f^{a_1} \tag{9–27}$$

Note the similarity with Eq. (9–8) for a tool-life line, where again the plot is a straight line on logarithmic coordinates.

Before proceeding further, there is a nagging question that sometimes arises here. Does this imply that in order to predict an accurate value of cutting force we must first conduct tests to obtain actual measures of what is to be eventually predicted? The answer is yes, and this should not be too disturbing if one pauses to consider the attainment and use of other engineering parameters. Consider the elastic modulus of *steel*, for example.

For decades, the numerical value of this property has been employed in design calculations, and its average value is usually taken as 30×10^6 psi (207 GPa). Long before calculations based upon atomic bonding forces were made and before experiments using single crystals of iron showed a variation in modulus from about 18 to 42 million psi, experiments with polycrystalline steels led to the average modulus value so commonly used by engineers. It would obviously be pure folly to continually repeat experiments involving this elastic modulus whenever it is used for predictive purposes; in a parallel sense, the use of cutting force expressions developed earlier *from experimentation* also find future use.

Now consider the plot on Fig. 9–37 and let us assume that as the feed rate was varied to produce the individual test values for F_t, the depth of cut and cutting velocity were maintained at 0.075 inch and 75 feet per minute, respectively. The straight line assumed to describe the behavior of F_t versus feed for all of the *fixed* values of the other seven parameters can be expressed as

$$F_t = K_1 f^{a_1} = 3800 f^{0.8} \qquad (9\text{–}28)$$

where the value of a_1 is the slope as shown and the value for K_1 is equivalent to the value of F_t when f equals unity. Obviously, a *physical* value of f equal to unity is ridiculous; remember it is only the *equation* of this line that is sought.

A good question at this point relates to the usefulness of Eq. (9–28). Just as we discussed the restrictions placed upon an individual tool life line, that is, Eq. (9–8), these also apply to Eq. (9–28). By again determining *families* of curves of cutting force as a function of feed rate, then depth of cut, and finally cutting velocity, plots such as those shown in Figs. 9–38 and 9–39 result. Using the same detailed procedures that lead to Eq. (9–17) (that is, the *generalized* tool life equation), the combined effects of feed, depth,

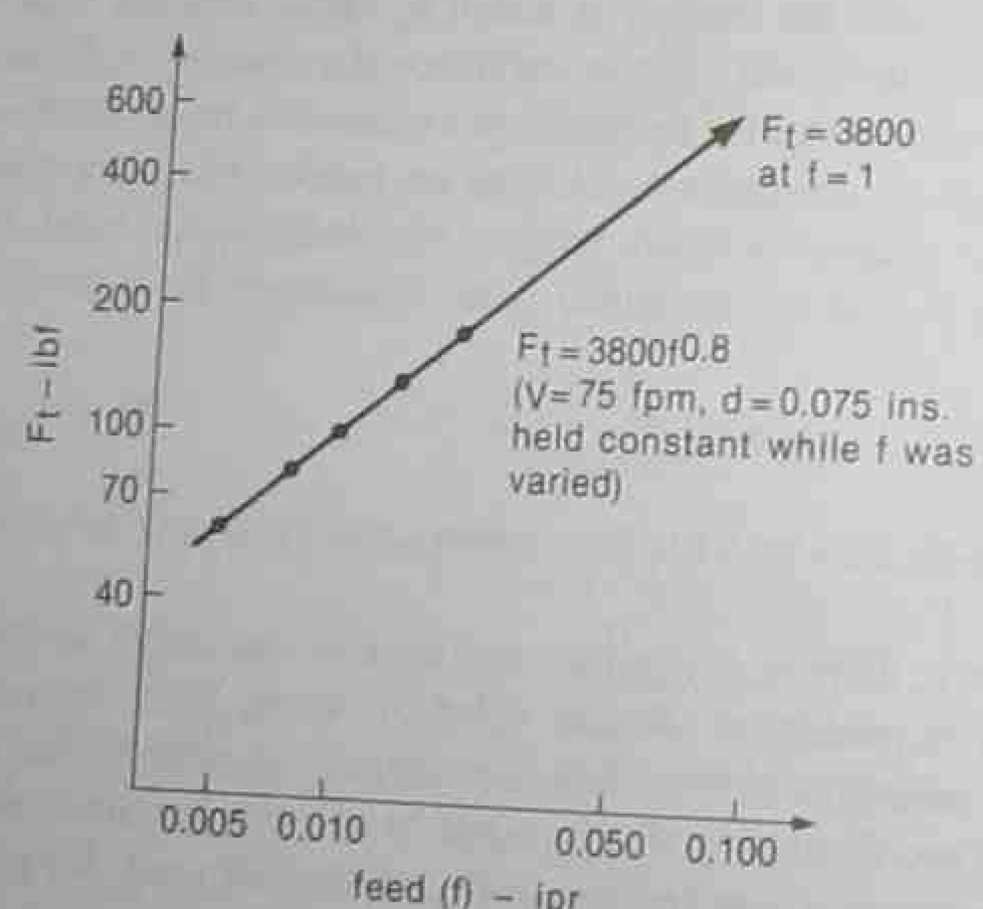


Figure 9–37 Typical variation of tangential or cutting force with feed rate using a constant cutting velocity and depth of cut.

and velocity on the cutting force, for individual materials, can also be expressed in a general form, in terms of the parameters controlled on the machine tool, as

$$F_t = K f^a d^b V^c \qquad (9\text{–}29)$$

Many published results have shown that the exponent c is approximately zero, that is, the value of F_t is little influenced by V over a broad range of cutting speeds.* For our purposes we shall take this exponent as zero *in all further discussions*.

The magnitudes of K, a, and b are dependent upon the work material, tool signature, and, to a much lesser degree, the cutting fluid and tool *material*. To date, no real success has been found in introducing these parameters into an equation such as (9–29). Again, as with tool life, recourse must be made to available graphical or tabular information; note that such information is not necessarily available for all such parametric variations. A little experimentation may then be necessary.

As to Eq. (9–29), we note the following:

F_t = tangential cutting force (in lbf, N, and so forth)
f = feed rate in ipr
d = depth of cut in inches
V = cutting velocity in fpm

* With many metals, for speeds less than 50 fpm (this is low for most operations), F_t may show an increase with speed, whereas at high speeds (say several hundred fpm), F_t may decrease as speed is increased.

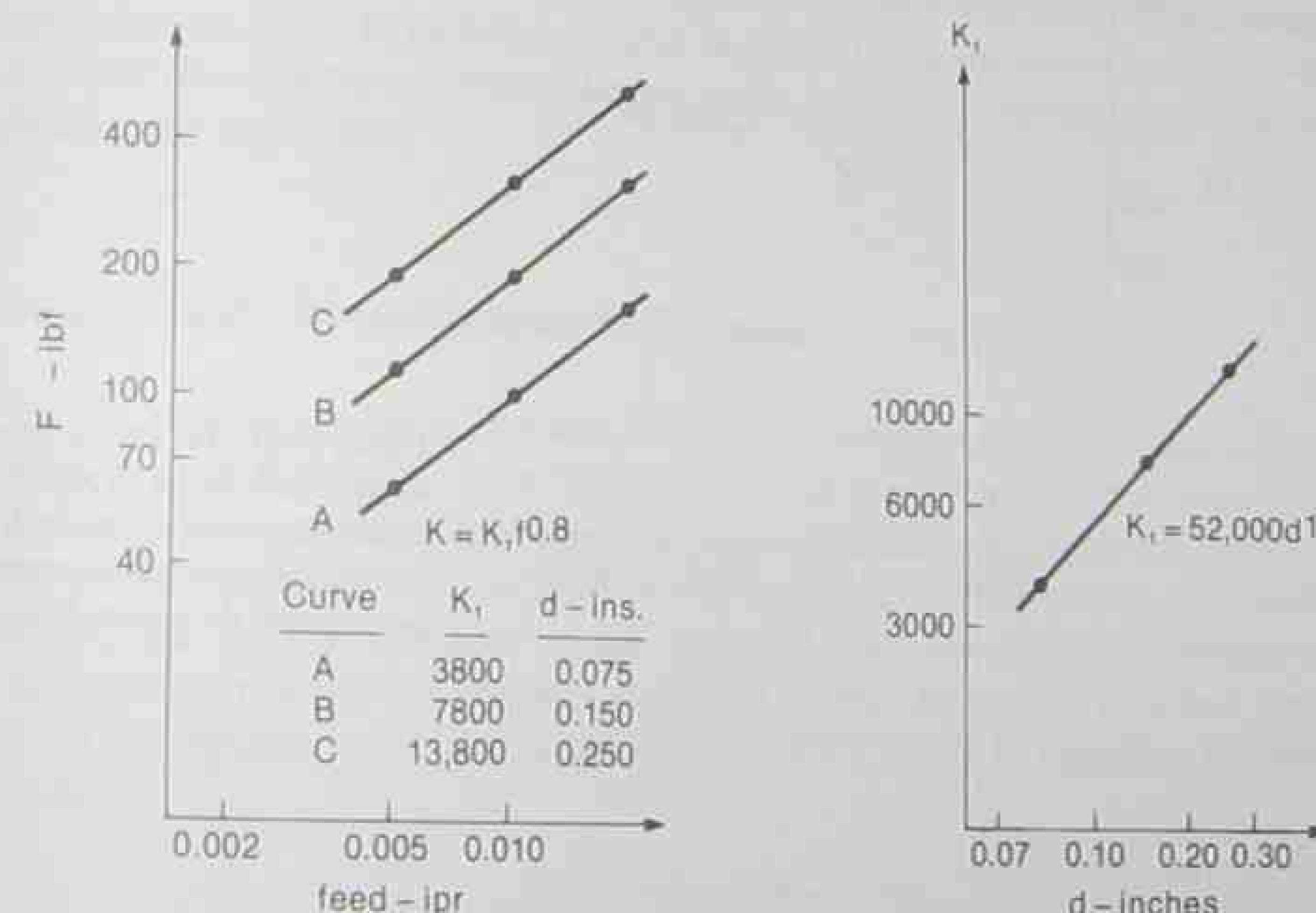


Figure 9–38 Family of curves of cutting force versus feed rate using three different depths of cut but a common velocity.

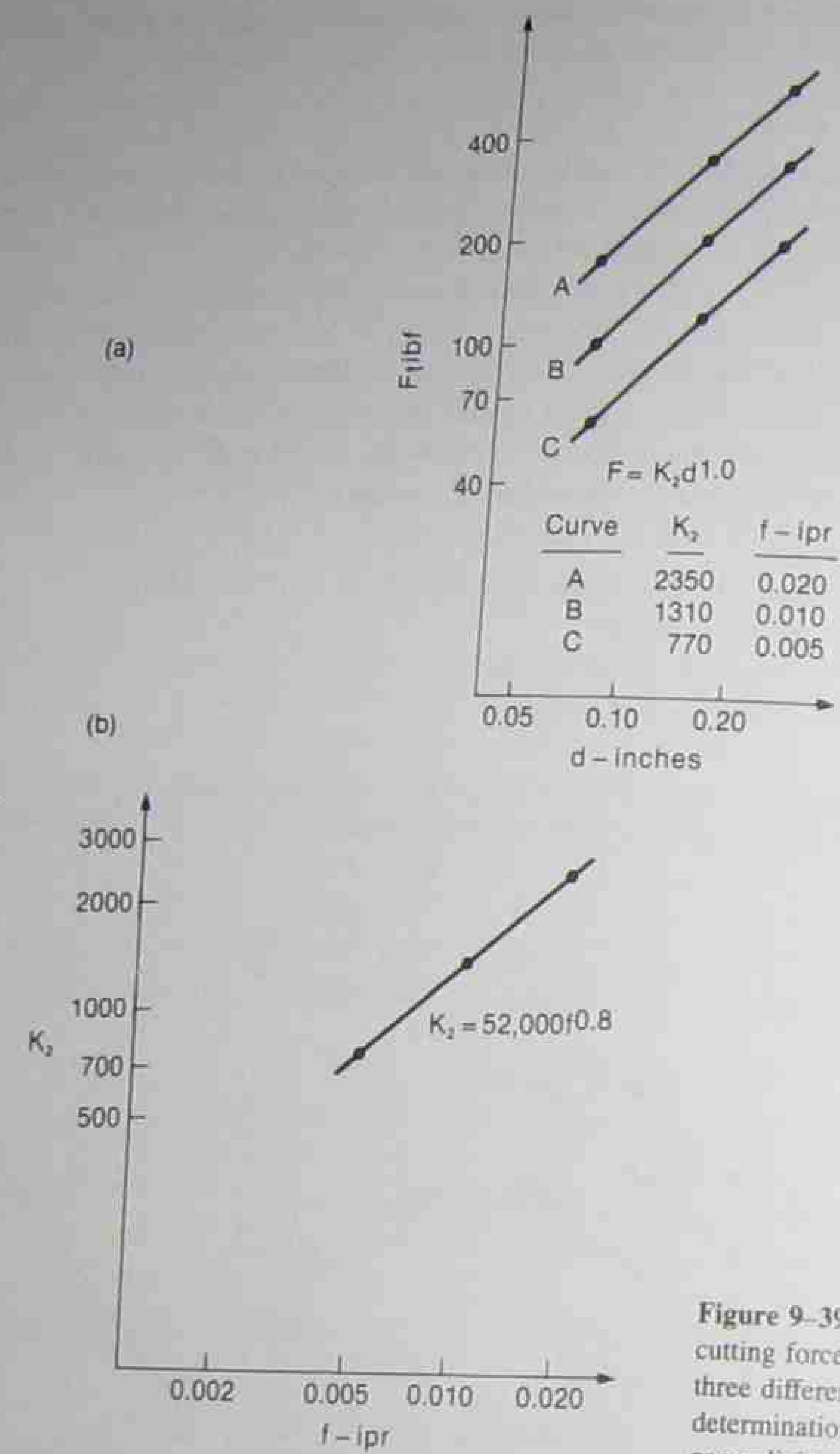


Figure 9–39 (a) Family of curves of cutting force versus depth of cut using three different feed rates and (b) determination of the parameter for generalizing the cutting force equation.

a = slope of plot of F_t versus f (typically 0.7 to 0.9)
b = slope of plot of F_t versus d (typically 0.9 to 1.1)
c = slope of plot of F_t versus V (typically zero)
K = coefficient whose magnitude depends primarily upon the work material and the tool signature.

Several qualitative comments are also provided, where for each parameter being discussed, all others are fixed:

1. *Increasing* the side rake angle causes a *decrease* in F_r.
2. *Increasing* the nose radius will usually *increase* F_r.
3. *Stronger* materials increase the value of K and, therefore, F_t (for example, steel requires larger values of F_t than does brass).
4. As tool wear proceeds, F_t increases, and near tool failure this increase can be drastic. We know of no *quantitative* correction that can be made to reflect the influence of tool wear on F_t. An increase in this force can be used as a diagnostic means to forewarn of impending tool failure.
5. Cutting fluids have relatively little influence on F_t.

All that was said about the method for determining F_t as a function of the many variables involved could be applied to the analysis of F_ℓ and F_r. To indicate typical differences or similarities that occur among these three components of the total cutting force it is most direct to include a number of plots. These are shown in Fig. 9–40. Table 9–3 includes a number of empirical expressions for force components for various materials and tool signatures.

9.10.2 Power Consideration in Turning

If for the time being we ignore that fact that no machine tool possesses a mechanical efficiency of 100 percent, it is then possible to concentrate on the horsepower that is demanded at the cutting tool for a particular set of cutting conditions. Since the previous section described how the total cutting force can be resolved into three useful components, the total horsepower at the cutter can be analyzed in a similar manner. It is common practice to consider the force in lbf, the velocity in feet per minute and the horsepower to be expressed as hP = $FV/33{,}000$. In regard to the three components,

$$hp_c = hp_t + hp_\ell + hp_r \qquad (9\text{–}30)$$

with the subscripts having the same meaning as earlier. In terms of specific components,

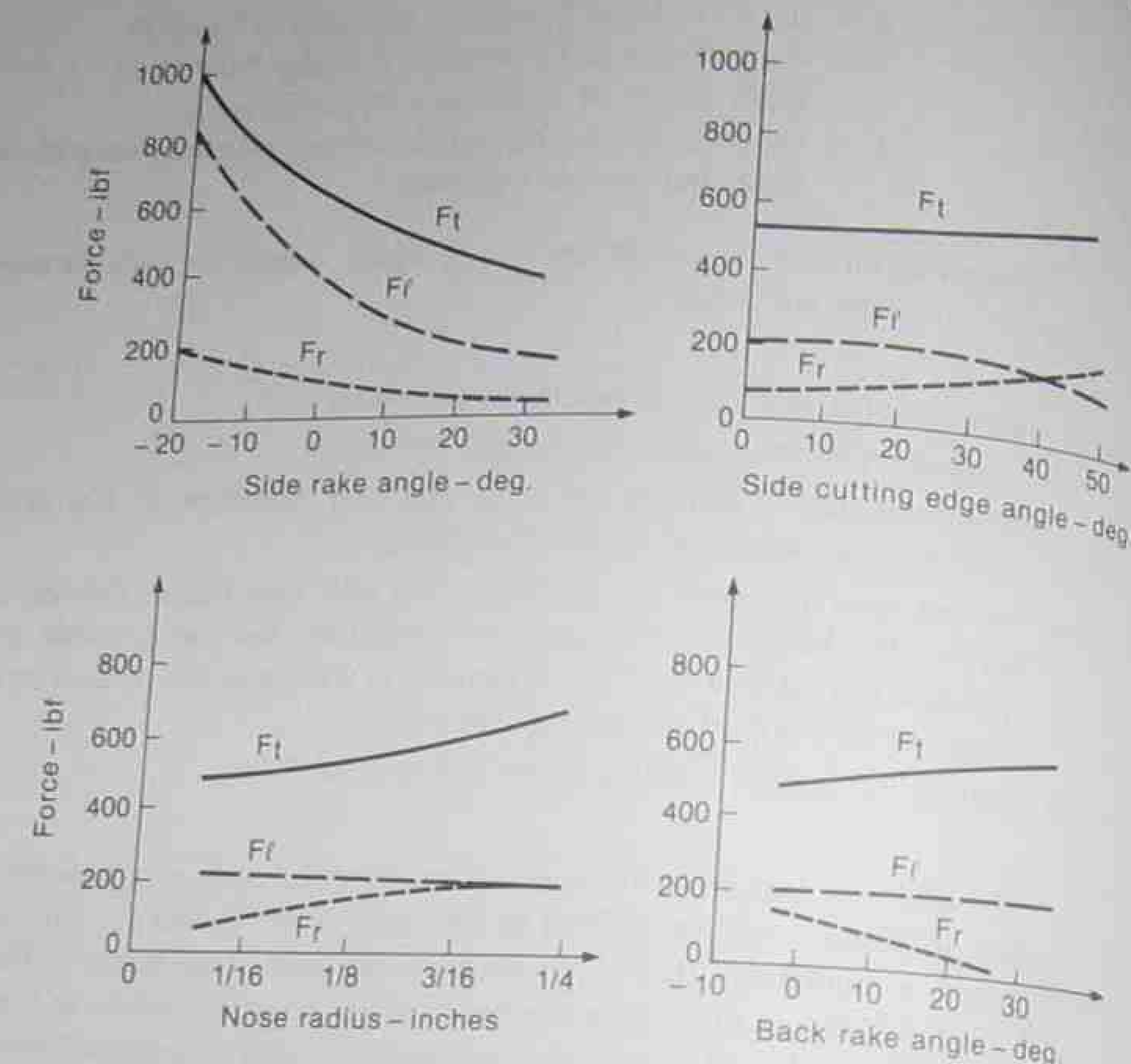


Figure 9–40 Empirical results illustrating how force components vary with tool signature.

$$hp_c = F_t V/33{,}000 + F_f f N/(12)(33{,}000) + F_r f_r N/(12)(33{,}000) \qquad (9\text{–}31)$$

Of course, for turning cuts, the feed in the radial direction (that is, f_r) is zero, so no power is consumed for that component. Note too that the feed f times the rpm N divided by twelve gives, in English units, the velocity in the feeding direction in feet per minute. A similar physical approach may be taken with other cutting operations and is, in fact, recommended in preference to memorizing numerous equations.

Before discussing Eq. (9–31) further, consider that the cutting velocity is related to the rpm by Eq. (9–2) as

$$V = \pi DN/12 \qquad (9\text{–}32)$$

If Eq. (9–32) is introduced into Eq. (9–31), there results

$$hp_c = (F_t \pi DN)/12(33{,}000) + F_f f N)/12(33{,}000) \qquad (9\text{–}33)$$

In most practical operations the product of πD is of the order of 200 to 1000 times larger

TABLE 9–3* TYPICAL VALUES FOR THE COEFFICIENT C AND THE EXPONENTS a AND b IN THE GENERAL EXPRESSION $F = Cf^a d^b$ FOR VARIOUS METALS. NOTE THAT HIGH-SPEED STEEL TOOLS OF DIFFERENT SIGNATURES WERE USED.

Work Material	Tool Signature†	F_t—Cutting			F_f—Feeding			F_r—Radial		
		C	a	b	C	a	b	C	a	b
Stainless steel (18-8)	1	151,000	0.85	0.96	23,000	0.48	1.26	46,500	1.0	0.7
1020 steel (hot rolled)	1	133,000	0.85	0.98	113,000	0.8	1.46	3,565	0.67	0.5
Brass (yellow)	1	83,400	0.81	0.96	95,100	0.9	1.43	9,860	0.97	0.4
Cast iron (126 BHN)	1	29,400	0.78	0.89	8,000	0.5	1.3	—	—	—
Cast iron (241 BHN)	1	64,000	0.78	0.89	13,500	0.5	1.3	—	—	—
Low C steel (100 BHN)	2	133,000	0.83	1.0	33,700	0.48	1.45	923	0.56	0
1020 steel (annealed)	3	156,000	0.88	1.0	51,000	0.42	1.58	2,020	0.69	0

*Adapted from O. W. Boston, *Metal Processing* p. 166 and *Manual on Cutting Metals* 2nd ed. (ASME, 1952), p. 274. Original data revised to above format.

†Tool Signatures were

1. 8,14,6,6,6,15,3/64.
2. 8,14,6,6,6,0,3/64.
3. 0,14,6,6,6,0,3/64.

than the feed f (for example, with a 3-inch-diameter bar and a feed rate of 0.020 ipr, the ratio is about 450 to 1) so as a sensible approximation, the horsepower at the cutter can be taken to be the power demanded by the tangential component only. *Note*, however, that this is because the *velocity* in the longitudinal direction is so small in comparison with the velocity in the tangential direction and *not* because of the relative magnitudes of these force components which may be fairly *similar*.

With the above observation, a reasonable approximation is

$$\text{hp}_c = F_t V_c/33{,}000 = F_t \pi DN/396{,}000 \qquad (9\text{–}34)$$

Now if the effects of machine parameters on F_t [that is, Eq. (9–29)] are used in Eq. (9–34), there results

$$\text{hp}_c = \frac{Kf^a d^b V^{c+1}}{33{,}000} = K_c f^a d^b V \qquad (9\text{–}35)$$

where K_c is, literally, K divided by 33,000 and the velocity exponent c is taken as zero for reasons discussed earlier. Equation (9–35) describes the influence of the parameters on the horsepower demanded at the cutter. An extremely useful quantity is called the *unit power at the cutter* and is defined as the power necessary to remove material at the rate of a cubic inch per minute, or

$$u \text{ hp} = \text{hp}_c/\text{in.}^3 \text{ per minute} = \text{hp}_c/12Vfd \tag{9–36}$$

Combining Eqs. (9–35) and (9–36) gives

$$u \text{ hp} = \frac{K_c f^a d^b V}{12Vfd} = C_c f^{a-1} d^{b-1} \tag{9–37}$$

As indicated in Eq. (9–28) and Table 9–3, the feed exponent is often about 0.8 while the depth exponent is unity. If, for a given material, *this were the case*, then Eq. (9–37) reduces to

$$u \text{ hp} = C_c f^{-0.2} \tag{9–38}$$

so the unit power varies inversely with the feed rate but is *independent* of cutting velocity and depth of cut. The magnitude of C_c is, of course, influenced by *all* of the parameters that affected K in Eq. (9–29). Physically, Eq. (9–38) indicates that it is more efficient from a *power* point of view to cut thick chips, that is, use heavy feed rates, and this is what is done in *roughing* type of cuts, as mentioned at the end of Sec. 9.9. Table 9–4 contains a number of values of unit horsepower at the cutter; in the absence of more complete information, these can be used as good engineering approximations in analyzing power and force predictions.

Three final observations pertain to Eq. (9–38):

1. u hp is independent of velocity if *and only if* the force F_t is independent of velocity [that is, the exponent c is zero in Eq. (9–29)].
2. u hp is independent of the depth of cut if *and only if* the force F_t varies directly with the depth of cut [that is, the exponent b is unity in Eq. (9–29)].
3. The rate of metal removal is reasonably described by $12Vfd$.

Example 9–5

For a certain metal, the magnitude of the tangential (or cutting) force is approximated by

$$F_t = Kf^{0.8}d^{1.0}$$

Suppose that, when $f = 0.008$ ipr and $d = 0.100$ in., F_t equals 400 lbf. How much horsepower would be required at the cutting edge (that is, hp_c) if the cutting conditions were $f = 0.010$, $d = 0.250$, and $V = 100$ fpm?

Solution Most students first find K via

$$400 = K(0.008)^{0.8}(0.100)^{1.0}$$

TABLE 9–4* TYPICAL VALUES FOR UNIT HORSEPOWER AT THE CUTTER (u hp) FOR VARIOUS METALS AND MACHINING OPERATIONS

Material	Hardness BHN (R_C or R_b)	Turning	Milling	Drilling
Plain C and low alloy steels	126–162	0.6–0.9	1.2	1.0
	(30–55)R_c	1.3–2.0	1.5–2.2	1.0–2.0
Stainless steels	135–275	1.2–1.5	1.1–1.6	1.0–1.4
Free machining steels	118–229	0.36–0.54	—	—
Cast iron-gray and nodular	110–190	0.8–1.0	0.8–1.0	0.8–1.0
	190–320	1.4–1.8	1.6–2.0	1.6–2.0
High-temp alloys (Ni and Co base)	200–360	1.6–2.0	2.0–2.5	1.6–2.0
Aluminum alloys	30–150	0.2–0.3	0.3–0.4	0.2
Magnesium alloys	40–90	0.2	0.2	0.2
Copper and copper alloys	(20–100 R_b)	0.6–1.2	0.6–1.2	0.5–1.1
Leaded brass	30–130	0.2–0.35	—	—

* Original data, revised in the above form, was abstracted from:

1. O. W. Boston, *Manual on Cutting of Metals*, (ASME, 1952), p. 282–295.
2. *Machining Data Handbook*, Metcut Research Associates, Inc., Cincinnati, Ohio, 1966, Section 2, p. 508.
3. ASTME, *Fundamentals of Tool Design* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1962), p. 53.

Note that these values will vary with differences in tool signature, sharpness of the cutting edge, and variations in microstructure for a given work material.

that is,

$$K = 400/(0.021)(0.1) = 190{,}480$$

The new value of F_t is then

$$F_t = 190{,}480(0.01)^{0.8}(0.250) = 1196 \text{ lbf}$$

Check this, using

$$400/(0.008)^{0.8}(0.1) = K = F_t/(0.01)^{0.8}(0.25)$$

without finding K explicity. Then

$$F_t = 400(0.01/0.008)^{0.8}(0.25/0.10) = 1196 \text{ lbf}$$

Now,

$$hp_c = F_t V/33{,}000 = 1196(100)/33{,}000 = 3.6 \text{ hp}$$

Recalling the earlier comment about the mechanical efficiency of the machine tool being less than 100 percent, consider now how one might determine the size of drive motor needed on a new machine tool. A useful approach has been to introduce a quantity called the tare horsepower, hp_{ta}, which is equal to the total power consumed to run the machine tool while "cutting air," and a measure of mechanical efficiency which describes that part of the total power from the motor that reaches the cutting tool under cutting conditions. In equation form this may be written as

$$hp_m = hp_{ta} + hp_c/\text{eff.} \tag{9–39}$$

where

hp_m = total horsepower delivered by the motor

hp_{ta} = tare horsepower (this may range from 0.5 to 4 hp with larger lathes requiring greater tare hp).

hp_c = horsepower demanded at the cutting tool for the size of cut under consideration [that is, Eq. (9–35)].

eff. = mechanical efficiency (On modern machines with antifriction bearings and efficient drive trains, this is of the order of 95 percent, while with older machines and low efficiency drive trains, it might be a slow as 60 percent).

As a final thought it should be realized that any of the parameters that affect F_t, and are not expressed as direct variables in the form such as Eq. (9–29), will have an effect on hp or u hp in a manner in which they influence F_t. Figure 9–41 illustrates such effects.

Example 9–6

For a particular metal, u hp = 0.8 when f = 0.010 ipr, d = 0.100 in., and V = 100 fpm. The cutting force for this metal, $F_t = Kf^{0.85} d^{1.1} V^0$.
If cutting is to be done where f = 0.005 ipr, d = 150 in., and V = 0.150 fpm, determine
1. The value of u hp for these conditions.
2. The size of motor needed if hp_{ta} is 1.4 and an efficiency of 85 percent is reasonable.

Solution

1. $u \text{ hp} = F_t V/33{,}000(12)Vfd = Kf^{0.85}d^{1.1}V/396{,}000(Vfd)$
so $u \text{ hp} = Cf^{-0.15}d^{0.1}$, where $C = K/396{,}000$
Now the new u hp can be found by first finding C. From the initial conditions,

$$0.8 = C(0.010)^{-0.15}(0.1)^{0.1}$$

so

$$C = 0.8(0.01)^{+0.15}/(0.1)^{0.1} = 0.505$$

Then, for the new conditions,

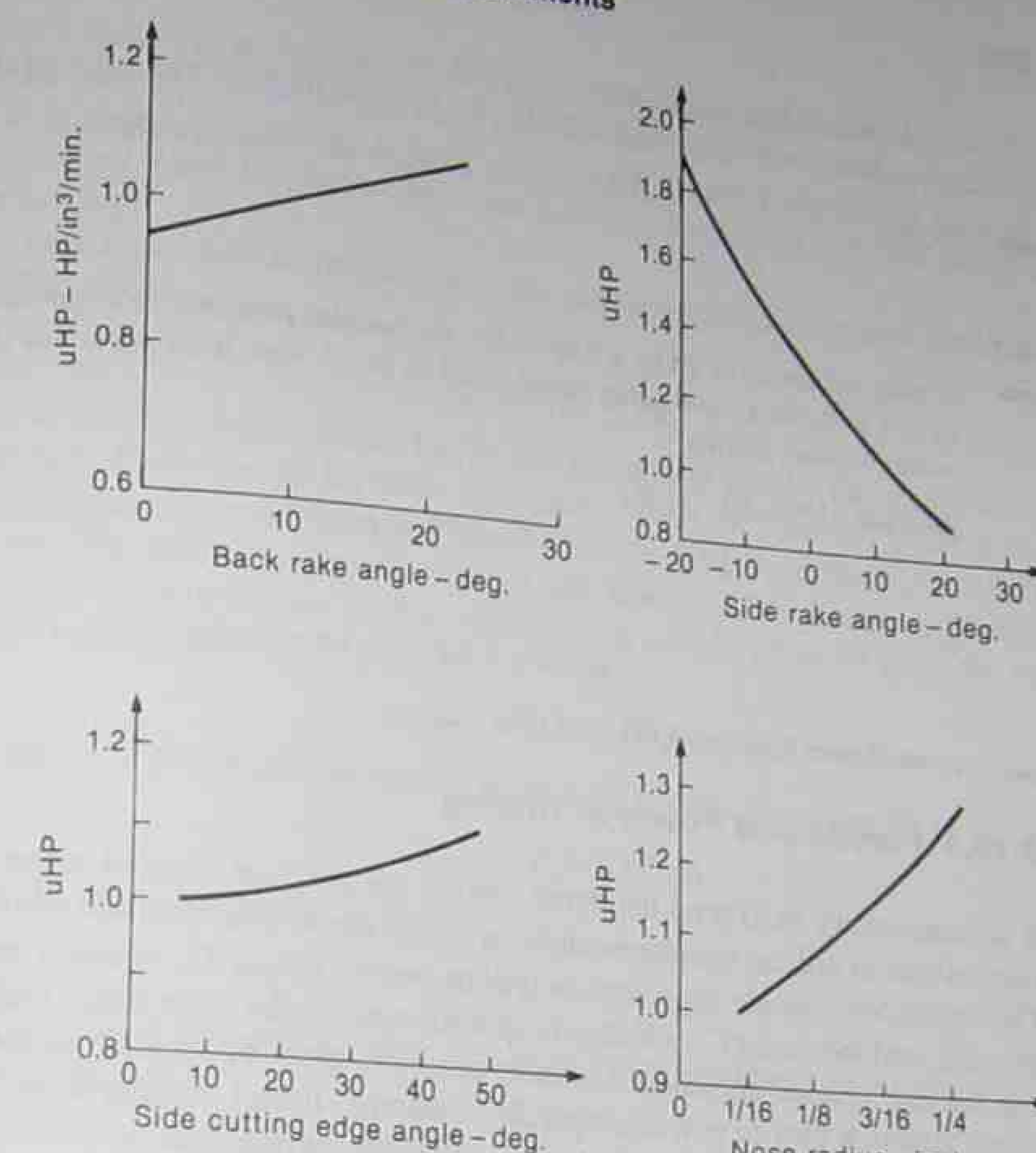


Figure 9–41 Empirical results showing the effect of tool signature on uhp; note that these are relative values.

$$u \text{ hp} = 0.505\,(0.005)^{-0.15}(0.150)^{0.1}$$

so

$$u \text{ hp} = 0.924$$

Another approach is as follows:

$$u \text{ hp}(f)^{0.15}\,d^{-0.1} = C$$

so

$$0.8(0.01)^{0.15}\,(0.100)^{-0.10} = u \text{ hp}(0.005)^{0.15}\,(0.150)^{-0.1}$$

or the new $u \text{ hp} = 0.8\,(0.01/0.005)^{0.15}(0.100/0.150)^{-0.1}$
so u hp = 0.924 *without evaluating C*.
2. With Eq. (9–39)

$$hp_m = hp_{fm} + hp_c/\text{eff.}$$
$$hp_m = 1.4 + (12Vfd)u\,hp/0.85 = 1.4 + 12(150)(0.005)(0.15)(0.924)/0.85$$

so $hp_m = 1.4 + 1.47 = 2.87$ (that is, a 3-hp motor would just suffice).

Example 9–7

A lathe spindle is driven by a 5-hp motor, the machine efficiency is 80 percent, and the tare hp is 0.5. If, for a metal being turned, the u hp is 1.0, what is the maximum possible rate of metal removal, (RMR)?

Solution From Eq. (9–39)

$$5 = 0.5 + hp_c/0.8$$

or

$$hp_c = 4.5(0.8) = u\text{ hp(in.}^3\text{/min)}$$

so maximum RMR $= 3.6/1 = 3.6$ in.^{3}/min

9.10.3 Forces and Power in Drilling

It is unnecessary to expend the detail used in the previous sections when analyzing the equivalence in drilling operations since an almost one-to-one approach could be followed. The major components of interest in drilling are the *torque* T_0, which is needed to shear the chip, and the *thrust* T_h, which acts opposite and parallel to the linear drill motion. Note that for this type of operation it is more useful to speak of a torque instead of a cutting force, although they have a corresponding meaning. In fact, we can show this correspondence if we reverse the procedure used earlier by starting with power requirements and note the force breakdown at the cutting edge as shown in Fig. 9–42. There the total force per edge is resolved into three components. The radial component F_r on each edge leads

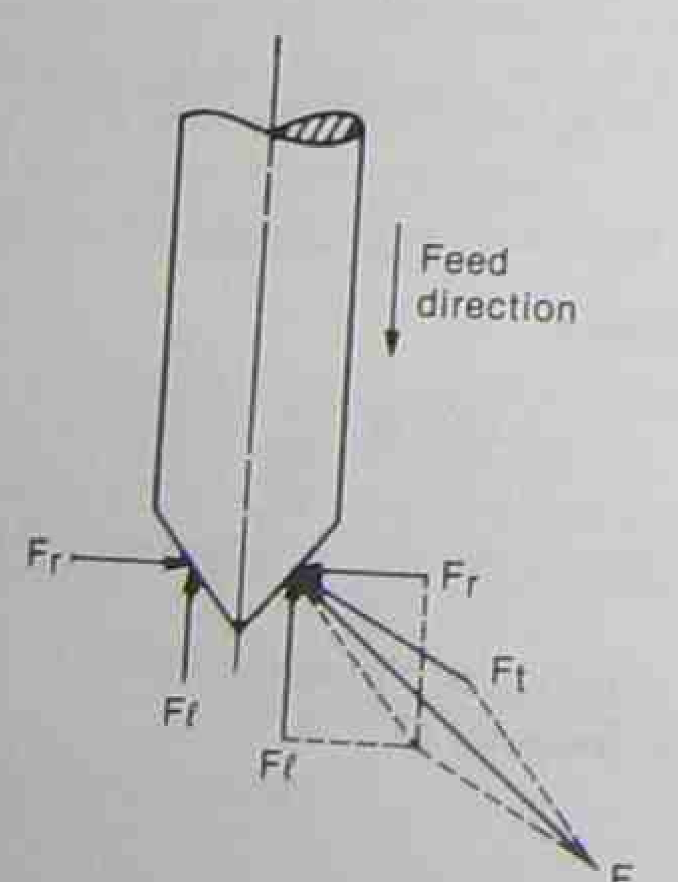


Figure 9–42 Schematic of forces acting on a drill.

to a net balance of zero and may be discarded from further consideration. The *thrust force* T_h is the sum of F_ℓ on each edge, while the force causing the torque acts normal to the edge and is assumed, for simplicity, to be concentrated at the midpoint on each of the two cutting edges (*Note:* this force component is directly analogous to F_t used earlier in single point turning).

Now if attention is concentrated at the cutting edge and if the same definitions apply as in Eq. (9–30),

$$hp_c = hp_t + hp_\ell \quad (9\text{–}40)$$

where in drilling, hp_t is demanded by the torque and hp_ℓ by the action of the thrust component. First consider the thrust component, where the symbol T_h will be used as the *total* thrust force opposing motion of the drill into the workpiece (that is, disregard F_ℓ from here on). The velocity in the direction of drill motion is the product of the feed rate in inches per revolution times the rpm; dividing this product fN by 12 gives the velocity in feet per minute; thus,

$$hp_{Th} = T_h fN/12(33{,}000) \quad (9\text{–}41)$$

Now the force producing the torque can be used to give the *torque hp* as

$$hp_{To} = F_t V/(33{,}000) \quad (9\text{–}42)$$

where the actual torque T_o, in *lb ft,* is

$$T_o = F_t(D)/(4 \times 12) \quad (9\text{–}43)$$

and D is the drill diameter in inches. The velocity corresponding to this point at which the force is indicated is

$$V = \pi(DN)/(2)(12) \text{ (in feet/minute)} \quad (9\text{–}44)$$

Introducing Eqs. (9–43) and (9–44) into Eq. (9–42) gives

$$hp_{To} = 48T_o(\pi DN)/(D)(24)(33{,}000) = (2\pi T_o N)/33{,}000 \quad (9\text{–}45)$$

This is the well-known relation between torque and rpm developed in most sophomore courses in mechanics.

Although we could have introduced Eq. (9–45) directly, it is hoped that the approach used above will point out that the torque effect in drilling is really equivalent to the tangential component in single point turning. Equation (9–40) may be expressed as

$$hp_c = (2\pi T_o N)/(33{,}000) + (T_h fN)/(12)(33{,}000) \quad (9\text{–}46)$$

As discussed after Eq. (9–33), although perhaps not as obvious, the portion of the *total horsepower* at *the cutting edge* due to the thrust component (that is, the feeding direction) is minimal as compared with that demanded by the torque. Again, this is because of the velocity or speed ratio and *not* because T_h is negligible. In fact, in every cutting operation, the major power demand comes from the *force component* associated with the *major velocity* involved in the operation and hp_c can, for all practical purposes, be defined in terms of that component only. For drilling, this would reduce Eq. (9–46) to

$$hp_c = 2\pi T_o N/33{,}000 \quad (9\text{–}47)$$

In a manner identical to that which led to Eq. (9–29) for turning, torque could be expressed as a function of machining variables by

$$T_o = Cf^m D^n V^p \quad (9\text{–}48)$$

Typical values for m, n, and p, for many metals, are 0.8, 1.8, and zero (note that T_o is independent of velocity in general just as F_t was in single point turning). Using these numerical values in Eq. (9–48) and introducing the result into Eq. (9–47) give

$$hp_c = 2\pi(Cf^{0.8}D^{1.8})N/33{,}000 \quad (9\text{–}49)$$

which is analogous to Eq. (9–35). In drilling, the rate of metal removal in cubic inches per minute is the area of the drilled hole times the feed and rpm; that is,

$$RMR = \pi D^2 fN/4 \quad (9\text{–}50)$$

The unit horsepower at the cutter for drilling is obtained by using Eqs. (9–49) and (9–50) to give

$$u\ hp = 2\pi CNf^{0.8}D^{1.8}(4)/33{,}000(\pi)D^2(fN) = C_1 f^{-0.2}D^{-0.2} \quad (9\text{–}51)$$

The similarity with Eq. (9–38) is apparent and for those materials having an exponent n = 2.0 in Eq. (9–48), the unit horsepower would again be dependent on the feed rate only.

Although the *thrust* horsepower is usually negligible, the thrust *force* itself can be extremely large and must be considered in fixture design and clamping analyses. As a function of the usual parameters, a typical expression for the thrust force would be

$$T_h = C_2 f^{0.8}D^{1.0} \quad (9\text{–}52)$$

It cannot be emphasized too strongly that the ratio of chisel edge to drill diameter can have a decided influence on the thrust, and a lesser extent on the torque. For the purposes herein, the typical values shown in Table 9–5 may be used for *first approximations.*

Example 9–8

For a certain drilling setup, the torque (T_o) and thrust (T_h) are expressed by

$$T_o = 600f^{0.8}D^2 \quad [T_o \text{ in lb ft}, f \text{ in ipr, and } D \text{ (drill diameter) in in.}]$$
$$T_h = 50{,}000f^{1.0}D^{1.2} \quad (T_h \text{ in lbf})$$

Using a ¾-inch drill, a feed of 0.010 ipr, and an rpm of 600, determine the horsepower demanded by the torque and thrust reactions.

Solution Torque: Using Eq. (9–45), we have

$$hp_{To} = 2\pi T_o N/33{,}000$$

or

$$hp_{To} = 2\pi(600)(0.010)^{0.8}(0.75)^2(600)/33{,}000 = 0.968 \text{ hp}$$

TABLE 9–5* TYPICAL VALUES FOR THE COEFFICIENTS AND EXPONENTS IN TORQUE AND THRUST RELATIONS IN DRILLING WHERE: TORQUE (T_o) = $C_1 f^m D^n$ and Thrust (T_h) = $C_2 f^x D^y$

Material	T_0 in pound-feet			T_h in pounds force		
	C_1	m	n	C_2	x	y
Aluminum alloys	550	0.83	1.9			
Leaded brass	418	0.73	1.9	48,000	1.1	1.2
1020 steel	1590	0.78	1.8	6,600	0.6	1.0
1045 steel	1590	0.78	1.8	40,000	0.78	1.0
4130 steel	1560	0.78	1.8	42,000	0.78	1.0
				43,000	0.78	1.0

* Selected values adapted from O. W. Boston *Metal Processing* (New York, John Wiley and Sons, Inc., 1951), pp. 337–338. *Note* that these can vary with the helix angle, the size of the chisel edge, the microstructure of a given work material, and type of cutting fluid. The feed rate f is in ipr and the drill diameter D is in inches.

Thrust: Using Eq. (9–41), we have

$$hp_{Th} = T_h fN/396{,}000$$

so

$$hp_{Th} = 50{,}000(0.010)^1(0.75)^{1.2}/396{,}000 = 0.009 \text{ hp}$$

This illustrates why the hp_c in drilling is, for all practical purposes, due to the *torque* demands as in Eq. (9–47).

Some final observations regarding the thrust force are made with respect to Fig. 9–43. To balance the drill it is essential to grind the point so that the angle and cutting edges are, essentially, evenly divided by the drill axis. This produces equal horizontal components and a hole that is just about equal to the actual drill diameter. The chisel edge is extremely *inefficient* as a cutting edge; rather, the material beneath the chisel edge is forced to flow by plastic deformation. It is for this reason that thinning of the chisel edge can drastically lower the thrust force, or that large holes are often produced *after* an initial small hole is drilled. If the major contact of the chisel edge is avoided, a tremendous

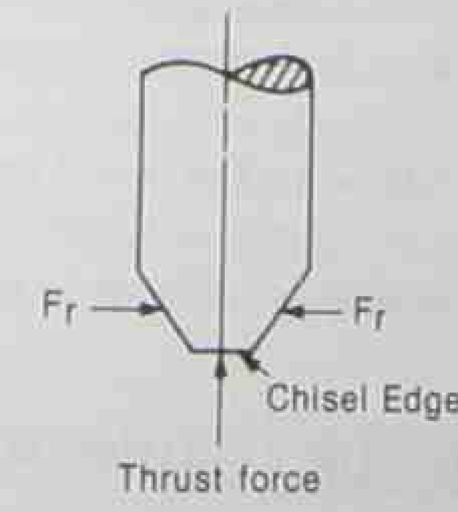


Figure 9–43 Schematic of forces on the cutting edges and chisel edge of a drill.

decrease in drill thrust follows. If the thrust force is used to hold the work in place on the drill table, *breaking through* of the chisel edge can lead to serious consequences, since this force is released for all practical purposes and the work will spin around with the drill. This is why the holding of a workpiece by hand is to be avoided when one is drilling materials that demand high torque for cutting. So long as the chisel edge exerts its influence, holding the workpiece may appear safe, but once the edge breaks through the bottom, the torque will tend to tear the work away from the operator. However, overthinning of the chisel edge can lead to a point whose structural weakening would necessitate very low feed rates to prevent the point from breaking. In practice, compromises must be made; what is important here is to understand some of the principal reasons why such approaches must be taken.

9.10.4 Closing Comments

If one understands the concepts presented in the previous two sections, then the use of available values of the *u* hp, can be applied readily to literally any machining operation if one is interested in determining the approximate magnitude of horsepower needed to produce a certain rate of metal removal. In addition, the level of cutting force F_t required can be easily computed once hp_c is found. Unfortunately, the *magnitudes* of the radial and longitudinal forces (which may be called by different names) cannot be deduced from such information. Assuming that such magnitudes are of interest, one must resort to a sensible literature search, hoping that, for the combination of conditions involved, earlier reported findings will be useful.

Example 9–9

A slab of steel, 6 in. wide by 4 in. high, is to be milled with a facing cutter to reduce the 4-in. height to 3.9 in. in a single pass. The 10-in.-diameter cutter contains 16 teeth; an average feed rate of 0.006 inches per tooth is desired, and the cutting velocity is to be about 130 fpm. If the *u* hp is about 1.2, estimate the magnitude of the cutting or tangential force expected.

Solution First, find the table feed rate to be set on the machine as found from Eq. (9–1),

$$F = fnN, \text{ where } n = 16 \text{ and } N = 12(130)/\pi 10 = 50 \text{ rpm}$$

then

$$F = (0.006)(16)(50) = 4.8 \text{ in./min}$$

$$\text{RMR(in.}^3\text{/min)} = Fwd, \text{ where } w = 6 \text{ in. and } d = 0.100 \text{ in.}$$

so

$$\text{RMR} = 4.8(6)(0.1) = 2.88 \text{ in.}^3\text{/min}$$

then

$$hp_c = u\,\text{hp(RMR)} = 1.2(2.88) = 3.46 = F_tV/33{,}000$$

so

$$F_t = 33{,}000\ hp_c/V = 33{,}000(3.46)/130 = 877 \text{ lbf}$$

We add a final comment about *u* hp for grinding operations. Such values are *much* higher than those associated with turning, as shown in Table 9–4. With grinding, values of 5 to 20 are not uncommon and, as related to Eq. (9–38), result because the size of cut (f,d) is exceedingly small. Even though the *u* hp is quite large, the RMR (in.^{3}/min) is very low, and as a consequence the *cutting* horsepower is also low. Grinding wheels operate at cutting velocities of the order of 5000 ft/min; thus, the torque or force requirements are extremely low. That, as noted earlier, is one of the major reasons why surface finish and dimensional control are superior with grinding as compared with turning, milling, or other *large* chip operations; abrasive operations induce much smaller forces, so deflection of the workpiece and distortion of the work surface are much less severe compared with large chip operations.

9.11 PLASTICS OR POLYMERS

Many of the observations and concepts covered earlier regarding metals find a reasonably direct correspondence with plastics. Rather than repeating the extensive coverage devoted to metals, a concise, qualitative summary, especially of any differences in machining behavior, will be presented. For specific recommendations of such factors as tool signature, size of cut, cutting velocity, and so forth, the reader can consult the text by Kobayashi* as an excellent starting point. In most instances, recommended differences in machining plastics as compared with metals, come about because of the significant differences in many of the properties of these two classes of solids. The following list includes these major differences:

1. Plastics have a much lower elastic modulus than metals. As an example, the modulus of steel is of the order of 100 times that of plastics. Because of this lower rigidity, plastics display greater deflections for certain machining operations; thus *elastic recovery* can cause problems.
2. For the same quantity of heat applied to equal volumes of a plastic and a metal, the temperature rise of the plastic is larger. This may lead to a softening of the plastic and a "gumminess" of the workpiece.
3. Properties such as specific heat, thermal conductivity, and coefficient of thermal expansion, are very different for these two classes of solids.
4. Plastics begin to soften at much lower temperatures.
5. Plastics have different toughness-to-strength ratios than metals. This can cause cracking as has occurred when plexiglas is drilled at too high a feed rate.

Because of these differences, cutting conditions that are fully acceptable with metals may have to be altered to provide successful machining of plastics. For example, turning

* A. Kobayashi, *Machining of Plastics* (New York: McGraw-Hill Book Co., 1967). That book contains a wealth of information, involving a number of common plastics as work materials.

requires relatively larger rake angles, although an upper or critical value must be maintained; beyond this, a degradation of surface finish occurs. Additionally, large relief angles are needed to minimize rubbing between the tool and finished surface. Such large angles, in both instances, are needed to alleviate the effects which can occur because of the large elastic recovery of the machined surface that results after the tool edge passes by that surface.

Rather serious problems can occur when drilling plastics if improper conditions prevail. As chips are removed along the flutes of the drill, the heat involved in cutting can cause the plastic to soften to the extent that the chips *weld* to the flute surfaces. Not only does this prevent easy removal of subsequent chips, but the necessary torque may reach a level that causes the drill to fail by literally twisting into two pieces. Large helix angles and an effective lubricant reduce the tendency of flute packing and, possibly, drill failure. However, care must be exercised in the use of a fluid. Many plastics are susceptible to structural degradation and, additionally, the phenomenon of *crazing*. A craze is a type of void that is filled with a web-like structure of fibrils. This comment indicates that sound suggestions from those knowledgeable about such chemical reactions must be sought. Due to their low elastic modulus and rather high coefficient of expansion, plastics have, in some cases, ended with drilled holes that are *smaller* than the drill diameter. This may present problems when attempts are made to back out the drill after a relatively deep hole has been machined.

As with metals, both continuous and discontinuous chips can occur; in general, continuous chips provide a better surface finish. However, if excessive heat is induced during cutting, the surface can actually melt; melted material, upon solidification, gives an extremely poor finish.

With regard to the magnitude of the cutting force, trends similar to those found with metals have been observed. Increasing the rake angle tends to reduce this force, whereas a larger size of cut (that is, feed and depth of cut) causes an increase in force.

Finally, with thermoset types of plastics, many of which are composites that contain quite abrasive filler materials, tool wear poses a serious problem. In certain instances, diamond cutting tools must be used for successful performance, although ceramics and carbide tools are often adequate.

The thrust of this section is to indicate that although many of the basic concepts related to the successful machining of metals show a useful correspondence when plastics are used, there can be significant differences. As a final comment, this section does not imply that machining is the operation used to *process* polymers in some initial configuration. Both forming operations (Chapter 8) and casting processes (Chapter 11) are used to produce various *shapes* of polymers. As with metals, machining is often used as a *subsequent* process.

REFERENCES

1. L. E. Doyle, C. A. Keyser, J. L. Leach, G. F. Schrader, and M. B. Singer, *Manufacturing Processes and Materials for Engineers*, 3rd ed. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1985.
2. E. P. DeGarmo, *Materials and Processes in Manufacturing*, 3rd ed. New York: The Macmillan Co., 1969.
3. H. W. Yankee, *Manufacturing Processes*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1979.
4. R. A. Lindberg, *Processes and Materials of Manufacture*. Boston: Allyn and Bacon, Inc., 1964.

PROBLEMS

9–1. A hole of 0.5 in. diameter and one inch depth is to be produced by using a drill of that diameter. If the drill is to operate at 100 rpm and a feed rate of 0.008 in. per revolution, determine
(a) The machining time (in minutes) after the full diameter is first reached.
(b) The rate of metal removal when the full diameter is being cut. Answer first in English, then SI, units.

9–2. A hole of 25.4 mm diameter is to be enlarged to 28 mm by boring. The single point boring tool is to operate at a feed rate of 0.076 mm/rev and the piece rotates at 250 rpm. If the length of the hole is 48 mm, find
(a) The machining time.
(b) The rate of metal removal

9–3. A plate having dimensions of 18 in. length, 6 in. width and 2 in. thickness is to be reduced to 1.75 in. thickness by using a shaper. The total length of stroke is set for 20 in., allowing a 1-in. overtravel at each end of the 18-in. dimension. If the *average* cutting velocity is to be 60 ft per min and the feed rate is set for 0.015 in. per stroke, find
(a) The machining time to shape the full surface.
(b) The rate of metal removal in both English and SI units.

9–4. For a particular work material, tool material, tool signature and cutting fluid combination, experiments indicate that for turning,

$$VT^{0.1}f^{0.7}d^{0.4} = 2.0$$

Assume that the desired tool life T is 30 minutes and the depth of cut d is 0.250 in. All parameters carry English units.
(a) If a feed rate is specified as 0.008 ipr, what cutting velocity V should be specified?
(b) What rate of metal removal results in a?
(c) If a feed rate of 0.005 ipr is used instead, what value of V should be used?
(d) What is the RMR for these new conditions?

9–5. Tool life information for a certain set of cutting conditions indicate
(a) $VT^{0.25} = 250$ when $f = 0.010$ ipr and $d = 0.200$ in.
(b) $VT^{0.25}f^{0.6}d^{0.35} = K$

A 5-in. diameter bar of steel is to be reduced to 4.4 in. in a single turning pass, with a feed rate of 0.010 ipr. If a 40-min tool life is desired, what cutting velocity should be used?

9–6. Machining studies for a certain work-tool material combination show that

$V_{60} = 1.6f^{-0.65}d^{-0.25}$ (V_{60} = velocity for a 60-min tool life)
$VT^{0.1} = 90$ when $f = 0.008$ ipr and $d = 0.150$ in.

If an operation will involve a feed of 0.012 ipr and depth of cut of 0.250, what cutting velocity should be used if the tool life is to be 25 minutes?

9–7. In turning, tool life can be described in terms of cutting velocity via $VT^n = C$, whereas the metal removal rate, RMR, comes from $12Vfd$. Develop an expression to express tool life in terms of cubic inches removed per tool failure. This should be expressed explicitly as a function of the parameters V, f, d, and C.

9–8. High-speed steel tools are used to turn steel bars from an outer diameter of two inches to 1.75 in. using a feed rate of 0.020 ipr. After turning two full bars at 300 rpm, a tool fails, whereas at 235 rpm, nine bars are machined before tool failure. The length of each bar is 10 in. Sintered carbide tools are to replace the HSS tools for this operation, and the velocity for a 60-min. tool life (that is, V_{60}) for carbides is 2.5 times V_{60} for HSS. An appropriate tool life equation for carbides is $VT^{0.23} = C$.
If carbides are to turn 50 full bars before failing, what rpm should be used if the same feed rate and depth of cut are maintained?

9–9. Large propeller shafts are to be turned on a lathe having 20 hp (maximum). The tare hp is 1.5 and the efficiency is 85 percent. Estimates of the *maximum* level of cutting conditions in any one combination are

$$f = 0.024 \text{ ipr}, \quad d = 0.250 \text{ in.}, \quad V = 150 \text{ fpm}$$

For the tool-work combination, it is known that,
(a) $F_c = K_1 f^{0.7} d^{0.85}$
(b) u hp $= 1.0$ when $f = 0.008$ ipr and $d = 0.050$ in.
Does the lathe possess adequate horsepower? Show supporting calculations.

9–10. Pertinent drilling information indicates

$$T_o = 4180 f^{0.7} d^{1.9}$$

where T_o is in lb-ft, f in ipr, and d, the drill diameter, in inches.

$$T_h = 6940 f^{0.6} d^{1.0}$$

where T_h is in lbf and f and d as above.

$$VT^{0.15} f^{0.65} d^{0.95} = 4$$

where V is in ft/min, T is in min, and f and d are as above.

A drill press to be used possesses a drive motor of 7.5 hp, an efficiency of 90 percent, and negligible tare hp. The operation will employ a ½-in.-diameter drill, and the desired tool life is 45 min. Two feeds, 0.010 and 0.020 ipr, are to be checked in the following analyses.
(a) Including *all* contributions to power demands, is adequate horsepower available for both feed rates?

(b) Regardless of your answer in part a, which combination of f, d, and V gives the highest RMR *if* adequate power were available?

9–11. Milling is to be used to reduce a plate from 305 mm length by 152 mm width by 75 mm height to 305 mm by 152 mm by 70 mm in a single pass. The 250-mm-diameter cutter has 20 teeth, and the *average* feed rate is to be 0.200 mm/tooth. A reasonable cutting velocity is about 0.5 m/s. If the unit horsepower for this operation is 0.9 hp/in.³/min, determine the magnitude of the cutting force.

9–12. A company must purchase a new lathe for turning large drive shafts. Both roughing and finishing cuts will be performed using the following conditions:

Finish turning: $f = 0.010$ ipr, $d = 0.100$ in., $V = 300$ fpm
Rough turning: $f = 0.030$ ipr, $d = 0.350$ in., $V = 100$ fpm

When the same material was machined on an existing lathe, the motor indicated the following horsepower demands:
(a) $hp_m = 1.5$ when $f = 0.010$ ipr, $d = 0.050$ in., $V = 150$ fpm
(b) $hp_m = 2.5$ when $f = 0.010$ ipr, $d = 0.100$ in., $V = 150$ fpm
(c) $hp_m = 4.0$ when $f = 0.020$ ipr, $d = 0.100$ in., $V = 150$ fpm
During these three tests a tare horsepower of 0.5 and machine efficiency of 90 percent were noted. It is also known that

$$VT^{0.2} = 400 \quad \text{and} \quad VT^{0.2} f^{0.55} d^{0.2} = K$$

pertain for this situation.
The manufacturer of the lathe to be purchased estimates that the tare hp will be about 5 while the efficiency should be around 95 percent. What size or capacity motor (that is, horsepower) should be specified?

10 joining processes

10.1 INTRODUCTION

The topics discussed in this chapter include the chemistry and mechanics of the following processes:

1. Welding and other processes that join parts by localized melting of the parent and filler metals.
2. Brazing and soldering, processes which bond solids by filling the joint with molten filler material, without melting the parent metal.
3. Cold bonding.
4. Diffusion bonding, sintering, and coalescence of ceramic, metal, and plastic powder particles.
5. Gluing and adhesive bonding.

Joining and bonding technologies are too numerous, too complex, and too varied to adequately describe in a general textbook. Furthermore, new developments are appearing at a great rate, so any short description of the subject would soon be out of date. Much information of the type needed by manufacturing engineers, namely, the cost effectiveness of these processes, the skills required to operate the machinery, and the production rate of the machinery, can be found in publications of the American Society for Metals and the Society of Manufacturing Engineering, as well as in the sales brochures and catalogs of equipment suppliers. The necessary arts and skills for training purposes are described in the well-illustrated instruction manuals and textbooks of industrial arts schools.

But reliable information on the *chemistry* and *mechanics* of many of the processes and on the mechanical strength of the product is generally not available. Most new developments of old processes and most new processes contain some proprietary or *secret* aspects, and many products are advertised with exaggerated claims of superiority over other processes, either in terms of low cost, ease of processing, or high product quality. A common practice in industry, in order to learn the real capability of processes, is to send parts to several purveyors of *new* processes for evaluation purposes. Should one or two of the processes be found useful, the next step is to determine whether the technology of the candidate processes can be successfully applied where needed. This depends on the expertise and resources of the vendors, which can best be estimated by asking questions on the fundamentals of joining and coating. Thus the emphasis in this chapter will be on the atomic aspects of bonding or joining of separate bodies, followed by a discussion of the problems and properties of the bonded system. With this approach all processes, old and new, may be better understood.

10.2 JOINING PROCESSES: AN OVERVIEW

Few products are made of a single piece; rather, most products are made of several pieces, which are then joined together in some way. One class of joining, which is not discussed in this book, is by the use of *mechanical* devices such as bolts or screw threads, rivets, stakes, bands, clips, cast-in inserts, press-fits, and so forth. The most important disadvantage of mechanical joints is their high cost, since the surfaces to be joined must usually be made fairly precisely, holes must often be provided, and several operations are required to effect the joining operation. The major advantage of mechanical bonding is that the strength of the joint can be *predicted* quite reliably through the use of design handbooks or handbooks of mechanical fasteners. Fewer data are available for the joining processes discussed below, mostly because the strength of the joint is strongly dependent on how the process was done. Thus, where reliability of a small production run of products is more important than economy, mechanical joining will often be specified.

The economic advantage of *adhesive* and *cohesive* bonded joints derive from the fact that imprecise surfaces can be used, and the processes are fast. (*Adhesion* is defined as bonding between dissimilar materials, and *cohesion* is defined as bonding between pieces of identical materials.) Their major disadvantage is that the joined parts are not readily separable for maintenance or other use. Think how awkward photography would be if the back of a camera had to be unsoldered and resoldered whenever a film cartridge is to be changed. However, for many other products, permanent joints are acceptable or even preferred. Examples are floor covering, pages in books, auto body parts, and electrical connectors.

Bonding processes themselves can introduce problems which require special care to minimize. For example, polymeric adhesive compounds produce stresses when they contract at a different rate and amount than does the substrate. This will produce distortion of the finished part, particularly if the cross section is thin. Welding processes also produce distortion and leave residual stresses, because they progressively melt and solidify small amounts of the part interfaces as the heat source moves along. All heat sources for the welding processes cause these problems; these include the electron and ion beams, light beams including lasers, electrical resistance, electric arc, plasmas, and hot gases. A second problem is entrained gases. For example, if an adhesive supply contains excessive amounts of gases and solvents from the mixing and formulation processes, these may produce bubbles and other voids in the solidified joint. Fusion welding processes that operate in air or under water also entrain oxygen and water vapor into the molten metal, thereby forming oxides and other defects.

A third problem with bonding is seen in welding, where there is often the problem of undesirable metallurgical phase change in the parent metal. There is always a region extending down into the substrate beneath the weld, known as the heat affected zone (HAZ). The temperatures in the HAZ will over age and soften a hardened aluminum, and the fast cooling that follows heating may be fast enough to produce brittle martensite in alloyed steel. Many welded joints will fail in the HAZ rather than *at the original interface.*

There are, therefore, two aspects of the integrity of a bonded joint, namely, the nature of the bond interface or thin region between the two separate parts, and the properties of the near surface region within the bodies near the joint. Both will now be discussed.

10.3 THE NATURE OF INTERFACE BONDS

One fact of nature is that if two atoms are brought close enough to each other, they will bond together. The spacing between atoms in the solid state and the strength of the bond depends on the electron structure of the atoms. For the primary solid bonding systems, namely, the ionic, the covalent, and the metallic systems, the operative distance is about 0.3 nm and the strength is very high, in the range of 1 to 3 GPa. For the vanderWaals and dispersion forces the distance is somewhat greater and the bond is weaker, of the order of 0.1 to 0.2 GPa.

If any two atoms will readily bond together, then two clean and flat solids, each being composed of atoms, will also bond together by mere *contact,* that is, without heat of pressure. This does not accord with common experience, however. Two *clean* coins, for example, are readily separated even though they are vigorously rubbed together. But the same two coins, when thoroughly cleaned in the vacuum of space, will bond or weld together by mere contact. It should be apparent that there is a difference between cleanness as commonly defined and the cleanness required to effect strong bonding. A perspective on cleanness will be given after the further consideration of bonding energy.

It is useful to view the bonding between atoms in terms of energy. For example, in order to form a crack in a solid, that is, in order to separate atomic bonds, external energy must be applied. This amount of energy (ignoring plastic flow at the crack tip) is the order of one MJ per square meter of new surface as measured by the methods of fracture toughness. Healing of such a crack involves the liberation of the same amount of energy, either as mechanical energy or as thermal energy.

Another method of accounting for energy of bonding is to measure the energy of adsorption of gases on solids. This begins by grinding a brittle material into a fine powder in a vacuum; the fine powder has a large ratio of surface area to volume. When a gas is admitted into the enclosure, the surfaces become covered with one or several layers of gas, and the temperature of the powder increases. The source of the heating is the high energy state of the free surface in the vacuum. When the gas adsorbs on the surface, the energy state of the gas and solid surface combination has a lower potential for bonding to yet other gases than did the solid surface alone. The energy difference is *given off* as heat, and this heat of adsorption is unique to each combination of gas and solid surface. For some combinations the gas can be driven off the solid by heating in a vacuum (note that most "vacuum" is only a reduced gas pressure, not the complete absence of gas). This is called *physical adsorption,* for which the heat of reaction is in the range of 0.05 to 0.5 eV (or from 4 to 40 kJ) per mol of the adsorbed gas (see Ref. [1]).

Very few surfaces remain truly clean, even if they can ever be made clean in some way. If a square meter of surface begins in the perfectly clean state in air at atmospheric pressure, it will be bombarded by gas molecules at a rate $\varnothing_0$ that can be estimated from the equation (Pirani):

$$\varnothing_0 = \frac{4.7 \times 10^{28} P}{\sqrt{(MT)}} \text{ molecules per second} \qquad (10\text{–}1)$$

where P = pressure in Pa (1 atm. $\approx 10^5$Pa)
M = molecular weight (big molecules move slowly)
T = temperature in degrees Kelvin

This is a very fast rate as compared with the fact that a square meter of surface can hold only about 6.2×10^{18} molecules of nitrogen. Actually the entire surface does not become covered at once. If a molecule strikes a surface but only remains attached at a site not previously occupied, then the rate of surface coverage at any instant is

$$\varnothing = \varnothing_0(1 - \varnothing) \qquad (10\text{–}2)$$

where $\varnothing$ is the fraction of uncovered surface. $\varnothing$ also equals N/N_0, where N is the number of molecules that had previously settled on the surface and N_0 is the maximum number that can occupy a unit of area. Then $\varnothing = (dN/dt) = N_0(d\varnothing/dt)$. Substitution yields $\varnothing_0(1 - \varnothing) = N_0(d\varnothing/dt)$, for which the solution is

$$\text{In } (1 - \varnothing) = -(\varnothing_0/N_0)t \qquad (10\text{–}3)$$

Calculations for nitrogen at 250°F (120°C) in a vacuum of 10^{-6} Torr. (1.33×10^{-4} Pa) show that the surface is covered progressively over several seconds of time, as shown in Table 10–1. These data have been verified in adhesion tests in vacuum. Atmospheric pressure is about 7.5×10^8 as dense as the pressure used in the calculation, and thus the rate of molecular bombardment of gases on a surface would be 7.5×10^8 as great were it not for fact that a molecule cannot travel a distance equal to its mean free path without colliding with others in a dense atmosphere. We can only estimate, therefore, that the rate

TABLE 10–1 RATE OF COVERAGE OF A SOLID SURFACE BY ADSORBING GAS

% of Surface Covered	Time, Seconds	
	in vacuum of 1.33×10^{-4}Pa	at atmospheric pressure
25	.8	multiply the previous column by 10^{-8}
50	1.7	
75	3.5	
90	6.0	
95	7.5	
99	12.0	

of coverage of a surface in atmosphere is about 10^8 that shown in the center column of Table 10–1. These are very short times, hardly perceived in ordinary experience.

The above calculation applies to the first adsorbed layer (excluding oxide layers for purposes of discussion) of gases, including water vapor. Subsequent layers condense more slowly, depending on the temperature and humidity. Once covered, the surfaces are not in a sufficiently high state of energy to bond firmly to other solid atoms. This layer of gas, which is about 80 percent as dense as liquid, is then the first fundamental problem to overcome in the joining of solid surfaces together.

Many adsorbed gases react chemically with the substrate to form a new compound, the most common of which is a metal oxide. These chemical reactions, which occur after physical adsorption, usually liberate between 10 and 100 Kcal/mol, and more. The total reaction is referred to as *chemical adsorption,* or *chemisorption,* and it is not reversible. Because the energy of chemisorption is very much higher than for physical adsorption, an oxide usually forms quickly in air. But the rate of oxidation is limited by the diffusion of ions and vacancies through the oxide film. The adsorption of additional gases upon the oxide film takes place at a higher rate than that gas can be diffused into the oxide. Thus the oxides are also covered with physically adsorbed layers of gas. Both layers are between 2 to 10 nm thick on technologically clean surfaces (see Ref. [2]).

Although the adsorbed films form quickly, there is a limit to their mechanical durability. Rubbing two solids together displaces small regions of the films and allows some microscopic solid bonding. In fact, the measurable resistance to sliding, called *friction,* involves some solid bonding. However, though locally strong, there are very few bond sites, and the total joint strength is hardly measurable in directions normal to surface. The adsorbed gases are one reason, but another is the nonconformity of two touching surfaces. This uneven topography usually consists of several dimensional scales, ranging from the microscopic to the visible or macroscopic level. The smaller scale protruberances or *asperities* produce *roughness,* whereas the larger-scale features are called *waviness.*

The consequence of roughness and/or waviness is that contact between surfaces is

made on only a few widely separated *points,* which in reality are very small areas. As combinations of normal and shear stresses are applied to these small areas by external loading, without sliding, the stress state rises, plastic flow begins, and the sizes of the small contact areas increase. When the external loading is decreased, the stored elastic energy beneath the contact areas produces an elastic recovery to a less deformed state, resulting in a decrease in the size of the small areas of contact. All of this usually occurs on the scale of the asperities.

During sliding some of the adsorbed film will be displaced from the contact regions into the *valleys* between. They remain in the system, limiting the amount of surface that can be *cleaned* and bonded by rubbing. But even in those regions that were bonded, much of the bonding would be fractured by elastic recovery when the stresses on the contacting region are reduced, leaving a scarcely measurable adhesion.

We can estimate the amount of technological surface in actual contact if we assume that the contact pressure in isolated asperity contact areas is about three times the yield strength of the materials in the asperities. If one cube of lime-soda glass (density of 2.2 gm/cc, YS of 150 MPa) were placed upon another, each of dimension 100 mm, the apparent area of contact would be 10^4 mm^2, but the accumulated minute contact area would be only 2.2 mm^2, a factor of 4500 less. Sliding of one glass surface over another will produce fracturing, but with metals the asperities will plastically deform into conformity with each other to increase the bond area.

In vacuum, practical weld strengths can be achieved by sliding, and such a process is known as *cold welding.* In air some truly cold (ambient temperature) welding can be achieved by squeezing two parts together so that the surfaces expand to expose areas upon which there is little adsorbed gas. A more practical process is to rub two surfaces together repeatedly or continuously until the majority of the surfaces have been plastically deformed. This reduces the effects of local elastic recovery. In addition, if the rubbing heats the surfaces, some physically adsorbed gases are removed and the yield strength of the materials is reduced, thereby enhancing the possibility of plastic flow. This process is often called *friction welding.*

So far the discussion has focused on the bonding of solids. Liquids also bond or attach to solids. This phenomenon is called *wetting.* As in the case of solid bonding, there are degrees or strengths of wetting, but these quantities are not expressed in units of stress. One way to express the tendency for a drop of liquid to spread over, or wet, a flat horizontal solid surface is in terms of a contact angle (after T. Young). For each combination of liquid and solid there is a specific contact angle β. Figure 10–1 shows how the contact angle is measured. Table 10–2 gives the values of β for four fluids on glass. Water with detergent is seen to wet glass better than does tap water, but the alcohol wets completely. In the same way a glue or adhesive compound may wet surfaces to different degrees, with consequences on the quality of the bonded joint; molten aluminum may be deposited on a solid surface of aluminum, but it may not attach well. In the latter case the

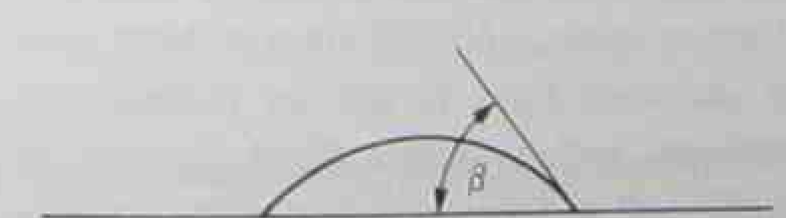


Figure 10–1 The measurement of contact angle.

TABLE 10–2 THE CONTACT ANGLE FOR FOUR FLUIDS

Liquid	β
tap water	110°
water with detergent	80°
furfural	30°
isopropyl alcohol	<1°

oxide on the solid aluminum impedes wetting of liquid metal to the solid with the effect of preventing good welding. With time, and in inert atmosphere, the liquid-covered oxide dissolves away, or with sufficient heating the oxide will melt and float away so that good wetting may occur. Thus wetting is a dynamic event.

In summary, the state of matter on earth is such that separate solid bodies do not readily bond together. This is fortunate because if human bodies were to stick to whatever they touch, all matter would soon become one coherent (or incoherent) mass. Joining and coating therefore require an intervention in the natural order, by introducing some special effort. These special efforts in the form of manufacturing processes will now be discussed.

10.4 WELDING

Fusion welding begins by placing two solid parts near to each other. Recall that two surfaces are never able to contact each other at every point though clamped tightly together. But in practice large gaps (of the order of millimeters) often exist between surfaces. For fusion welding, the acceptable range of gap size is about one-fourth to one-half of the diameter of the *welding rod* used, which in turn is usually no larger than the thickness of the thinnest member being welded. These ranges are wide and depend upon the circumstance and the experience of the people doing the work.

Welding proceeds by melting a small volume of metal at the surface of each part. If these two pools of molten metal can be contained, that is, prevented from falling away, they may mix together. If they then solidify, a solid bridge between the two parts has been effected, and the parts have been joined or welded together. Very localized fusing of two parts together is called *tack welding* and is sometimes done in preparation for a final weld. The final weld will usually, but not always, require the addition of *filler material* because of the gap that usually exists between surfaces to be joined. Filler materials can be of the same composition as the metal parts, or perhaps a material that will add some desirable property to the joint.

Welding methods are often classified by the heat source used to melt metal; there are five major heat sources. One is a flame, which is achieved by the oxidation of the carbon and hydrogen in fuel gases. Carbon releases 60,900 J/gm of heat, and hydrogen releases 254,600 J/gm, but these are awkward fuels to use by themselves. Combinations or compounds of carbon and hydrogen, such as butane (C_4H_{10}), propane (C_3H_8), and acetylene (C_2H_2) are easier to handle and will produce hot enough flames for melting most

materials. The hottest flame is achieved by using oxygen with acetylene, which produces 27,720 J/gm. This flame, if adjusted properly, will just melt a small piece of tungsten (MP = 3410 ± 20°C). If acetylene or any other gas is burned with air, the flame temperatures are a few hundred degrees cooler than when burning oxygen, because of the need to heat up the large fraction of useless nitrogen in air. It should be noted that the combustion of fuel gases produces CO_2 and H_2O as well as complexes of nitrogen or other gases in the vicinity of the flame, some of which become entrained into the molten weld metal.

Another common heat source is the *electric arc*. An arc is a high-density flow of electrons, across a voltage, from a cathode to the anode, with a flow of ions in the opposite direction. Welding is done at voltages between 30 and 75 volts with arc lengths of about 2 to 10 mm. The technology of arc welding, incidentally, also includes getting the arc started. An arc would not spontaneously begin by applying 75 volts over a gap of 2 mm. An arc is therefore "struck" by momentarily touching the cathode to the anode, which induces a flow of electrons. After electron flow begins a useful arc can be maintained, and the arc may *carry* several hundred amperes.

The actual temperature of the arc is a fictitious property, but its effect may be estimated in terms of a temperature. An arc in air has a *surface temperature* of the order of 10,000°K, while one drawn between iron electrodes, as in the majority of welding, has a *surface temperature* of about 6,000°K, because the iron vapor has higher electrical conductivity than does air. (The arc loses about 20 percent of its heat by radiation when welding is done with a short arc and a current of 100 A.) A metal arc is therefore sufficient to melt all materials by radiation and convection alone, but electron bombardment of the anode also heats the surface. The cathode is correspondingly cooled by evaporation of electrons. In arc welding the workpiece is usually the anode, upon which the impinging power density is about 10^7 to 10^8 W/m^2, and the filler material becomes the cathode. Industrially, this arrangement is called *straight polarity*. In some cases, such as when thin sheet is welded, *reverse polarity* is used to prevent melting the sheet before the filler material is melted.

In and around the arc there is a large magnetic field which causes a flow of gases and molten metal to attain velocities of the order of hundreds of meters a second, generally from the cathode to the anode. Some of the metal will be heated to the point of evaporation, which, because of expansion, produces a net flow of material out of the arc. Up to 10 percent of the metal may be lost in this way, and it condenses as dust on everything in the vicinity of the welding operation. In addition, there is a *spatter* of liquid metal out of the arc. Between 10 and 20 percent of the filler metal is lost from the weld by spatter, and this metal is distributed around the weld as moderately firmly attached spheres.

The evaporation of metals proceed somewhat selectively according to the boiling point of the metal. A table of melting points and boiling points of various metals is given in Table 10–3. The end result of selective boiling of metal is to change the composition of the weld.

In practice, arc welding involves a number of choices in order to achieve the desired joint properties most reliably. These include:

TABLE 10–3 THE MELTING POINTS AND BOILING POINTS OF SOME ELEMENTS AND OXIDES AT ATMOSPHERIC PRESSURE

Element	Density g/cc	MP°C	BP°C	Th.cond.	Oxide	ΔH, cal/m	MP°C	BP°C
Ag (silver)	10.50	961.9	2212	4.29(W/cmK)	Ag_2°	−7,740	230	
Al (aluminum)	2.70	660.4	2467	2.36	Al_2O_3	−404,080	2072	2980
Au (gold)	19.32	1064	2807	3.19				
Be (beryllium)	1.85	1278	2970	2.18				
B (boron)	2.34	2079	2550(s*)	0.32				
Cd (cadmium)	8.65	321	765	≈0.9				
C (carbon)	1.8–2.3	3550	3367(s*)	0.01 − 26				
(diamond)	3.15–3.53							
Cr (chromium)	7.19	1857	2672	0.97	Cr_2O_3	−274,670	2266	4000
Co (cobalt)	8.9	1495	2870	1.05				
Cu (copper)	8.96	1083	2567	4.03	CuO	−37,740	1326	
					Cu_2O	−40,550	1235	1800
Fe (iron)	7.87	1535	2750	0.87	FeO	−65,320	1369	
					Fe_3O_4	−268,310	1594	
					Fe_2O_3	−200,000	1565	
Mg (magnesium)	1.74	648.4	1090	1.57	MgO	−144,090	2852	3600
Mn (manganese)	7.3	1244	1962	0.08				
Mo (molybdenum)	10.22	2617	4612	1.39				
Ni (nickel)	8.9	1453	2732	0.94	NiO	−57,640	1984	
Pb (lead)	11.35	327.5	1740	0.36	PbO	−52,800	886	
Si (silicon)	2.33	1410	2355	1.68	SiO_2	−210,070	1723	2230
Sn (tin)	5.75	231.97	2270	0.5–0.7	SnO_2	−68,600	1630	1800(s*)
Ta (tantalum)	16.65	2996	5425	0.57	Ta_2O_5	−492,790	1872	
Ti (titanium)	4.54	1660	3278	0.23	TiO_2	−125,010	1825	
V (vanadium)	6.11	1890	3380	0.31				
W (tungsten)	19.3	3410	5660	1.77	WO_3	−201,180	1473	
Zn (zinc)	7.13	419.6	907	1.17	ZnO	−84,670	1975	
Zr (zirconium)	6.5	1852	4377	0.23	ZrO_2	−262,980	2715	

*sublimes

1. Selecting the method by which the weld joint will be protected from the effects of entrained gas bubbles and inclusions, and the system for preventing gases and inclusions in the weld.
2. Selecting a type of filler material.
3. Setting the velocity of welding.
4. Setting the voltage and the limiting current provided by the power supply for the operation.

Protection against gas entrainment may be done either by keeping harmful gases out of the weld region, or by removing whatever gases and products of reaction enter the weld metal. The first may be done with the use of an inert gas such as argon. (Nitrogen is not sufficiently inert in the welding environment, and the other noble gases are very expensive.) The flow rate of inert gas must be controlled to keep harmful gases away without

carrying excessive amounts of heat away. In some cases a small amount of readily ionizable gas must accompany the inert gas to control ionization of gas in the plasma or arc. Oxygen is one such gas, and it *stabilizes* the arc. A stable arc is one in which the region on the cathode, from which electrons depart, and the region on the anode, at which they bombard, do not dance about wildly, making welding difficult or impossible to do. But primarily, the inert gas minimizes the amount of oxides and nitrides that can form in the weld metal.

The second method of control is to feed a solid substance, containing several complex compounds, into the arc area. These compounds are called *fluxes*, and they provide some combination of inert gases when heated, some atomic elements to lower the electrical resistance of the gases in the arc, and some atomic elements to add desirable properties to the weld joint. Such compounds are often contained in a coating on the *electrode* or filler material for hand-welding operations. In the older automatic filler-wire feed machinery, the fluxes are in a granular form that drop around and cover the moving arc. The arc is submerged under the flux, and the method is appropriately referred to as *submerged arc welding*. Granules of flux may be applied overhead or to vertical surfaces by blowing them into the arc, but it is expensive. Newer automatic machinery uses hollow filler-wires or tubes, filled with flux, which are more useful for all positions of welding.

Filler materials may be steel, brass, aluminum, or other materials. It is not necessary to use a filler material of the same composition as the pieces to be joined. Joint composition may be adjusted by adding alloys to the fluxes as well as to the filler metal. Thus for welding steels and irons, a low-cost filler material may be used. Chromium may be added to the flux to impart some corrosion resistance to the weld; nickel may be added to retain ductility of the weld joint at low temperature, and particles of tungsten carbide (WC) may be added to impart wear resistance to the weld.

Engineers in manufacturing are usually concerned with both the cost to weld and the final properties of the weld. For both purposes the arc voltage, arc current, and velocity of arc movement must be selected. High arc power (voltage × current) heats and melts large quantities of metal in a short time, and high velocity distributes the power over a large area. There are practical limits imposed on velocity by the ability of people or machinery to properly direct the path of the weld operation, but there are no real limits on the arc power.

Joint strength is achieved by control of *weld penetration* and to a lesser extent by control of *bead height*. Both of these terms can be defined with the aid of Fig. 10–2. The penetration is the depth to which the plates shown in Fig. 10–2 are fused together. To achieve this penetration a section of each plate is melted, and the size of the molten pool is determined by the rate of heat input versus the rate of heat conduction into the plates. The heat enters at the surface, by radiation and convection, and also by electron bombardment if the plates are the anode or positive pole.

Theoretical equations relating penetration to the arc power, arc length, velocity of welding, and polarity of the arc are very cumbersome. It is thus more common to develop empirical equations by measuring the penetration that results from exposing a surface to a moving arc of varying powers. The results will vary over a wide range, depending on the polarity of the arc if it has DC applied, and whether the arc is a metal arc or one drawn

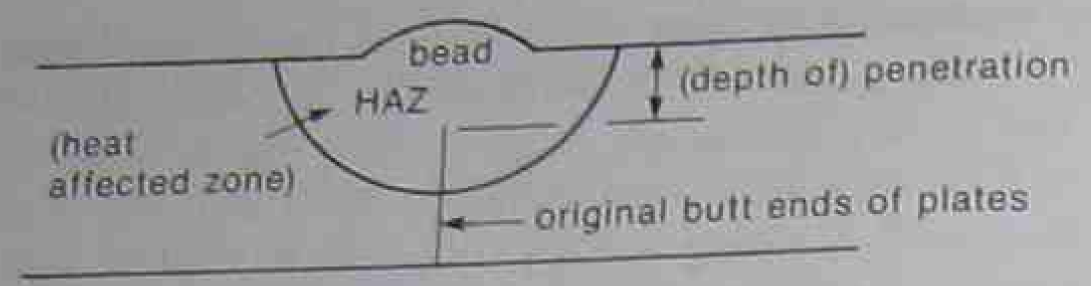


Figure 10-2 Sketch of the cross section of a weld, joining two thick plates.

by a tungsten or carbon electrode. The common form of such equations is

$$p = k\frac{E^a I^b}{v^c} \tag{10-4}$$

where E is the voltage across the arc during welding (usually in the range of 25 to 75 volts, *not* the open circuit voltage), I is the current in amperes through the arc (usually in the range of 75 to 1000 amperes, not the limiting current set on the machine), and v is the welding speed or velocity (usually in the range of ⅓ to 5 meters per minute). The exponents and the constant k are characteristics of the metal as well as the process, and the functions of the thickness of the plates being welded, the diameter of the filler rod, the atomic elements passing through the arc, the arc length, and other parameters. Typically for welding steel at moderately high speeds, where p is to be given in mm, $k \approx 10^{-4}$, $a \approx 0.5$, $b \approx 1.5$, and $c \approx 0.33$.

Precise values for careful prediction of welding costs should be obtained from handbooks on welding or should be obtained by tests using the actual production equipment. The performance of various combinations of equipment and welding materials vary as much as 5 to 1. Modern production equipment for arc welding no longer depends on the characteristics of generators and transformers. Rather, the equipment can produce virtually any desired electrical waveform, such as pulses of various voltages, polarities, and durations, depending on the need. Each waveform will have different equations for penetration and bead buildup. The manufacturer of the equipment and vendors of welding supplies have some data of this type available.

The bead height is as difficult to predict as the penetration. Again, these should be empirically determined if this dimension is important. Generally, the bead height is about one-fourth of the penetration, for metal arc welding with straight polarity. For reverse polarity the bead height approaches one-half of the penetration, and for welding with an alternating voltage the bead height is halfway between one-fourth and one-half.

Three other heat sources are available beside the electric arc and gas flame: these are the *plasma-arc*, the *laser*, and the *electron beam*. The plasma-arc system passes gas mixtures (for example, 75% A_2 + 25% H_2 or other combinations with helium or nitrogen) through or near an arc, where the gas is ionized. The gas absorbs most of the energy in the arc. This stream of ionized gas can be directed toward any cool surface, including a nonconducting material, where the ionized gas evolves heat and reverts to molecular gas.

Lasers produce electromagnetic radiation which is coherent and therefore can be propagated as nearly parallel beams. Radiation that impinges on a solid surface is re-

flected, absorbed, and transmitted in various proportions according to the *optical* properties of the target substance and the wavelength of the radiation. A common type of laser is the helium-neon laser, which produces red light of 632.8 nm wavelength, and another is the argon laser, which produces blue light of 488 nm and green light of 525 nm. These do not heat surfaces very efficiently. Infrared (invisible) radiation is better in that it is mostly absorbed by metals, which causes heating. The CO_2 laser, which produces infrared radiation in the range of 1000 nm or 1 μm, is the best commercial type for heating most solid surfaces.

The electron *beam* is noncoherent radiation of wavelengths in the range of 0.01 to 10 nm. Because of its short wavelength, over 98 percent of the radiation is absorbed when the electrons *bombard* an anode. Electrons are emitted from a large surface of heated thorium-tungsten or other source. They are directed toward another surface, and given energy, by applying several thousand volts between the source, which is the cathode, and the target, which is the anode. (Higher voltages produce X-rays.) The beams can be focused by electric or magnetic fields upon a small region on the work surface.

It is difficult to rate these sources relative to the others in terms of their temperatures, but figures are available on the maximum practical power density available on the surfaces upon which these heat sources impinge. For the plasma it is between 10^9 and 10^{10} W/m^2; for the CO_2 laser it is between 10^{10} and 10^{11} W/m^2, and for the electron beam it is between 10^{11} and 10^{12} W/m^2. The rate of heating for each of these sources can be increased at will, but a limit is reached when the molten pool is yet very shallow when the surface of the metal evaporates.

The plasma is a stream of ions, which revert to molecular gases when cooled by the target surface. The gases can be inert, which expel oxidizing gases, but unfortunately they can also become entrained into a weld and produce bubbles. Both the electron beam and the laser beam are "clean," and both beams can be shaped and focused precisely and quickly, but both are expensive. The laser is very *inefficient* in the use of energy, using only about six percent of the power taken from the electrical source. The electron beam is about 75 percent efficient, but it must operate in a vacuum of about 10^{-3} to 10^{-4} Torr. (0.1 to 0.01 Pa). The ordinary arc welder is also about 75 percent efficient in terms of the use of power to effect deposition of the weld metal. Pulsed arc welders are somewhat more efficient, whereas the plasma arc system is somewhat less efficient.

In principle, if the edges of two parts are melted progressively and allowed to fuse together, eventually the two parts should be joined. However, a major problem arises because of the progressive nature of solidification. This may be seen by holding a bar by the ends in a very stiff vise, as shown in Fig. 10-3. If the rod is heated, with no heat

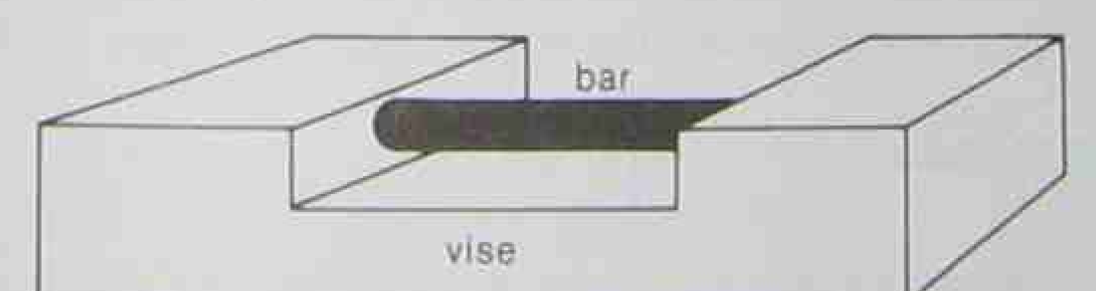


Figure 10-3 Sketch of a bar held in a vise.

transfer to the vise, the rod will expand by an amount equal to $e = \propto \Delta T$, where $\propto$ is the thermal coefficient of expansion and ΔT is the temperature rise. If the vise resists this expansion completely, as if it were rigid, the rod will not change in length. Rather, its expansion would be exactly countered by a compressive strain, and the stress in the rod would rise to

$$s = eE = E\propto \Delta T \tag{10-5}$$

The linear increase in stress as a function of temperature rise is shown graphically in Figure 10–4 as line segment *a*. If the bar is cooled from any value of ΔT along line *a*, the stress in the bar will return to zero. But if heating continues to a temperature at which the yield strength of the material in the rod is reducing, the change in stress then follows a different course, as shown by line segment *b*. In this region the bar is still expanding, but it is also being *plastically* compressed. If now the temperature decreases after plastic flow occurs, the stresses will again decrease linearly with temperature, but will follow curve *c* in Fig. 10–4. If the rod were originally loosely inserted into the vise, it will separate from the vise at the temperature at which the stress becomes zero, and perhaps fall out. If the rod had been bonded to the surface of the vise, further cooling would develop a tensile stress at the interface between the rod and the vise as shown by line *d* in Fig. 10–4.

The rod in the vise is a useful analogy for a nugget of molten metal in a weld area if the rod is heated to the melting point. The stresses in the rod are seen to pass through a maximum and decrease to zero at MP. Solidification and cooling now involve contraction only and the (attached) rod is left in a state of tensile stress, as shown by line *e*. These stresses cannot increase linearly because temperature affects the flow stress of the material. Thus line *e* is curved as shown. The stress state now remains or *resides* in the rod, and it is referred to as a *residual stress*. The magnitude of the residual stress, in the case of a rigid vise, is seen to be dependent on the amount of differential heating and on the properties of the material of which the rod is made. In the more practical elastic case, the vise would have strained (open) so as to reduce the stresses in the rod upon heating. In a weld there is no sharp dividing line between the heated and upheated regions. Thus it is best to simply recognize that the residual stresses arise from differential strains that arise from local heating, as well as from localized phase changes and from local plastic flow due to mechanical applied stresses.

There are two practical considerations that arise from the inevitability of localized

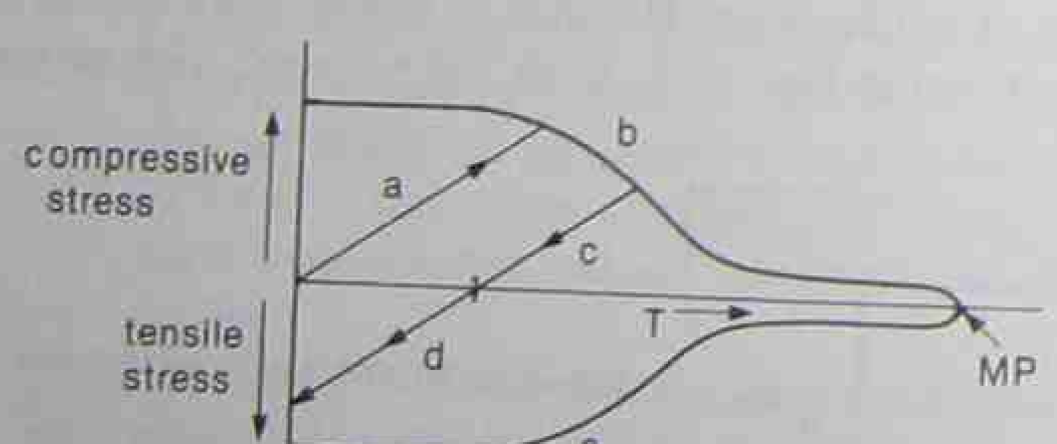


Figure 10–4 Stresses in a bar that is held in a vise and heated.

and differential strains due to conventional welding. The first is that these strains can exceed the fracture strain of such brittle materials as cast iron, primary martensite, and ceramic materials. These materials are then likely to crack during cooling of the weld, wherever they are located, in the HAZ, outside of the HAZ, or in the bead. There are several ways to prevent such cracking. These include:

1. Weld with a filler material that has a yield strength significantly lower than the fracture strength of the joined materials.
2. Heat a large area around the joint before welding so that the stress gradients from the weld region outward are small, that is, heat the vise in the above example.
3. In the case of steel, heat the area around the weld above the M_s (martensite start) temperature to prevent the formation of the brittle martensite when the weld cools (see Chapter 7).

The second problem is that welded parts and structures inevitably become distorted during welding, and distort again when machined or heated after welding. Figure 10–5(a) shows an example of two plates, perpendicular to each other, and welded in the corner. A molten bead will have been placed in the corner and then cooled. During cooling it contracts. If the plates are held rigidly, the finished and cooled weld will be in a state of tensile residual stresses. If the plates are not held in place, the solidifying bead will alter the angle between the plates by as much as 2 to 4 deg. In fact, if the rigidly held plates were released after welding the angle would again change, but by a smaller amount. One solution to the distortion problem is to plastically bend the joint to an angle of 92 to 94 deg and release it. Elastic recovery may return the angle to 90 deg. A better solution is to set the unwelded parts to an angle of 92 to 94 deg, and then weld without restraint. Contraction of the bead will pull the parts to an angle near 90 deg. The proper angles for setting the parts can be calculated if enough of the relevant details of the welding process are known, but it is more quickly and accurately done experimentally.

The above example shows distortion across the *width* of the weld. There is also distortion due to tensile residual stresses along the length of the weld. This can be seen by passing an arc along the edge of a steel bar. The result will be a curved bar, as shown in Fig. 10–6. The radius of curvature can be estimated by assuming that the arc has melted the top layer of the bar of height *t*, to a depth *of penetration p*, and that the molten layer has contracted when it solidified and cooled. If the bar were held straight during this contraction, then the top layer would have a tensile stress in it, and the lower part of the bar would have a compressive stress so that the forces in the entire bar would be equal. Calculation of thermal contraction and the observation that welding with brittle filler materials causes cracking in the weld beads suggests that the strain in the top layer is in

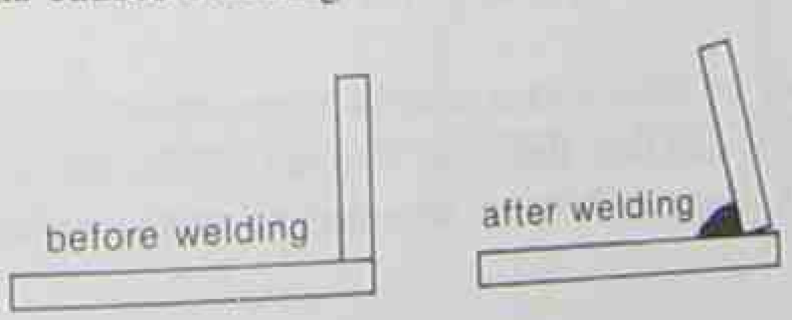


Figure 10–5 Two plates welded at an angle.

Figure 10–6 A bar, bowed by welding all along the top edge.

the range of 0.001 to 0.003. Thus for convenience, let us assume that the tensile stress in the top layer is equal to the yield strength S_y of the metal, acting at a distance $p/2$ from the top face. The force exerted by the top layer, per unit width of the bar is $F_t = S_y p$. A moment M of magnitude $S_y p$ times the moment arm, $t/2 - p/2$, is required to hold the bar straight. Thus, $M = S_y p(t - p)/2$. If the external moment is removed, the bar will bend to a radius R, which may be calculated from $R = EI/M$. I, per unit width of the bar, is $t^3/12$ so that

$$R = \frac{Et^3}{6S_y p(t - p)} \qquad (10\text{–}6)$$

It may be seen that where p is zero, and where $p = t$, R is infinite; that is, the bar is straight. R is minimum when $t = p/2$, or when the depth of penetration is half the height of the bar. For this condition, and for a material where $E/S_y \approx 500$, $R \approx 333t$. Thus for a bar of $t = 1$ cm, $R \approx 3.33$ m. By this method one could make a hoop of steel 6.66 m in diameter. The same can be done by localized heating above the recrystallization temperature without melting. By this method large beams may be made curved for bridge decks, and by this method a moderately bent beam may be straightened.

In the design of welds there should be a strategy for accommodating the distortions that occur. As mentioned above, distortion can be minimized by beginning with parts that are oriented with a negative distortion. Welding will simply straighten the parts, but leave residual stresses in them. If now material is removed from one face of the welded part, for example, by grinding, the residual stress distribution will be altered and the part will distort again. The distortion of welded structures such as ships can be minimized by assuring that the welding along the left side of the ship keeps pace with the welding along the right side. A bent part can be straightened by plastic bending in the opposite direction, or it may be straightened by *peening* on the surface containing the tensile stress that is the cause of the undesirable bend.

10.5 "CUTTING" OR PARTING BY FOCUSED HEAT SOURCES

All of the heat sources that will melt metal quickly can be used for "cutting" as well as for joining. The simplest approach is simply to melt a slot from a piece of metal to make two parts.

The cutting of steel may be done with a special oxyacetylene torch by heating an edge of the steel to about 850°C and then directing a stream of oxygen toward the heated region. As the iron oxidizes, enough heat is liberated to heat up the layer beneath.

The oxide and some metal melt and are blown away by the oxygen stream, and the process is self-sustaining. In fact, the original flame can be shut off, leaving only a stream of oxygen to accomplish the cutting. This process depends primarily on two factors; the heat of oxidation must be high, and the thermal conductivity of the metal must be low. Few metals fulfill these conditions as well as does iron, as may be seen from Table 10–3. For some that do not, iron powder can be injected into the gas or plasma stream to aid the process. It should, in principle, be possible to cut cast iron in this manner as well, but the flakes of graphite (Chapter 11) deflect the oxidizing stream away from the metal.

10.6 BRAZING AND SOLDERING

In many cases, metals and ceramic materials can be joined without melting the parent material. The casting of a soft metal into the gap between two parts can produce a joint provided the soft metal fills the gap, wets the parent surfaces, and finally bonds to the parent materials. Gap filling is best achieved by heating the filler metal to a temperature where it has low viscosity.

Wetting is achieved by cleaning the parent surfaces, usually with fluxes. A flux typically contains some elements or compounds that will, with the oxide of the parent material, form a new compound that has a lower melting temperature than that of the filler metal. This new compound will usually have a lower density than that of the joint metal, so it will float on the molten metal, which now readily wets the solids.

The success of cleaning can readily be checked by observing the contact angle between a drop of molten filler metal and each of the surfaces to be joined, as shown in Fig. 10–1. Or, if the parts have been joined but later they become easily separated, one can observe the extent to which the filler metal was drawn into the gap. The extent to which the filler metal *should have been drawn* into the gap can be estimated by using the equations for calculating the height of rise of a liquid in a vertical capillary with one end set in the liquid. Examples of good and poor wetting are shown in Fig. 10–7. For the

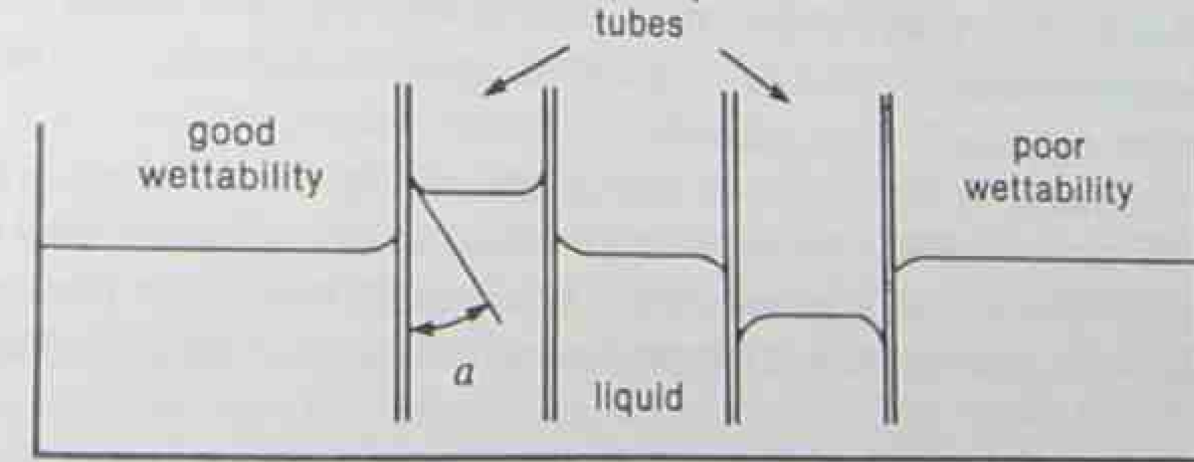


Figure 10–7 Sketch of the effect of wetting forces on liquid height in a small tube.

case of liquid rising in the tube, the lifting force F of the surface film is the vertical component of the *surface-tension* force T acting around the inner surface of the tube, or $F = \pi\, dT \cos a$. (Here a is proportional to the contact angle of a drop on a flat surface.) This is countered by the weight of the fluid, $\pi\, d^2 h\rho g/2$, where ρ is the density of the fluid and g is a gravitational constant. Thus

$$h = \frac{4T \cos a}{d\rho g} \tag{10-8}$$

It may be seen that for poor wettability, that is, where $a > 90$ deg, the fluid level in the capillary will be lower than that of the outside level, and in joining, the filler metal will resist entering the gap.

Good wetting and a strong bond are effected where some constituent in the filler metal is soluble in some constituent in the parent surface. For example, only very small amounts of lead will dissolve in iron, and thus lead does not wet an iron surface very well. But lead is a very useful and low-cost constituent in solder. To make a versatile solder, therefore, tin, antimony, silver, copper, and indium are usually included. None of these metals have oxides that interfere with the joining process. Aluminum would seem useful because of its low cost and low melting temperature, but its oxide forms very quickly and it is hard to remove.

Brazing and soldering are different from each other in that brazing uses a predominantly copper filler alloy, whereas soldering uses filler alloys made chiefly of lead and tin. Usually the parts to be joined are materials that are much stronger than the filler metal. Though the filler metals are soft and weak, a joint may be made very strong by effecting a proper geometry. The three geometric variables that control strength are joint contact area, thickness of filler, and whether the stress applied to the joint is a shear stress or normal stress. Figure 10–8 shows two ways to join parts, by a butt joint or by a lap joint. It can be seen that the load-carrying cross section of the filler material is restricted in the butt joint, but potentially unlimited in the lap joint. The lap joint could therefore be made stronger than the butt joint, but it has two disadvantages. The large step is rather conspicuous, and a large tensile force will cause bending, because the forces in each part are acting in different planes. Both of these problems can be overcome by a double lap joint as shown in Fig. 10–8. Another difference between the two is that the butt joint applies a tensile stress upon the filler metal, whereas the lap joint applies a shear stress. Shear strength is usually about half the tensile strength of metals, but in addition, a thin joint in tension can benefit from *triaxiality*.

Triaxiality can be shown in the sketch of Fig. 10–9, where a soft solid joins two much stronger solids. The parts are loaded in tension, and the stress in the joint is largely normal (that is, little imposed shear). As the tensile load on the parts increases, the stress state in the joint approaches the yield strength of the filler material. The filler material will

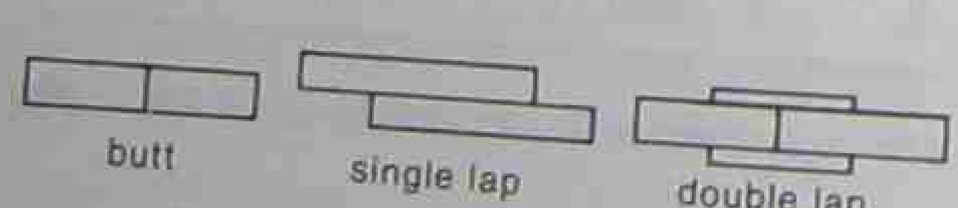


Figure 10–8 Sketch of butt joint and lap joints.

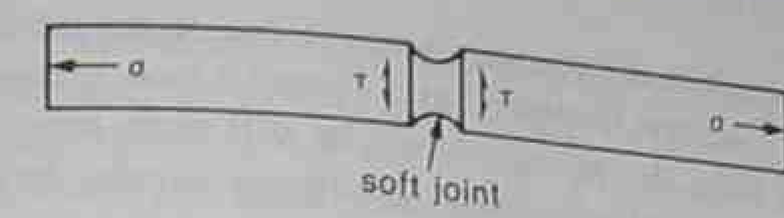


Figure 10–9 Triaxiality effects in a thin joint.

deform plastically with the effect of diminishing its horizontal dimension. However, the parent material will constrain the filler material from changing dimension by exerting an outwardly directed shear on the surfaces of the filler material as shown. This has the effect of imposing a horizontal tension on the filler material. By the principles of plasticity, additional vertical strain in the filler material requires a stress equal to the plastic flow stress *plus* a stress approximately equal to the horizontally imposed stress. The thinner the joint, the more complete the constraint, and thus thin joints appear to have high strength. Generally, a joint of approximately 10^{-4} mm shows the highest strength. Perhaps thinner films would be even stronger, but at this dimension it is difficult to determine what the film thickness is. There is little commercial interest in very thin films because of the high precision required in making the adjoining parts.

Soldering in particular, but brazing also, may be done by mass production methods because of the use of low-melting-point fillers. One can assemble complex structures, wires, brackets, and the like with flux and filler material properly placed. The entire assembly can then be heated in a furnace, or in a large flame shield, or by induction, or by infrared radiation. Alternatively, one can heat each joint in the assembly in turn by a fast-moving laser beam or an electron beam, programmed to focus briefly on each joint. These technologies are readily automated, but inspection is not as readily automated.

10.7 SINTERING, DIFFUSION BONDING, HOT-ISOSTATIC COALESCENCE

A growing technique for making solid parts is to press and coalesce powders into structural shapes. Parts ranging from styrofoam cups to roller bearing races are made in this way. Powders range in size from about 1 μm upward, and virtually every material is available in this form. Whereas in former days one chose materials from among those microstructures available on phase diagrams, it is now possible to mix a variety of powders together to achieve a wide range of properties. This is sometimes called making parts by *nonstochiometric chemistry*. The variations are infinite in number, far more than with the choices from phase diagrams.

The success of powder technology lies in compacting of the powder to near full density, and diffusion of the particles together. In the compacting operation the major problem is to avoid density gradients. Because of friction between the dies and the powder it is difficult to achieve uniform compaction. For example, if a long tube is filled with powder, and then if a rod were inserted into each end of the tube to compact the powder, the powder would be more dense at the ends and least dense in the middle. More uniform compacting, with lower applied force, can be done if some oil is mixed with the powder. Good compaction can also be aided by die design. Generally, the best part shape is such that, for example, if a die plunger moves 1 cm, it should compact the powder everywhere by the same volume fraction. But if a powder were compressed from all sides rather than from one side, the powder would be still more dense, and there would be fewer planes of weakness that arise from shearing the powder during uniaxial pressing.

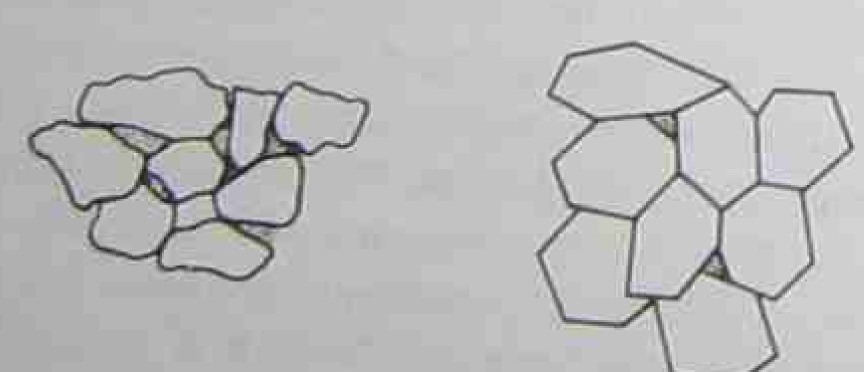
Pressing a material from all sides is very difficult to do. One could press a small amount uniaxially, on three orthogonal surfaces in turn, but at some risk of fracture. In practice a powder is first compacted a small amount, usually uniaxially, to develop some *strength*. Then it is encased in a fluid-tight container, inserted into a chamber, and pressurized hydrostatically.

Compressed powder has little physical strength. The strength of this powder is known as *green strength*. A useful increase in powder strength can be achieved by heating without hydrostatic pressure or with very little pressure. This process joint is called *sintering*. The powder particles are bonded together over a fraction of their surfaces, as shown in Fig. 10–10(a). The maximum strength of a powder is developed by heating powder particles while under pressure, so that they conform to each other and bond together. This is known as *hot isostatic pressing*, or HIP. When well done the particles are completely bonded, leaving porosity of less than one percent, as shown in Fig. 10–10(b). The pressures for this process range from 100 MPa to 300 MPa, and temperatures as high as 2000°C can be reached.

An important attribute of the finished powder product is the remaining porosity, or its converse, density. Almost any density may be achieved, up to about 99.999 percent of the theoretical density. This is done by heating to about 50°C above the solidus temperature of some major constituent in the powder and maintaining a hydrostatic pressure of the order of 250 MPa while hot. An interesting mechanism of bonding is seen in the processing of Si_3N_4 powder. If a small amount of oxygen is admitted into the enclosure, an oxide forms in conjunction with the Si_3N_4 to form a low-melting-point compound, which holds the powder together quite well.

10.8 GLUING AND ADHESIVE BONDING

Adhesive bonding is done by filling the gap between two parts with a liquid which will solidify by one method or other. There are two important aspects to the strength of adhesively bonded joints, and these are the strength of the material in the joint and the strength of the interface.



a. High porosity b. Low porosity

Figure 10–10 Sintered and HIP'd particles.

The strength of the material in the joint depends on the material, the geometry of the joint, and the methods used to apply the adhesive material, as discussed previously in soldering. There are many types of adhesives, and most of them are too difficult to describe by molecule structure and reaction mechanisms. The manufacturers of adhesives have been vigorous in marketing their products, so they supply detailed information on the properties, curing methods, and precautions to observe when using adhesives. The major generic types of adhesives are listed below.

Major types of adhesives and some of their properties

1. *Epoxies* are liquids that polymerize, or cure into a *thermosetting* solid, to hardness up to 90 ShoreD. They are available with a wide range of cure rates. Some are available in two "parts" which must be thoroughly mixed, and as soon as they are mixed, curing proceeds, with some heat evolution. Some are available in one part, that is, where the two parts are already together but formulated to cure very slowly at 20°C. These must be heated in order to cure in a few minutes. Epoxies are the strongest and most reliable of the adhesives and are available in forms that maintain strength up to 250°C. They are not recommended for joining plastics, because they attack the polymers chemically.
2. *Acrylic adhesives* are available in two major forms:
   a. The *anaerobic* form in the liquid state is prevented from curing when oxygen "terminates" or prevents free radical polymerization or hardening. Curing occurs when the normal exchange of oxygen becomes one way, namely evolution only. This occurs by containing some of the liquid between impermeable solids and exposing very little of the surface of the liquid to oxygen. This process is hastened by heating the liquid. Curing also requires the presence of metal ions, which are not available in most plastics. Where metal ions are not available, such as on surfaces, a "primer" can be applied to the surface (of the parts to be bonded) to supply the needed ions.
   b. The *cyanoacrylates* are thermoplastic; that is, they soften at elevated temperatures and become brittle at low temperature. The ethyl form is made to bond plastics and rubber, and the methyl form is made for bonding metals. These soften in the range of 50°C to 100°C. There is also an allyl monomer which retains strength up to about 220°C. All of these adhesives polymerize by an ionic mechanism initiated by weak basic ions, water being the most common. Adsorbed water on all surfaces is one source of initiator.
3. The *urethanes* are as strong as epoxies, and have greater toughness than many adhesives. However, they are formed by the reaction of a hydroxyl group and an isocyanate group, which produces toxic fumes.
4. The *silicones* are weak but resilient. They consist of Si and O in the chain. Some cure by liberating acetic acid and should therefore not be used on zinc, copper, or concrete. Others liberate the less reactive alcohols or amides, these are more expensive but are more widely useful.

5. The *methacrylate* adhesives are as good as the epoxies and cyanoacrylates, and they have the advantage of bonding plastics together.

An important part of the technology of adhesives is the design of the joint. Generally, polymeric adhesives should not be used to fill large gaps, one reason being that low-viscosity liquids will not stay in large gaps. Another reason, at least for the acrylic adhesives, is that curing is initiated at the interface of the liquid adhesive and the parts being bonded. A thick joint requires much time to cure, and a thick joint is exposed to oxygen over more of its surface than is a thin joint; this inhibits some curing. For the adhesives that evolve volatile chemicals during curing, the diffusion distance to the outer surface limits the rate of full cure. For the adhesives that cure by combination of their constituents, these considerations are not important.

Most of the adhesives provide joint strength of the order of 10 to 20 MPa, which usually refers to the shear strength. The joint strength in simple tension may be about half the shear strength, and the peel and/or cleavage strength is considerably less. Joints are usually more durable where one of the bodies is made of a material of low elastic modulus. Adhesively bonded joints will usually fail in the adhesive if it is inadequately cured; otherwise the interface usually fails. It is useful to visualize the nature of the bond sequence through the interface.

Virtually all metal surfaces are covered with an oxide. Metals such as aluminum, titanium, and chromium containing steels have relatively thin (3–5 nm) and adherent oxides. Iron, copper, silver, and many others have thicker (5–10 nm) and weaker oxides. Some oxides crumble and become detached when they are very thick. These should be removed before an adhesive is applied. However, an oxide layer will inevitably form, and most such layers are invisible. The instructions to remove oxides must refer to the loosely attached oxide. Next upon the oxide in a clean atmosphere is a layer of adsorbed water, as mentioned above, about 10 nm thick. In the presence of humans and machinery, however, most surfaces will also accumulate oils and greases. These may form under or over the water film. The major difficulty with these contaminants is that they constitute a thick fluid layer, which diminishes joint strength. The oily films are usually removed by a basic chemical solution, usually sodium compounds. These include carbonates, bicarbonates, sesquicarbonates, hydroxides, tripolyphosphates, tetrapyrophosphates, metasilicates, and orthosilicates.

The purpose in cleaning is to reduce the thickness of the fluid layer. If a liquid wets two parallel flat plates, the force required to pull the plates apart normally is inversely related to the thickness of the fluid film. At the edge of the plate there will be a meniscus of liquid, curved inward if the liquid wets the plate material. Surface tension in the fluid surface will tend to produce a flat meniscus, and thus a curved meniscus implies that there is a higher pressure outside of the mensicus than within the liquid. We may estimate this pressure difference from the simpler case of the pressure difference between the inside and outside of a liquid drop or a bubble of gas in a liquid. The equation for this pressure differential is $p = 2\gamma/R$. Here γ is the surface energy and R is the radius of the drop or bubble. In the case of liquid between two flat plates the separation of the plates h is about $2R$. The differential pressure from outside to inside the liquid, acting over the plates, will

produce a force F needed to separate the plates and the magnitude, $F = 2A\gamma/h$. If $\gamma = 30$ dynes/cm, $A = 1$ cm^2, and if there is a thick contaminant film about $h = 10^{-5}$ cm thick, then $F \approx 6 \times 10^4$ dynes, or about 60 newtons of force. But if the surface is "clean," that is, if the adsorbed gas film is $h = 10$ nm (10^{-6} cm) and the gap between two such adsorbed films on two surfaces is filled with a hardened polymer, the force required to separate the surfaces normally—that is, to draw the adsorbed water inward—will be 600 N/cm^2 or about 600 MPa. This is the approximate strength of many adhesive joints. It can readily be seen that an additional contaminant film would lower the joint strength.

Admittedly, most adhesively bonded joints do not fail in tension, but rather in shear. In this case the fluid films might flow and readily fail. However, in a thin joint the asperities of the surfaces, that is, the surface roughness, will be much higher than the thickness of adsorbed water, thus preventing easy shear.

REFERENCES

1. A. H. Cottrell, *The Mechanical Properties of Matter*. New York: John Wiley and Sons, 1964.
2. L. Holland, *The Properties of Glass Surfaces*. Chapman and Hall, 1964.
3. R. Pirani and R. Yarwood, *Principles of Vacuum Engineering*. Reinhold, 1961.

PROBLEMS

10–1. Assume a cube of steel of 100-mm dimension, set upon a steel table. Both have a YS of 80 MPa. Calculate the real area of contact between the two due to the weight of the cube alone and due to the addition of your own weight.

10–2. If one second of time is required to melt a steel surface with an electric spark, and 10 seconds are required when an oxyacetylene torch is used, what is the approximate energy density on the surface due to the latter?

10–3. Calculate the penetration due to arc welding steel at a velocity of 3 mm/s.

10–4. Sketch the effect of an elastic vise instead of a rigid vise on a figure of the type of Fig. 10–6.

10–5. With reference to the data in Table 10–2, how far would each of these liquids rise in a capillary tube of 0.5 mm diameter?

11
casting

11.1 INTRODUCTION

In the most general sense, casting includes all processes where a material in liquid form is introduced into a cavity (or mold) so that upon complete solidification, the surfaces of the new solid product have the opposite shape of the mold. Consider the pouring of water into an ice-cube tray; the cavity is basically a cubical hole and upon solidification, a solid cube of ice has formed. This is the simplest example of casting that involves all of the major fundamentals involved, and shows that the process entails a liquid-to-solid transformation to produce the desired shape. Thus, unlike the processes called forming and machining (cutting) discussed in earlier chapters, where end products are produced from one solid state to another, casting usually involves more complex possibilities with regard to the microstructures that may result upon solidification.

Although certain processes which use the word *molding* fit the definition of casting, others do not, and this could lead to confusion. For our purposes, any molding process that involves conversion from the liquid to solid will fall under the heading of casting. Those that are processed from one solid to another are classified differently. In this way, the fundamental nature of the process takes precedence over the name given to the process itself.

Most of the *analyses* developed for those processes that involve the altering of a starting solid into some other solid configuration were based upon the subject of solid mechanics, and applied forces or induced stresses were of primary concern. In contrast, the major disciplines involved in casting include fluid mechanics, heat transfer, concepts of solidification, and thermal contraction. Consider an example of involving *most* of these topics. A hot fluid is being pumped through a pipe made of stainless steel, and the

temperature of the fluid at various locations along the pipe is to be predicted. If such factors as the thermal conductivity of the steel, the rate of flow, and the likely effects of both conduction and convection are known, a reasonably straightforward analysis can be made by using an appropriate background in fluid mechanics and heat transfer. Of course, the geometry involved is relatively simple, and solidification is not considered. The great majority of cast products are far more complicated in shape, and even those who possess advanced technical backgrounds in the topics of fluid flow and heat transfer would have a difficult time to provide a similar, straightforward analysis concerning temperature gradients. Because of this, we intend to cover such situations from a physical rather than a mathematical point of view. After all, manufacturing engineers are not expected, for example, to be involved with details of mold design; rather, they are often confronted with the processing of cast products that have been produced by others.

Here we distinguish between two broad categories of cast products. The first includes those that are produced to a particular shape that is as *close* to the desired product configuration as possible. As an example, the crankshafts shown in Fig. 11–1 would require only a limited number of subsequent grinding operations to become a finished product. Similarly, if one considers a carburetor body, which has an extremely complicated shape, only some final machining operations are needed to produce the finished part. For products like these, it is *always* desirable to first cast the part with as little *excess* material as possible, that is, to leave only such excess as is needed to be certain that final dimensions can be achieved by subsequent finishing operations. The second category

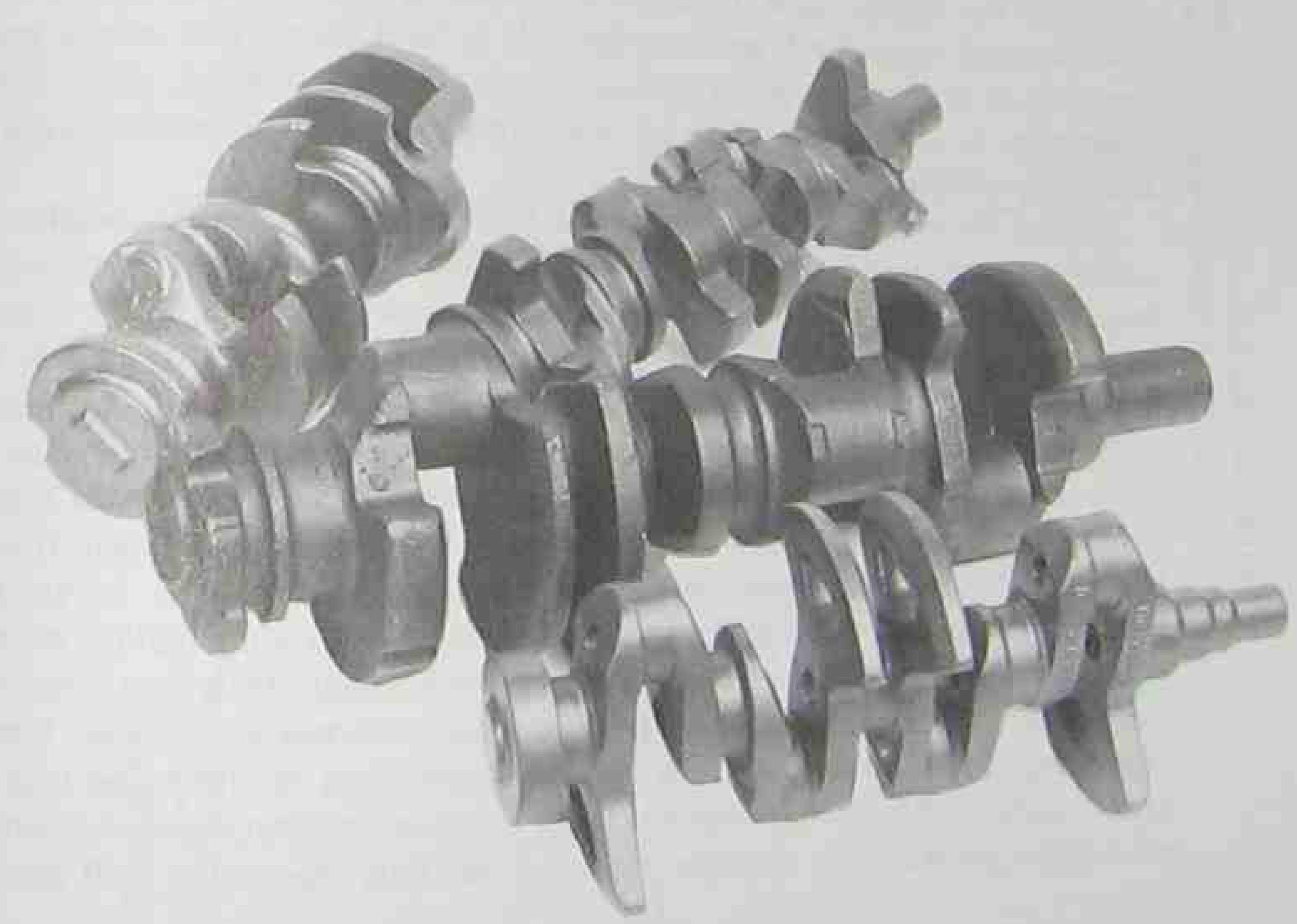


Figure 11–1 Photograph of several cast crankshafts.

includes those materials that are cast to an initially simple shape (often called an *ingot*) which is then subsequently processed by other methods to produce various forms that are then used in other processing modes. For example, a cubical billet of steel may be cast via a series of operations, as shown in Fig. 11–2. It is then subjected to one of a number of possible forming operations that lead to a wide variety of shapes such as flat sheets, round bars, wire, and the like; note the great variety of possibilities shown on Fig. 11–2. These operations produce a *wrought* structure which possesses a greater uniformity of finer grains, far fewer internal voids, and greater ductility than did the initial cast structure. Thus, ductile steel sheets used in subsequent forming operations were *initially* cast to some very different shape; this is true for the many shapes used in a variety of other processing operations. In this context, it could be said that casting is the truly primary process, since forming and cutting operations are performed on parts that were initially cast to shape. Continuous casting, a more recent process, is discussed in Sec. 11.8.

11.2 ASPECTS OF THE BASIC PROCESS

Consider Fig. 11–3, where a simple cube is to be produced; for now let us pay no attention to the material which forms the mold. Since the solidified product must be eventually removed from the mold, the latter is constructed of two separate parts (*cope* and *drag*) which meet along the *parting line* when assembled. The molten material is fed into the cavity via the *pouring basin*, from which it travels through a channel (called a *runner*), fills the *mold cavity* itself, and then proceeds to fill a *riser*. (This simply provides a reservoir of molten material, whose purpose is discussed shortly.) Using this simple illustration, we now address the variety of occurrences that take place prior to, during, and after final solidification results.

First, the molten material must possess adequate *fluidity** so as to completely fill all cavity of the system before solidification begins anywhere. The term fluidity involves the combined effects of the fluid flow *and* heat transfer aspects involved. If, for example, the molten metal had the consistency of molasses as compared with water, the flow rate into the mold would be extremely slow. As it contacts the walls of the runner and early sections of the mold cavity, heat transfer into the various contact surfaces begins, and the temperature of the leading zone of molten material decreases. Once this reaches the freezing temperature, solidification begins; if the flow rate is low enough (because of inadequate fluidity), the leading zone solidifies across an entire section, thereby preventing further flow. As a consequence, the entire cavity, from pouring basin to riser, would never be fully filled and an incomplete casting would result. Techniques involving initially higher fluid temperature, mold materials of lower thermal conductivity, better mold design, and, in some cases, additions of particular elements into the initial fluid may be used to alleviate this problem. Some of these, however, can introduce other problems if not handled properly. For example, increasing the fluid temperature will increase the

* Particular test methods have been developed to measure the relative fluidity of molten materials. See R. A. Flinn, *Fundamentals of Metal Casting* (Reading, Mass.: Addison-Wesley, 1963), pp. 87–95.

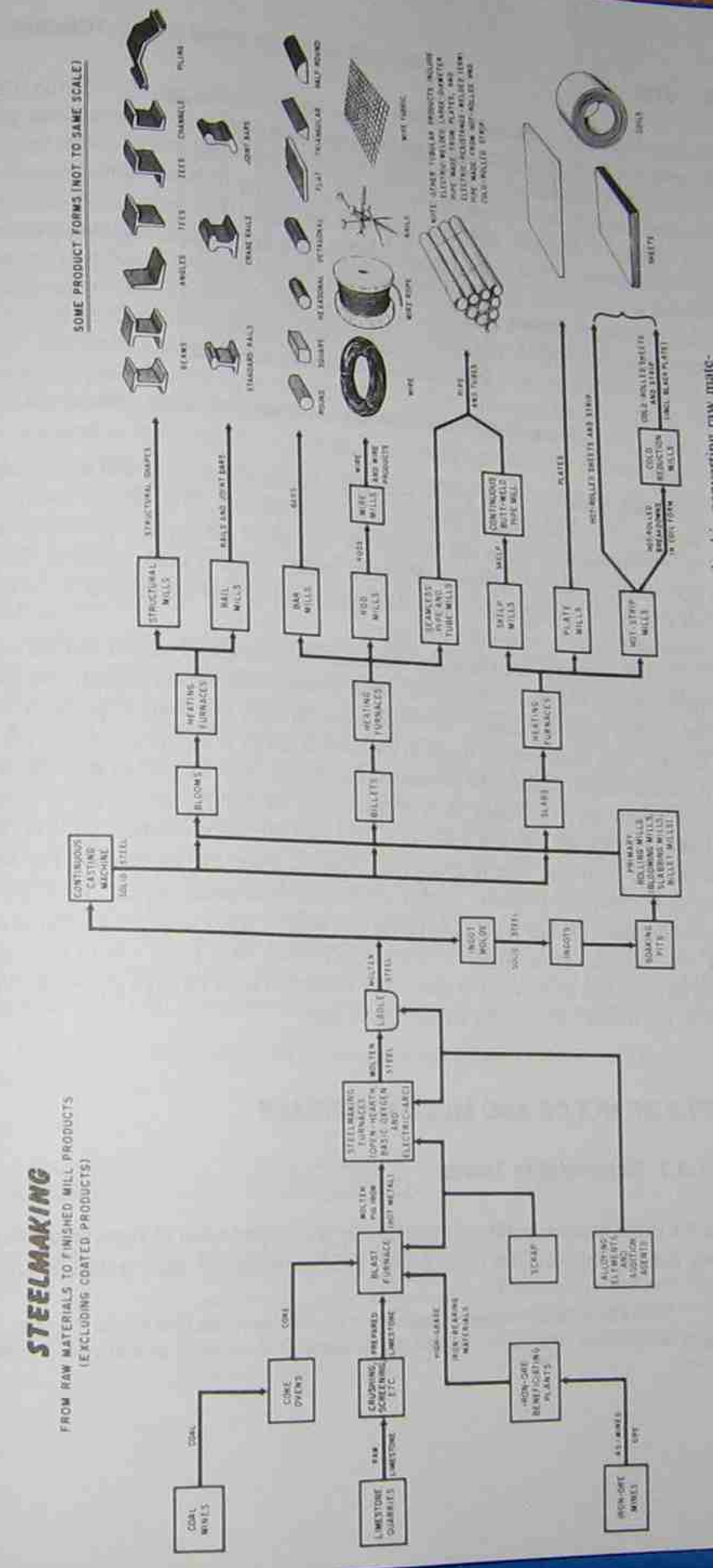


Figure 11–2 Flow diagram showing the principal process steps involved in converting raw materials into the major product forms, excluding coated products.

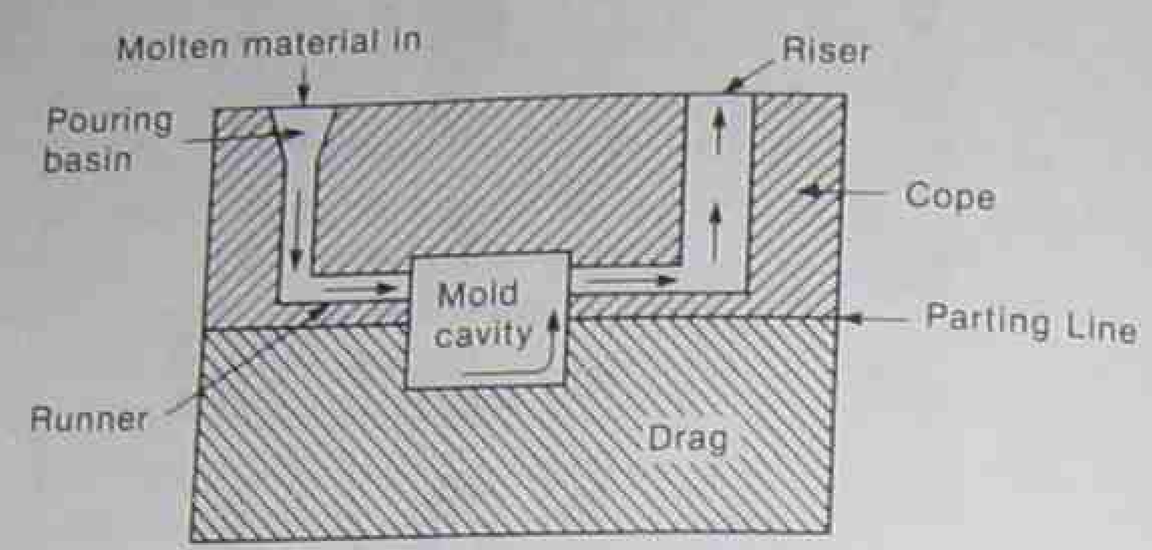


Figure 11-3 Sketch of the major components related to a simple casting.

solubility of the liquid for gases, such as hydrogen and oxygen, and will produce a greater amount of *porosity* (discussed later) in the solid product. With certain metals, the addition of silicon leads to improved fluidity *but* can cause subsequent problems in machining operations, since the presence of silicon oxide, which is quite abrasive, leads to higher wear rates of cutting tools. These examples indicate that the manner by which fluidity is controlled cannot be wholly arbitrary.

Now assume that the entire mold has been completely filled and the cubical cavity (still molten) begins to cool down. Some thermal contraction begins, and the volume of the molten cube decreases. To compensate for this, the riser, if properly designed and located, then *feeds* molten metal to the cubical cavity to compensate for such shrinkage.* As solidification of the cube continues, the change from a liquid to solid causes additional contraction due to the latent heat of fusion. (In the liquid form we are at a higher energy level as compared with the solid form, so as solidification occurs, and is accompanied by a volume decrease, heat is given off.) Finally, when the entire cube has solidified and cools to room temperature, further solid-state contraction occurs. Near the end of full solidification, the last regions of liquid that solidify must undergo a volume decrease. As a consequence, small voids or holes result. They may form within the part or on the surface and can be reduced to acceptable minimum sizes through proper casting design or with the use of *chills;* both are discussed later.

11.3 TYPES OF MOLDS AND MOLD MATERIALS

11.3.1 Expendable Molds

As the name implies, molds of this category are broken apart to remove the cast part; thus, they are expendable. The most widely used material for such molds is *sand,* which is

* The size of the riser is crucial here, since if a section *across* the riser freezes too soon, it can no longer supply molten metal to the cavity. A good rule to observe is never to try to feed thick sections through thin sections.

usually mixed with clay, filler material such as cereals, and water. The combined effects of clay and water provide the binding and necessary strength (to maintain a desired shape of cavity) which would be absent if only dry sand were used. Cereals, as they burn off, provide one method to allow gases, air, or even steam to permeate through the mold and away from the casting as it solidifies. Both the size and shape of the grains of sand influence the surface finish of the end product. In Fig. 11-4, a schematic of the interface between a number of grains and the molten material is shown. As heat is transferred from the molten material, the grains tend to expand; if they were packed together as indicated, they would tend to heave (much like a concrete highway without adequate expansion joints). The surface finish that results is directly influenced by such action. Using grains of finer size and adequate filler material, which after burning off leaves spaces for freer expansion between adjacent grains, can alleviate this problem and improve surface finish. In a comparative sense, however, typical sand castings produce the poorest surface finish of all casting processes. If this poses a problem, such surfaces can always be machined to produce smoother surfaces; of course, this introduces added cost. There are, however, several advantages in using sand. It is inexpensive and plentiful, and, because it can be exposed to high temperatures without gross deterioration, practically any material can be sand cast. In addition, there is really no restriction on the size of the part that can be made by this method.*

Shell molding utilizes fine-grained sand that is mixed with and bound by a thermosetting resin. In essence, the pattern (usually metal) of the part to be cast is coated with the sand-resin mixture and cured at an elevated temperature to produce a thin shell which is then simply lifted away from the pattern. Making two such shells (each effectively duplicates half of the full pattern) and then clamping them together produce a cavity that duplicates the outer shape of the initial pattern. As compared with sand castings, parts made by shell molding produce much better surface finish and less dimensional variation (that is, closer tolerances). These characteristics result because the finish of the shell cavity itself is much smoother than cavities produced in sand castings.

Plaster molding involves the use of plaster of Paris and additives such as asbestos and silica, which improve the strength of the base material. These ingredients are mixed with water to form an almost pastelike composition, which is then poured over the pattern of the part to be cast. Once the mold material sets to a reasonably hardened consistency,

* Sand casting is often subdivided into three categories, namely green-sand mold, skin-dried mold, and dry-sand mold.

Figure 11-4 Schematic illustrating how expansion of sand grains causes heaving, which results in surface variations of the cast part.

the two halves (similar to those discussed under shell molding) are stripped from the pattern and dried at an elevated temperature to provide adequate strength. The halves are assembled together, and the complete internal cavity produced is then ready for accepting the molten material. Although such molds produce castings of good surface finish, close tolerance variation, and excellent detail in terms of surface reproduction, they cannot be used with molten materials having high melting temperatures, since the plaster mold will rapidly degrade at temperatures in excess of about 1200°C. This problem can be avoided if the mold material contains certain ceramics instead of plaster of Paris. Procedures involved to produce two halves of the mold are quite similar to that discussed under plaster molding. Because the major ingredients differ, however, the latter process is called *ceramic-mold casting*.

With *investment casting*, also called the *lost-wax process*, the pattern is made of wax, which is dipped into a mixture of refractory material (such as fine silica) and liquid. The wax is coated by this *slurry* (that is, silica plus liquid) and when the casting has dried, redipping follows. This is continued until the full coating possesses adequate thickness to provide the necessary strength for handling. After heating the entire unit at a few hundred °F in an inverted position, the wax melts and runs out of the mold (thus the lost-wax concept), which is then heated to between one and two thousand °F. This removes any fluid from the mold material, which in turn, hardens and sets. In general, excellent reproduction of surface details, good surface finish, and close dimensional control result with parts cast by this method. Unlike plaster molds, the materials used to form the mold by investment casting can handle materials of higher melting temperatures.

To summarize the major concepts in this section, a satisfactory mold material must be capable of receiving the molten material and permit it to solidify without the mold's degrading. Thus the melting temperature of the mold must be greater than that of the material to be cast. In addition, the smoother the internal surfaces of the mold cavity, the better will be the surface finish and dimensional control of the cast product.

11.3.2 Permanent or Reusable Molds

When cast parts are to be produced in large volumes, it is beneficial, where possible, to use molds that need not be broken up to remove the casting. Instead, such molds can be used over and over; thus the word *permanent* is applied. As compared with those discussed in Sec. 11.3.1, permanent molds provide much greater production rates. Again, the mold material must possess the ability to avoid degradation from contact with the molten material to be cast, and thus such molds are made from various metals, refractory materials, or graphite; the choice to a large extent depends upon the temperature of the molten material as it is introduced into the mold cavity.

Permanent-mold casting is one category of such processes. In essence, the mold halves are clamped together and the molten material is introduced into the cavity; flow occurs by the effect of gravity. After solidification, the halves are separated and the casting is removed, often with the aid of ejector pins. When metal molds are used, the cooling rates involved are much faster then those associated with the methods discussed

in Sec. 11.3.1, since the thermal conductivity of metals is much higher than sand, plaster of Paris, and the like, so the resulting microstructure of the cast product can differ accordingly. Besides high production rates, this process produces a product having a good surface finish and close dimensional control. Of course, the initial cost of equipment is higher than expendable molds, but if the quantity of production is great enough, the initial investment can be justified economically. It is also noted that permanent molds do not permit the permeation of gases through the mold walls as do sand molds, so special venting techniques must be used when such molds are designed.

The second major category is called *die-casting*, and the principal difference between this process and permanent-mold casting is that the molten material is fed into the mold cavity under a pressure (rather than gravity). Depending upon the size of the cavity and the type of molten material, pressures up to thousands of pounds per square inch may be required, and the resulting forces acting on the cavity walls tend to separate the die halves. Consequently, high clamping forces must be applied to prevent such separation; this is not as serious a problem with permanent-mold castings. Die casting of relatively low-melting-point materials is usually referred to as *hot-chamber* die-casting; as illustrated in Fig. 11–5, pressures of a few thousand psi are typical. With higher-melting-point materials, the process is called *cold-chamber* die-casting; this is shown in Fig. 11–6, and pressure requirements are perhaps ten times higher than those used with the hot-chamber process. As with permanent-mold casting, high production rates, good surface finish, and close dimensional control result with die-casting. Because cavities are filled more quickly, due to the pressure, die casting is somewhat faster.

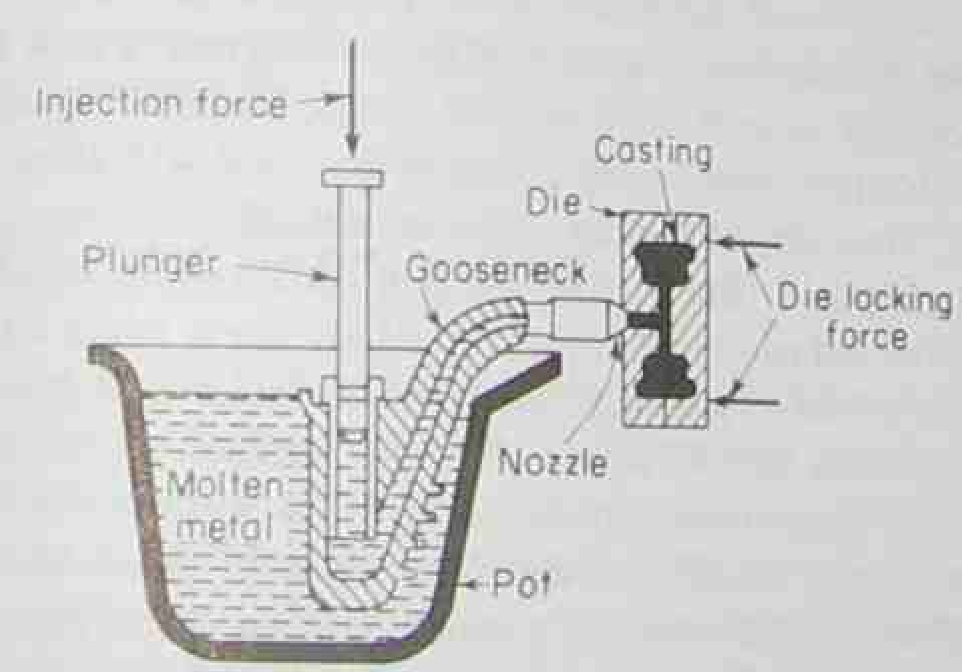


Figure 11–5 Schematic of hot-chamber die casting.

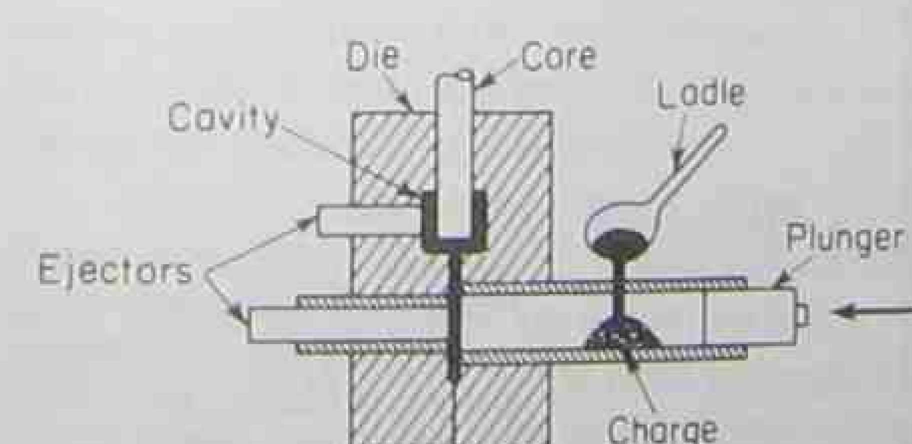


Figure 11–6 Schematic of cold-chamber die casting.

11.4 CASTING OF METALS

As *pure* metals are poured into a mold cavity, nucleation of solid grains begins along the interface of the liquid and mold walls. Conversion of material from liquid to solid occurs at a constant temperature, since the composition of any remaining liquid at any instant remains essentially constant. Thus, from the start to the end of full solidification, liquid and solid phases of the same composition coexist throughout. In most cases, a layer of fine, equiaxed grains first forms in the region of the interface, where the rate of heat transfer is highest and there is little resistance to uniform grain growth from the inner liquid. Further solidification generally produces large *columnar dendrites* which, as they grow away from the equiaxed region, find no resistance to growth towards the center. Because adjacent dendrites tend to interfere with each other's lateral growth, a type of *preferred* orientation results in the formation of long, columnar grains. Figure 11-7 illustrates a typical cast structure of a pure metal.

With *alloys*, solidification from beginning to end does not occur at a constant temperature. Consider two elements called A and B, where the melting temperature of A is substantially higher than B. Further assume that these elements are fully soluble in each other in both the liquid and subsequent solid states and that we begin with 50 kg of each element as solid bars. As the melting temperature of B is reached, that mass will liquify first; eventually at a higher temperature, A becomes liquid and the two elements form a homogeneous solution of liquid. As the temperature is lowered, solidification begins, and the first solid to form will be richer in A (the higher-melting-temperature element). In a sense, this is the exact opposite of what happened when the two solid elements were heated, since B melted first. As discussed in Chapter 7, if the cooling rate is slow enough (that is, nearly equilibrium cooling) as more solid solution forms, the effect of diffusion will produce a homogeneous structure whose *composition* at any time will show a lowering of element A by percent. Thus when the entire mass has solidified, its composition will be 50 percent A and 50 percent B. But to arrive at this final structure demands a cooling rate low enough to allow diffusion to be effective. Many times when alloys are cast, the initial solid forms at a rate of cooling that is much faster than that called equilibrium; in addition, the cooling rate varies widely from surface to center. As a consequence, the composition of grains in regions adjacent to the surface can be quite different from those nearer the center; in addition, the composition of grains themselves vary from surface to center. This is called *coring* or a cored grain structure. Figure 11-8

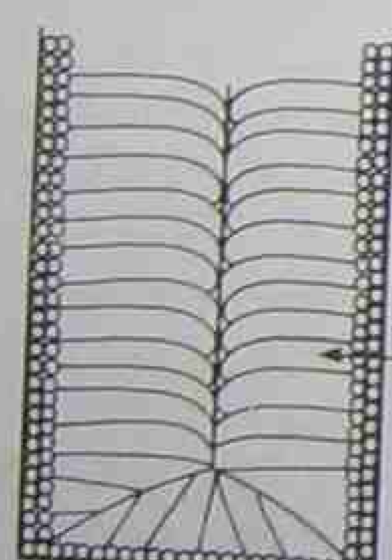


Figure 11–7 Schematic of the cast structure of a pure metal showing equiaxed grains at the surface and columnar dendrites in the interior.

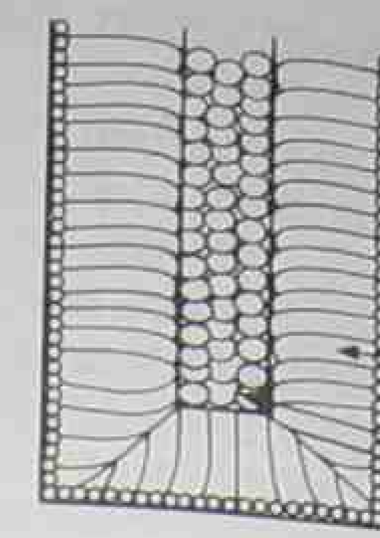


Figure 11–8 Same as Fig. 11–7 except for an alloy.

is a schematic of an alloy casting; it is noted that the central region contains some equiaxed grains and lesser-sized dendrites on the whole as compared with pure metals. These differences arise because (unlike the freezing of pure metals, where only the *masses* of solid and liquid change during solidification) as alloys solidify, not only do the masses of solid and liquid change, but the *compositions* of the two phases change as well. Several methods can be used to decrease the nonhomogeneity of cast metal alloys. One is to use an *additive* (such as sodium in aluminum alloys or ferrosilicon in cast irons) which induces nucleation in regions across the section rather than preferentially at the mold walls. The avoidance of columnar dendrites and a more uniform grain size result. To produce a more uniform structure in terms of *composition*, reheating of the casting to elevated temperatures for relatively lengthy time (often called *soaking*) accelerates diffusion effects in the solid state. A more homogeneous structure results and is somewhat similar to that expected *if* equilibrium cooling rates had prevailed during initial solidification.

11.5 CASTING OF POLYMERS

Because of the names given to many processes that produce polymers in the form of rods, sheets, and so forth, strict categorization poses some problems. For example, *extrusion* sounds like a forming process, and to a major extent it is; many thermoplastics are processed this way. Solid pellets of the material are fed from a hopper into a chamber containing a heating zone which softens the pellets. Often a screw-type conveyer then applies pressure to the softened pellets and forces them through a die whose particular configuration governs the shape of the outlet product. Since the pellets, although heated, are not poured into a mold cavity to solidify, this can't be called a casting process as we have defined it. Instead it best falls under the category of a forming process. Figure 11–9 shows the essentials of this process.

Another widely used process, using either thermosets or thermoplastics, is called *injection molding;* this has all of the characteristics of a casting process and is so defined here. Pellets (or granules) are fed from a hopper to a heating chamber, where they become molten. Under pressure, either from a plunger or a screw drive, the melt is forced to flow into a cavity (usually a split mold). Upon solidification, the mold is opened and the part

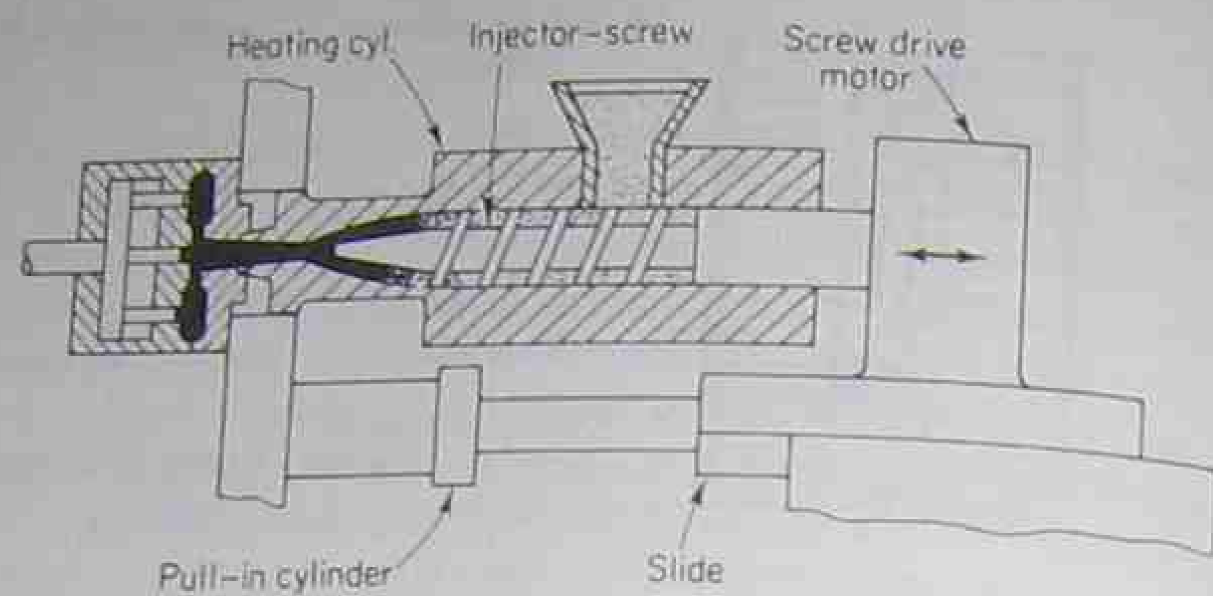


Figure 11–9 Essentials of screw extrusion.

is removed; often ejector pins assist in separating the part from the split mold. This process is quite similar to die casting, especially the hot-chamber type; Figure 11–10 is a schematic showing various details of this process.

Rotational molding, used to produce relatively large, hollow parts such as drums or tanks, is another process that poses difficulties in terms of categorization. A two-piece mold, usually metal, is designed in such a way that after the mold is clamped to form a unit, it can be rotated about two axes perpendicular to each other. The polymeric material, either in the form of a fine powder or liquid, is introduced into the mold cavity, which itself is heated. As rotation commences, particles are thrown against the mold walls by centrifugal force, whereupon they fully melt and coat the cavity wall. After cooling to form a solid, the part is then removed. Since this process essentially involves the cooling of a molten material to produce a shape governed by a cavity, we classify it as a casting process.

Both thermoplastic and thermoset materials are produced to shapes such as rods or tubes, with techniques similar to those discussed in Sec. 11.4. With thermoplastics, the monomer is mixed with ingredients such as additives, heated to a molten condition, and poured into a mold to solidify. Items such as wheels and gears are processed in this way. Procedures used with thermosets are similar.

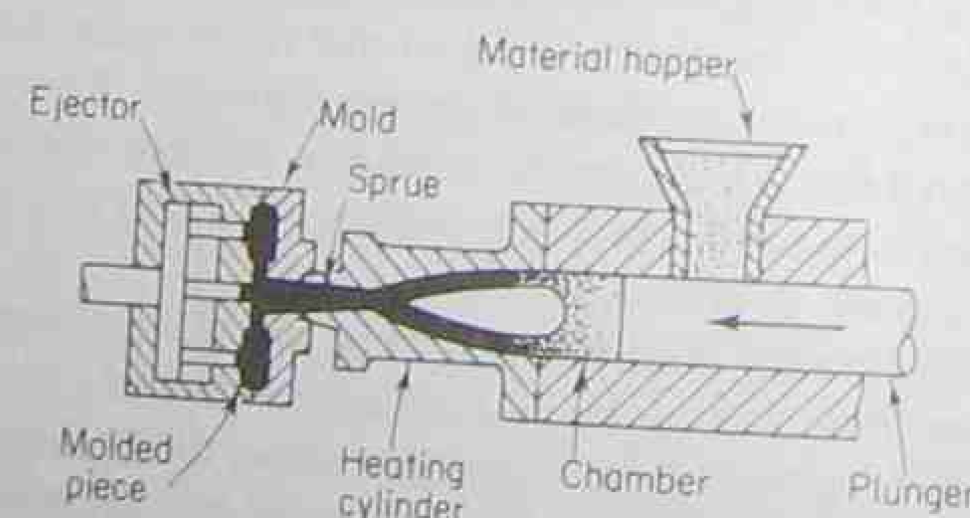


Figure 11–10 Essentials of injection molding.

11.6 POTENTIAL DEFECTS AND POSSIBLE CURES

Undesirable occurrences such as holes and cracks can result in castings for several reasons. Since the solubility of *gases* is much greater in a liquid than solid material, upon solidification from the melt, such gases can end up forming small holes. Hydrogen, nitrogen, and oxygen typify gases that can cause problems. Because the solid structure contains these porelike holes, the term *porosity* is often used to describe this action; in effect, the entire cross section is not sound. This type of porosity can be alleviated by providing a means for these gas *bubbles* to vent to and from the surface of the melt. Another technique is to *purge* the melt. This involves flushing the melt with an inert gas such as argon or helium. There, for example, hydrogen in atomic form diffuses into the bubbles of inert gases, which are then carried to the surface and out of the molten material. Pouring in a vacuum is done on occasion; this reduces the amount of dissolved gases but requires special equipment.

Another, often severe effect of porosity occurs in the internal region of the molten material where the last bit of metal solidifies; this was mentioned in Sec. 11.2. Since the last liquid to solidify must undergo a volume contraction, a void, hole, or porous region results. Figure 11–11 is an illustration; this is called *shrinkage*. This type of porosity can be reduced or avoided by improved casting design or the use of *chills*. In either case, the fundamental problem to be solved involves a higher rate of heat transfer from those sections which would solidify last (that is, reduce temperature gradients). As shown in Fig. 11–12, a metal chill, which could be a small piece of steel, is located in a sand mold *before* pouring proceeds. When the cavity is filled, the largest section now cools at a faster rate than it would if the chill were not present; the rate of heat transfer through the steel is much greater than through the sand mold. This tendency to reduce the volume of *hot spots* leads to more uniform cooling across the section, thereby lessening the size of the porous region that would normally occur. Using chills against external surfaces is also helpful. A simple illustration of improved casting design is given in Fig. 11–13; the radial shape leads to a more uniform cross section, thereby reducing temperature gradients

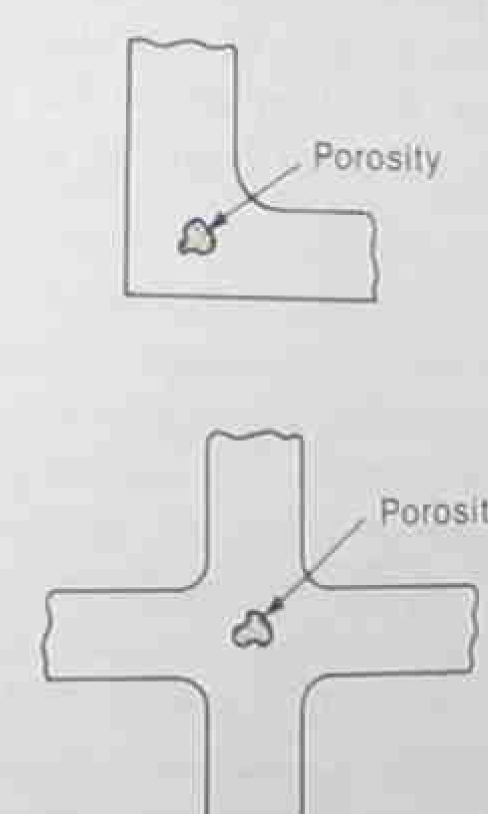


Figure 11–11 Illustration of shrinkage cavities.

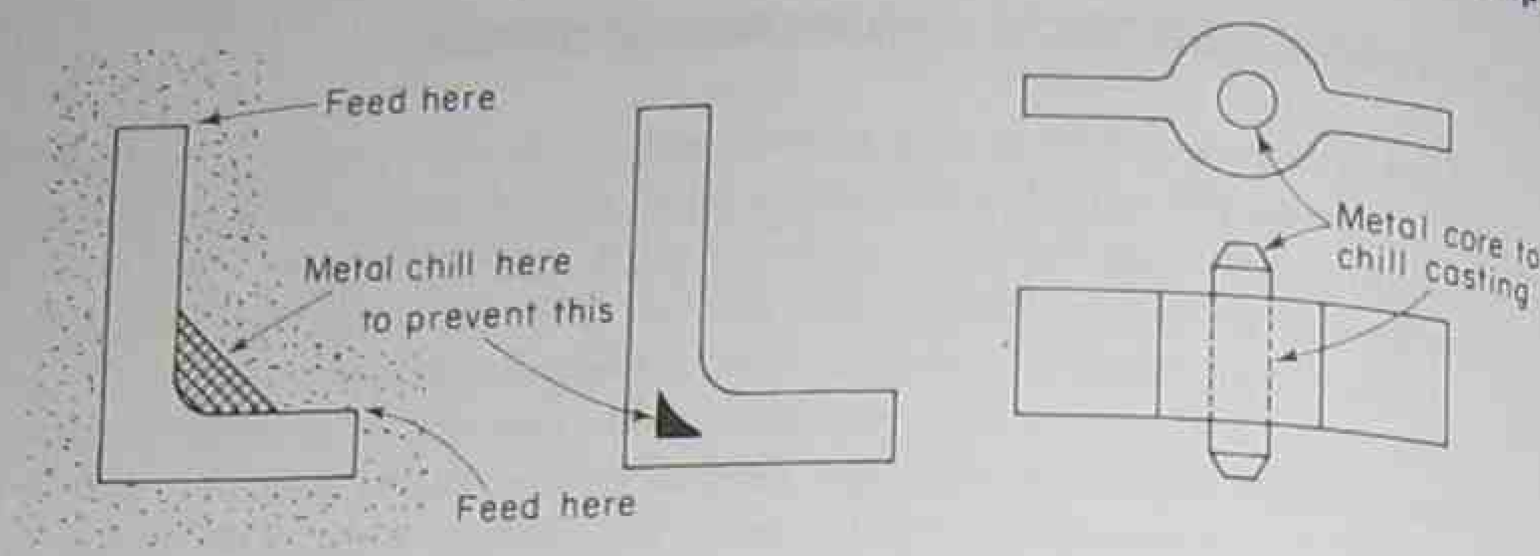


Figure 11–12 Examples of the use of chills.

across the corner. In general, except in large ingots, porosity due to gases leads to small, spherical holes, whereas shrinkage causes larger voids with rougher surfaces.

Another possible defect of this type is called *interdendritic* shrinkage. In regions where dendrites have formed as three-dimensional solids containing branchlike growth, adjacent dendrites eventually interlock and trap still molten material within the interlocked branches. As this material solidifies and shrinks, small voids or holes result, since additional liquid cannot be supplied to compensate for such shrinkage. This type of porosity, in terms of pore size, is generally quite small and in many cases is not as serious as the voids left by the type of shrinkage cavities discussed above. Of course, any type of porosity is undesirable from a strength viewpoint, since the effective area supporting loads or stresses is smaller than that based upon external dimensions. Stress concentration effects may also be more severe. Also, if the cast product is to contain a gas or fluid under pressure, there is a greater problem with porous structures, since the pressurized gas will have a greater tendency to permeate through the porous structure compared with one that contains few if any pores.

Hot tearing is another serious defect that often occurs; in essence, an actual crack (or hot tear) results in the cast part. Figure 11–14 indicates how this comes about. There the cavity, shaped like an I-beam, has ends of somewhat larger thickness than the center portion. As solidification occurs, the center section will completely solidify before the ends do. With further cooling, the center begins to undergo thermal contraction, but

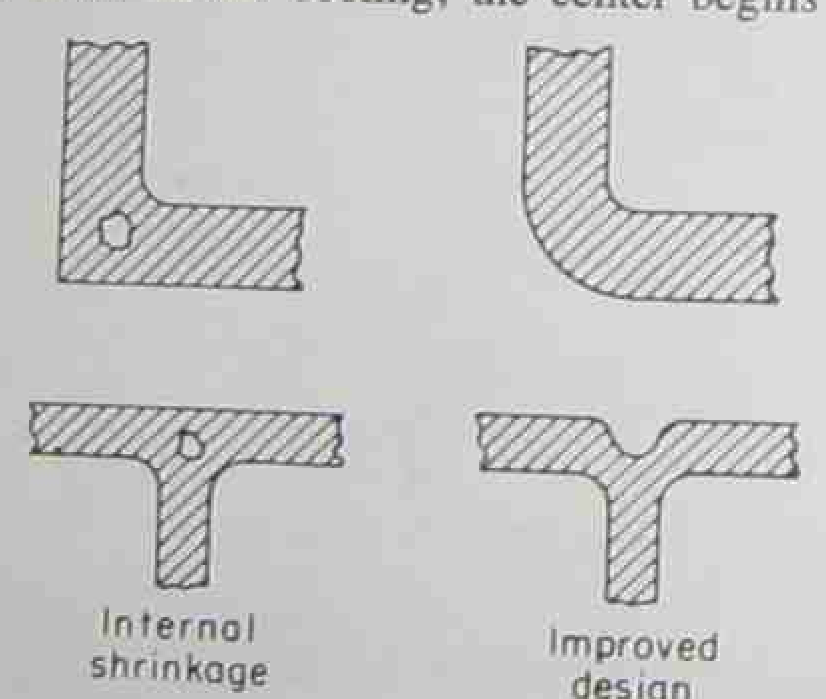


Figure 11–13 Design alterations to reduce or eliminate shrinkage.

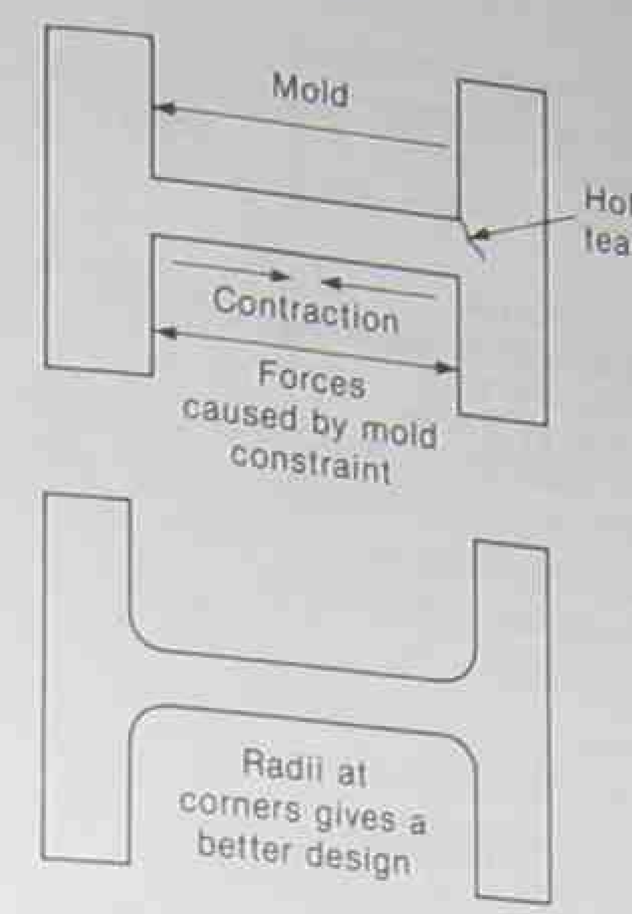


Figure 11–14 Schematics to illustrate the possible cause of hot tearing and an improved design to reduce such a possibility.

constraint from the mold walls prevents free contraction and induces forces to act against the ends as shown. If these forces reach a high enough level, the induced stresses cause a hot tear or crack to occur as indicated. Such a result must obviously be avoided. Better casting design, as shown in Fig. 11–14, can alleviate this problem. Another solution is to literally break up the mold once the entire *outer surface* has solidified. Thus, removing the ability of the mold to cause excessive constraint negates the possibilty of hot tearing. Finally, a change of material is another possibility; using a material that can tolerate greater strain to fracture could solve this problem.

Although numerous other kinds of defects (mainly at the surface) can be categorized, those we have discussed are often the most serious and, for purposes of illustration, provide an adequate introduction to this topic.

11.7 RESIDUAL STRESSES

Whether this topic should fall under Sec. 11.6 or be treated separately is a matter of preference. Since cast products will *nearly always* contain residual stresses and in many instances perform acceptably, we treat this separately from the topic of defects. The primary cause of residual stresses is the nonuniform thermal gradients that typify the process of solidification of a molten material. Consider Fig. 11–15, which illustrates a component consisting of two thick round selections (top and bottom) connected by a thin-walled annular section. Because of the different thicknesses, the annular section will solidify first, and then attempt to undergo solid-state thermal contraction as the temperature decreases. Uninhibited contraction is restrained by the thicker sections. Now the top section will fully solidify next and as it begins to shrink in the radial direction (due to contraction in the solid state), it applies radial forces to the upper portion of the annular section. A similar sequence of events follows as the bottom (thickest) section

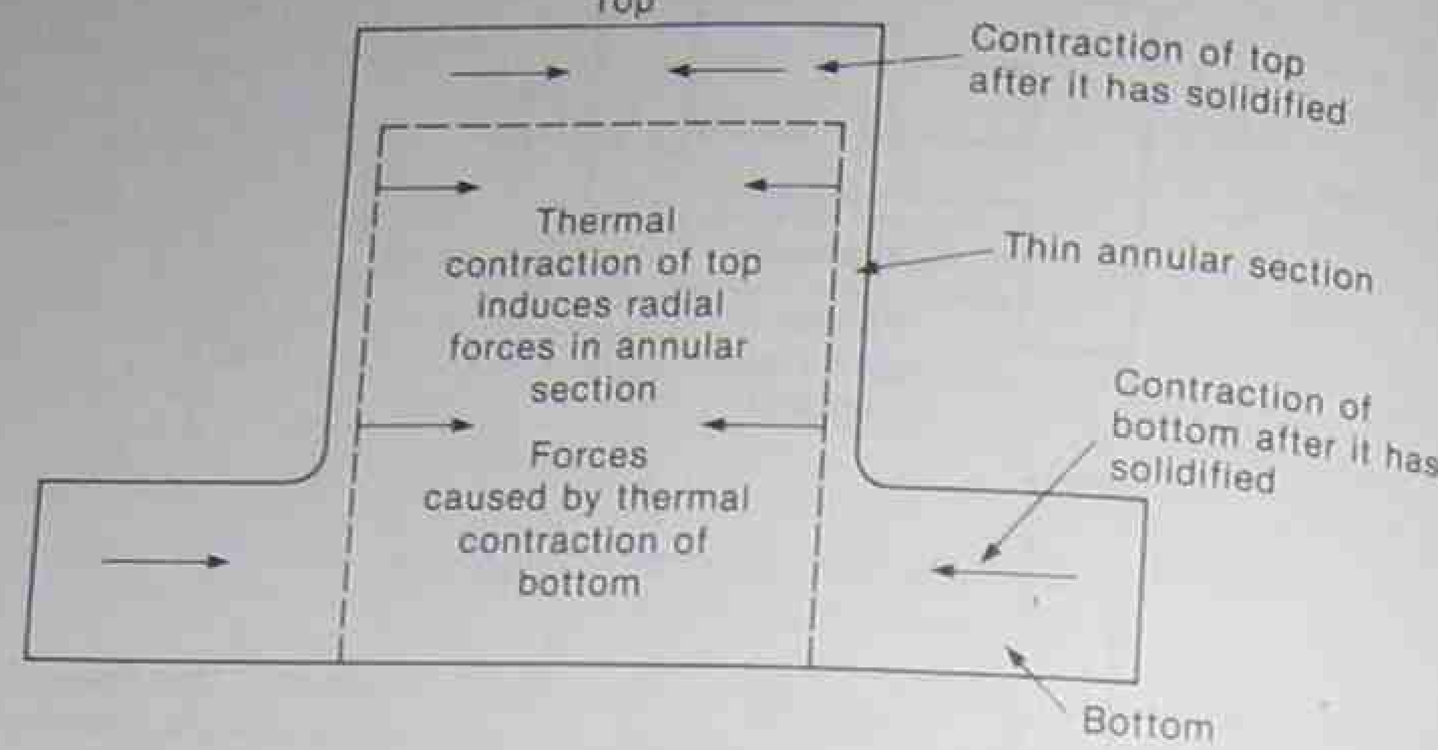


Figure 11–15 Illustration of a cast component which may contain residual stresses after solidification.

solidifies and contracts, except that the lower section of the annular section experiences forces induced by the contraction of the bottom section. This complicated series of events can produce residual stresses of varying magnitudes in different regions of the final casting.* There are several potentially serious consequences that can result. First, when this component is placed in service, the applied loads, forces, torques, pressures, and the like induce stresses that are superimposed upon the residual ones, so the onset of yielding, necking, or fracture could occur at lower levels of *applied* stresses than would happen if the residual stresses were not present. Another detrimental effect can result if such a casting were subjected to subsequent machining operations, since they induce forces (thus, stresses) as the cutting action takes place. For example, suppose the full height of the annular specimen is to be turned † to produce a particular dimension and improved surface finish. When the stresses caused by the turning operation are superimposed upon the residual stresses, fracture in the form of a crack could occur around the periphery, where the annular section meets the thickest section at the bottom. Several solutions could alleviate these potential problems. First, the casting could be subjected to a stress relief by proper heat treatment; this entails an added cost. Next, the thickness of the annular section could be increased; this would lessen the degree of thermal differences as the various sections solidify and undergo thermal contraction. As a consequence, the magnitudes of residual stresses would be lowered. Of course, this design change increases the weight of the casting; this can be reduced by machining off some of the material considered as excess. However, this increases material costs and adds a cost for machining. Finally, a substitute material possessing greater inherent strength and resistance to crack formation and propagation could be considered. Among

* If they lead to hot tears as discussed in Sec. 11.6, the stresses are effectively relieved *but* a cracked part results.

† Discussed in Chapter 9.

other problems that may result because of residual stresses are distortion that accompanies subsequent machining, difficulty in holding desired tolerances, and long-term dimensional instability.

11.8 BATCH VERSUS CONTINUOUS CASTING

Although a variety of metallic alloys are cast to nearly final shape, we include this section to illustrate two broadly different techniques that are used to produce cast structures that are further processed to a variety of shapes such as flat sheets, round bars, and the like. In addition, the discussion is restricted to low-carbon steels; although this is arbitrary, it is noted that the tonnage of such alloys used for industrial purposes makes them one of the most widely used materials.

Batch casting involves the production of an individual ingot cast in a large mold. Once solidified, the ingot can then be subjected to a variety of operations, discussed earlier in connection with Fig. 11–2. The types of steels of concern here are broadly classified as *rimmed* or *killed*.* Rimmed steels are not deoxidized before solidification occurs, and as the ingot freezes, dissolved carbon and oxygen react to cause a violent evolution of CO bubbles; this is the *rimming* effect, and it leads to a stirring of the molten metal. One of the consequences is the segregation of carbon towards the center of the ingot. The net result is that the surface regions of the ingot contain a carbon content lower than the overall average composition. If processed into sheets, for example, the rolled sheet will likewise have surfaces of low carbon that tend to be free of carbide particles; in essence, those regions are nearly pure iron.

With *killed* steels, elements such as aluminum and silicon are added to the melt. They combine more readily with oxygen than does carbon, thereby forming oxides (that is, the steel has been *deoxidized*) and the rimming action is prevented. Solidification occurs without violent bubbling; hence the term "killed." Not only is segregation of carbon to the center avoided, but the oxides often float to the top of the melt. As a result, the composition across the ingot is quite uniform and largely free of porosity. One drawback is the extremely large shrinkage cavity (called a *pipe*) that forms at the top of the ingot; this must be cut off or *cropped* before further processing, which leads to a larger material loss than that connected with rimmed steels; Fig. 11–16 illustrates these differences.

A more recent method for producing various shapes involves *continuous casting*. Figure 11–17 illustrates different methods that are used. Essentially, the molten metal flows through cooling chambers that act as molds. With proper cooling, a solid skin forms and acts to support the partially cast structure which, with further cooling, completely solidifies. As it exits in a solid form, it can be cut to desired lengths or subjected directly to forming operations that produce shapes of desired contours. Compared with ingot

* *Semi-killed* is a further subdivision that falls between these two. These are partially deoxidized and have less porosity than rimmed steel and little or no pipe as with killed steels. Rimmed steels are often "capped" to control the rimming action and reduce the degree of porosity that would otherwise result.

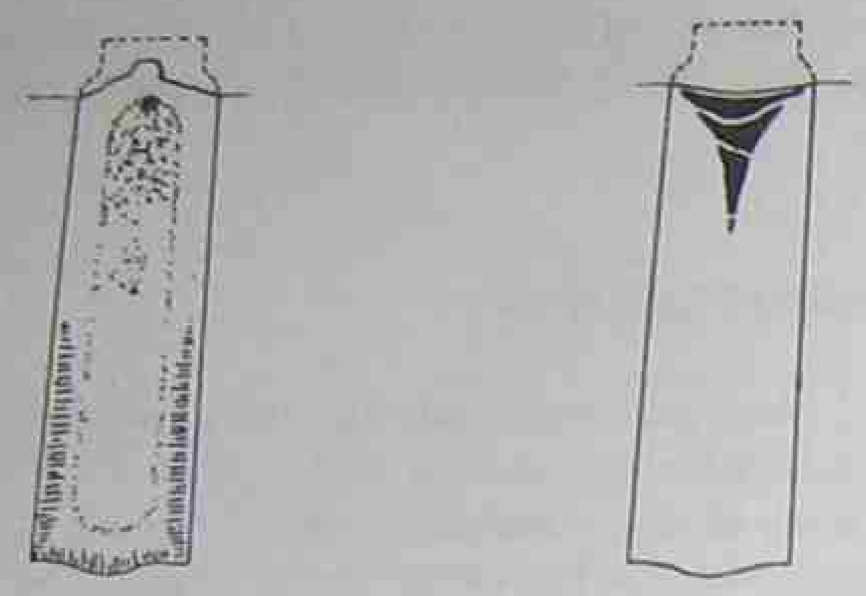


Figure 11–16 Sketch of the cross sections of ingots made from (a) rimmed and (b) killed steels.

casting and subsequent processing, this technique provides far greater production rates. Of course, the specialized equipment is expensive.

11.9 CAST IRONS

In our experience, many students who have completed an introductory course in materials science have heard of *cast irons* but have decided misconceptions about these metals. First, they seem to think that these are the only metal alloys that are cast to some shape and, in addition, that these alloys are inherently brittle. From what has already been discussed in this chapter, the first idea is clearly wrong, but the second needs some attention. Cast irons are basically alloys of iron and carbon; however, they differ from steels in two major aspects regarding *composition*. While steels seldom contain carbon in excess of one percent, cast irons, as a class, contain carbon from about two up to four percent. In addition, and of extreme importance, while silicon is usually controlled in steels at low levels (~0.2 percent), it is deliberately added in cast irons in amounts up to 3.0 percent.

There are three *primary* types of cast iron; they are white, gray, and nodular. The term "primary" as used here indicates that these irons can be produced directly from the melt. In Sec. 7.4.1, various steel microconstituents like ferrite and pearlite were discussed. These same micros occur in cast irons, and their relative amounts depend upon both the initial solidification rates and any subsequent solid-state heat treatment. In cast irons there is the additional possibility of cementite (Fe_3C) *dissociating* into iron and free carbon (*graphite*). A combination of a *slow* cooling rate and a *high* percentage of silicon provides a greater chance for this dissociation to occur.

White cast iron (white iron) contains relatively lower carbon and silicon than do the others, and to produce this material requires relatively high cooling rates. As a consequence, no free graphite results, and all of the excess carbon (other than the small amount in any ferrite) exists in the *combined* form of cementite (Θ) In general, the structure of most white irons usually consists of pearlite and rather large masses of free cementite. Because the carbon content is higher than with typical steels, the *amount* of

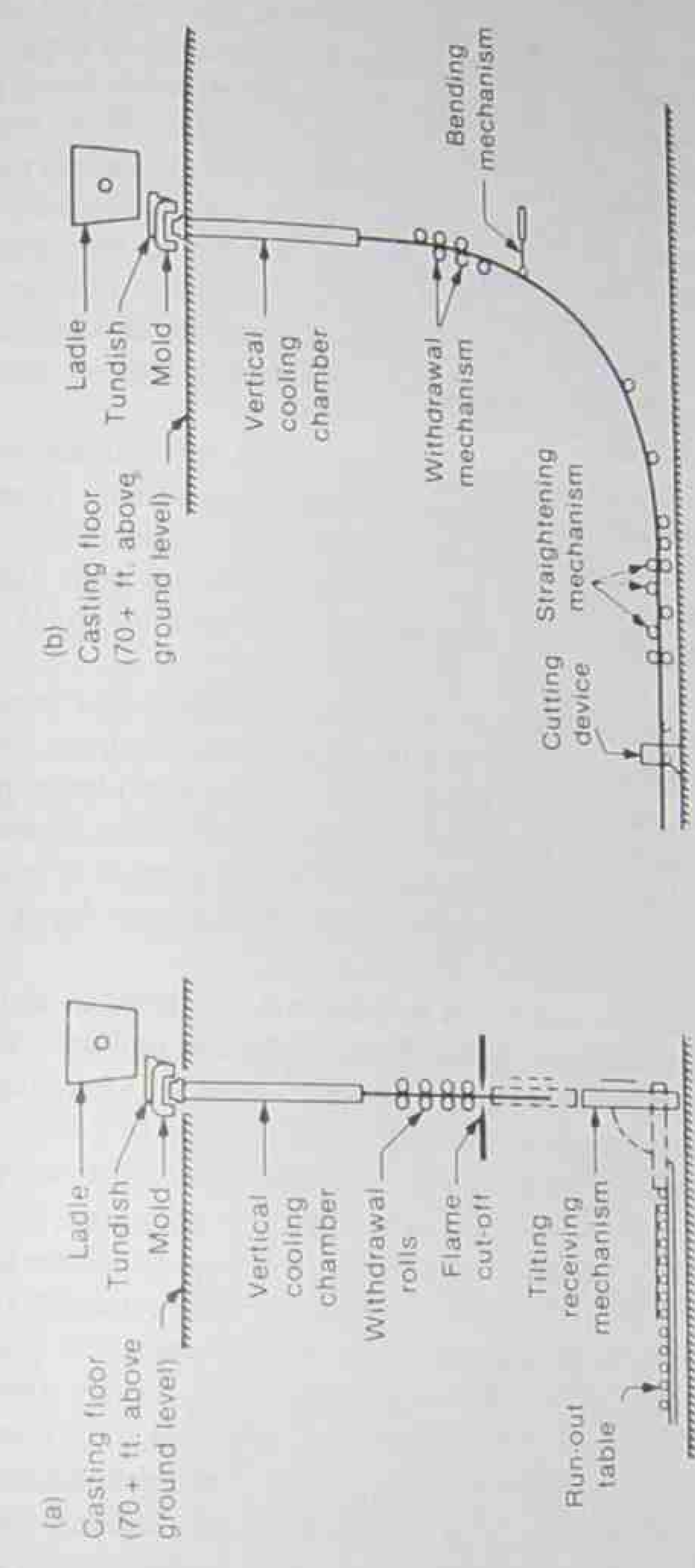


Figure 11–17 Three methods used in continuous casting of steel.

free cementite is much greater. Since Θ is extremely hard but quite brittle, the ductility of white cast iron, as measured in a tensile test, is negligible; it does, however, provide excellent wear resistance.

Gray cast iron (gray iron) contains larger percentages of carbon and silicon than does white. As these metals solidify and cool to room temperature, a large amount of dissociation occurs and free graphite, in the form of sharp-ended *flakes*, forms part of the final structure. These flakes act as stress risers such that the stresses at the flake tips are much higher than the nominal stress (that is, load divided by area). The net result is that gray irons display little tensile ductility. If these metals are given subsequent heat treatments, it is possible to reduce the *size* of the flakes, which increases the tensile strength but, since the *shape* of the free graphite remains as sharp-ended flakes, ductility cannot be improved to any meaningful degree. In any event these cast irons are substantially stronger in compression than in tension, since compressive loading tends to "close" the flakes and the stress-raising effect is avoided.

Nodular cast iron (also called *ductile* or *spheroidal*) has compositions similar to gray iron but also contains the elements cerium and/or magnesium. These elements influence the formation of free graphite to produce spheres or nodules, rather than flakes. Since the stress concentration effects are greatly reduced, nodular irons display much greater tensile ductility than does gray iron; reduction of area up to 25 percent is possible.

Figure 11–18 shows typical photomicrographs of these three cast irons; note the obvious difference in the shape of the graphite in the gray and nodular irons. We add here that, in a comparative sense, the gray and nodular irons involve much larger percentages of silicon and slower cooling rates as compared with white iron. In fact, if one wished to produce a structure of white iron, the cooling rate *must* be fast enough to prevent dissociation of iron carbide. Consequently, white iron cannot be fully produced throughout thick sections.

For completeness, *malleable cast iron* is included and is referred to as a *secondary* type. It is classified this way because, unlike the others, it *cannot* be produced directly from the melt. One must form white iron (note that this therefore limits the section sizes of malleable iron). The white iron is then heated to about 1750°F for up to two days and then slowly cooled to room temperature. At the elevated temperature the massive regions of cementite dissociate to form free graphite that has a shape which is unlike that in gray or nodular. Further dissociation can occur when the austenite attempts to transform to pearlite at the lower critical temperature. With slow enough cooling rates, the cementite (which must exist if pearlite is to result) may break down into ferrite and additional free graphite. Such a structure would produce a *ferritic malleable iron*, as shown in Fig. 11–19. Although greatest ductility can be produced with a ferritic nodular iron, malleable irons possess greater ductility than either the inherently brittle gray or white irons. It is noted, however, that the use of malleable irons is on the decrease, while nodular irons have replaced them in many applications. This is probably due to two factors. First, much thicker section sizes can be made directly from the melt with nodular irons; in contrast, since white irons are restricted to thinner sizes, due to the need of higher cooling rates, malleable irons are so restricted. Next, the need to heat treat white irons at elevated temperatures for rather long time periods

(a)
(b)
(c)

Figure 11–18 Photomicrographs of (a) white, (b) gray, and (c) nodular cast iron.

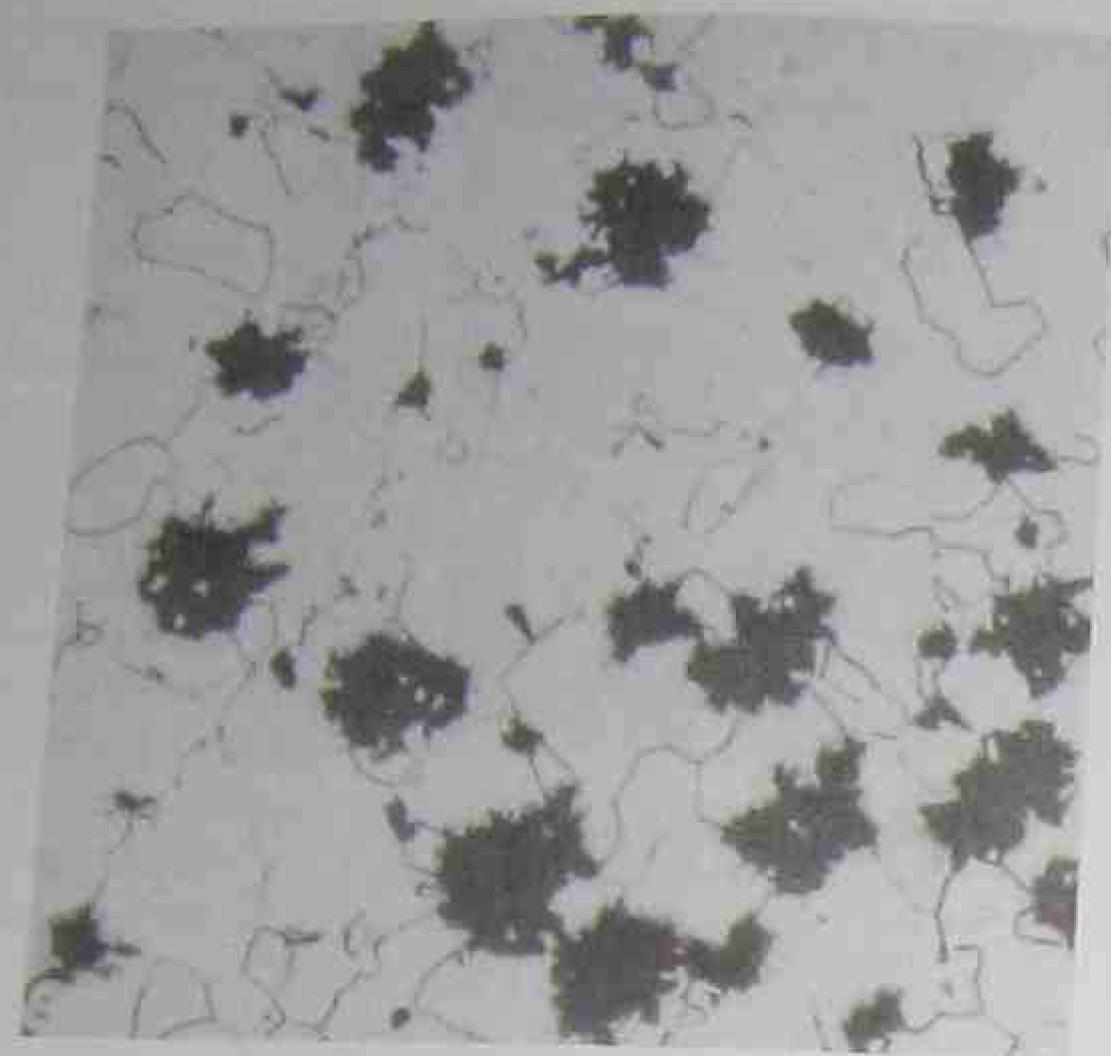


Figure 11–19 Photomicrograph of ferritic malleable cast iron.

adds a cost in producing malleable irons that is usually higher than similar costs associated with the nodular type.

Compacted graphite iron has a composition similar to nodular iron but lower amounts of magnesium. Probably because of lower magnesium, the shape of the graphite in these irons is not as spheroidal as in nodular irons, nor are the flakes that form as sharp-tipped as in gray irons. Instead the graphite tends to form as a type of flake with somewhat rounded tips. As a result, the mechanical properties of compacted graphite irons fall between those displayed by gray and nodular irons.

11.10 DIRECTIONAL SOLIDIFICATION

There are several important industrial applications that require the preferential solidification in a single direction; in essence, single crystals are produced so that the external surface is the entire grain boundary. Semiconductor materials, such as germanium, are made by the *crystal-pulling* method illustrated in Fig. 11–20. There, the molten metal is made to contact a small seed in the form of a single crystal. As solidification occurs at the interface, the seed is rotated and slowly pulled away from the melt, and the new solid duplicates the structure of the original seed. Alloying elements, called *dopants*, are sometimes included in the liquid metal to produce special properties.

A second method, called the *floating-zone* technique, is used with higher-melting-temperature materials such as silicon. There, a piece of polycrystalline material contacts

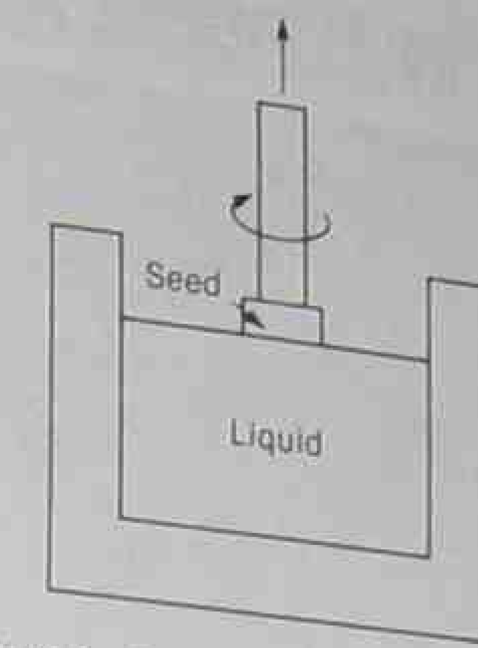


Figure 11–20 Schematic of the crystal-pulling method to produce single crystals.

a single crystal of the material, and then a radio-frequency (RF) heat source is applied while moving upward at a slow rate. In effect, the polycrystalline rod is converted to an ever-increasing length of a single crystal. The net result is a single crystal having a single orientation; see Fig. 11–21. The large single crystal is then sliced to provide thin wafers that are the *chips* used in many electronic devices.

Another important application of directional solidification is in the manufacture of turbine blades used in high-temperature devices such as jet engines. There, by carefully controlling the rate of heat transfer from the molten metal, solidification occurs in a unidirectional manner. As shown in Fig. 11–22, long, parallel grain boundaries depicting columnar grains can result. It is also possible to produce the entire blade as a single crystal whose only grain boundary is the external surface. Such blades provide superior high-temperature properties (such as resistance to creep and thermal shock) compared with conventionally cast blades shown in the same figure.

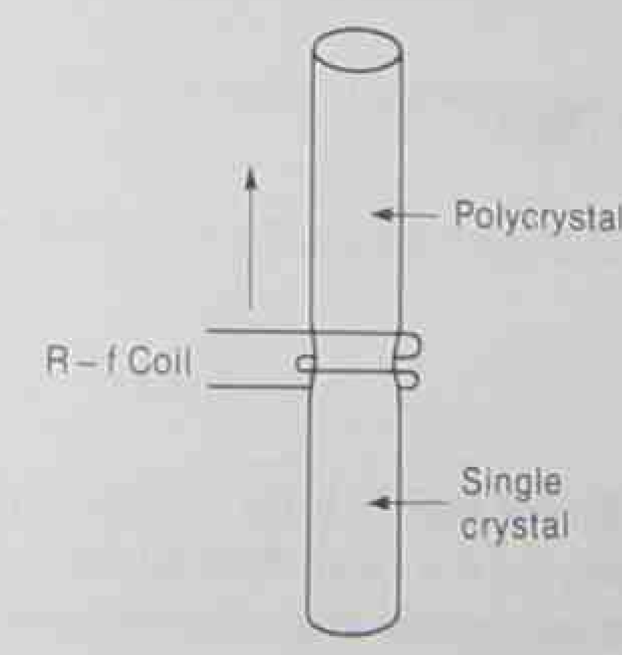


Figure 11–21 Schematic of the floating-zone method to produce a single crystal.

ADVANCES IN TURBINE AIRFOIL MATERIALS

Conventional casting (a) | Columnar grain (b) | Single crystal (c)

Figure 11–22 Turbine blades produced by three methods: (a) conventionally cast, (b) directionally solidified to produce columnar grains, and (c) a single crystal.

11.11 SOLIDIFICATION TIME

From a heat transfer analysis,* the time for solidification of a molten metal may be given as

$$t = B(V/A)^2 \tag{11–1}$$

where t is the solidification time, V is the volume of the shape of the part, A is the total *surface* area of the part, and B is called the *mold constant*. As such, B is a function of such parameters as density and specific heat of the metal as well as the thermal expansion and

* See, e.g., R. A. Flinn, *Fundamentals of Metal Casting*, pp. 32–38.

thermal conductivity of the mold material. Work by Chvorinov* showed a good agreement between Eq. (11–1) and experimental results for shapes of relatively simple geometries. It is best to consider this equation as a reasonable guide that shows a general tendency, since the freezing times of more complex shapes and certain metals cannot be analyzed adequately with such a simple equation.

PROBLEMS

11–1. Figure P11–1 shows a cavity in the form of a tapered plate where the direction of the flow of the molten material is indicated. After solidification is complete, voids along the centerline result, as shown in that figure. Is this defect due to a lack of adequate fluidity, poor mold design, or a combination of both. What do you recommend should be done to avoid this *centerline shrinkage*?

11–2. A tee section, shown in Fig. P11–2, is to be made by sand casting a metal. The only concern here has to do with the most sensible location of the riser, noting the different dimensions of the cavity. Where would you suggest the riser be located? Explain why.

11–3. Coarse-sized grains of sand are used in a particular sand mold. Suppose that at one location of the mold surface, two adjacent grains are in contact and that the constraint from regions adjacent to these grains is rather large. The cavity is then filled with a molten metal. Discuss what happens as the grains become heated and what result is likely to occur on the surface in that region when solidification takes place. How might this result differ if very fine grains of sand had been used?

11–4. In terms of fundamental considerations, why might shell molding be used instead of plaster molding?

11–5. A particular metal component may be produced by die casting or sand casting. Regarding the grain size at or near the surface of the part after solidification, what differences are apt to be found with parts cast by these two methods? Why does this occur?

11–6. Dendrites are found in many castings. Why does such grain formation result and why is this often disadvantageous?

11–7. Alloys of copper and nickel are completely soluble in the liquid or solid state for any composition; consider an alloy of half of each element by mass as shown in Fig. P11–7. Under *equilibrium* cooling, an α solid solution of 50 Cu − 50 Ni results in all individual

* N. Chvorinov, *Proceedings of the Institution of British Foundry*, **32** (1938–39), p. 229.

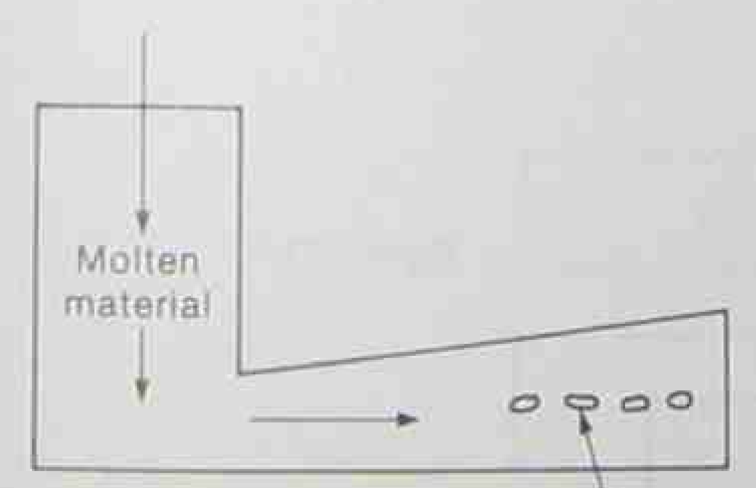


Figure P11–1

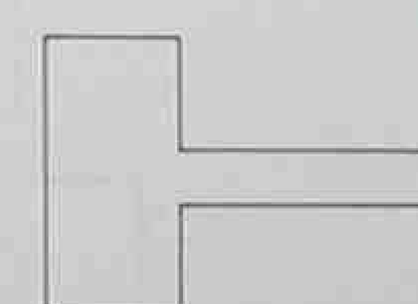


Figure P11–2

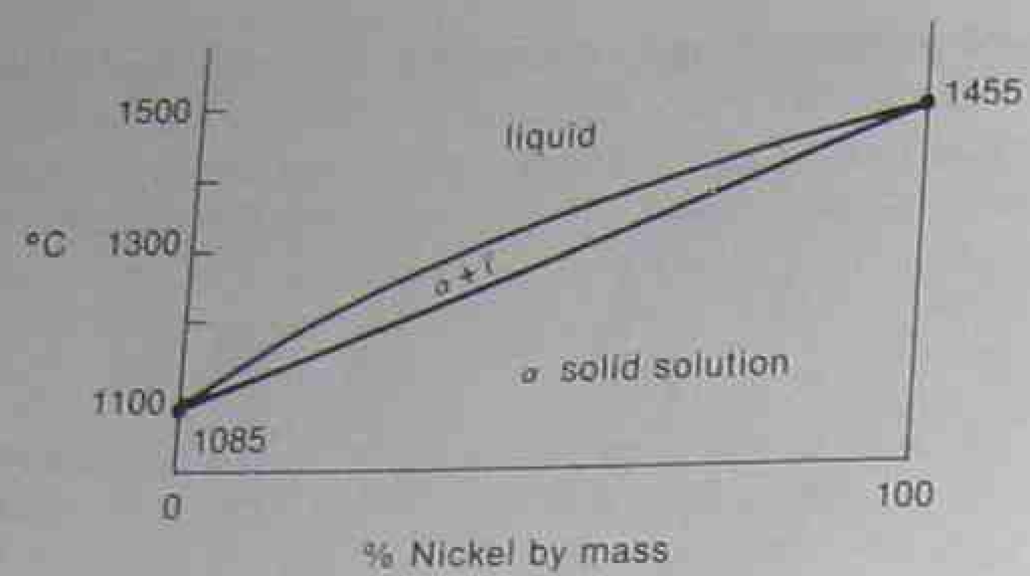


Figure P11–7

grains. If this molten mass were poured into a mold cavity, the composition of solid grains is found to vary from one location to another. Why does this occur?

11–8. If you section a cast structure, you are apt to see a number of voids in the interior. How would you distinguish between those due to gas porosity as compared with shrinkage cavities?

11–9. Figure P11–9 shows a section of an angle plate that is to be cast.
- **(a)** At which of the lettered locations would a *hot spot* be most likely to occur. That is the region to solidify last.
- **(b)** Assuming that an unsatisfactory casting results, specify any changes you would recommend to alleviate this problem.

11–10. The boxlike shape shown in Fig. P11–10 is to be produced with a sand mold.
- **(a)** At which location is a hot tear most likely to occur?
- **(b)** Explain what might be done to avoid that possibility.

11–11. Concerning residual stresses in castings,
- **(a)** Explain why they are usually disadvantageous.
- **(b)** Explain how they might be handled (that is, avoided or removed) after solidification.

11–12. Both pearlitic white and pearlitic gray cast iron are essentially brittle.
- **(a)** With regard to the microstructure of each, why does this brittleness occur?
- **(b)** Using only heat treatments (that is, you *can't* change the compositions), specify how you could bring about a *decided* increase in tensile ductility of these two structures.

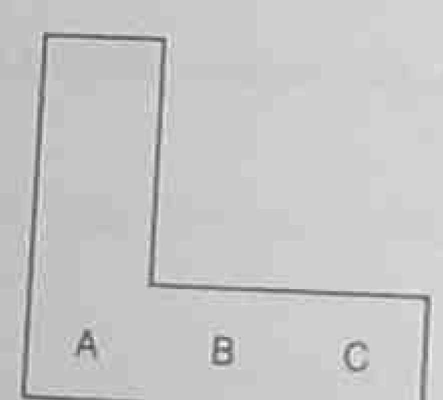


Figure P11–9

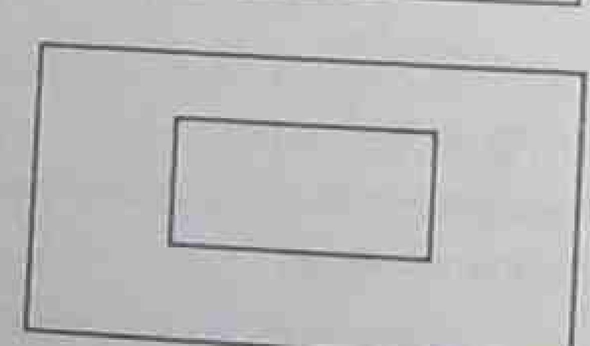


Figure P11-10

11–13. Why might ferritic nodular cast iron be chosen over ferritic malleable iron for a certain part? Consider all possible reasons.

11–14. You are to compare the solidification *times* for a *square* cross section whose height is three times its lateral dimension with a *cylindrical* section of the same height whose diameter equals the side of the square section. Repeat the above if the *volumes* and heights of these two shapes are equal. Equation (11–1) should be used.

12 surface processes

12.1 INTRODUCTION

There are relatively few products on the market that are single components and made of homogeneous materials. Examples include nails, cups made of foamed styrene, concrete blocks, steel beams, and rope. It is instructive to visit a shopping center to see how few such products there are.

The great majority of products are assemblies of two or more obvious and separable components, each selected to fulfill some of the desired attributes of the assembly. For example, a durable shoe is, in essence, a composite structure consisting of a wear-resisting sole attached to a flexible upper segment. The versatility of such products is limited only by the imagination of the designer's knowledge of materials and knowledge of ways to attach the separable parts together. The availability of such products is limited by economics, however, mostly by the high cost of joining materials together. Thus there have always been efforts to achieve desirable properties in single components by making the "surface" different from the substrate. The substrate is usually expected to provide mechanical strength, ductility, conductivity, and several other functions. The surface is expected to perform very different functions, namely, to resist wear and corrosion, and to have an acceptable appearance, among other things. This chapter discusses surface processing, where the intent is to achieve properties different from that provided by the substrate. This chapter does not include methods of surface finishing for achieving texture or topography, but it does include such surface finishing processes as painting.

Surface processes can be classified in terms of *surface treatment, surface modification,* and *surface coating*. Short examples in each of these groups are listed below, with longer discussions following:

1. *Surface treatments* are the processes by which surface properties are changed separately from that of the substrate. Perhaps the most common example is found in steel. A piece of 10100 steel can be annealed throughout to achieve a hardness of 250 VPN (Vickers Pyramid Number). The surface can then be heated to 730°C by a flame or a laser to some shallow depth and cooled quickly to produce martensite of 800VPN hardness. There is no change in chemistry, only a difference in hardness due to heat treatment.
2. *Surface modification* processes are those that change the chemistry of the surface to some shallow depth, ranging from one μm to about three mm. One old method adds carbon to the austenitic form of low carbon steel, by diffusion. When the entire part is then cooled quickly (quenched in water), the substrate remains tough and the surface becomes hard because of the difference in carbon content. A newer method implants nitrogen and other ions into metals with the effect of distorting the lattice structure near the surface, thereby hardening it.
3. *Surface coating* processes *build up* the dimension of some region of a surface. All types of metals, polymers, and ceramics are used as coatings, and they are applied to all types of substrates.

Surface processes are many and varied, and are applicable to virtually all materials. Data on prices and properties for purposes of evaluating these processes cannot be put into a convenient table; available information for specific production problems should be obtained from vendors of the machinery and supplies available for such processes. Unfortunately, surface processes are often advertised in the same manner as one hears for laundry soap, including testimonials from shop foremen and sundry purchasing agents. An interested process engineer should assess processes by testing them on actual production materials. Before such tests, however, it is well to become aware of the fundamental events that take place in each process. These are described in the next sections.

12.2 SURFACE TREATMENTS

Virtually all processes that change bulk properties will also change only the surface properties, if properly applied. The properties of some materials are changed by heat treatment; the properties of others may best be changed by plastic flow. A partial list of the surface treatments is given in two groups, namely, those that use heat and those that plastically deform.

1. Heat treatment is effected by heating at any convenient rate, but by cooling at controlled rates. The major heat sources are listed below in order of potential increasing surface heating *rate*. The higher the rate of heating, the thinner will be the heated layer, where the goal is to reach some specific surface temperature. A thick layer will resist wear and indentation longer than will a thin layer, but a thin layer will produce less part distortion than does a thick layer. Note that processes are often given names that only partially describe what takes place. For example,

laser hardening of steel implies that a laser hardens steel. In fact, the laser only heats the steel, after which fast cooling (usually in water) effects the hardening.

a. *Flame hardening* uses a gas-fired flame, usually oxygen-acetylene, propane, or other high-temperature fuel. This process can be quickly installed, but it is not as readily automated as some others, and it can not be focused upon very small regions on a surface.

b. *Induction hardening* is done by placing a metal into a loosely fitting coil, which is cooled by water and in which an alternating high current (60 Hz up to radio frequency) flows. The current in the coil induces a magnetic field in the metal, which because of magnetic reluctance causes heating in the metal, mostly in the surface at the higher frequencies. The coil current is shut off and cooling water is applied to the part at the appropriate time. This process is clean and readily automated, but it is restricted in its ability to heat specific regions on a surface.

c. *Laser hardening* uses a laser for heating a surface. The usual wavelength is in the infrared, in the range longer than 1000 nm or 1 μm. The CO_2 laser is commonly used. A laser system is expensive to install, but the beam is easily steered or directed along any path on a surface by automatic control of mirrors, even into regions that are out of sight.

d. *Electron beam hardening* uses a stream of electrons to heat a surface. The electron accelerating voltage is usually held below 25 kV when X rays are to be avoided. The beam can be steered by magnetic lens but only in line of sight. Conventional electron beam systems require that the part being processed should be placed into and removed from a vacuum chamber. This usually requires some time and skill to operate and obviates the use of fluids to cool a heated part. At higher cost, one can purchase an electron beam system which directs a beam from the vacuum enclosure through an orifice into the atmosphere where part handling and cooling can be done conveniently. This beam cannot be steered, and thus the part must be moved about under the beam.

Where cooling of a surface is required, after heating, in order to cause a phase change, it may be necessary to do so by quenching in liquid or by spraying liquid on the hot surface. However, a very thin *layer* of heated material will also cool quickly by conduction to the substrate, if the temperature gradient and the thermal conductivity are high enough. For example, the conduction cooling that follows heating by a laser or by the electron beam can be sufficient to produce martensite in 1040 steel, but this will not occur when the surface is heated by a flame.

2. Some plastic flow processes include the following:

a. *Burnishing* involves pressing and sliding a hardened sphere or (usually) roller against the surface to be hardened. It is a rather crude process which can leave a severely damaged surface. Lubrication reduces the damage.

b. *Peening* is done either with a heavy tool that strikes and plastically indents a surface, usually repeatedly, or by small particles that are flung against a surface with sufficient momentum to plastically dent the surface. The latter is called *shot*

peening if the particles are metal of the size of ballistic shot. The velocity of shot or other particles may be as high as 35 m/s: it is, therefore, a very noisy and dangerous process.

c. Skin pass rolling is done with spheres or (usually) rollers, of a diameter and loading such that the surface to be hardened is plastically indented to a small depth. Large rolls will plastically deform thin plate or sheet throughout the thickness, but skin pass rolling can be controlled to plastically deform to shallow depths.

The local plastic flow that occurs in these processes expands an element of material laterally and ''thins it,'' with the effect of developing a compressive residual stress in the surface. A bar that has been shot peened, for example, will bow so that the peened surface will be on the *outer* radius.

The hardness of a surface that has been severely plastically deformed depends on the original ductility of material. Generally, the hardness of an annealed material can be increased by local indentation by a percentage that is twice that of the percentage reduction of area of an annealed specimen in a tensile test.

12.3 SURFACE MODIFICATION PROCESSES

Surface modification processes are those that change the chemistry of existing materials in the surface of the original material. These include the following:

1. *Carburizing* is done to increase the carbon content of steel. The maximum hardness of a piece of steel is related to the carbon content. For structural purposes a steel of less than 0.4 percent carbon is desired for toughness, but for wear resistance and indentation resistance a carbon content of about 1 percent is desired. The carbon content of steel can be increased only when the steel is in the austenitic or face-centered cubic state where the maximum solubility of carbon is about 2 percent (at 1130°C: see Chapter 7). Thus when steel is heated in an atmosphere rich in carbon, some of the carbon will diffuse into the steel. A carbonaceous atmosphere is achieved by using CO, by burning fuel gas with inadequate O_2, or by heating chips of gray cast iron (which usually contain over 2.5 percent carbon). A very rich carbonaceous atmosphere will usually produce a steep gradient of carbon content in the part, which results in large stress gradients and possible cracking during heat treatments. A lean atmosphere adds carbon slowly. The proper depth and thickness of carburized layer is controlled by temperature and atmosphere. However, precautions must always be taken to prevent oxidation, hydrogen diffusion, grain growth of the steel, and undesirable migration of alloying elements in the steel. Carburizing layers of any thickness can be obtained, but the usual thickness is in the range of 1 to 3 mm.

2. *Carbonitriding* may be done either in a gas atmosphere of ammonia diluted with other gas, or it may be done by inserting a piece of steel into a *salt bath,* which is a molten cyanide salt or compound. The cyanide supplies both carbon and nitrogen for diffusion into iron, which itself must be in the austenitic state. The role of the

carbon is described above. The nitrogen that diffuses into the steel forms nitrides with such alloys as aluminum, chromium, molybdenum, vanadium, and nickel, producing a hardness between 900 and 1000 VPN.

3. *Ion implantation* is done in a vacuum of the order of 10 mPa. Many types of ions may be inserted into a wide range of surface materials in this process, but the easiest to describe is nitrogen in iron. Nitrogen gas is ionized in an electric field gradient of 10^5 volts/mm. The ions are then propelled to a high velocity in a field of the order of 100 KeV toward an iron surface held electrically negative. The usual area rate of impingement of ions is of the order of $10^{15}/mm^2$. As ions enter the iron surface, several iron atoms are evaporated from the surface, and a channel of atoms is displaced to receive the nitrogen. The nitrogen concentration builds up to about 15 to 20 atomic percent with a peak concentration at a depth of about 0.7 mm.

   An implanted surface is in a compressive state of stress, which will usually increase the fatigue life of the surface. The surface is also harder, but very thin. Implantation affects the corrosion properties of metals and increases wear resistance for some forms of mild wear.

12.4 COATING PROCESSES

A very significant industry has developed which offers as many as 60 coating processes. Most of the processes can be broadly classified as given below. No attempt is made to name the processes, because in most cases the process is named after the machine that applies the coating, or is given the name of the inventor. In the following paragraphs several processes will be described in terms that will lead to an understanding of the vital information an engineer needs concerning a process, namely, the quality of the product. Information on cost must be obtained from the suppliers of coating service. There are very many suppliers, ranging from the substantial industries to the part-time home-based operation. The broad categories of processes include the following major processes:

1. Weld *overlaying* is done with all of the heat sources mentioned above, but most often by arc and by gas flame. Welding produces very strongly adhering layers, which may be built up to any desired thickness. For corrosion resistance the *filler* or coating material may be a stainless steel, and for wear resistance the *filler* may incorporate nitrides and carbides. Soil-engaging plow points and mining equipment are often coated with steel filler materials containing particles of two forms of tungsten carbide, WC and W_2C, which have a hardness of the order of 1800 VPN.

2. Spraying of molten and semi-molten metals and ceramics is done in air or in *vacuum*. The durability of the product depends strongly on the strength of the bond between the coating and the substrate, which in turn depends on how much of the adsorbed gases, oxides, and contaminants found on all commercial surfaces are removed or displaced so that the sprayed material can bond to the substrate of the target material. Several processes are described:

a. Molten metal, usually aluminum, is sprayed to coat steel pipe and tanks exposed to weather and to coat engine exhaust systems. The metal doubtless begins to travel from the "gun" to the target in the molten state, but some of the droplets cool to the 2-phase region of the equilibrium diagram before they reach the target. This transaction is not instantaneous, because a phase change entails the evolution of some heat. In any case the spray travels at various speeds, usually less than 30 m/s. If the spray is solid, the particles would bounce off the target. Liquid would wet a solid surface and solidify, but 2-phase droplets partially flatten against the target surface and remain attached partly by wetting forces due to the liquid phase of the spray. A "wet" snow ball hurled against a wall behaves the same way. Upon solidification some other bonding mechanisms must be involved, however. Recall that all solid surfaces are covered with adsorbed gases. The hot sprayed metal, upon striking the target surface, will cause desorption of some of the water. A bond is therefore effected between the sprayed metal and the oxide on the metal substrate. Later the sprayed metal contracts and doubtless produces high residual stresses at the bond interfaces, which will limit the adhesive strength of the film to the substrate. But practically, sprayed coatings are fairly durable against very mild abrasion. Their effectiveness against corrosion depends on their continuity. Here again, one can pile drop upon drop from the spray, but the drops must fit tightly together to prevent the incursion of acids and other corrosive substances. Each drop will bond to another through an oxide film, and there will be high residual stresses because of differential contraction from one drop to another.

b. The coating of surfaces for wear resistance is a fast growing industry. One process uses a spray which is produced by feeding a powder into the flame of a gas fired or plasma torch. The powder can be a mixture of dozens of available metals, ceramics, and intermetallic compounds, selected both for cost and effectiveness for resisting wear. The spray velocity is in the range of 150 to 500 m/s, and the sprayed material reaches the target surface again in the semi-molten state. The firmness of attachment, or stress to separate the coating is of the order of 70 MPa, which is adequate for many tasks, but not for severe abrasion. One process achieves a velocity as high as 1300 m/s of particle impingement, by detonation of a fuel gas in a tube containing a powder of the coating material. The high-velocity particles from such a device apparently remove a large amount of adsorbed water and other contaminants. Perhaps there is also an effective packing of particles in the layers of coating. This type of coating appears to have a strength of attachment in excess of 140 MPa, which makes it much more suitable than other processes for abrasion and erosion resistance.

c. Paints and polymers are in a class of coatings usually used for appearance and also for mild corrosion protection, but not for wear resistance. These materials are applied to a surface by spraying, wiping, or rolling of liquid. For effective bonding the surface to be coated must be clean and the liquid coating must wet the solid surface. The coating is then expected to solidify, either by the evaporation of a solvent or thinner from the coating, or by other mechanisms of polymerization of the molecules.

d. Surfaces can be coated by *electroplating*, usually in the range from 0.5 μm to about 0.25 mm thick. The common coatings are chromium, nickel, copper, zinc, cadmium, tin, and molybdenum. Some coatings are hard and provide wear resistance. Some are soft and provide protection against scuffing, while others are well suited to protection against corrosion. The process is done in an acid (electrolyte) bath containing a salt of the metal to be plated (for example, a nitrate, a sulfate, or others). A few volts are applied with the part to be plated as the cathode (−). The plating ion concentration, the bath temperature, and the applied voltage must be carefully controlled to avoid poor adhesion of plating to the substrate, spongy plating, or large crystals in the plating. Overvoltage must be avoided because it produces hydrogen, which usually embrittles some metal. In addition, since the plating thickness is proportional to the current density, some care must be taken in part design, anode geometry, and shielding to make the plating of the proper thickness in all areas.

e. *Electroless plating* is a process that is named such because it was developed to overcome some of the difficulties of electroplating. Coatings of nickel-phosphorous or nickel-boron alloys may be applied to a wide range of metals and alloys. Plating occurs by hydrogenation of a solution of nickel hypophosphite, usually available commercially with proprietary buffers and reducing agents. Coatings of any thickness can be applied. The applied coating has a hardness of ≈500 VPN, and the hardness increases to ≈900 VPN when heated to 400°C for one hour.

f. *Impregnated* coatings are not strictly coatings but are usually classified as such. They are formed by direct contact of the surface to be coated with a solid, liquid, or gas of the desired element. An alloy forms in the surface of the part to be coated, which has different properties than that of the substrate. The catalog of such processes is large, including *calorizing* (Al), *carburizing* (C), *chromizing* (Cr), *siliconizing* (Si), *stannizing* (Sn), and *sherardizing* (Zn).

g. Another process that is not strictly a coating involves the melting of a thin layer of a metal part, and then sprinkling TiC or other hard compounds into the molten layer. Upon solidification the TiC becomes firmly bonded and serves to increase wear resistance.

h. *Physical vapor deposition* (PVD) is a process that is done in a vacuum of about 10 mPa. The coating material is heated and evaporated (boiled). This vapor fills the enclosure and condenses on cooled surfaces, including the part to be coated. Coatings of any thickness up to about 100 μm may be applied. The adhesion to the surface (often called the substrate) depends on the cleanliness of the surface, but PVD coatings are readily rubbed off unless the coated part has been heated for some time, allowing diffusion of some of the coating into the oxide on the part surface.

i. *Chemical vapor deposition* (CVD) takes place in a "vacuum" of about 10 to 100 mPa. The enclosure also contains a gas, which includes ions of the type to be

deposited on the part surface. There usually is sufficient chemical reaction of the coating with the part to effect a bond. Chemical reaction occurs at the surface of the base metal M′, with deposition of the coating metal M. There are three types of reactions:

(1) When the coating medium or vapor is a chloride (for example),

$$MCl_2 + M' >< M + M'Cl_2$$

(2) By catalytic reduction of the chloride at the base metal surface when the treating atmosphere contains hydrogen

$$MCl_2 + H_2 <> M + 2HCl$$

(3) By thermal decomposition of the chloride vapor at the base metal

$$MCl_2 <> M + Cl_2$$

The last reaction appears the simplest, but thermodynamically it is often not possible or very economic. Specialists in these processes should be consulted on such detail.

PROBLEMS

12–1. List 20 items each of the three classifications, surface treated, surface modified, and surface coated among consumer products.

12–2. List 10 items in hardware stores that are single-component items and made of one homogeneous material.

12–3. If austenite of 0.4 percent carbon will transform to martensite at a cooling rate of 600°C/s, what should the minimum temperature gradient be in a bar of steel so that conduction heat transfer will accomplish the formation of martensite?

FOREWORD TO PART C

Here we attempt to tie in costing with particular processes. In Chapter 13, examples are given to illustrate how one can proceed to determine the cost of specific processing operations. This involves the individual costs of labor, materials, overhead, and the like. Attention is also paid to the manner by which such costs might be reduced. In Chapter 14, the idea of integrating design and the method of processing to make the designed part is put forth. A natural consequence of this integration leads to the need to consider alternative methods of processing. In essence, different methods are often available, and a comparison of the costs of each will often dictate which should be used. Two major points are stressed here. First, the most economical method may depend on the use of equipment that is not available, so a decision involving the expenditure of money as an investment in capital equipment enters the picture. Such decisions go well beyond the jurisdiction of the manufacturing engineer; all he or she can do is to present the most accurate cost factors possible, and then it is up to upper management to decide if new investments are to be made. The second major observation leads to the general conclusion that there is no absolutely ''best'' method that provides greatest economy. The quantity or number of parts to be produced almost always plays a key role in deciding which of several methods is most economical.

13 costing of manufactured parts

13.1 INTRODUCTION

This chapter concentrates on procedures for estimating the costs to manufacture parts, restricting our attention to material, labor, tooling, and what is loosely known as overhead. Four methods are discussed, with some industrial examples given.

1. Simple calculations are made of the costs in *cost centers*, where a prescribed manufacturing technique is employed to produce or process a component. The effects of varying the production parameters (speeds of processing, for example) in such cost centers are then investigated, leading to considerations of optimizing the processing conditions through that cost center.
2. The cost of cutting materials with a tool is calculated, and found to be dependent on a balance between the high cost in labor when cutting slowly, and the high cost in tooling when cutting fast.
3. The economics of alternative methods of manufacturing the same shape in the same materials are studied, particularly comparing the modern metal forming process which produces *near net shape*, with the traditional metal removal methods.
4. Costs of competitive routes to manufacture either the *same* overall design of article, or articles of *different* design that will all perform the same function, are illustrated by examples.

Some of the examples below use data obtained several years ago, without converting old data to present-day values. It is a useful exercise for the student to insert only one or two modern cost figures into the given examples in order to gain experience in predicting the effect of possible future variations in prices. For products made in other countries one must consider different divisions of costs of material, labor, and overhead in different countries; also, there are different tax laws, direct and indirect government subsidies, and so on across the world.

13.2 PROCESS PLANS: COST CENTERS

The cost of producing a component is usually estimated from process plans. The complete plan is a detailed list of *what* equipment and material is to be used, and *how much* equipment time, material, and man-hours are required to produce the designed product (or perhaps to perform a service for another cost center). The standards for many products are usually set by industrial engineers who have data both from handbooks and from previous experience of the company. They decide what production rate is reasonable to achieve with the equipment available, with the level of workforce skills, and with the quality of material currently available.

Note that estimates are usually developed for the *existing* available equipment and the way it has been used in the past, and not necessarily for the optimum equipment and tooling. There is wisdom in this practice in some instances, but sometimes it is vital for economic survival to apply specialized knowledge of materials and processes to develop new ways to make old products.

As an example of a process plan for a simple cost center, consider the production of steel plate in a hot rolling mill. The function of the cost center is typically to convert slabs 100 to 150 mm thick into plate 12 mm thick and then to shear the plate into 15-m lengths. After some years of experience, manufacturers develop *cost standards* for the amount of material required to produce a particular amount of prime or salable product, allowing for scrap and scale (mostly iron oxide), and for the equipment time or man-hours required to produce that amount of salable product. Slightly different standards will apply for different width of plates; the wastage in scrap and scale is a smaller proportion, the wider and thicker the plate. There is about 18 percent wastage for 750 mm wide × 12 mm thick hot rolled plates, but only some 6 percent wastage for a 2 m wide × 12 mm thick plates. For the narrow plate 1.18 tons of ingoing slab yields 1 ton of salable plate, whereas 1.06 tons are needed to yield 1 ton of wide plate. Once standards are established, the quantities actually used are compared with the standard quantities. In this way unified comparisons may be made day by day, or between several mills. Note that although the material standards are expressed here in tons, there is often some benefit in using a dollar-equivalent value, particularly if two mills in different parts of the country use different sources for raw materials and the prices vary somewhat independently.

No cost center is in production 100 percent of the time. In the case of a rolling mill, for example, *downtime* or *outage* occurs when rolls must be changed, or when there are mechanical and electrical breakdowns. Operators are allowed time for personal needs and rest time, and there will be time involved for setting the job up, in starting up, and in finishing the process. These factors may have a negligible effect in some cases, but a

significant influence on the overall costing in others. Setup time in complicated machining operations or in welding is usually relatively greater than in other operations such as plate rolling.

Production standards for a rolling mill may either be expressed as allowed hours per ton of salable product (no credit is given for the production of scrap) or, alternatively, its reciprocal, which is production rates (tons per hour). Then the *production performance* of the cost center can be assessed in terms of the ratio of the *earned standard hours* (that is, the standard allowed hours/ton multiplied by the quantity produced) divided by the actual number of hours taken to do the job. Thus, for some given plate thickness and width, if the standard operating practice is 0.007 hour/ton, it should take (0.007 × 500) = 3.5 hours to roll 500 tons of salable plate. If it actually takes only 3.0 hours to do the job, the "production performance" is given by the quotient of 3.5/3.0 = 117 percent.

To determine the standard product cost, the value of the starting material is added to the cost of the labor required to perform the operation. Taking the 1.18 tons of slab at \$105/ton to yield 1 ton of salable rolled product the gross material cost is 105 × 1.18 = \$123.90. But the scrap and (oxide) scale from the 1.18 tons can be recovered and sold for \$12.00, so the net material cost is \$111.90 to produce 1 ton of salable plate. For labor, if the standard operating time is 0.007 hour/ton, and the cost of several employees operating the mill is \$245 per hour, then that cost is \$245 × 0.007 = \$1.72 per ton. The product cost is therefore \$111.90 + \$1.72 = \$113.62 per ton of salable product. Note that labor is only about 1.5 percent of the direct cost of the product. A rolling operation is therefore not *labor intensive* but is *capital intensive*.

To this, one can add the overhead charges (both direct and indirect) associated with the product. For convenience, these are often quoted as a percentage of the labor costs of a batch (assuming some standard production plan), but note that things like factory rent, interest on money borrowed to buy machines, and the like remain fixed when a plant is not operating at full capacity; these moneys have to be found even if there is zero production.

13.3 MACHINING ECONOMICS

In previous chapters, the words *desired tool life* were used, and it is now appropriate to discuss why one particular tool life might be better than another. Practical interest is centered around one of two viewpoints, namely,

1. The combination of parameters (cutting conditions) that leads to a tool life associated with the minimum *cost per piece*.
2. That tool life which leads to the maximum production rate *or* minimum *time per piece*.

At first glance, the above distinction may seem contradictory, but as we will now show, the conditions leading to greatest economy are *never* equivalent to those providing maximum production.

First, consider the viewpoint of cost per piece. The total cost in *dollars per piece* may be designated as C_p and is defined as follows*:

$$C_p = C_i + C_m + C_c + C_g + C_s + C_r \tag{13–1}$$

where

C_p = total cost in dollars per piece.

C_i = idle cost. This includes loading and unloading of the workpiece, plus any other machine handling time.

C_m = cutting or machining cost. This involves *only* the time (and, therefore, the cost) when material is being machined or cut.

C_c = tool changing cost. In essence this involves the time when a tool must be replaced with a freshly ground tool after it has been used to machine a number of pieces. As such, it must be prorated over the number of pieces produced prior to the change.

C_g = tool cost per grind. This includes the cost to regrind a worn tool (the original grind is included) and involves the depreciation of the initial cost of the cutting tool.

C_s = setup cost. This involves the cost to get the machine tool ready for operation. If handled in this manner, this cost is prorated over the total number of pieces produced on the existing setup.

C_r = cost of raw material.

Note that most authors ignore C_r and C_s, while others combine C_c and C_g. Our preference is to break the cost C_p down into as many specific items as possible.

Now consider each of the components individually. On any operation, noncutting time costs money, since operators are being paid hourly and since the department in which the operation is carried out must have some type of *burden*† or *overhead* rate that must be accounted for. Let R_m be the sum of the operator's rate *and* burden rate of the department; its units will be dollars per *minute*. Suppose that the total idle time t_i is the number of minutes required to handle the loading and unloading of the workpiece and, perhaps, alter the feed rate or spindle speed setting of the machine. Then

$$C_i = t_i R_m \tag{13–2}$$

The machining cost is simply the product of R_m and the time it takes to complete the cutting operation; this is t_m and its units are minutes. Then

$$C_m = t_m R_m \tag{13–3}$$

If it is assumed that the same operator changes tools when a worn tool must be replaced, the total cost of changing the tool is simply the product of the tool changing time t_c in minutes and R_m. This total cost must be distributed over the total number of pieces that were produced during the life of the tool. Now the number of pieces produced per tool failure (that is, up to the time the tool is changed) is simply the tool life in minutes divided

* *Each component* has the units of dollars per piece.

† Burden includes maintenance, depreciation, and indirect labor. Data on these items are usually readily available in a given company. Details may be handled differently, *but all industrial* organizations utilize some factor of this type.

by the time in minutes required to machine one piece. In essence, then, we are assuming that a fixed *fraction* of the tool is *used up* to machine each piece. Thus if the tool life T were 60 minutes and the machining time per piece t_m were 2 minutes, then 30 pieces would be machined before tool changing would occur and 1/30 of the total tool changing cost would be attributed to each piece. In symbolic form:

$$C_c = t_c R_m (t_m/T) \tag{13–4}$$

With regard to the tool grinding cost, it is best to think of this in terms of a cutting tool whether it has one edge or many. One aspect of this cost is that associated with the actual grinding of the tool; this requires so many minutes t_g. Since hourly operator and overhead rates may be different for grinding than for the machine operator and his department, the symbol R_g will be used, although its components have the same meaning as used with R_m.

In addition to the actual cost to grind a tool, the original cost of the cutting tool should be prorated over the number of cuts the tool makes, rather than to regard it as part of overhead costs. Each time the tool is reground, some of it is used up; if the initial tool cost is divided by the number of permissible grinds, then this cost D_g added to the actual grinding cost gives the total cost involved to replace a worn tool. Again, as with Eq. (13–4), this cost must be sensibly distributed over the number of pieces produced per tool failure, so

$$C_g = [t_g R_g + D_g](t_m/T) \tag{13–5}$$

Note that the terms inside the bracket can be considered as the cost in dollars associated with a ground tool (that is, \$/tool) where the ratio t_m/T is the fraction of a tool used per piece; thus, the final units are still in dollars per piece.

With regard to the setup cost, consider that it might cost one hundred dollars to assemble and set up all components required to complete a certain operation. Say that after ten thousand parts are produced, the existing setup is broken down and the machine tool prepared for a new operation. The setup time t_s multiplied by the labor and overhead rate, say R_s, would lead to the one-hundred-dollar cost. If that is divided by the *total* pieces machined on that setup N_s, the setup cost per piece becomes

$$C_s = \frac{t_s R_s}{N_s} \tag{13–6}$$

The raw material cost per piece C_r is self-evident. Inserting Eqs. (13–2) through (13–6) into Eq. (13–1), we obtain

$$C_p = t_i R_m + t_m R_m + t_c R_m (t_m/T) + [t_g R_g + D_g](t_m/T) + \frac{t_s R_s}{N_s} + C_r \tag{13–7}$$

Consider now a turning operation where

$$t_m = L/fN \tag{13–8}$$

With $N = 12V/\pi D$ and $VT^n = C$ or $T = (C/V)^{1/n}$ substituted into Eq. (13–7), we have

$$C_p = t_l R_m + R_m(\pi D L V^{-1})/12f + t_c R_m(\pi D L V^{1/n-1})/12fC^{1/n} + [t_g R_g + D_g](\pi D L V^{1/n-1})/12fC^{1/n} + t_s R_s/N_s + C_r \quad (13\text{–}9)$$

To find the cutting velocity that will yield a *minimum cost per piece*, we perform the operation $\partial C_p/\partial V = 0$. This leads to the following form for this velocity V_{cm}

$$V_{cm} = (CR_m^{\,n})/[(1/n - 1)(t_c R_m + t_g R_g + D_g)]^n \quad (13\text{–}10)$$

Since $VT^n = C$, Eq. (13–10) may be easily revised to give the *tool life* for minimum cost per piece. This is T_{cm} or

$$T_{cm} = (1/n - 1)(t_c R_m + t_g R_g + D_g)/R_m \quad (13\text{–}11)$$

If we are interested in maximum production rate (*minimum time per piece*), the expression equivalent to Eq. (13–1) is

$$t_p = t_l + t_m + t_c(t_m/T) + t_s \quad (13\text{–}12)$$

Note that the terms associated with grinding and raw material in Eq. (13–1) have no counterparts in Eq. (13–12). Taking $\partial t_p/\partial V = 0$, one finds

$$V_{tm} = C/[(1/n - 1)t_c]^n \quad (13\text{–}13)$$

and

$$T_{tm} = (1/n - 1)t_c \quad (13\text{–}14)$$

Note that in Eq. (13–10) if the cost connected with grinding approaches zero, V_{cm} approaches V_{tm}. A similar comment applies regarding T_{cm} and T_{tm}. Of course, tools and grinding are never free, and that is the key reason why the most economical tool life (or velocity) will not also provide the maximum production rate. Figure (13–1) shows a qualitative plot of the components for minimum cost per piece. Since Eq. (13–14) will *always* predict a shorter tool life than will Eq. (13–11). The optimum velocity and cost for maximum production will be higher than that for the minimum cost condition.

Now a practical word about the most sensible use of the ideas and results just presented. Whether we use Eq. (13–11) or Eq. (13–14), it is usually a relatively simple

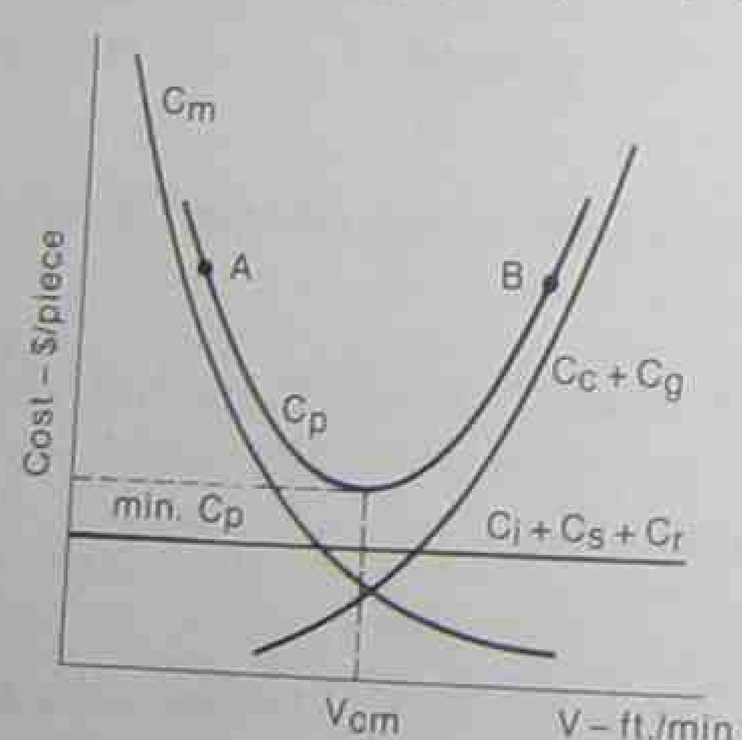


Figure 13–1 Components in the analysis of minimum cost in cutting.

task to obtain the necessary *standards* related to t_c, R_m, t_g, R_g, and D_g for a given operation, and a reasonable estimate can be made for the exponent n since much data are available. However, the machine operator does not set the tool life; rather he sets the rpm which produces a cutting velocity. Thus it is Eqs. (13–10) and (13–13) that would be most useful *if* the value for the constant C were available.* Often, this is not the case and instead of making wild guesses, it is best to use the actual production machine as an experimental tool while it is physically producing parts. For example, suppose that the value of T from Eq. (13–11) is found to be 75 minutes. By starting with some reasonable velocity, parts can be produced and the tool life of the operation can be found. Suppose the tool lasts only 30 minutes. A modest lowering of spindle speed should be made (remember how sensitive tool life is to small changes in velocity) and the new tool should be timed to failure. In a few such tests one can find the most appropriate spindle speed available that will produce a tool life close to the desired 75 minutes. Because many machine tools *do not* possess variable speed spindle drives, but rather a number of discrete values of the spindle rpm, it may be impossible to obtain the exact velocity to produce the desired tool life. Instead, one wants to get as *close* to the optimum speed as possible and not be operating at speeds that are much too high or low, as indicated by points A and B on Fig. (13–1). After all, the exactness of the value used for n is not without flaws, and the reality of this fact should be remembered at all times.

Note from Fig. (13–1) those components that are unaffected by changes in velocity but which must be considered in Eqs. (13–7) or (13–9) when the *actual cost* per piece is desired. Finally, if more than one tool cuts at the same time, the *tool life for minimum cost* increases in comparison to that which prevails for one tool cutting. In effect, the total tool cost increases with multitool operations and, as we see in Eq. (13–11), this necessitates longer values of T_{cm}. Although such operations are difficult to analyze directly, the above principle will hold and should be kept in mind.

Example 13–1

1. The cutting velocities for minimum *cost* per piece and minimum *time* per piece are to be determined for a turning operation where a feed rate of 0.008 ipr and depth of cut of 0.075 in. are involved and high-speed steel tools are to be used. For this operation, the following are known:

Machine operator rate = $13.00 per hour
Machine department overhead = $15.00 per hour
Tool grinding rate = $10.00 per hour
Grinding department overhead = $12.00 per hour
Cost of a 0.5 in. by 0.5 in. by 3 in. long HSS tool = $ 3.75
Number of regrinds per tool = 15
Tool regrinding time = 3 minutes
Tool changing time = 2 minutes
Tool life exponent = 0.10

* This would come from an equation such as Eq. (9–8) for the particular feed and depth *being used*. Alternatively, Eq. (9–17) may be used.

$$VT^{0.1} = 175$$

Idle time t_i = 1.5 minutes

Solution For minimum cost per piece, Eq. (13–10) is used. There,

$$R_m = (13.0 + 15.00)/60 = 0.467 \text{ \$/min}$$
$$t_c = 2 \text{ min}, \ t_g = 3 \text{ min}$$
$$R_g = 3(10.00 + 12.00)/60 = \$1.10/\text{edge}$$
$$D_g = 3.75/15 = \$0.25/\text{edge},$$

so

$$V_{cm} = \frac{175\ (0.467)^{0.1}}{[\{(1/0.1) - 1\}(2 \times 0.467 + 3(1.10) + 0.25)]^{0.1}} = 112 \text{ fpm}$$

For minimum time per piece, Eq. (13–13) is used, so

$$V_{tm} = 175/[((1/0.1) - 1)2]^{0.1} = 131 \text{ fpm}$$

2. Now suppose a *throwaway* sintered carbide tool is used. One type is a thin blank of carbide that is ½ in. by ½ in. square and ⅛ in. thick.* At each corner, a nose radius is usually ground, so with four corners and two faces (top and bottom), eight cutting edges are provided by this tool as purchased. The blank is clamped in a special tool holder such that the form of the holder produces the tool signature. When one cutting edge fails, the blank is unclamped, rotated 90 degrees to provide a fresh cutting edge, and then reclamped. After all eight edges are used up, the blank is *literally* thrown away—thus the name. Note there is *no* regrinding done on such a tool. Here, the machine operator and burden rates are the same as in part 1 and

Cost of throwaway = \$2.50
Number of edges = 8
Tool changing time = 0.75 min

$$VT^{0.2} = 400$$
$$R_m = 0.467 \text{ \$/min as before}$$
$$t_c = 0.75 \text{ min}$$
$$t_g = 0, \ R_g = 0$$
$$D_g = \$2.50/8 = 0.3125 \text{ \$/edge}$$

Then

$$V_{cm} = \frac{400\ (0.467)^{0.2}}{[\{(1/0.2) - 1\}(0.75 \times 0.467 + 0.3125)]^{0.2}} = 283 \text{ fpm}$$

and

$$V_{tm} = \frac{400}{[\{(1/0.2) - 1\}(0.75)]^{0.2}} = 321 \text{ fpm}$$

* Such throwaway tools are also made in triangular and circular cross sections.

The fact that the velocities using carbide tools are much higher than their HSS counterparts indicates why production rates (or RMR) are much greater when carbides are used. Note that the cost of the special toolholder is included in the overhead rate in this analysis. Since such a toolholder can be used for many years, it is simplest to handle it in this way. The same comments apply to the toolholder used with the high-speed tools.

Example 13–2

Using the necessary information from Example 13–1, determine the minimum cost in dollars/piece for both situations if

C_r (material cost/piece) = \$0.85
C_s (setup cost/piece) = \$0.05
D (work diameter) = 2.5 in.
L (length of cut) = 10 in.

Solution 1. For HSS tools:

$$N = 12V_{cm}/\pi D = 12(112)/2.5\pi = 171 \text{ rpm}$$
$$(T_{cm})^{0.1} = 175/V_{cm} = 175/112 = 86.74 \text{ min}$$

With Eqs. (13–1) or (13–7),

$$C_i = t_i R_m = (1.5)(0.467) = \$0.7005$$
$$C_m = t_m R_m = 10(0.467)/(0.008)(171) = LR_m/fN = \$3.4137$$
$$C_c = t_c R_m\ (t_m/T) = 2(0.467)(10)/(0.008)(171)(86.74) = \$0.0787$$
$$C_g = [t_g R_g + D_g](t_m/T) = [3(1.1) + 0.25](10/(0.008)(171)(86.74) = \$0.2992$$
$$C_s = \$0.05$$
$$C_r = \$0.85$$
$$C_p = \$0.7005 + \$3.4137 + \$0.0787 + \$0.2992 + \$0.05 + \$0.85$$
$$C_p = \$5.39 \text{ per piece}$$

2. For the throwaway carbide:

Only C_m, C_c, and C_g differ from part 1, since they are the only terms influenced by the different velocities.

$$N = 12(283)/2.5(\pi) = 432 \text{ rpm}$$
$$T^{0.2} = 400/283 \text{ or } T = 5.64 \text{ min}$$
$$C_p = 10(0.467)/(0.008)(432) = \$1.3513$$
$$C_c = 0.75(0.467)(10)/(0.008)(432)(5.64) = \$0.1797$$
$$C_g = (2.5/8)(10)/(0.008)(432)(5.64) = \$0.1603$$
$$\text{So } C_p = 0.7005 + 1.3513 + 0.1797 + 0.1603 + 0.05 + 0.85$$

or

$$C_p = \$3.29 \text{ per piece}$$

Note that with HSS tools there is a higher C_m (machining) and higher C_p (grinding), but a lower C_c (tool changing).

13.4 COSTS OF JOINING METHODS

13.4.1 Arc Welding Costs

Arc welding processes are fairly consistent, in that the time and material required to produce a given length of weld, by a given welding method, is reasonably constant between different jobs. Thus it is easy to establish standard costs. The cost of producing a weld is essentially that of materials, labor, and overhead. Material costs in welding include the cost of electrodes,* fluxes, shielding gases, and other consumables. The cost of electricity for arc welding is either included with the ''consumables'' or it may be included in the overhead charges, depending on the accounting procedures of different companies. (About 1.75 kWhr of electricity is required to deposit one pound of weld, that is, 3.85 kWhr/kg).

The cost of welding is conveniently expressed in terms of dollars per unit run of weld (formerly per foot, nowadays per meter). Thus we need to know

1. The time to produce a given length of weld (which, with the labor rate, gives the labor cost per length).
2. The amount of electrode used together with the electrode cost per unit weight (formerly per pound, nowadays per kilogram) to give the electrode cost.
3. The quantity of flux and/or shielding gas required, together with relevant unit costs (if the flux is separate from the electrode).
4. The overhead rate, expressed usually as a percentage of the labor cost. As with other metal working processes, it is difficult to make generalizations on overhead as it depends on local accounting methods, but figures ranging over 50 percent to 400 percent of the labor cost per job are found between different organizations.

In these types of calculations it is necessary to bring in the efficiency in the use of materials and labor:

1. Not all the electrode is actually deposited as weld metal; some electrode weight is lost in spatter and vaporization. Some portion of the weight of electrode for manual welding is the flux, and the unused or ''stub ends'' of these electrodes are discarded. The ''deposition efficiency'' is the percentage of the weight of original electrode actually deposited as weld metal. Typical values of the factor for various welding processes are

Stick-electrode welding	65 percent
Self-shielded flux-cored welding	82 percent
Arc welding with shielding gas	92 percent
Submerged arc welding	100 percent

2. An operator does not, in practice, weld all the time that he is paid. This leads to the so-called ''operating factor'' defined simply as the ratio of the arcing time to the total time for which the welder is being paid. It is expressed as a percentage usually, but also as a decimal (<1.0) in the costing calculations which follow. The operating factor depends on the particular welding process being employed, and also on the management and organization of the fabrication shop. The usual range is between 20 percent and 60 percent, with even lower percentages being found for constructional welding in the field and some higher values for automated welding in the shop. Reasons for the different values are as follows:

   a. *Manual welding*. The operating factors range from 10 percent to about 45 percent; 65 percent represents the extremes of human achievement. In addition to the usual personal breaks, there are additional breaks caused by setting up equipment, preparing the job for welding, shifting of the working positions of the operator, changing electrodes, and chipping off slag (flux) to check the weld or to prepare for a second welding pass. Since deslagging can take much time, there are considerable advantages in using electrodes of easy slag detachability.

   b. *Semi-automatic welding*, for which the typical range of operating factors would be 25 percent to 60 percent. Generally, the biggest advantage of the semi-automatic process is the absence of electrode changing and reduced need to shift operator position. These factors will increase the operating factor by between 5 percent and 15 percent. For many applications, a higher welding current and increased travel speed can be utilized; with bare wire inert gas welding, there is no deslagging to be done, which can significantly improve the operating factor.

   c. *Automatic welding*. With fully mechanized automatic welding, it is possible to work at an operating factor of over 90 percent, provided the setup time and deslagging times are very small.

In all of these processes, poor quality welding reduces the operating factor considerably. Poor welds are repaired by grinding away sections of defective weld and rewelding, which requires much time and is chargeable to the cost center.

For common designs of weld in most metals, handbooks and sales and product brochures list the time required to run the weld, the cross-sectional area of the weld, and hence the weight of electrode required to make the run, and the quantity of solid flux and cover gas that will be required for the job. Where data are not available, the deposition rate may have to be taken on test welds.

Calculation of cost may be done with the following equations and the symbols listed in Table 13–1.

$$\text{cost of electrode/ft of weld} = (WE)(CE)$$
$$\text{cost of flux/ft of weld} = (WF)(CF)$$
$$\text{cost of gas/ft of weld} = (VG)(CG)$$

* Most arc welding processes supply filler metal in wire form, where the arc is formed between the filler wire and the surface to be welded. In manual arc welding the wire is available in short rods or sticks, frequently coated with *solid flux, and usually referred to as electrodes.*

TABLE 13–1 Symbols used in calculations of the cost to weld.*

A	=	Cross-sectional area of the weld, (in.^{2})
CE	=	Cost of electrode (or filler wire), ($/lb)
CF	=	Cost of flux, ($/lb)
CG	=	Cost of shielding gas, ($/cu ft)
CL	=	Cost of labor and overhead, ($/ft or weld)
CM	=	Cost of consumable welding materials, ($/ft of weld)
CR	=	Cost of labor and overhead, ($/hr)
D	=	Deposition rate, (lb/hr)
DE	=	Deposition efficiency, weight ratio of weld metal deposited to electrode consumed
OF	=	Operating factor, ratio of productive time to total time required by the welder
OH	=	Overhead
M	=	Melt-off rate of the electrode, (lb/hr)
S	=	Speed of electrode travel, (ft/hr)
T	=	Time required to weld, (hr/ft)
VG	=	Volume of shielding gas required per linear foot of weld, (ft^3/ft), or (ft^3/hr)
WE	=	Weight of electrode (or filler wire), (lb/ft)
WF	=	Weight of flux required, (lb/ft)
WW	=	Weight of weld metal deposited, (lb/ft)

* Taken from the *Lincoln Electric Welding Handbook*.

so that the materials cost per foot run of weld is*

$$CM = (WE)(CE) + (WF)(CF) + (VG)(CG) \tag{13–15}$$

It is convenient to combine the labor rate and overhead into one parameter *CR* (when overhead is based on labor charges), in which case

$$CL = (T)(CR)/(OF) \tag{13–16}$$

where *T* is the time required to run a foot of weld and *OF* is the operating factor. We note that $T = 1/S$, where *S* is the welding speed. Thus as an alternative expression,

$$CL = (CR)/S(OF) \tag{13–17}$$

It is important to be aware that in the case of multiple-pass welds, the total time required to complete the joint may be obtained from the "average speed," which is the *harmonic* mean of successive passes, that is,

$$S = \frac{n}{1/S_1 + 1/S_2 + \ldots} \tag{13–18}$$

where S_1, S_2, . . . are the speeds of the first, second, . . . welding passes, after *n* total passes over the same length of run. Hence

$$T = \left(\frac{1}{S_1} + \frac{1}{S_2} + \ldots \frac{1}{S_n}\right) \tag{13–19}$$

*Note that the term *(VG)(CG)* would be zero in submerged-arc or self-shielded electrode welding, since no shielding gas is used. Similarly, the term *(WF)(CF)* would be zero with all welding processes other than submerged-arc. Note also that, as an alternative, *(VG)* may be given as cubic feet per hour rather than per foot of weld.

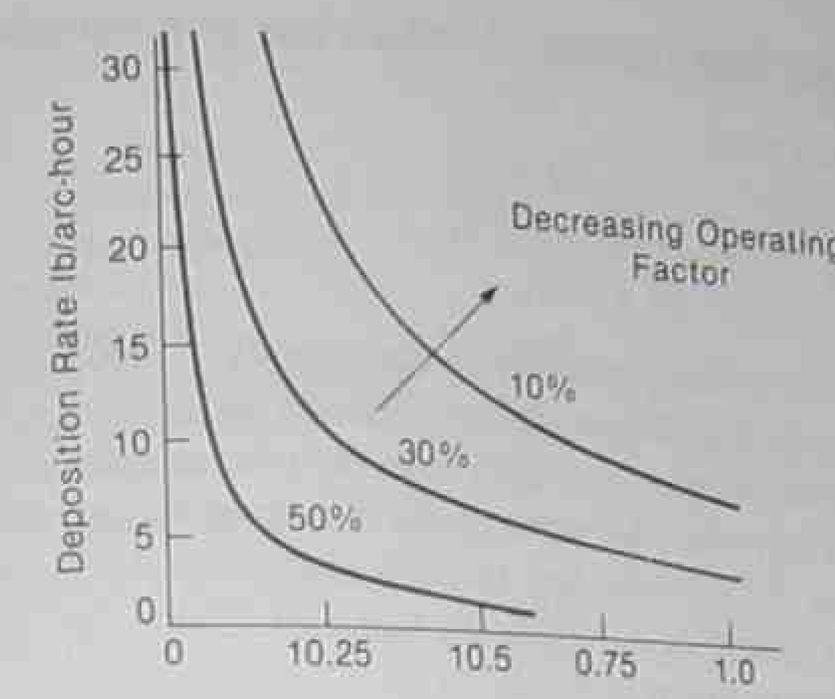


Figure 13–2 The relationship between metal deposition rate and labor cost.

Owing to the inverse relation between welding speed and time to effect a weld, the relation between metal deposition rate and labor costs takes the form shown in Fig. 13–2; the deposition rate (which depends directly on welding current and electrode size) is given by *S(WW)*. At high deposition rates, the labor costs for each pound of weld metal deposited are relatively small; at low deposition rates the labor costs are relatively higher, and the effect of a low operating factor becomes important.

Although the amounts of consumables required to produce a given weld will vary with different processes, it turns out that the cost of filler metal does not vary widely between different welding processes when the deposition efficiency is taken into account. Table 13–2 shows the cost of consumables per foot of weld making a ¼-in. fillet by various methods, and it is seen that all are about 5 cents/foot, in 1973 prices. Consequently, estimations of *comparative* welding costs may be quickly made by using Eq. (13–17) relating to labor, overheads, welding speed, and operating factor.

Equations are also available to calculate costs. We may take as an example a ¼-in. horizontal fillet weld, using a 7/32 in., E7024 electrode. For a speed of 85 ft per hour and

TABLE 13–2 ESTIMATED COST* OF CONSUMABLES FOR FOUR WELDING PROCESSES

Process	Type of Electrode	Electrode Cost $/lb	Deposition Efficiency	Cost of Deposited Metal	
			%	$/lb	$/ft
Shielded metal arc	E7024	0.191	65	0.29	0.045
Self-shielded flux-cored	E70T-G	0.315	82	0.38	0.059
Submerged arc	EL12	0.199 + flux	≈100	0.34	0.053
Gas metal-arc	E70S-3	0.243 + flux	92	0.31	0.048

*The volume of shielding or cover gas ranges from 25–45 cu ft/hr, and the maximum amount of solid flux is a weight about equal to the weight of deposited metal.

an operating factor of 30 percent, with labor plus overhead rate of \$7.00 per hour, we have

$$CL = (CR)/S(OF)$$
$$CL = (7.00)/85(0.3)$$
$$CL = \$0.274/\text{ft} \qquad (13\text{–}20)$$

The use of equations gives a welding engineer an opportunity to study the effect of changes in the fixturing and material handling in the job. Perhaps the operating factor can be increased to 40 percent by such changes, in which case the cost as calculated by Eq. (13–20) becomes \$0.206/ft. But there may be some doubt that an operating factor of 40 percent can be achieved by using stick electrodes, so semi-automatic processes may be proposed with the expectation of achieving an operating factor of 50 percent. The resulting costs from the same equation are given in Table 13–3.

As an example of the influence of different types of filler metal supply, let us compare the use of E8018-C3 (a low-hydrogen stick electrode coated with iron powder) and 9000 C-1 (a flux-cored wire) to weld low-alloy steels. The relevant data and conclusions are shown in Table 13–4, where it may be seen that the lower cost of stick electrodes is seriously offset by the cost of labor.

Figure 13–3 gives a compounded chart which demonstrates the relationship between various processes, electrode sizes, deposition rates, operating factors, hourly labor rates, and labor costs. The welding processes are divided into manual, semi-automatic, and automatic processes; the approximate ranges of operating currents for each electrode size are also incorporated. To make cost calculation more accurate, the deposition rates and welding current relations used for the various processes should be taken from accurate graphs.

Figure 13–3 consists basically of three parts, which can be used separately or all together:

1. *Welding processes.* The left-hand section demonstrates the variation of the deposition rates of the various welding processes with electrode size and current. The different vertical columns are the range of deposition rates for particular sizes of

TABLE 13–3 ESTIMATED COST TO WELD FOR VARIOUS PROCESSES

Process	Arc Speed (ft/hr)	Operating Factor	Labor and Overhead (\$/hr)	Weld Cost (\$/ft)
Shielded metal-arc	85	0.30	7.00	0.274
Shielded metal-arc	85	0.40	7.00	0.206
Gas metal-arc semi-automatic	80	0.50	7.00	0.175
Self-shielded flux-cored, semi-auto.	100	0.50	7.00	0.140
Submerged arc semi-automatic	110	0.50	7.00	0.127

TABLE 13–4 COMPARATIVE COSTS OF USING TWO TYPES OF FILLER STOCK

Factor		E8018 Electrode and Manual Welding	9000C-1 Wire and Semi-automatic Welding
Labor and overhead rate	(*CR*)	\$20.00 per hour	\$20.00 per hour
Deposition rate	(*D*)	3.11 lb/hr	5.00 lb/hr
Operating factor	(*OF*)	0.30	0.45
Cost of labor and overhead to deposit weld		\$21.43/lb of weld	\$8.88/lb of weld
Electrode cost	(*CE*)	\$0.667 per lb	\$1.019 per lb
Deposition efficiency	(*DE*)	0.68	0.78
Cost of deposited metal		\$0.98/lb	\$1.31/lb
Shielding gas flow rate	(*VG*)	none	40 cu ft per hour
Shielding gas cost	(*CG*)		\$0.09 per cu ft
Cost of shielding gas			\$0.72/lb dep. metal
Cost of welding		\$22.41/lb dep. metal	\$10.91/lb dep. metal

electrode, the different shading representing the different processes; the numbers alongside these columns are the maximum and minimum currents for the electrode size. The deposition rates are the points of mild steel weld metal deposited in one arc-hour, taking into account spatter and stub end losses and the like.

2. *Welding speeds*, shown in the right-hand section, relate the joint completion speeds (in./min) and times (arc-minutes for each foot of joint) to the deposition rates for varying weights of weld metal required for the joint. The lower horizontal scale (arc-minutes/ft) is the reciprocal of the upper scale, multiplied by 12.
3. *Labor costs*, shown in the bottom right-hand section, are related first to the operating factor and second to the hourly labor rate.

An example of utilization of this chart is given by the dotted line *a-b-c-d-e-f*. By using a ¼-in.-diameter conventional manual metal-arc electrode with solid flux attached at about 320 amperes, a deposition rate (point *b*) of 6 ¼ lb/arc-hour will be obtained. For a weld operation which requires 0.6 lb of weld metal per foot (point *c*), the welding speed (or joint completion speed in the case of multipass welds) will be just over 2 in./min, or alternatively, the joint completion time will be 5 ¾ arc-minutes per foot (point *d*). With an operating factor of 20 percent (point *c*) the time to complete 1 foot of weld is 28.75 minutes or 0.48 hours; the labor costs at \$8 an hour will be \$3.83 per foot of weld (point *f*).

As mentioned earlier in this chapter, there is often a conflict between the equipment *available* in the factory to do a job and *more desirable* equipment which would be better for the job. Thus elementary cost calculations of the sort given in this section may suggest that one process is far cheaper than another, but equipment for the cheapest process may not be available and would have to be bought. One means of making more reliable comparisons between welding processes is to add to the direct welding costs that portion of the capital cost attributable to depreciation of equipment.

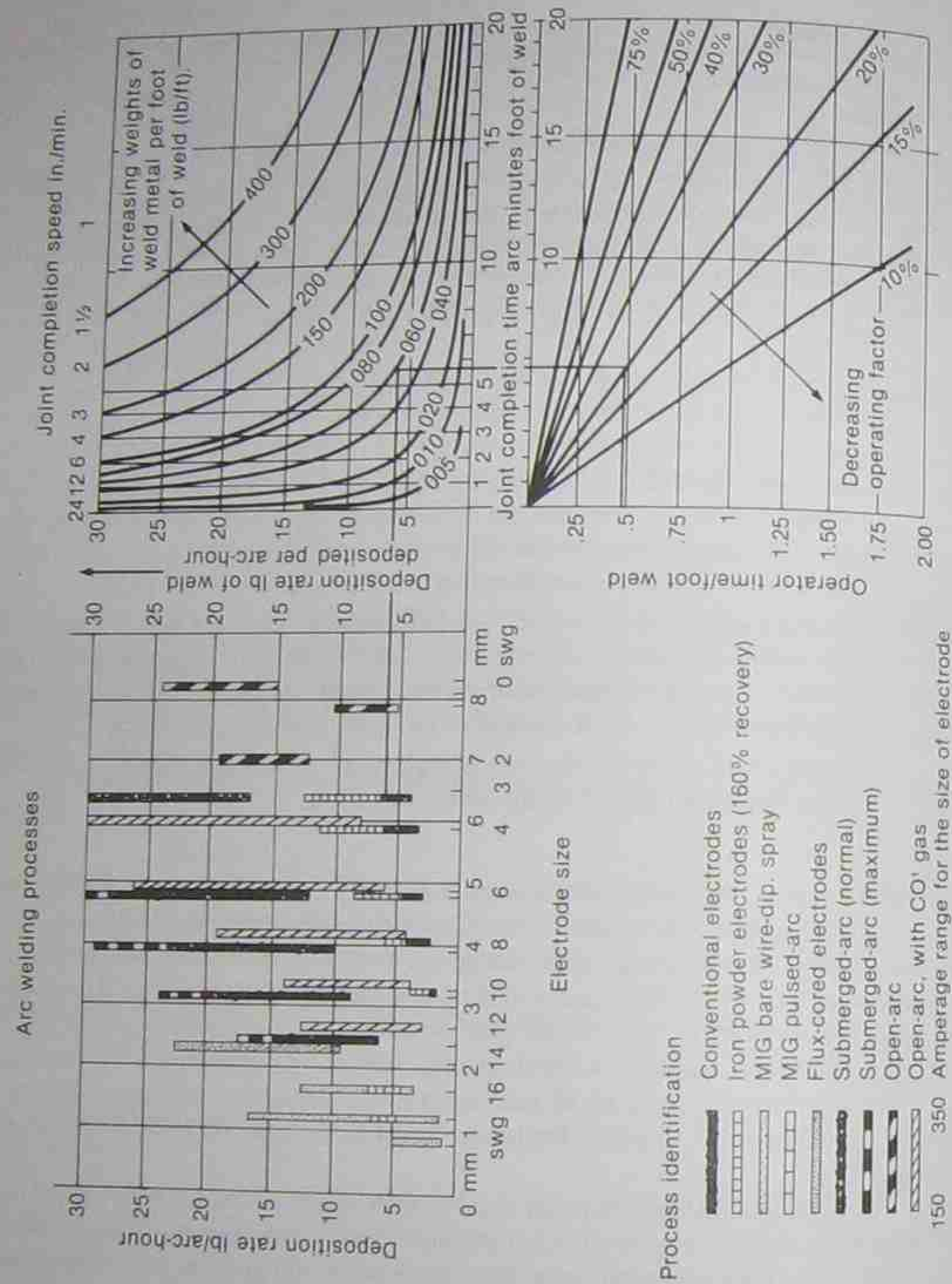


Figure 13-3 Weld cost comparison for welding mild steel.

13.4.2 Adhesive Bonding Costs

Half of the total costs in manufacturing relate to costs of assembling parts together, which fact encourages innovation in joining methods. A great deal of assembling is now done with the use of "structural" or "engineering" adhesive compounds (hereafter referred to simply as "adhesives"). Apart from providing desirable product attributes, there are often cost savings to be gained by using adhesives, because parts to be assembled with adhesives need not be made to high accuracy.

For example, in the assembly of automobile gear shift levers, a metal sphere with a hole in it is slid onto a steel shaft and held in place by an adhesive, thereby replacing forging and machining operations. One can also choose between types of adhesives, as in one example of taillight assemblies. A hot-melt adhesive bonding process replaces an acrylic adhesive, which increases the production rate by about a factor of ten because the hot melt holds the parts together immediately upon cooling, whereas the acrylic adhesive sets by a more time-consuming evaporation of solvent. This reduces the labor costs and produces considerable overall savings, even though the hot melt adhesive costs about 30 percent more than the acrylic.

Another study involves the making of a box overarm for a small horizontal milling machine. Competing proposals were to make it from (1) cast iron, (2) welded steel, and (3) channels and I-beams adhesively bonded together. Section sketches of the three proposals are shown in Fig. 13-4. The traditional method of casting has several inherent disadvantages, one of them being that the section thicknesses must be governed by casting technology rather than by strength requirements. Casting is an old industry; it produces low-cost products, but it is slow and relatively inflexible. The manufacture of machine structures by building up of readily available bars and plates, and so forth, has some advantages, but it was not considered seriously in the past because of difficulties in fabricating the built-up overarm. For instance, welding produces residual stresses which causes structural distortions, a major drawback when accuracy is of prime importance. These residual stresses may be reduced by heat treatment, which is an unwelcome extra production operation, especially for large machines. Other fabrication techniques such as bolting and riveting are also impracticable and cause unequal stress distribution in the

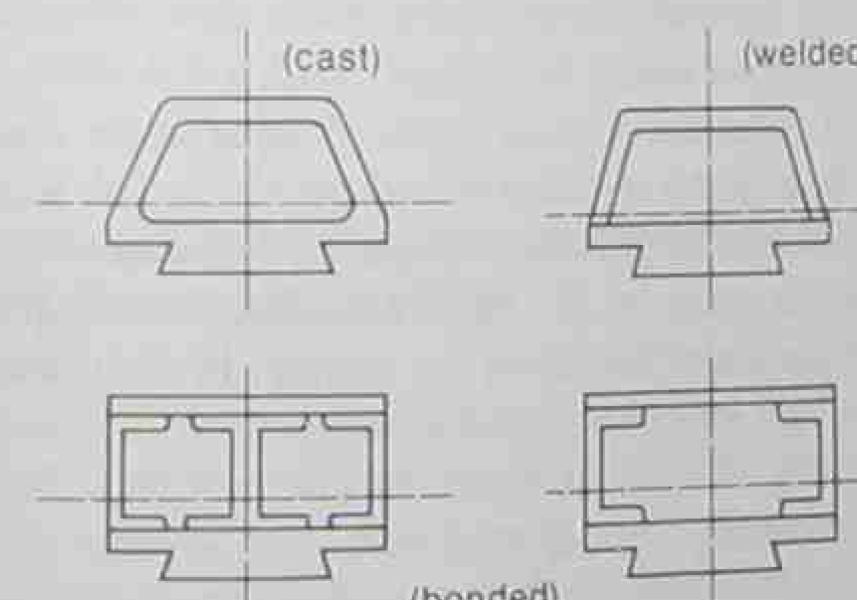


Figure 13-4 Cross sections of different overarms.

TABLE 13–5 ESTIMATED COST TO MAKE OVERARMS BY THREE METHODS

Overarm	Cast Iron	Welded Sections	Bonded Sections
Cost	$85	$75	$44

joint. However, adhesive bonding can be used to good advantage in the manufacture of machine tool structures, and when it is combined with other manufacturing techniques, new possibilities in fabrication arise.

Realistic costs for innovations are difficult to find, since costs of all products depend on the scale of production. Estimates of the production costs of single overarms of the three types, as would be done by a shop that is not equipped for multiple production, are given in Table 13–5.

These 1975 figures cover material, labor cost, and overhead. For all cases the capital cost of equipment for efficient volume production—that is, jigs, fixtures, and the casting pattern—was excluded. When the cast-iron overam is produced in large quantity, its cost drops to $36. The cost of the welded and adhesively bonded types is not very much affected by numbers, since these processes are highly labor-intensive, in which relatively little can be saved by special jigs or tools. These cost figures cover production only. They do not take into consideration any interest to be "paid" on the cost of stock levels required for avoiding delays in delivery, which are particularly high for cast components and low for bonded ones. It would appear that bonding should be particularly attractive when the components are very large and/or when a large number of them is required at short notice.

13.5 COSTING OF NEW PROCESSES: ORBITAL FORGING

When new methods of manufacturing are proposed, accurate cost calculations may not be possible, particularly if related processes are not now used. An example is making pulley blanks from a cropped billet (short cylinder) by "orbital forging." In this process a cylindrical "slug" is progressively worked outwards to give the general shape shown in Fig. 13–5. A more conventional method would be to press blank a disk with a large hole in it, into which a hub is pressed. The attraction of orbital forging is the elimination of cleating the inner surface of the hole in the disk and then pressing the hub into that hole. The major disadvantage is the slowness of the orbital forge. The example also illustrates a not-uncommon situation in which many of the basic data for the new process were lacking at the time of a study, in particular the life of the tooling. In such circumstances,

Figure 13–5 The pulley blank used in the example of orbital forging.

a "base case" is set up with some assumptions, and sensitivity analyses are carried out to investigate the significance of the various assumptions made.

Total material costs (including the hub cost and cost of the cleating operation in the traditional route) are less for the orbital forging method, as the following calculations show for one particular size pulley:

	Gross	Scrap	Net
1. *Blanked and cleated*			
(a) Weight (*g*) (disc only)	228	85	143
(b) Material cost ($/ton)	225	55	
(c) Unit material cost (¢) (a) × (b)	5.14	0.46	4.68
(d) Hub (purchases) (¢)			8.01
(e) Cleating (¢)	0.7		
(f) Total cost (¢)			13.24
2. *Orbitally forged*			
(a) Weight (*g*) (includes hub)	271	49	222
(b) Material cost ($/ton)	231	55	
(c) Unit cost (¢)	6.26	0.26	6.00

There is thus a saving per unit of (13.24 − 6.00) = 7.24¢ with regard to material for this particular size pulley.

Set against the reduced material cost, however, is the cost and life of the tooling on the orbital forge, and the cycle time of the operation on the forge. In general terms, if tool lives are short and orbital forge cycle times are long (that is, low production rates), the savings on the materials side will have to be large to justify the new method, whose costs must include depreciation of the new equipment and so on. On the other hand, if tool lives are long and cycle times are short, the materials savings can be less. Clearly it is required to establish the relationship between tool life and cycle time for various assumed materials savings costs. In this way, options for different size pulleys can be accommodated. Figure 13–6 plots the results of extended calculations for other assumed tool lives.

For the particular size pulley for which material savings of 7.24¢ were calculated above, it appears that orbital forging will be viable under any tool life/cycle time combinations to the upper left of a contour about midway between the 7¢ and 8¢ contours on the graph.

13.6 A METHOD OF ACCOMMODATING PART COMPLEXITY: SIMPLE FORGED PARTS

There are many manufacturing processes that are too complicated to analyze in useful detail. For such operations seemingly illogical cost figures in handbooks are given in terms such as ¢/lb. An example to be discussed is that of forgings. The process begins with a piece of metal, known as the billet, which is of crude shape and the proper weight. It is set onto a "lower" die, which is a piece of metal into which a shape has been cut and

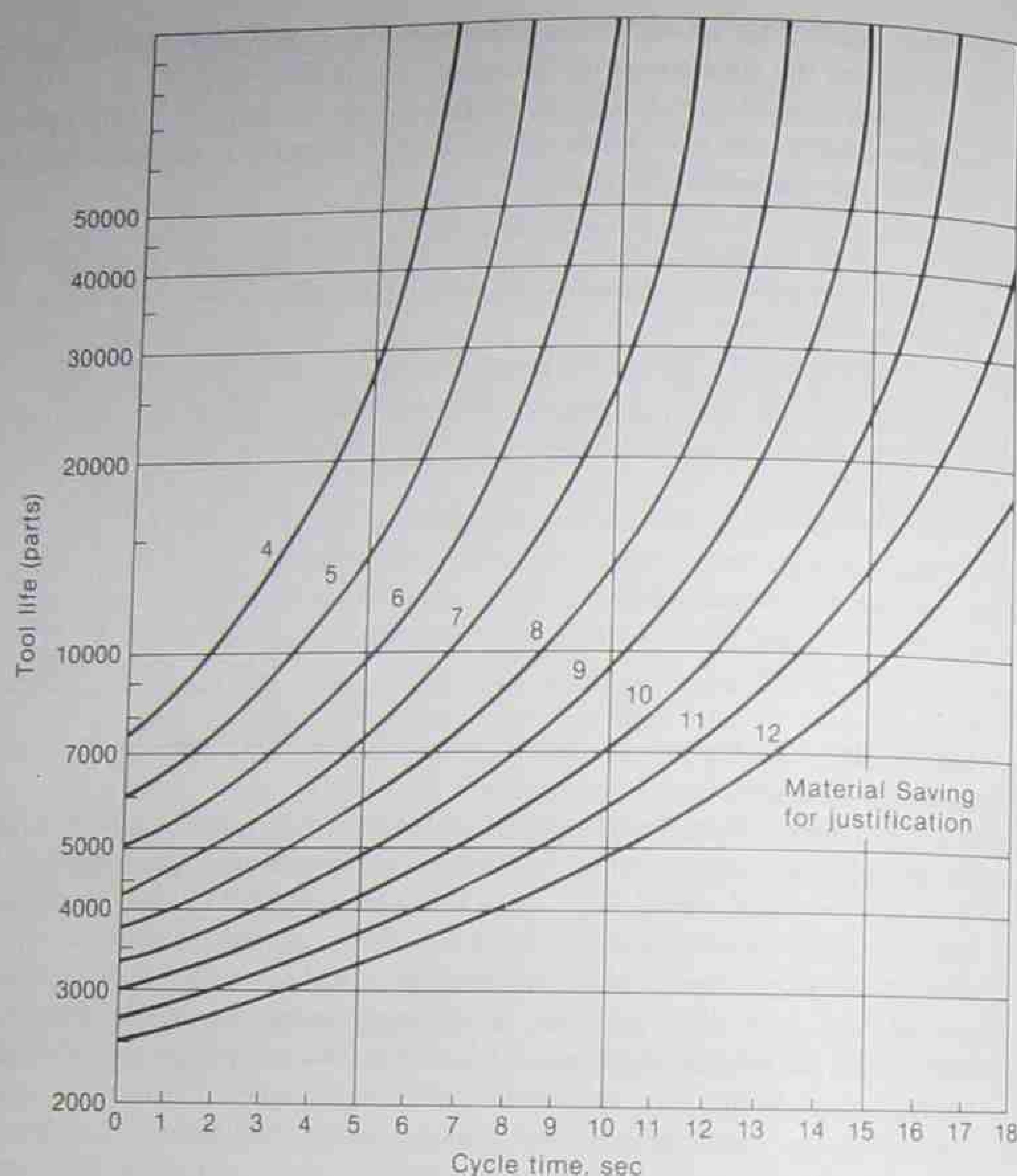


Figure 13–6 Economic factors in orbital forging.

ground. The upper die also contains a carefully made cavity. The closed die will have the shape of some desired part. The upper die may be forced against the billet slowly and with great force in a press, or it may be attached to a weight and dropped onto the billet. With continued squeezing or with repeated hammering, the billet is deformed into the shape of the cavities in the die segments. However, the billet is never perfectly located in the die, so there may be excess metal in some places and insufficient metal elsewhere. The solution to this problem is to use a billet with excess metal and to provide an area in the die set for excess metal to flow in all directions. This excess is called the *flash* or flashing, and it must be removed before final processing of the product.

It can be seen that forgings can be made in many different materials and in a wide range of complexity. Thus, a single cost to forge in terms of ¢/lb may not be realistic. An

inexperienced product designer would have some difficulty in estimating a realistic cost for forgings. To aid in this exercise Poli and Knight* have developed a ''classification of component features,'' and ''materials to be used,'' to assess the relative costs implied in alternative design and processing routes for forged components.

The elements to cost in their forging model relate to those listed earlier:

1. material (including wastage)
2. capital charge on equipment
3. direct labor
4. heating and furnace costs
5. handling
6. set up costs
7. die costs
8. overheads

On average the material cost represents approximately 50 percent of the cost of forged parts, die costs around 15 percent and direct labor 15 percent. Material costs will increase, for a given workpiece material, with the complexity of the part, as more wastage (in the form of flash and so forth) will be necessary to produce the required shape. Die costs will increase as the number of dies is increased and the die life is reduced, which will be influenced by the complexity of the shape and the material being forged. Direct labor costs will increase with the size of the crew required to operate the press or hammer (that is, as the capacity of equipment increases). The precise proportion of each of these cost contributions to overall forging costs will vary from plant to plant, dependent to some extent on local practice and methods.

Since, in forging, setup time is a relatively small part of component cost, it is neglected, and the calculations condense down to costs (per unit produced piece) of

(i) material (allowing for scrap), K_m
(ii) ''equipment operating costs'' (that is, items 2, 3, 4, 5, and 8 above grouped together to represent the ''most usual practice,'' with appropriate furnace handling equipment and crew size), K_p
(iii) die costs, K_D

Using their symbols for convenience, we can express these components as

$$\text{(i)} \quad K_m = V\rho C_m K_i \qquad (13\text{–}21)$$

where V is the volume of the part, ρ the material density, C the material cost per unit mass, and K_i the ratio of gross-to-net weight for the particular forging in question.

$$\text{(ii)} \quad K_p = C_p/P_R \qquad (13\text{–}22)$$

where C_p is the operating cost of the equipment per unit time and P_R is the production rate, that is, (\$/time) / (pieces/time) = \$/piece.

*C. Poli and W. A. Knight, *Design for Forging Handbook*, (Amherst, Mass.: University of Massachusetts, 1981).

(iii) K_D = die cost/no. of parts in die life (13–23)

where n/D is the number of sequential dies required to do the job at the forging cost center. The costs and lives of a series of dies employed sequentially in the complete forging operation may all be different, which complicates matters, so in order to simplify calculations it is customary to assume an average die cost C_D and an average die life N_L. Then

$$K_D = n_D C_D N_L \tag{13–24}$$

The total unit cost to manufacture is the sum of $(K_m + K_p + K_D)$.

The detail of these calculations relies on information about blow rates, factors that influence initial die cost, and factors that influence die life, among other things, as discussed in the following:

For presses, which are comparatively slow-acting, the number of blows required is normally the same as the number of operations required, excluding flash removal. For hammers several blows per operation are used and an average of around three blows per operation is usual. The relative blow rate of the machines should be the effective values, that is, reflecting the handling of the forgings between operations, rather than the ratios of the maximum cycling rates. Thus for very large presses and hammers the effective rates will be governed largely by the speed with which large parts can be manipulated. Hammers tend to be used in preference to presses when large changes in workpiece cross section are required.

Die costs can be considered to be made up of

1. The material of the die blocks
2. The cost of machining and finishing the die cavities.

In general, for small dies, the latter makes up the major proportion of total cost, whereas for large dies, the material cost is more significant. Consequently, as the size of the part is increased, the proportion of the total cost attributable to the die material increases. It is reasonable to assume that the quantity of die material will be influenced predominantly by the size and material of the forging, and the effect of increasing the complexity of the part will be to alter mainly the machining cost of the dies.

Die life is influenced by a number of factors, including

1. The material to be forged.
2. The shape of the part.
3. The tolerances applied to the part.
4. The forging equipment to be used.

The cost of manufacturing a simple disc-shaped part from low-carbon steel has been chosen as the yardstick against which all other costs are referenced. Using the additional subscript 0 to refer to this basic part, we have the total cost made of component costs

$$K_{T0} = V_o\rho_o C_o K_{fo} + C_{po}/P_{Ro} + n_{Do}C_{Do}C_{Do}/N_{1o} \tag{13–25}$$

Alternatively, Poli and Knight* refer to the relative values of K_m, K_p, and K_D taken separately. That is, for materials

$$K_{mR} = \frac{V\rho C_m K_f}{V_o \rho_R C_m K_{fO}} = \frac{V\rho_R C_R K_{fR}}{V_O} \tag{13–26}$$

where K_{mR} is the cost of material in the forging of interest relative to the cost of material in the basic disc-shaped part, ρ_R is the density ratio of the (possibly different) materials, C_{mR} is the material cost relative to that of the reference low carbon steel, and K_{fR} is the relative ratio of gross-to-net weights.

Likewise

$$K_{pR} = \frac{C_p P_{RO}}{C_{pO} P_R} = \frac{C_{pR} N_p}{R_{pR}} \tag{13–27}$$

where C_{pR} is the relative operating cost compared with the equipment used to produce the reference disc part, N_p is the number of blows on the forge required to produce the part, and R_{pR} is the relative blow rate of equipment used compared with that of the forge used to produce the reference disc part.

Again

$$K_{DR} = \frac{n_D C_D N_{LO}}{n_{DO} C_{DO} N_L} = \frac{n_{DR} C_{DR}}{N_{LR}} \tag{13–28}$$

where n_{DR} is the relative number of dies compared with the standard operation, C_{DR} is their relative collective cost, and N_{LR} is their relative collective life.

If, on the basis of fully costing out the basic part in a given factory on given equipment, it appears that the breakdown of total costs K_{TO} is 50 percent for material costs, 30 percent operating costs, and 20 percent for die cost, then the total cost K_T of any other forging is given by

$$K_T = K_{mR}K_{mO} + K_{pR}K_{pO} + K_{DR}K_{DO}$$
$$K_T = K_{mR}(0.5K_{TO}) + K_{pR}(0.3K_{TO}) + K_{DR}(0.2K_{TO})$$
$$K_T = (0.5K_{mR} + 0.3K_{pR} + 0.2K_{DR})K_{TO}$$

that is, the total relative cost per piece is

$$K_T/K_{TO} = 0.5K_{mR} + 0.3K_{pR} + 0.2K_{DR}$$

(for the assumed initial 50 percent/30 percent/20 percent breakdown of K_{T0}).

An example of the application of these ideas relates to the part shown in Fig. 13–7. It may be demonstrated that the ribs are of such proportions as to be considered closely spaced and consequently of relatively high forging difficulty. Altering the design so that the ribs are less severe or so that a peripheral rib only is used for stiffening would lead to a reduction in forging difficulty. Table 13–6 shows the effect of these design alterations on the relative forging costs estimated by using the described procedure. In this way the

* Poli and Knight, *Design for Forging Handbook.*

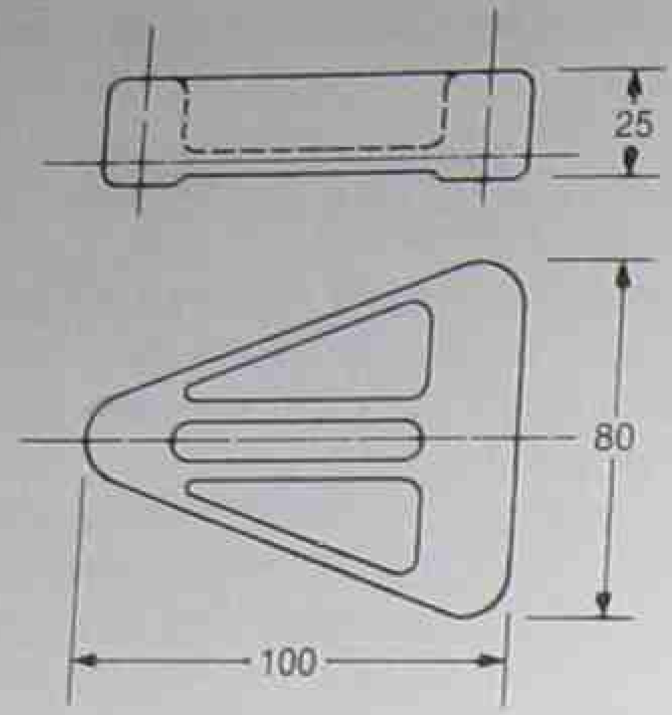


Figure 13–7 Flat nonround part with severe ribs—shape code 282.

product designer is given an indication of the relative forging costs of the different designs and hence the cost benefits of various design improvements.

TABLE 13–6 ESTIMATED FORGING COSTS FOR VARIATIONS IN DESIGN OF THE PART

	Relative Material Cost	Relative Production Cost Hammer	Relative Production Cost Press	Relative Die Cost
Original design	0.41	9.0	8.8	58.0
Redesign with less severe ribs	0.38	6.0	6.3	25.0
Redesign with peripheral ribs	0.36	5.6	6.3	18.5

14 the economics of integrated design and manufacturing

14.1 INTRODUCTION

In this chapter, we discuss how the original design of a product can influence the choice of what manufacturing process is most sensible. Ideally, design, materials, and processing should be considered as complementary and integrated activities.

Established rules, by which a component or assembly of components, may be designed to perform a desired function are well established. These relate, for example, to stressing, shape and proportioning, choice of material, environment in which the component will perform, and the like. The degree of complexity of this sort of design analysis depends on circumstances, ranging from elementary strength of materials calculations for simple components, to complicated finite element/finite difference calculations for crucial components in nuclear plant applications. Many books cover the functional performance aspects of design, and we shall not repeat the basics here. We shall assume that appropriate rules are being followed and that these rules may be different for components made from metals, polymers, or ceramics. But functional aspects of design are only part of the full picture. It is not much good if a brilliant design, incorporating some splendid material, cannot be *made, assembled,* or *fabricated,* especially in an economic sense. It seems less well-known that there are rules and procedures that should be followed so that the component can be *successfully* made and/or assembled.

How a part is going to be made, and how it may fit with another part should not be left until the last step of the overall design procedure. Exact reproduction of a part as-designed may prove to be a difficult and costly, if not impossible, task. Some slight alteration in the original design can often permit easier manufacture without detracting from the functional performance requirements. This pertains at both intermediate stages of manufacture, when semi-finished products are being produced, and at finishing stages, regardless of the process to be used.

The overall process of design should involve proper consideration of functional design, materials, and manufacturing method, as shown in Fig. 14–1. It is an interactive situation, but with few exceptions, these different areas have usually been treated separately both in industry and in academia. But there is a growing realization that this is not the best approach, and they are being brought together in more and more establishments.

There can be many combinations of the three elements of design, materials, and manufacture without there being one *best* route from the drawing board to the market-place; it depends on circumstances. If one, or even two, of the three elements are fixed (for example, a fixed design with fixed materials), there may still be a number of acceptable alternative manufacturing routes to produce the desired design; this may depend on the number of parts to be made, the availability and quality of raw material, the number and capability of existing machines in the factory, and, of course, the resulting cost per part. Alternative designs for different manufacturing procedures are usually possible. Design changes may be required if CNC manufacturing methods rather than more traditional ones are used.

The in-service properties of the manufactured design may be influenced strongly by the production route employed. For example, a complicated shape, machined from the solid, will display the properties of the initial solid, whereas the same shape produced by forging will usually have properties that differ from the raw material, especially if cold working is involved. In many cases, desired in-service properties may be attainable only by certain thermomechanical processing sequences. Sometimes, it may be nearly impossible to manufacture an article using particular materials owing to production difficulties. Materials which are extremely difficult to machine or weld can introduce severe problems.

Clearly, there must be some interplay between what the designer specifies and what can actually be achieved economically by certain processes. Specifications of tolerances and surface finishes required, for example, must be considered when different manufacturing methods are to be compared as alternative methods of production. Linked to such considerations is the question of overall economics. Savings arising from one or two cheap operations early in a production sequence may be lost altogether if expensive finishing operations are required later.

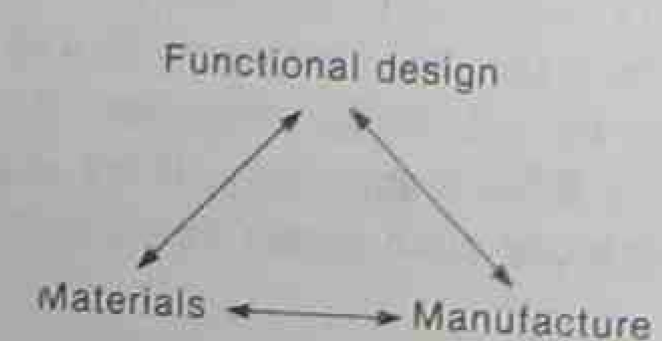


Figure 14–1 Interactive considerations in the design and manufacture of a product.

14.2 CUTTING OR FORMING: NEAR NET SHAPE

Expensive machining operations involve grinding and contour milling, and there will be clear advantages, in setting up production schedules, if the need for these two types of operations are avoided. On the other hand, drilling and turning typify less costly machining operations. Thus, in casting, for example, it is often more economical to drill holes subsequently instead of using cores for holes in the basic casting, since this can be a more expensive operation. Yet in the general field of design for production of metallic components, there is growing awareness that *all* metal removal processes are really wasteful in terms of time, money, and energy consumption. One example relates to the production of engineering components of complex shape from expensive raw material, where massive machining can involve the removal of as much as 98 percent of the metal in producing gas turbine components made from nickel alloys. Figure 14–2 shows typical types of advanced engineering components which have to be machined from solid stock having appropriate wrought metallurgical structure. Their shapes are so complicated (and they may "twist" in and out of the paper) that it would be impossible to forge them between closed dies owing to the difficulty of extracting them from the dies after the forging is completed. Open-die forging would not achieve the awkward shapes of these types of components with the precision needed for subsequent finish machining. Hence they are made with massive machining operations. In the case of some highly stressed parts of

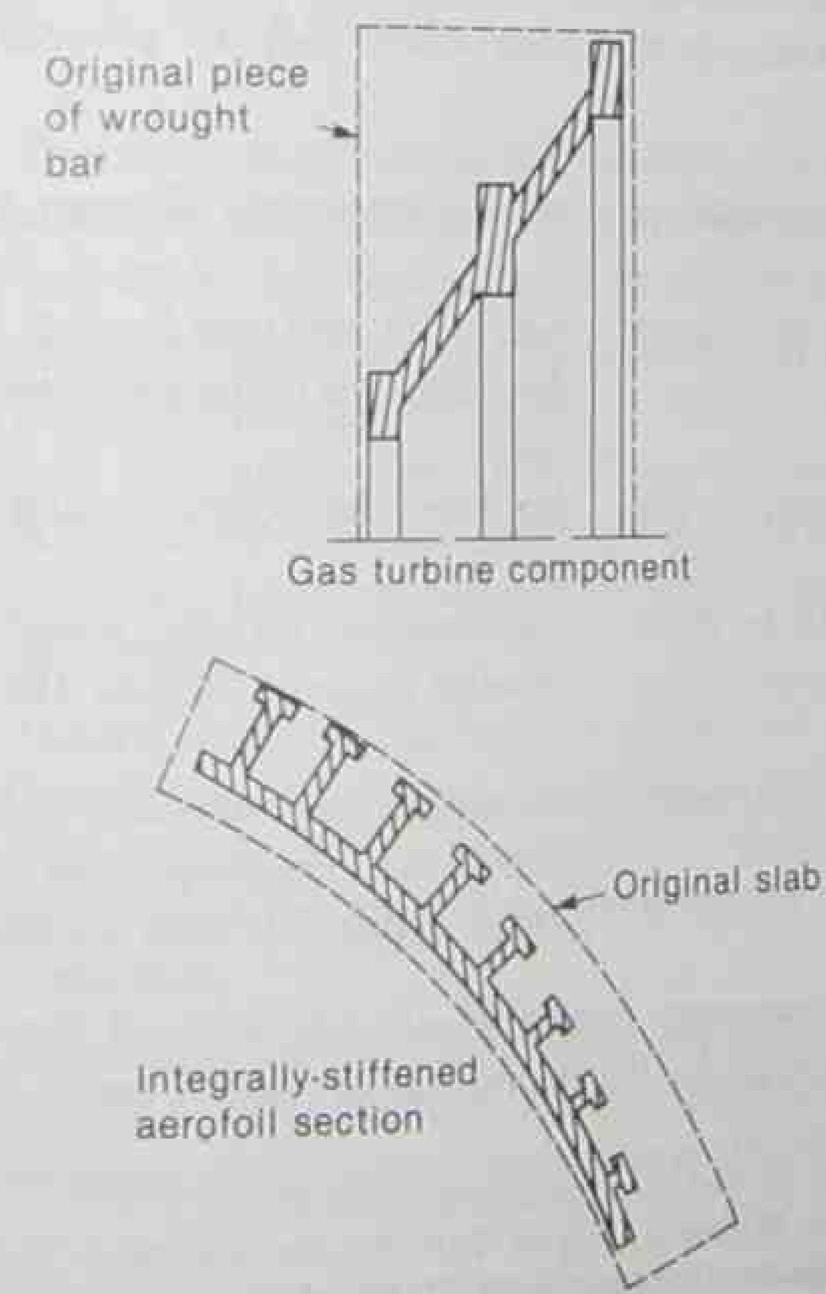


Figure 14–2 Types of components that require machining from solid stock.

TABLE 14–1* TYPICAL ENERGY CONSUMPTION AND MATERIAL UTILIZATION FOR VARIOUS PROCESSES

Process	Energy Required—MJ	Material Utilization—%
Cold or warm extrusion (forging)	41	85
Hot, closed-die forging	46–49	75–80
Process casting	30–38	90
P/M pressing and sintering	28.5	95
Metal removal processing	66–82	40–50

*Source: K. H. Kloos, VDI Berichte, Nr 277 193 (1977).

complicated shape which are not excessively large, such as turbine blades and connecting rods, precision casting processes may be used, if the desired metallurgical structure can be achieved. This is not always possible, so machining has to be used.

Forging and casting processes are examples of so-called *near net shape* processes, while true precision casting and powder metals methods are called *net shape* processes. Material wastage is greatly reduced, as illustrated in Table 14–1, which also includes information on the energy required per tonne* of finished product. For reference purposes, comparable data are included for traditional metal removal processes. It can be seen that conventional cutting methods consume about twice as much energy and at the same time display only half the material utilization of the forming methods.

These efficient net shape forming methods have been usefully classified into five categories.† The relevant chapters in this book where they are covered are also indicated:

1. Forming by solidification of molten or semi-molten metal	casting processing (Chapter 11)
2. Forming from solid billets by material displacement	hot, warm, and cold forging, extrusion, wire drawing, tube making and ring rolling (Chapter 8)
3. Forming from sheet material by displacement or perforation	blanking, deep drawing, and ironing (Chapter 8)
4. Forming from powder metal by pressing, sintering, and (sometimes) deformation	conventional powder metallurgy, coining, powder forging (Chapter 10)
5. Forming by building up components on a mandrel or former	electroforming, chemical and physical vapor deposition, spraying, spray peening, weld metal deposition with forging (Chapter 12)

Note that category 5 incorporates comparatively new ideas of *deposition forming*, in which a component is built up by the deposition of molten metal followed by the

* Note that tonne equals 1000 kg mass, or approximately 2200 lbm.

† Michael Neale and Dennis Waterman, Report TRS 289 for Science & Engineering Research Council/Department of Trade and Industry. (Farnham, Surrey, U.K.: Michael Neal and Associates, 1982).

immediate hot working of the solidified deposit. This idea is clearly the antithesis of metal cutting.

It will be evident from later coverage that there may be a cutoff or breakeven point, in terms of quantity of parts produced, beyond which it may be cost-effective to employ metal forming methods in place of metal removal methods. Such an alternative would be used even more widely were forming processes more accurate in meeting design requirements, thereby involving only limited finish machining or even satisfying the necessary precision directly. Questions of tolerances and surface finish, both desired and attainable, are discussed in Sec. 14.5

14.3 ELEMENTARY CONSIDERATIONS OF DESIGN FOR MANUFACTURE

Since metal-cutting methods have been with us for so long, one might think that designers would know by now how to design parts that are amenable to machining with no complications. But this is not the case, and parts to be machined are sometimes designed without regard to the limitations of the machine tools intended to be used. A number of considerations should be borne in mind, among which are (1) allowing ample space for the cutter, (2) adequate provision for chip removal, and (3) keeping the number of separate cutting operations as small as possible, all of which can reduce costs. Strasser* made the following points:

1. Tool holders, noncontacting peripheries of spinning cutting wheels, and other unavoidable obstructions in machine tools must clear the workpiece. Parts designed with ample tool clearance can be machined rapidly and without damage.
2. Machined chips, whether discontinuous or continuous, must be removed from the cutting tool/workpiece interface. Often, a coolant or a stream of air can be used to help the chips escape. Detail features of some components, such as long or blind holes, do not lend themselves to easy chip removal, and provision must be made in the design to accommodate a buildup of chips in order not to reduce the efficiency of cutting.
3. Material handling can become a significant part of the cost of a machined product. Scheduling and shifting a workpiece from one machining operation to another adds to the cost of the finished product. To keep machining costs down, the part should be designed to minimize the number of separate operations on the workpiece.
4. The volume of material removed and the distance the cutting tools travel should be kept as small as possible, in order that time and material are saved. In castings and forgings, these objectives may be achieved by including recesses.

* F. Strasser, *"Designing Parts that are Easy to Machine"*, Machine Design, (Aug. 1975), p. 65.

The Smallpiece Education Trust* published a list of do's and don'ts and a number of interesting examples of "before" and "after" designs which well illustrate "design for production" problems. All the component designs were the work of experienced engineers; however, they lacked manufacturing experience. While the examples were acknowledged to be rather elementary, they were intended to be just that because there is much more scope for redesigning for production, simple items which are usually produced in much larger quantities than more complicated parts. Furthermore, complicated parts and assemblies often incorporate numerous simple components.

In many cases, relatively minor changes from the original design, which do not in any way affect functioning of the component, facilitate machine setting, operation, and inspection to such an extent as to enable at least half the machining cost to be saved.

A number of related examples of changes in design to facilitate easier manufacture are given in the 1974 edition of the Fulmer Materials Optimizer.† The point is made that it is sensible to see whether a given massive part can be subdivided into components that are individually more easily made. For example, Fig. 14–3 shows a die-cast pulley wheel which, had it been made "solid," would have meant either casting a solid rim (bad practice owing to large changes in cross section) followed by machining out the V-groove, or the manufacture of a complicated die in which the pulley groove would have been formed by a withdrawable-mold section. Instead, the same result can be achieved more economically by forming two dished castings which are then riveted together with integrally cast pins. Advantage can also be taken of the possibility of welding two or more castings or forgings together to give final components. This is clearly of value when the size of the component exceeds the capacity of the forge or foundry, and smaller parts can be designed appropriately to make up the final article.

A specialist casting operation called "insert work" can simplify manufacturing procedures and eliminate alternative expensive routes. Figure 14–4 shows a universal ball joint made by *in situ* casting of a zinc alloy housing around an existing sintered steel ball; the ball is loosened in the housing by a blow as part of the trimming action after casting. Complicated precision-machining of the housing has been avoided.

An example which illustrates the need for the designer to be thoroughly familiar with different methods of manufacture, and their associated costs, is that of the pro-

* In *Machinery and Production Engineering*, 312, (Feb. 1968).
† Fulmer Research Institute, 1974, Slough, U.K.

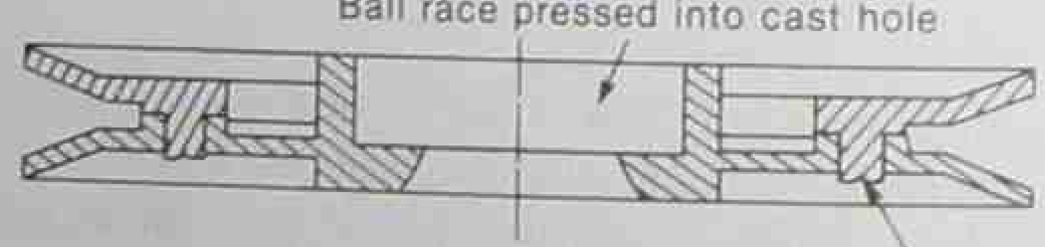


Figure 14–3 Schematic of a pulley wheel.

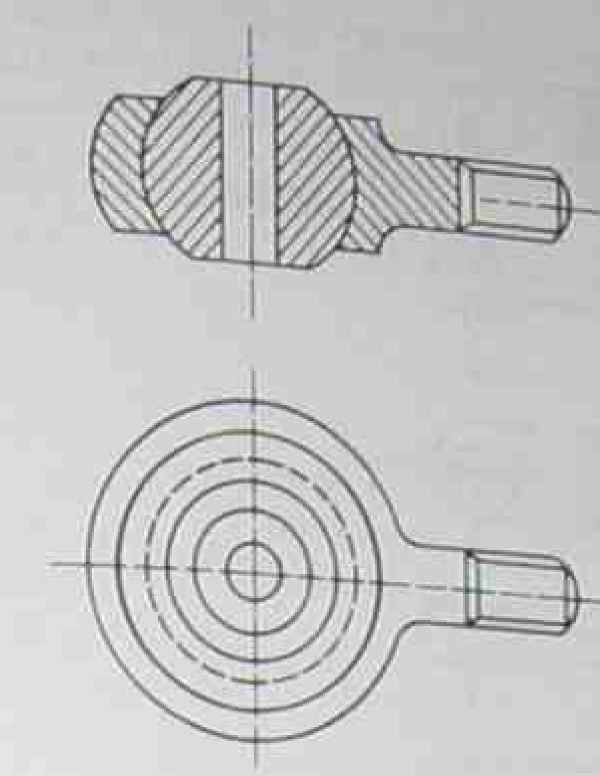


Figure 14–4 Universal ball joint made by casting a zinc alloy around a steel ball.

duction of a simple rocking arm by various methods.* Figure 14–5(a) to (d) shows different detail arrangements which permit manufacture by (a) casting, (b) drop stamping, (c) hand forging, and (d) welding. Although particular economic conditions are involved, if we *presume* that all of the capital equipment required for the separate methods of manufacture are *available*, welding is cheapest for batches up to 75, casting for batches of 75 to 700, and drop stamping for batches greater than 700. Figure 14–5 brings out an important point, namely that the cost curves for different processes *intersect*, so there are different ranges of numbers over which different methods are most economical. As we shall discover, this is a recurring theme in assessments of alternative methods of manufacture.

A number of particularly interesting examples of producing the same shape design by different methods concerned the employment of the fine-blanking process, in place of ordinary blanking combined with subsequent machining which is necessary to transform the sheared edges up to adequate standards. Fine blanking avoids the jigging and clamping inaccuracies which can arise between operations by the conventional route.

Table 14–2 gives details of the times taken to manufacture, by various blanking methods, the plunger shown in Fig. 14–6. Progressive fine-blanking tooling eliminated the time-consuming shaving operation required to achieve cleanly sheared surfaces. As the chamfers are coined before shearing, the costly milling operation could also be excluded. From Table 14–2 it can be seen that approximately 456 minutes of manufacturing time were saved in the production of 1000 parts.

The extra tooling costs (780 hours for building the progressive fine-blanking tool) were amortized after the production of 100,000 to a maximum of 200,000 parts, depending upon the hourly rates involved. As at least 500,000 plungers were required each year, the change of working method paid for itself in a short period. A second example concerned a pair of handed yokes, illustrated in Fig. 14–7. Conventional and

* See C. Ruiz and F. Koningsberger, *Design for Strength and Production* (London: Macmillan, 1970).

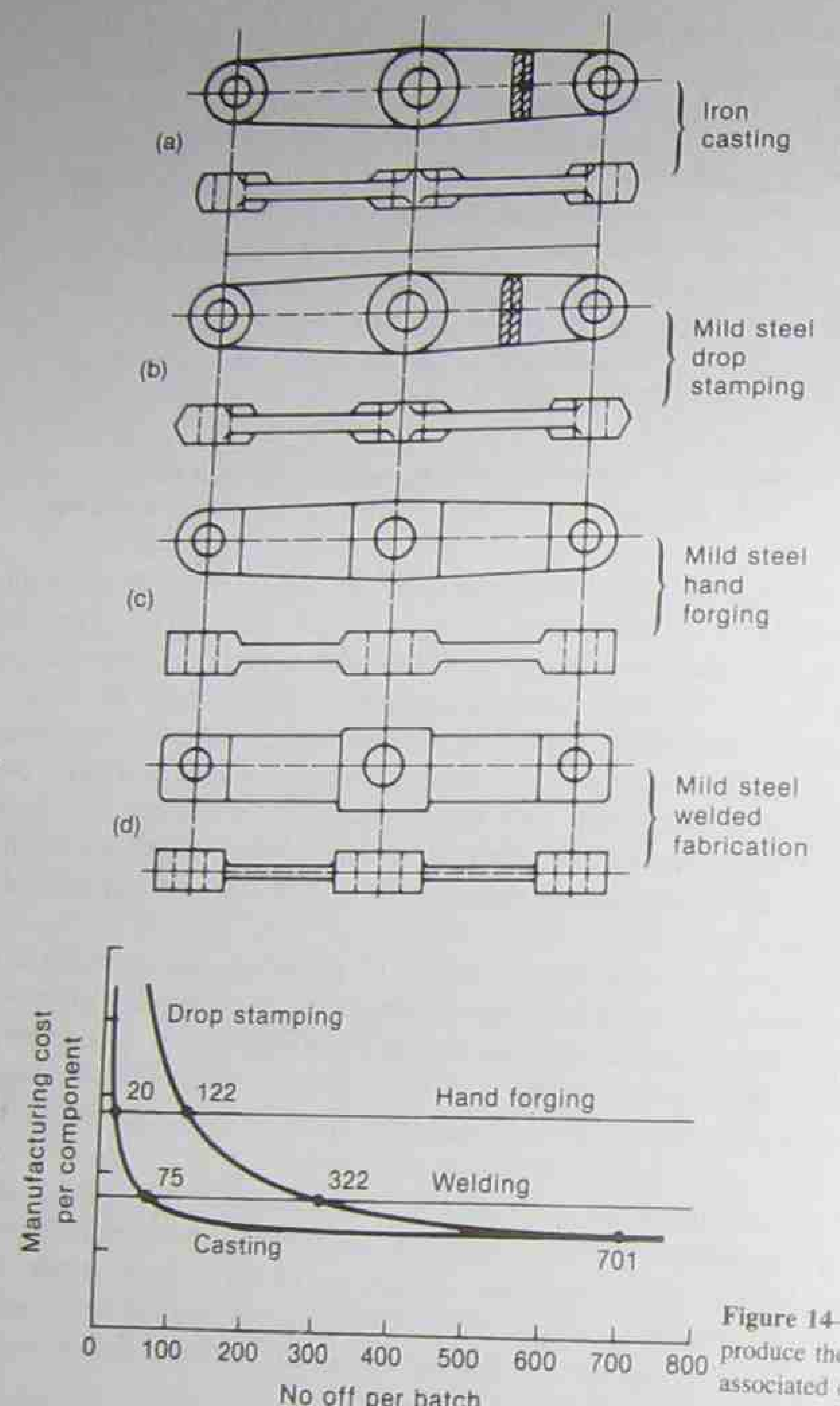


Figure 14–5 Alternative methods used to produce the same part, where relative associated costs are indicated as a function of the number of parts produced.

fine-blanking methods were again used. Expressed in time per 100 yoke components, 5234 minutes were saved, and the tooling costs for the old method of production were 82 hours higher. Servicing the tooling for the conventional method, and the preparation prior to production were also *more costly* than the newer method. The cost comparison shows that fine-blanking provided greater economy in producing these yoke components.

TABLE 14–2 COMPARISON OF PRODUCTION RATES AND TIMES FOR TWO BLANKING METHODS USED TO MAKE THE SAME PART

Operation	Conventional	Fine Blanking
Blank (single punch)	20	Fine blank (twin punch) 14.3
Shave	150	Deburr 30
Mill	300	
Deburr	30	
	500 min/1000 parts	44.3 min/1000 parts
Production Time for Tools		
Blanking tool	150	Fine blanking tool 1100
Shaving tool	120	
Milling jig	50	
	320 hours	1100 hours

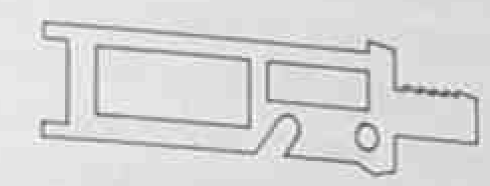


Figure 14–6 Plunger made by blanking methods detailed in Table 14–2.

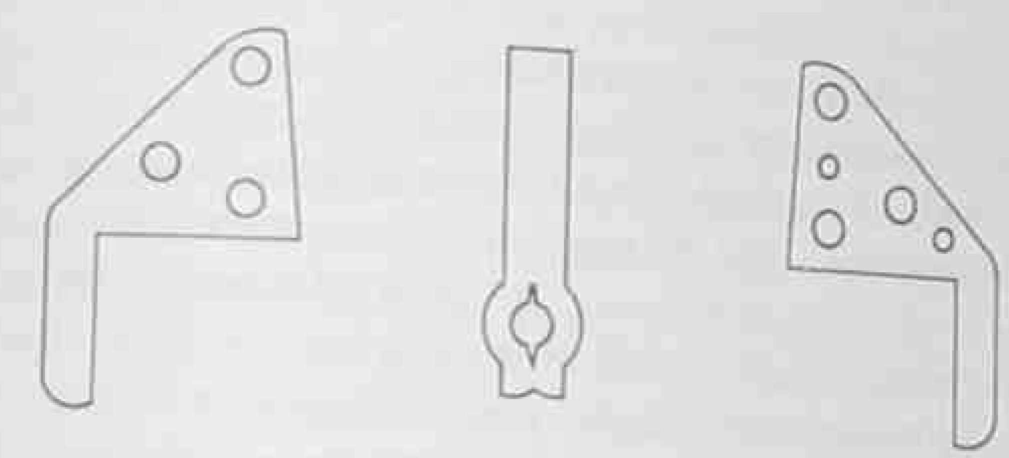


Figure 14–7 Yokes made by blanking.

In an entirely different field, the detailed design of weldments has a marked effect on manufacturing costs as well as strength performance. Consider the various alternative designs of T-joints in 25 mm plate shown in Fig. 14–8. The full-penetration butt weld at (a) requires initial preparation of the vertical plate, and the root runs made at relatively low deposition rates to gradually build up the joint. Additional work would include (1) backcutting of the first side root and (2) NDE* to ensure that full penetration was achieved; this adds materially to the total fabrication cost. A twin fillet weld of equivalent static strength, Fig. 14–8(b), would have approximately the same volume, but considerably lower costs by the elimination of an angled preparation, backcutting, and inspection. Relatively high welding currents could be used for all runs. A simple twin fillet weld of equivalent fatigue strength, transferring the site of failure from the weld throat to the weld toes, would need to be much larger, as in Fig. 14–8(c), but might still be more economical than a full penetration butt. A considerable reduction in weld volume could be achieved,

* Short for nondestructive evaluation.

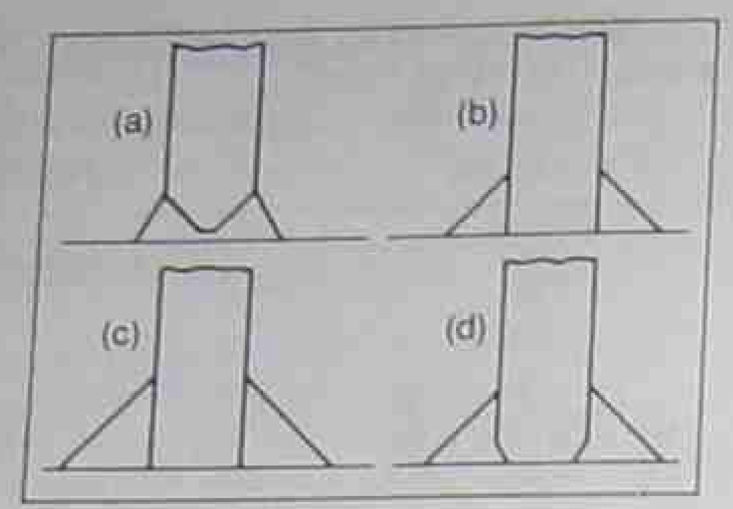


Figure 14–8 Different methods of welding used to produce a T-joint: (a) full penetration butt weld, and (b), (c), and (d), twin fillet welds of different sizes to produce joints of different strengths.

however, as indicated in Fig. 14–8(d) by simply chamfering the plate or ensuring that the root run was made by a deep penetration process.

For noncritically stressed joints, it is often convenient to specify *partial-penetration* butt welds, but the unfused edges can act as stress concentrators or cracks, thereby lowering both static and fatigue strengths. Of course, if a full-penetration weld is specified but not achieved, we have so-called lack of penetration (see Chapter 10).

At this stage, questions of *joint preparation* have to be addressed. It is possible to make full-penetration butt joints between adjacent sheets of material up to some 3 mm thick by using manual metal arc or Mig processes, and up to 12 mm with submerged arc welding, since the penetrating capacity of the arc is at least as deep as these sizes. For thicker materials, the edges must be *shaped* so that the abutting faces are no thicker than the limiting sizes just mentioned. A variety of joint preparations is shown in Fig. 14–9. Single V-bevels of about 60 degrees total included angle are the simplest type, followed by double-V chamfers used in thicker plate, say, > 18 mm, in order to limit the excessive amount of weld metal required to fill a single V cavity. In plates thicker than about 75 mm, the large amounts of weld metal required to fill even the double-V cavity can be reduced by using U-shaped joint preparations. Similar remarks apply to T-joints. Of course, use of double-V and double-U preparations necessitates that the joint be accessible from both sides, so that the component can be turned over.

It is clear that most preparation is concerned with access and that all the metal chamfered away has to be replaced subsequently. This affects overall welding productivity, so, in addition to increasing metal deposition rates as one means of optimizing

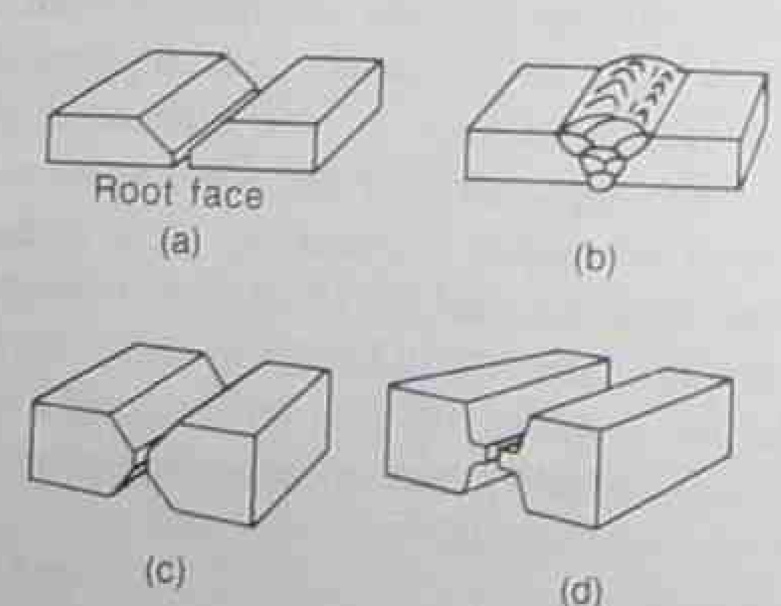
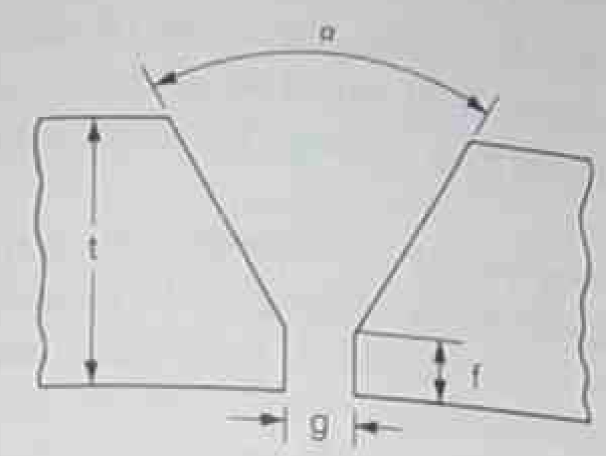


Figure 14–9 Various ways to produce a butt joint by welding.



f	g	α	Relative volume, %
0.25t	0	60°	100
0.125t	0	60°	117
0.25t	0	70°	121
0.25t	0.25t	60°	158

Figure 14–10 Effect of joint preparation on the volume of metal deposited to produce a satisfactory joint.

fabrication methods (Section 13.4.1), reduction in joint volume is also a worthy target. The majority of commonly used preparations have been developed for use with low penetration manual metal arc welding, for which it is essential that the arc impinge on the side wall at a minimum angle of 30 degrees if lack of fusion is to be avoided. Thus, a joint included angle of 60 degrees has become traditional. When tolerances on preparation angle, root face, and gap are taken into account, much greater angles are often encountered. The effect of these changes on joint volume, and therefore on welding time, can be considerable, as shown in Fig. 14–10.

A related area concerns the physical *size* of weld specified by a designer. In order to make the structure safer, welds larger than really needed may be specified, and costs in labor, materials, and inspection can get out of proportion. For example, a fillet weld of 4 mm leg thickness has a weight of weld metal of 73 grams per meter and requires an arc time of 2 min per meter; a 6-mm leg has 174 grams per meter and requires an arc time of 3.92 min per meter; an 8-mm leg has 304 grams per meter run and requires an arc time of 7.64 min per meter.

It was pointed out in Chapter 10 that, unlike many other manufacturing processes, which do not drastically affect the basic mechanical properties of components and assemblies of components, welding affects the metallurgical structure of the materials being joined and can also produce *variations* in properties around the joint region. Fabrication by welding is thus an area in which design, materials, and manufacture not only come together but are inseparable. Owing to large temperature gradients and the presence of pools of molten metal in welding, problems of distortion and residual stress have often to be anticipated. Welding can, for example, induce far higher internal stresses than does the cooling of whole components as in casting or hot working, and the designer should consider such difficulties.

14.4 BASIC DESIGNS FOR FORMING IN PLACE OF CUTTING

Much traditional design anticipates manufacturing routes which involve substantial metal cutting. Indeed, the familiar shape of many common engineering components is a direct result of such types of processing. It does *not* follow that, if a component were to be designed for manufacture by forming in place of cutting, the shape of the component would be the same as before. Alternative shapes, well suited to manufacture by metal forming, may be as good or better than those made by cutting. These considerations are relevant to near *net shape forming*.

It is valuable in this context to divide engineering components into *sections* which will experience bending moments, torques, tensile loads, rubbing contacts, and so on, and to consider what shape or particular feature is appropriate to meet each functional requirement. By considering a range of simple components, each of which carries one of these functional requirements, it is possible to see what would be an ideal shape for such a component if it were to be produced by a metal-forming process. This can then serve as a guide towards the application of this approach to more complicated components. It can also reveal where manufacture to finished size by metal forming, instead of cutting, can enable improved design of components as well as lower costs.

14.4.1 Tension Links

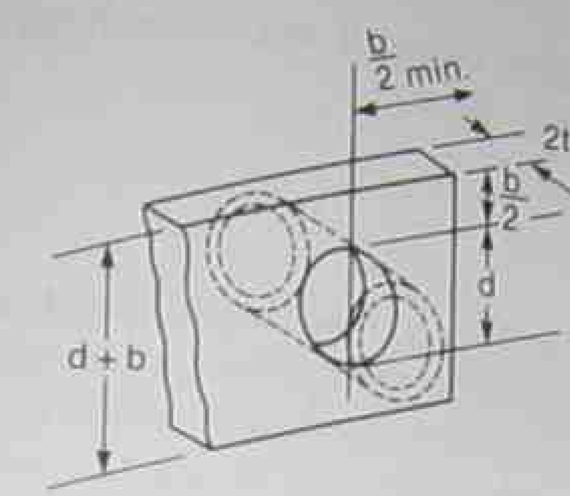
The shape and size of the main body of a tension link must provide sufficient cross-sectional area, in conjunction with the tensile yield strength of the material, so that the required load can be carried without the member's distorting. However, the shape and size of the *end fittings*, which usually connect to a pin or some form of bearing support tube, are more affected by the method of manufacture, as illustrated in Fig. 14–11. The various tension links in that figure are also generally suitable for compression, provided the body is short enough to avoid buckling. They are suitable for applications where the pivot pins at the ends, or the link itself, can be withdrawn.

If they cannot be withdrawn, the end connection must be split. Since an inserted bearing in the end of the link needs to be fitted with an interference fit, the clamping arrangement across the split needs to provide large controlled clamping forces. This is conveniently obtained from screw-threaded connections, as in a conventional engine connecting rod. The big end housing also needs to be rigid and of substantial radial thickness, because oil film thicknesses are small and large loads are generated. Due to these factors, the design of items like connecting rods would be unlikely to change even if they were formed to size. For lighter duty, they could be formed to size at low cost with a suitable integral bearing material, and some design changes might be possible.

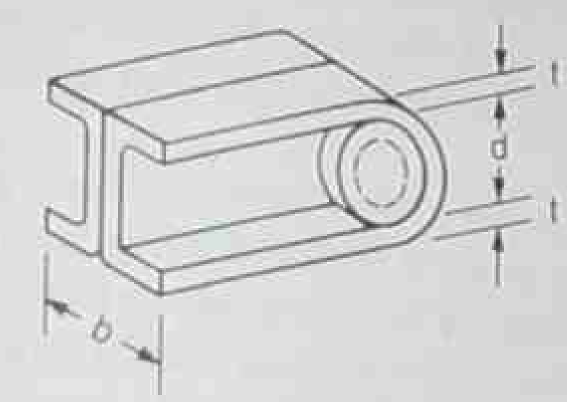
14.4.2 Torsion Shafts

Shafts are usually of circular section, since this is an efficient shape in terms of torque capacity per unit mass, and is particularly efficient if the section is a tube. Also, shafts of circular section have a surface which remains in a fixed position relative to the shaft axis when rotating; this enables bearings and seals to operate directly on this surface.

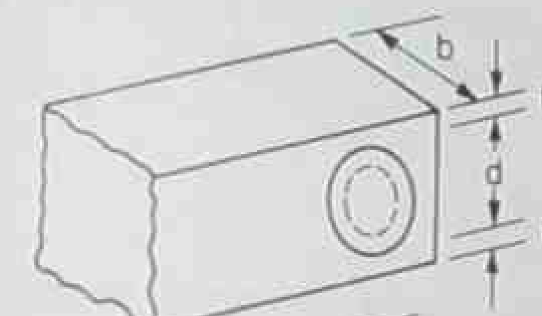
The choice of an appropriate shaft cross section in terms of torque capacity is a straightforward application of strength-of-materials theory, but the design of *end connections* to transfer the torque to other components is critical, particularly when these con-



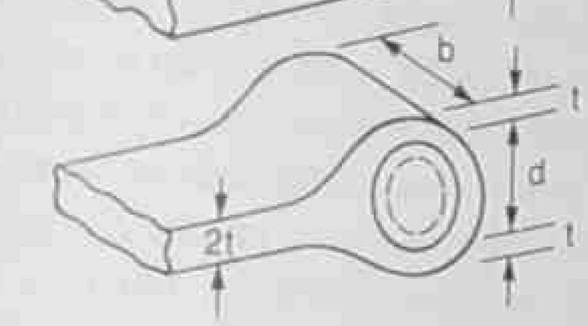
A simple plate link with a drilled hole at the end has to be large in order to have adequate tensile strength 'above' and 'below' the hole. Also the load on an inserted bearing support tube is concentrated at its center, with risk of distortion.



A formed link, as shown, is much more compact and distributes the load more evenly via an end strap into a bearing support tube.

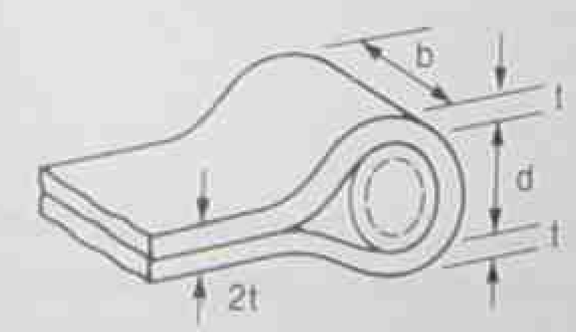


A link made from bar is heavy and again, like the plate link, the main body is understressed compared to the ends.



A link with a reduced center section is lighter and more economical in the use of material. Conventionally this requires a special forging and/or machining.

A precision powder forging or cold forging could be formed to final shape.



For short links an alternative formed-to-shape link could be made from a flattened tube.

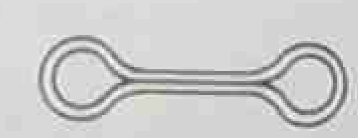


Figure 14–11 Various types of tension links of equivalent strength.

nections need to be made demountable. Some common end connections with their advantages and disadvantages are shown in Fig. 14–12.

14.4.3 Torsion Discs

Many engineering components include a number of disc-shaped members which carry torsional loads. Driving flanges, gears, sprockets, and turbine or compressor discs are typical. Nominally, for constant torsional shear stress, such discs would be expected to be made with a thickness which decreases toward the outside. In practice, however, due to attachment loads on the driving flanges, to the required width of gear or sprocket teeth, or blade roots, the periphery of such discs has to be made thicker than torsional shear stresses alone might require. Constant thickness discs are therefore commonly used in many cases. Only where centrifugal stresses are important in high-speed machines are profiled discs normally used.

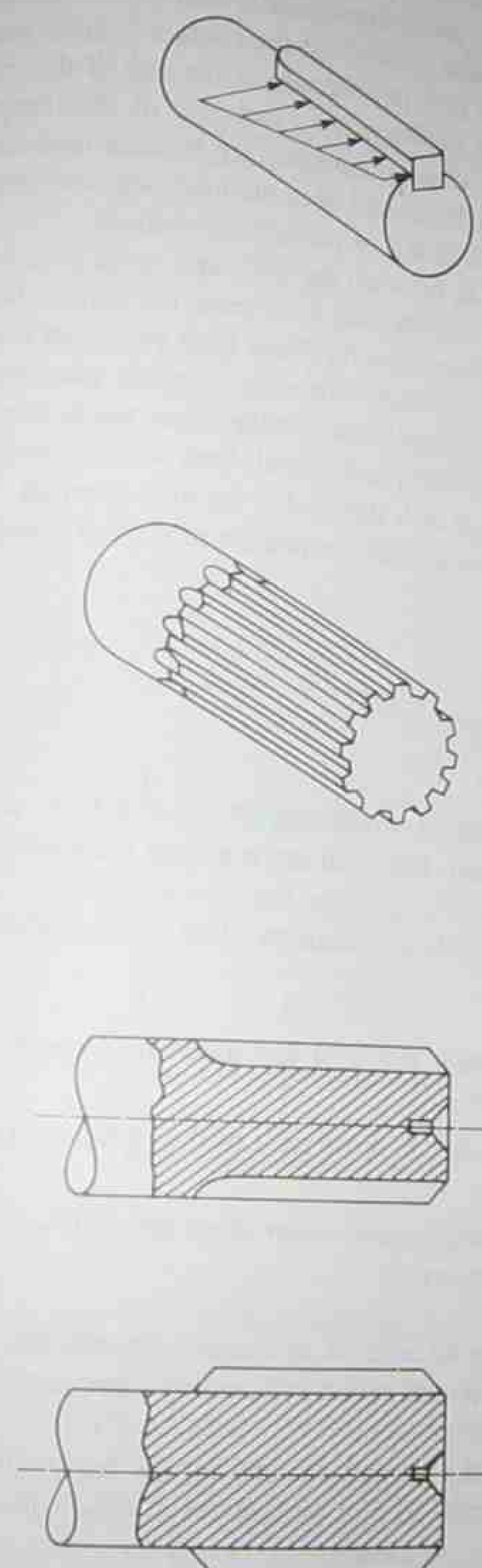
With gears, the loads are transmitted across a small number of teeth at one time. On sprockets, although the number of teeth in action is larger, loading is still carried by a limited number. Consequently there are substantial bending loads on teeth of this type and the meshing profile contact regions experience high contact stresses. Teeth which are cold formed rather than cut will tend to have superior strength. However, many gears still have to be cut and ground, owing to severe requirements on tooth profile accuracy, associated surface finish, and the relative peripheral location of the teeth.

14.4.4 Beams and Plates in Bending

For a given bending load, flat plates and beams of solid rectangular section are not as light or economical in the use of material as a section such as an I-beam or rectangular hollow section. Such sections have material remote from the central or neutral axis to carry tension and compression forces together with a connecting web or webs to carry shear forces. Sections like these are usually produced by forming processes such as hot rolling or extrusion. This is a direct example of net shape forming, but it is noted that the tolerances required are fairly lax, since they strike a balance between adequate strength and excess weight. While structural sections are formed to final shape, the shape is not particularly precise.

14.4.5 Springs

Springs are usually in the form of coils, which deflect by torsion, or beams, cantilevers, or diaphragms, which deflect by bending. They usually need to be made to some fairly precise stiffness, and since this varies as the cube of the material thickness, this dimension has to be accurate owing to the multiplying effect of this third-power relationship. Precision forming methods such as wire drawing or cold rolling usually give tolerances of sufficient accuracy and, additionally, elevate the yield strength, thereby giving a somewhat larger elastic strain range.



Keys and keyways are the most commonly used method since they are relatively easy to manufacture by metal cutting operations, although the accuracy of fit is critical.

The torque capacity of keys and keyways is limited by the fact that the torque is carried across a single region of the shaft periphery. Also due to shaft torsional elasticity the transmitted forces tend to be further concentrated towards the inner end of the key.

Splines involve more complex metal cutting but have a higher torque capacity because the transmitted forces are distributed around the periphery of the shaft.

They can also be made by metal forming processes.

To enable spline manufactured from simple shafts they are usually cut in below the existing shaft diameter, although ideally they should be cut into a raised boss at the shaft end. This, however, requires more extensive machining.

Figure 14–12 Various types of end connections on shafts as used to transmit torque.

14.4.6 Tribological Contacts

Tibological contacts are of two types called *conforming* and *nonconforming*. With the conforming type, the contact takes place over a considerable area such as between a plain bearing and shaft, or between a piston and cylinder. If the contact is lubricated, the film thickness tends to be of the order of 25 μm, but because of the size of the contact area, there are risks of edge loading due to misalignment, and also of the trapping of dirt between the surfaces. To deal with this situation, one surface is made hard and the other relatively soft. In service the soft one tends to bed in to the hard one, and therefore, it is the finish and accuracy of the hard surface that is particularly critical.

In nonconforming contacts such as between the balls and races of a ball bearing, between a cam and tappet, or between two involute gear teeth, the contact occurs over a very small area, and high contact pressures result. These high pressures cause the formation of elastic flats on the surfaces, together with a considerable increase in the viscosity of any lubricant which may be present. The resultant film thicknesses are of the order to 2.5 μm. To carry the loads without plastic yield, both surfaces need to be hard and, owing to the low film thicknesses, a high standard of finish is required. Where there is a line of contact, accuracy *along* that line is important, but in other directions it may not be as critical.

14.5 TOLERANCES AND SURFACE FINISH REQUIREMENTS

In Chapter 3, the typical surface finishes which can be achieved by using different manufacturing processes were discussed. Just what are *required* for certain applications and what will the *cost* be?

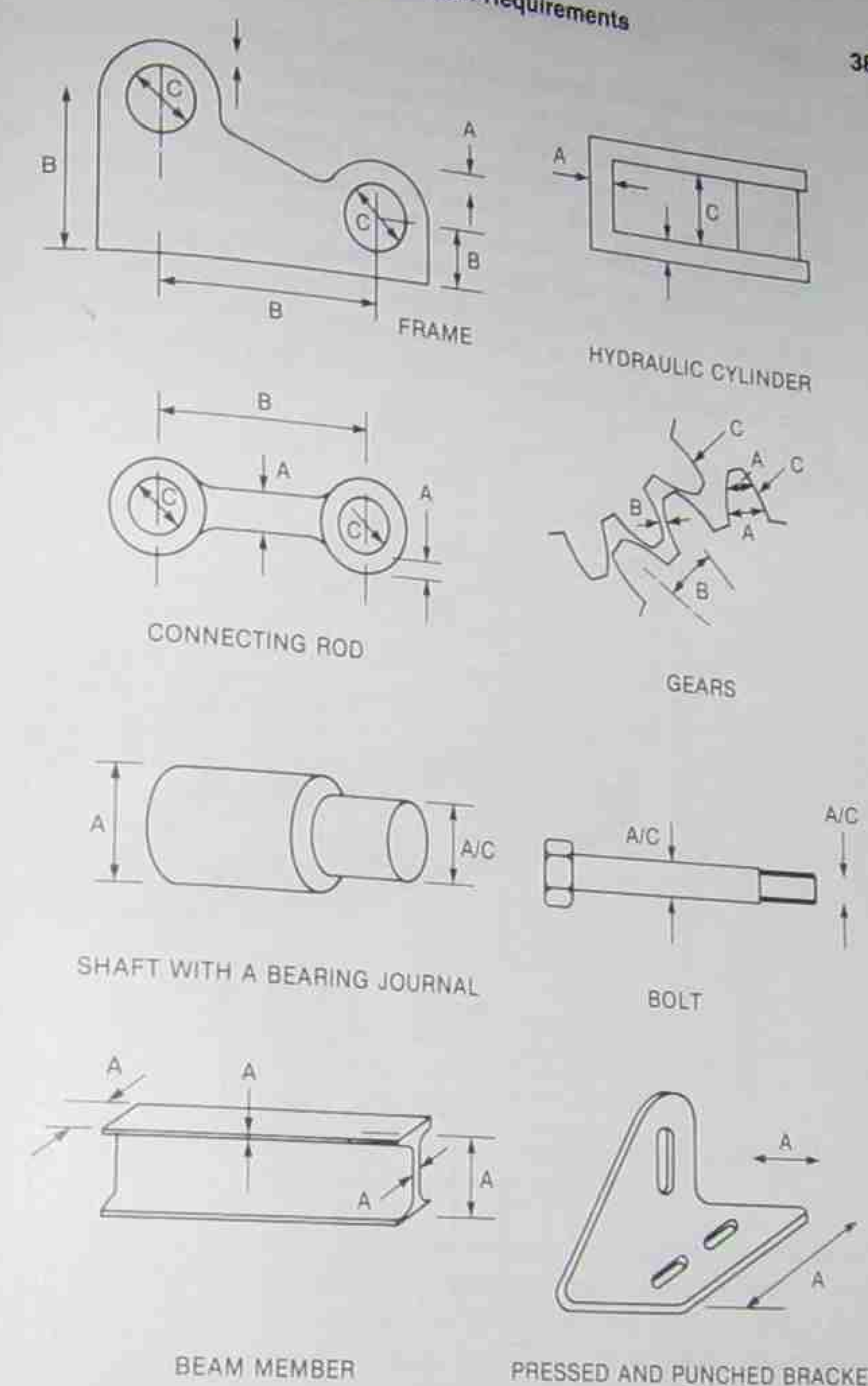
In broad terms, there are three levels of tolerances which are important in designed products.* These are

1. *Cross-sectional* tolerances to meet the need for adequate strength and stiffness without excessive weight or volume.
2. *Positional* tolerances on the location of those parts of the component surface which mate with adjacent components.
3. *Fit* tolerances on the actual mating surfaces where there are tribological interactions or precise fits for static connections.

These three requirements are in an order of increasing dimensional accuracy. Examples of their application to a selection of varied engineering components are shown in Fig. 14–13.

With regard to surface finish, the greatest demand for the finest finish arises with tribological surfaces owing to the small oil film thicknesses at which these surfaces are

* Taken from Neale and Waterman, Report TRS 289 for SERC/DTI, 1982.



TOLERANCES ON TYPICAL COMPONENTS

CROSS-SECTIONAL TOLERANCES ⟷ B
POSITIONAL TOLERANCES ⟷ A
FITS AND SMOOTH FINISHES ⟷ C

Figure 14–13 Comparative tolerances on dimensions of typical engineering components.

intended to operate. Sometimes, highly stressed components are required to have a high standard of surface finish to avoid the risk of rough surfaces giving rise to stress raising discontinuities. Some components used in applications in food processing require smooth finishes to make them easy to clean so as to maintain their sterility or appearance.

Figures 14–14 and 14–15 show how tolerances relate to component size and surface finish for different processing methods. Thus, in Fig. 14–15, open die forging is typically used on components with a characteristic size bigger than about 20 mm to over a meter; tolerances obtained are from 2 to 20 mm. Ordinary powder forming is used for objects between 5 and 100 mm in size, with tolerances between 10 to 100 μm. Other processes can be analyzed in a similar manner. It is seen that a number of processes can directly

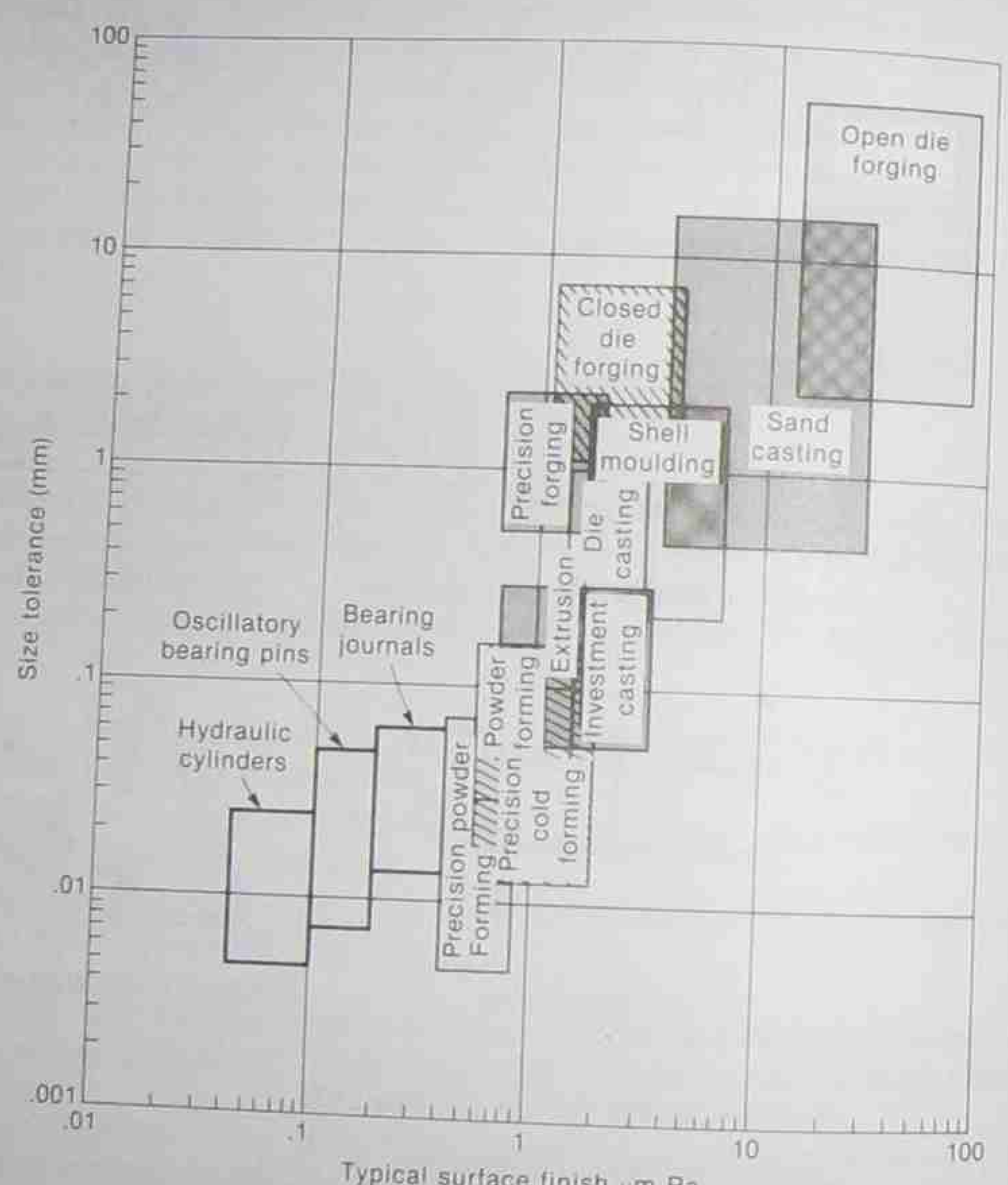


Figure 14–14 Interraction of tolerance, component size, and various manufacturing processes.

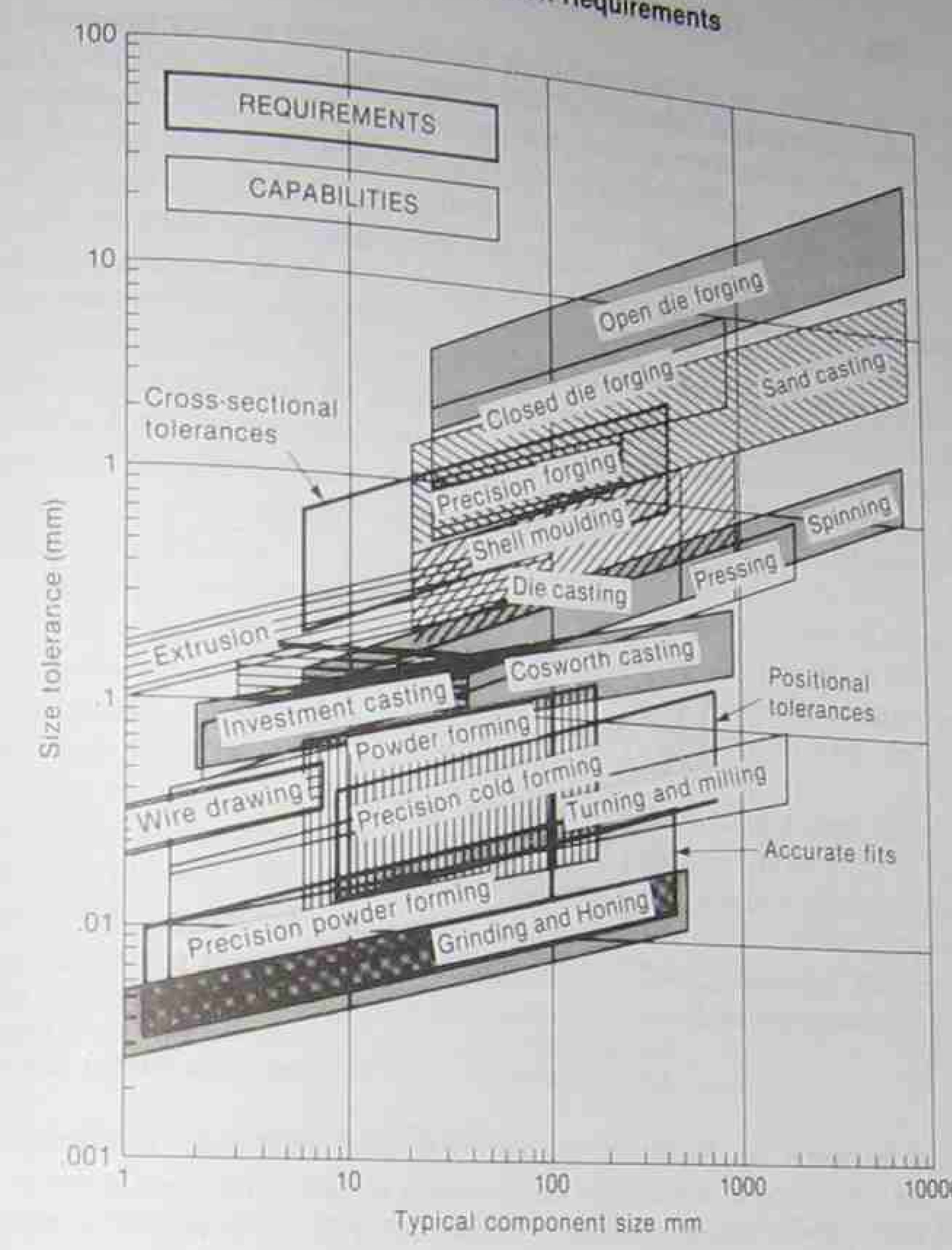


Figure 14–15 Surface finish and tolerances obtainable from various processes. These are compared to requirements for some sample components.

satisfy some of the tolerances demanded, whereas second operations are required to attain closer tolerances. Thus, cross-sectional tolerances can be achieved directly by casting or closed die forging, positional tolerances by precision cold forming, and accurate fits by precision powder forming or grinding. Table 14–3 gives approximate relative costs of achieving different grades of surface finish by different cutting operations. The standard of comparison in the table is a very rough machining operation, which is taken as having unit cost.

The increase in cost of attaining a greater degree of accuracy and finer surface finish is illustrated by the curve rising to the right in Fig. 14–16. However, when a manufactured component consists of an assembly of parts, the costs of assembly and fabrication will usually be *reduced* if more accurate parts are used. This is reflected in the curve falling

TABLE 14–3 SURFACE FINISH AND COST COMPARISON FOR VARIOUS CUTTING OPERATIONS

Surface Finish	RMS, um	Relative Cost
Very rough	50	1
Rough	26	3
Semi-rough	13	6
Medium—Machined	6.5	9
Semi-fine	3.2	13
Fine	1.6	18
Coarse	0.8	20
Medium—ground	0.4	30
Fine	0.2	35
Super fine—lapped	0.1	40

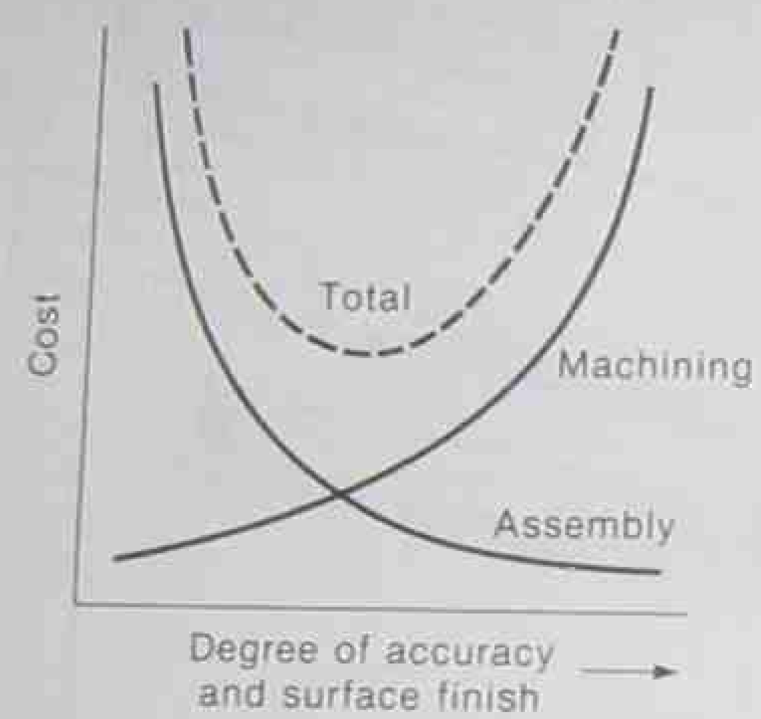


Figure 14–16 Typical type of cost curve showing the individual costs of machining and assembly as the degree of accuracy of finer surface finish is increased.

to the right in Fig. 14–16. The combination of the two effects leads to an optimum, where the overall cost of manufacture and assembly is least. Note the similarity of this curve with optimum values in machining economics, Fig. 13–1. A similar type of optimum occurs regarding reliability, in that the cost of design and manufacture must increase with increased reliability of the product, but the costs after delivery (to do with warranties, and the like) fall with improved reliability. It does *not* follow, of course, that manufacturers can always work at the optimum, as it may be necessary, for component operational design requirements for example, to have degrees of accuracy to the right of the minimum in overall costs.

14.6 DESIGN OF PRODUCTION ROUTES

14.6.1 Introduction

Within a given general manufacturing method, there will usually be a choice for the sequence of operations performed and what takes place in each operation. For example, in machining from a rough casting, how many cuts are to be taken and what speeds, feeds,

and depths of cut are proper for each? Appropriate conditions applying to repetitive mass-production machining have been given in Sec. 13.3 relating to optimum cutting combinations to achieve either maximum production rates or minimum cost per piece. In solid forming processes, what sequences of forward or backward extrusion, upsetting, heading, piercing, die angles, reductions, and so on are possible, and which is best? Or in sheet forming, what sequences of cupping, redrawing, ironing, blanking, and with what tooling and so on are possible and best? Similar thoughts apply to other manufacturing and fabrication processes.

The *best* sequence is that which is most economical but which does not prejudice the in-service performance of the component. Yet even for relatively simple components the number of feasible alternative sequences can be great, and laborious calculations are often involved. Computer-aided calculations can help greatly in these circumstances. In this section we explore options available in forging and how sequences of operation are designed to achieve certain aims.

Forging operations can produce a wide range of shapes and sizes of components. Even with simple articles, a variety of sequences of operations to obtain the required shape and size are nearly always possible. As shown in a subsequent example, the ease of production and final part cost can be markedly influenced by early decisions on component and process design.

The determination of forging schedules is still based to a great extent on the experience, imagination, and intuition of skilled personnel. In the past, it has been difficult to propose rules for other than simple shapes. But instead of relying totally on the skills of experienced individuals, component and process design can be approached systematically. Such an approach is sometimes called group technology. The core of the method is to identify groups of related parts by means of an appropriate classification system.

14.6.2 Some Concerns in Forging

One book* builds on and systematizes earlier techniques of classification used by a variety of groups in the international forging industry, and is intended to give guidance for schedules for easy manufacture. The attributes of a part which may cause difficulties in forging are indicated by a coding system with the aim of avoiding undesirable features in the design finally decided upon. Most forgings are (1) *compact*, with similar dimensions in the three orthogonal directions, (2) *flat*, such as discs, or (3) *long*. Of forged parts manufactured, less than 10 percent are compact, about 20–30 percent are flat, and the rest are long. Each type is characterized by a basic sequence of operations. Compact parts are made from billets or blanks and, in general, have a simple forging sequence, but a proportion of parts within this category may require multiple-action machines. Flat parts are also produced from billets or blanks with upsetting (causing material flow) being predominant, and with a significant proportion of parts requiring pierced-through holes.

* C. Poli and W. A. Knight, *Design for Forging Handbook*, (Amherst, Mass.: University of Massachusetts, 1981).

Long parts are produced directly from bar stock and generally require initial elongation and drawing-down stages, prior to the impression die forging sequence. One or more bending stages may be necessary if the longitudinal axis of the part is not straight. In general, more deformation is required in going from compact to flat to long parts.

The major factors which influence forging difficulty in a particular case are the complexity of the part and the material used. These factors are strongly interrelated, since a shape that is relatively easy to forge from one material may be difficult or impossible to obtain in another.

As an illustration of the use of the *Design for Forging Handbook* coding system, Fig. 14–17 shows two flat round forgings, together with the code numbers allocated. One is a simple part which has a low-valued code number and is consequently relatively straightforward to produce as a forging. The second part is increased in complexity by the addition of closely spaced radial ribs, which are difficult to produce by forging. This requires additional forming stages, with reduced die life and greater material losses, and

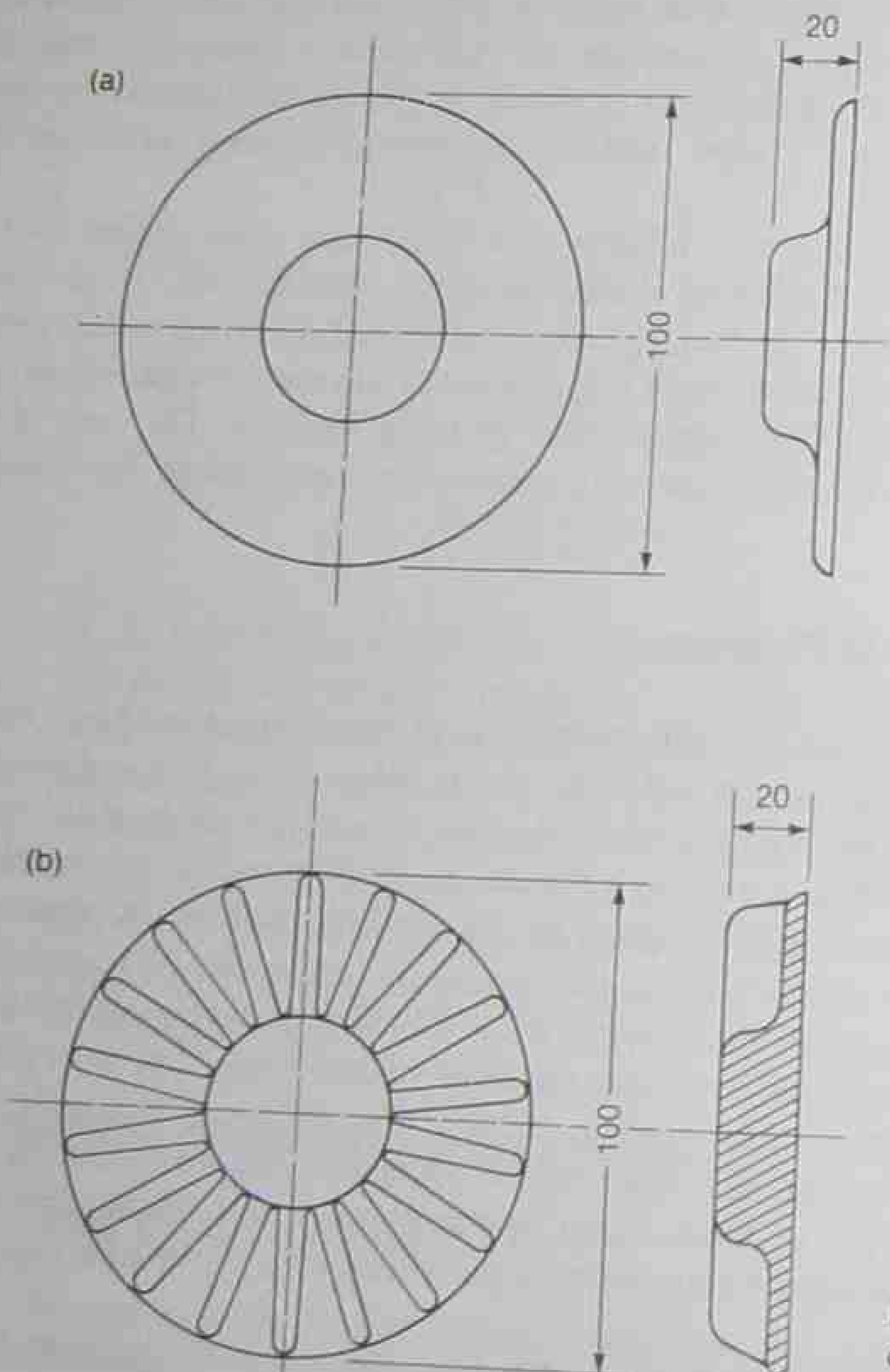


Figure 14–17 Two forgings described by code numbers from the *Design for Forging Handbook*. (a) Code 102, (b) Code 181. All dimensions in mm.

is given a higher-value code number within the classification. Allocation of a code number enables relevant data sheets to be consulted, which contain recommendations for forging design features appropriate to various classes of parts.

For the various classes of parts defined by coding systems, standardized alternative processing sequences can be developed, with the number of stages determined by the basic design limitations. Consideration of load capacity, limiting strains, need for interstage heat treatment, and the like are involved. Once likely alternative operations are established, the subsequent laborious calculations required to optimize the processing sequence are ideally suited for the computer. This gives the designer and the manufacturing engineer more scope to ensure that the sequence selected is best for the given conditions. Computer-aided-design (CAD) of this sort permits rapid comparisons to be made between alternatives under conditions prevailing at the time, and also rapid "sensitivity" assessments of the effects of future changes in material costs, energy costs, labor rates, and other cost factors.

14.6.3 Alternative Routes in Forging

Calculations of this sort have already been presented in Chapter 13 for *prescribed* manufacturing forming sequences. But what about the assessment of a large number of alternative sequences under greatly varying conditions? One such example follows and relates to Fig. 14–18. First, a number of alternative sequence of operations was devised which appeared feasible for this particular geometry. To retain a reasonable clarity only thirteen alternatives are shown in Fig. 14–18; this is only a small proportion of the total studied.*

Depending on the level of work hardening, the second annealing and lubrication might not be required for alternatives 9 and 13. Similarly, the annealing of the slugs would not be necessary if the bar from which they are made is supplied in a soft condition. Of the sequences illustrated in Fig. 14–18, 11 and 12 utilize hollow machined slugs, while solid slugs produced by a number of alternative methods are used for the rest. There are also alternative methods for slug production from the solid, such as cutting off by machining, sawing, cropping or, if the length/diameter ratio of the slug is less than a specified value, cropping followed by axial compression or blanking from sheet metal. Furthermore, depending on the actual value of angle α in the particular case, it might be better to extrude partially at the optimum die angle and obtain α by an additional operation. The thirteen alternatives indicated in Fig. 14–18 represent well over a hundred different sequence of operations, although the differences in some cases might be the addition of just one further operation to an otherwise unchanged sequence.

Each sequence is characterized by a combination of operations carried out in a specific order and, for it to be feasible, each operation must be possible in the order corresponding to that sequence. Every operation is subjected to a set of constraints which must be satisfied before proceeding to the next operation, as shown in Fig. 14–19. Typical

*B. Lengyel and T. Venkatasubramanian, *Proceedings of the 18th Machine Tool Design and Research Conference* [MTDR] **153** (1977).

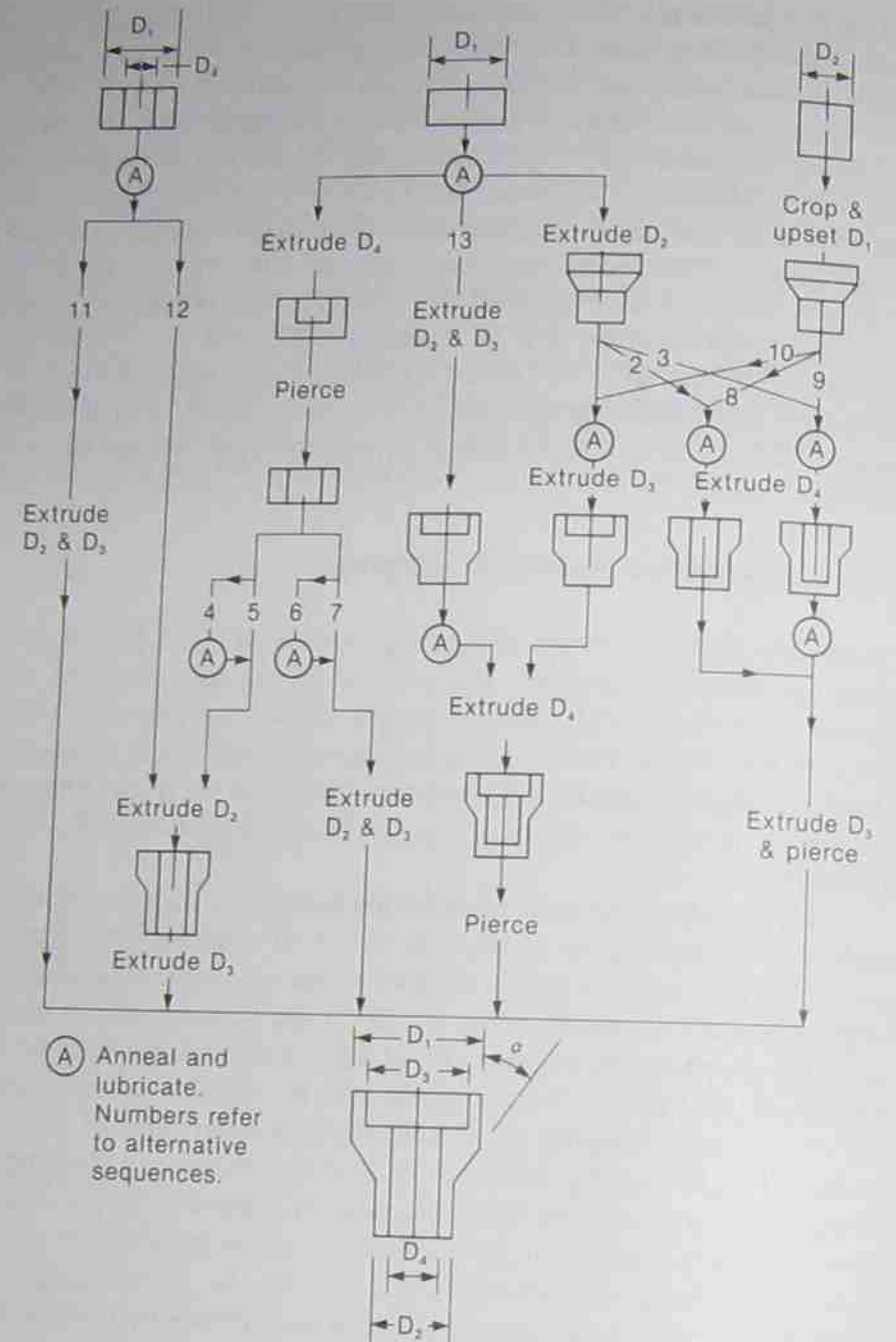


Figure 14–18 Alternative sequence of operations for a certain component shape.

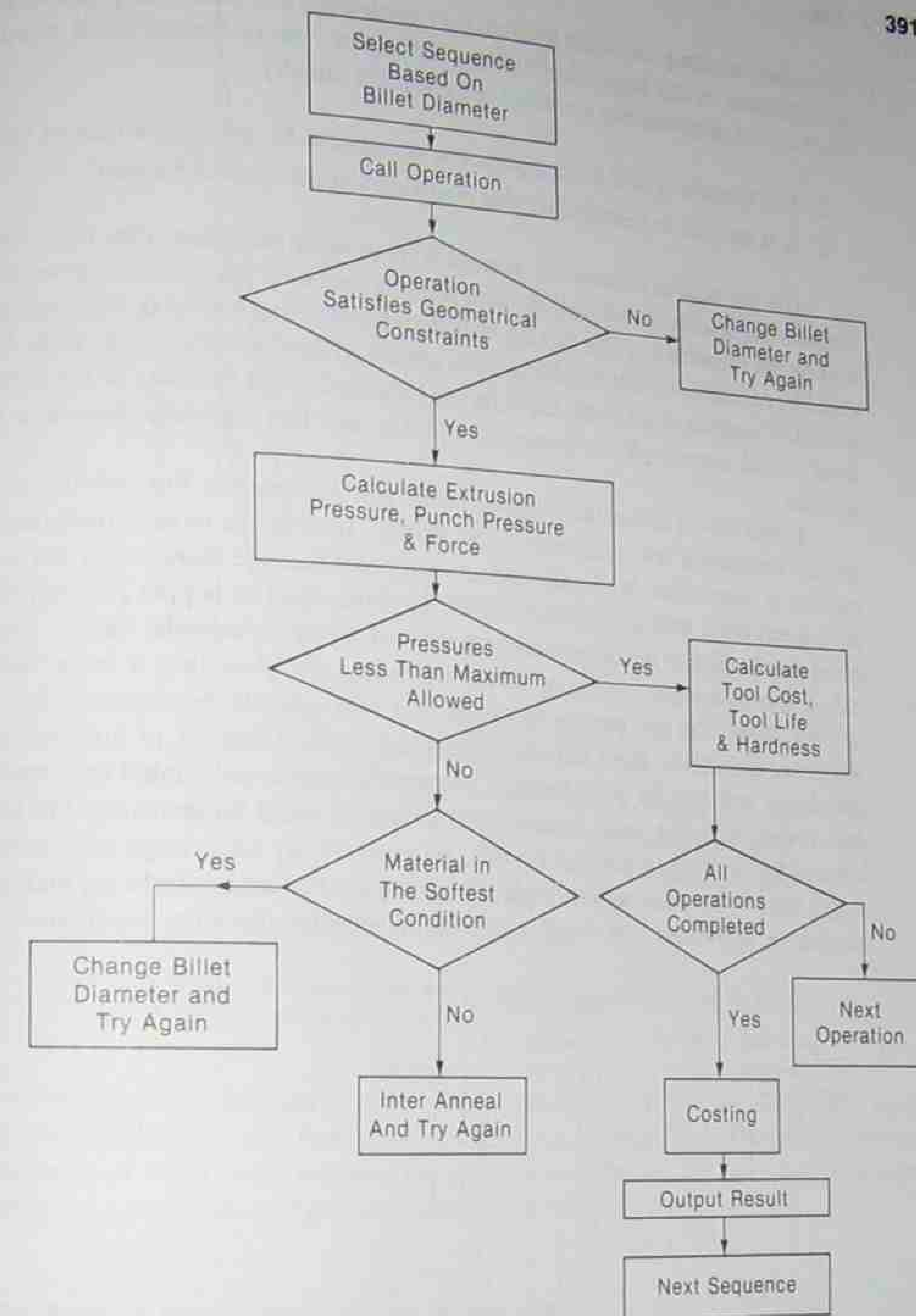


Figure 14–19 Steps to determine the optimum sequence for a part such as shown in Fig. 14–18.

constraints are the permissible level of work hardening, the maximum practicable extrusion pressure, or the largest feasible length/diameter ratio of the extrusion punch. As a result of this approach two questions are answered initially.

1. Is it feasible to cold forge a particular component by applying a certain sequence?
2. Is it possible to manufacture that component at all by cold forging?

It is essential to estimate the average level of work hardening after each operation, because that determines whether the next forming operation may be carried out and what will be the properties of the forging at that stage. Whether the next forming operation could be carried out or not is influenced by prior work hardening because, if the hardness exceeds a predetermined limit, the deforming material could fracture, or the permissible level of tool stress could be exceeded; in either case that particular sequence must be rejected.

Figure 14–20 shows two versions of the part presented in Fig. 14–18, while Fig. 14–21 illustrates a few comparative cost results. In order to have a reasonably small number of alternatives in this illustration, computed data are shown only for a vertical mechanical press with a single station tool, for either hand or hopper feeding. Costs are given as a function of the number of components required per calendar month. For clarity, only a few of a large number of curves are shown, but even then a large variation is obvious, both from one process to another and with quantity requirements for any one sequence. In general, those sequences requiring the least number of operations tend to give lower unit cost for small batches, because both labor and capital cost contents are then lower. Also, for small batches, tool utilization could be increased and tool costs lowered by allowing higher tool stresses, for example, by using large deformations in a single stage or by planning for combined operations. For large quantities, increasing the number of operations was found not to incur penalties, because automatic feeds and

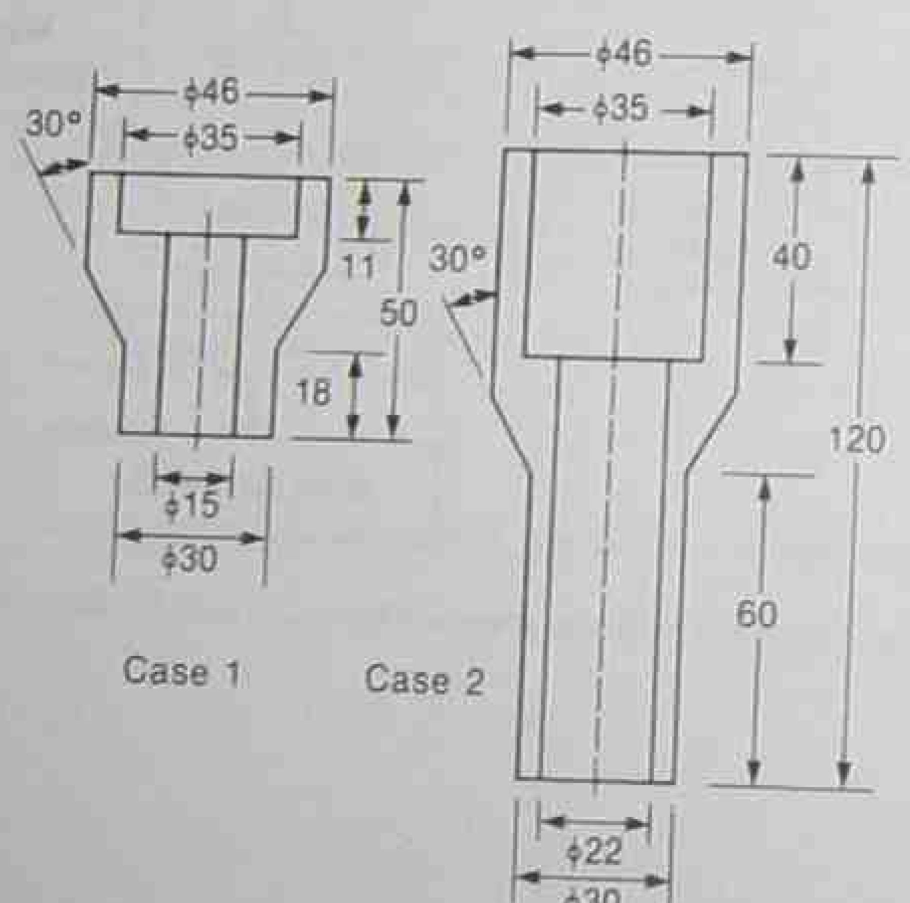


Figure 14–20 Alternative dimensions for the part in Fig. 14–18.

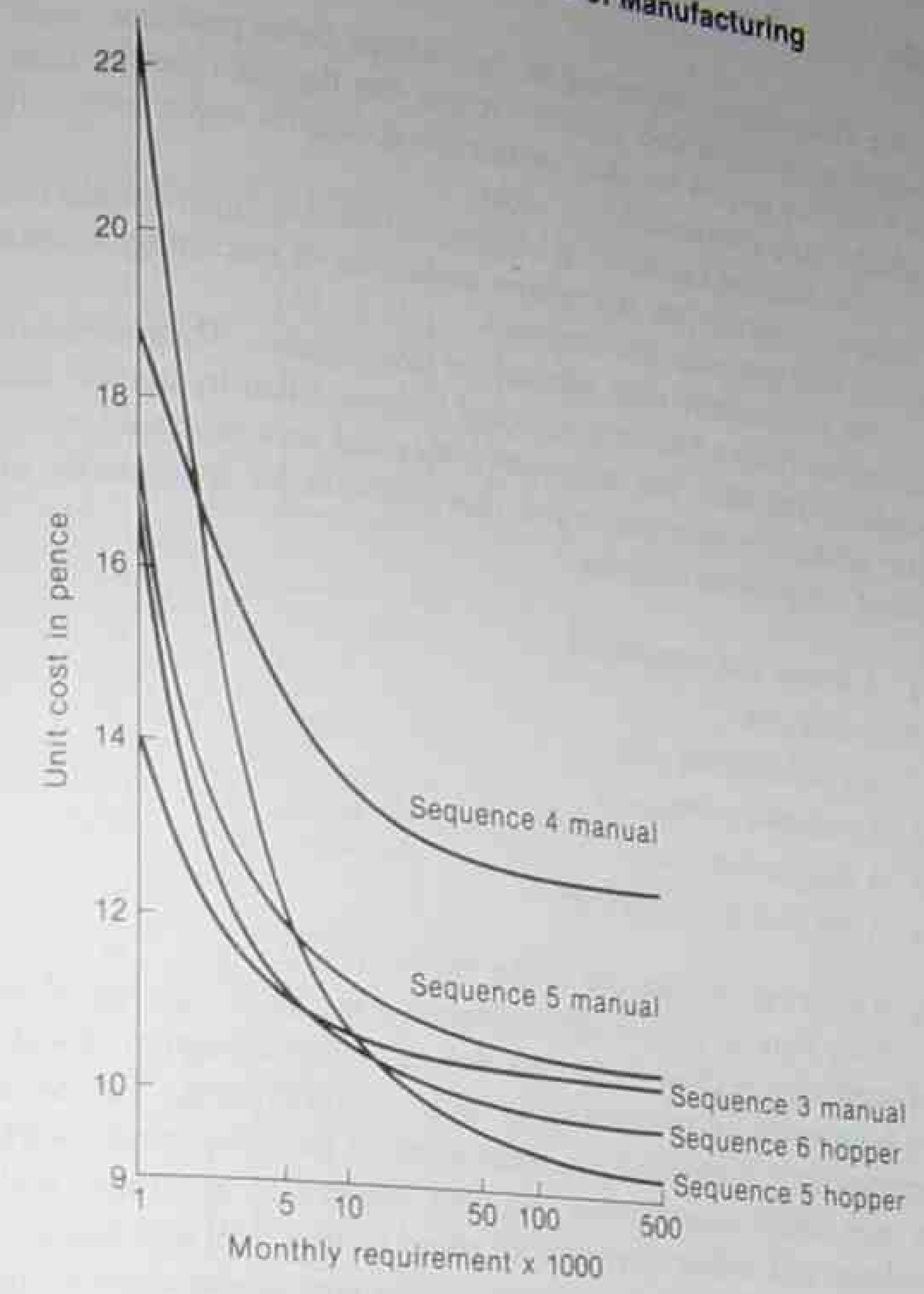


Figure 14–21 Comparative costs for alternative methods of producing the part in Fig. 14–18 as influenced by batch size.

transfer tools could be used to ensure better equipment utilization and lower labor costs. At large batch sizes, tool lives must be maximized by reducing tool stresses; otherwise, frequent tool replacement increases the unit cost by decreasing the utilization of the more expensive equipment owing to extensive machine downtime.

14.7 ECONOMICS OF ALTERNATIVE METHODS OF MANUFACTURING

In Chapter 13, the procedures for costing articles were outlined and examples were given for a variety of different components manufactured by different prescribed routes. What, though, if there are alternative manufacturing methods to achieve the identical or essentially the same result? Are some processes always more expensive or cheaper than others, or are there ranges, depending on the numbers being produced, over which one process is more economical than another? A hint that the latter can be true was shown in Fig. 14–5. In this section we shall explore the conditions under which different methods are possibly more economical than others.

The National Engineering Laboratory (NEL) in Great Britain commissioned in 1965 a pioneering survey into the relative economics of machining methods versus combinations of cold extrusion and forging.*

Six components were selected for investigation. Their choice centered on the fact that *originally* they had been designed for production by normal machining methods in reasonable quantities, but *subsequently* they had been produced by cold forming. Appropriate production information was thus available for both modes of manufacture. The selected components included

1. A double can component
2. A spur gear
3. A synchronizing gear
4. An annular component
5. A flanged component
6. A stepped can component

Two of these components are shown in Fig. 14–22.

They were all typical light engineering products, capable of manufacture by standard production machine tools; none was of an unusual asymmetrical shape, which might have been expected particularly to favor cold extrusion. The sequences of extrusion operations used in producing the final forms of the components were not necessarily the only ones which were feasible, but were selected by qualified production engineers in accordance with ruling conditions of plant, tooling, and operational experience.

Figure 14–23, which shows the relative cost of manufacture against the number of the synchronizing gears produced, typifies the type of findings. A breakeven, or crossover, point occurs, and we see that cold extrusion is more expensive at low batch sizes but improves relative to machining with increase in batch size. Similar curves are given in the NEL Report for the other components, with the crossover batch quantities in Table 14–4.

The crossover batch quantity at which extrusion becomes cheaper is generally low, with the exception of the annular component. That particular component used tube as the starting stock with consequent little material waste, but in all other cases extrusion makes large savings in raw material and, at large batch numbers, this outweighs all other factors, such as power costs for annealing furnaces, chemicals for lubrication, and the like. At low production numbers, of course, these savings are offset by the increased prorated cost of tooling; indeed, tooling cost is the only significant reason for the economic disadvantage of forming processes compared with traditional machining at these low quantities. If parts were to be made on a lathe, an entire small production run could be finished in the time it takes to design, manufacture, and deliver dies suitable for cold forging. Even so, it is

* B. P. Clapp *Economics of Cold Extrusion*, NEL Report No. 195, East Kilbride, Scotland, 1965.

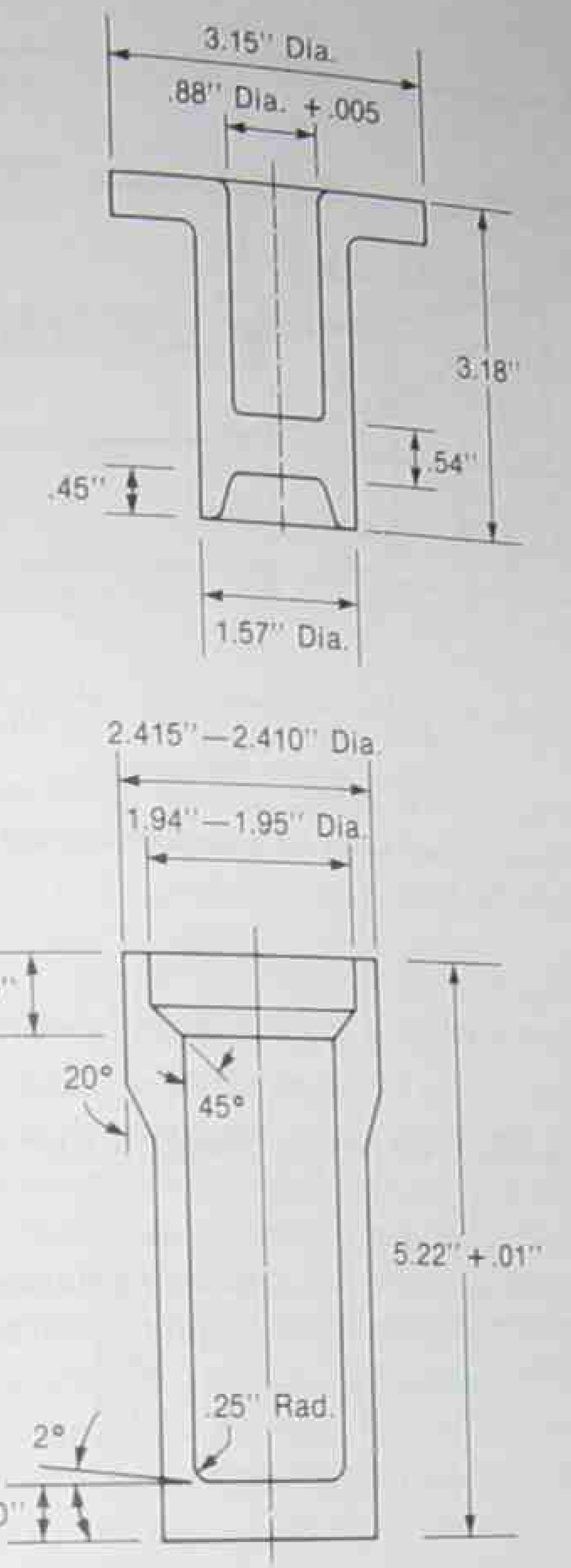


Figure 14–22 Two of six components selected for comparison of the relative economics involved in machining or cold forming these parts.

clear that under appropriate conditions, manufacturing methods involving a large proportion of cold extrusion/forging are highly competitive. For all six components, no further improvement in cost advantage occurred for batch sizes greater than about 30,000, and savings over machining settled to the figures shown in Table 14–5.

Another study conducted at NEL in 1969 involved a steel bearing race sleeve as shown in Fig. 14–24;* this is a hollow axisymmetric part well suited to cold forging. Manufacturing costs were calculated on the basis of

* B. P. Clapp, A. M. Evans, and T. H. Pasteur, *Cold Forging Study*, NEL Report No. 468, East Kilbride, Scotland, 1969.

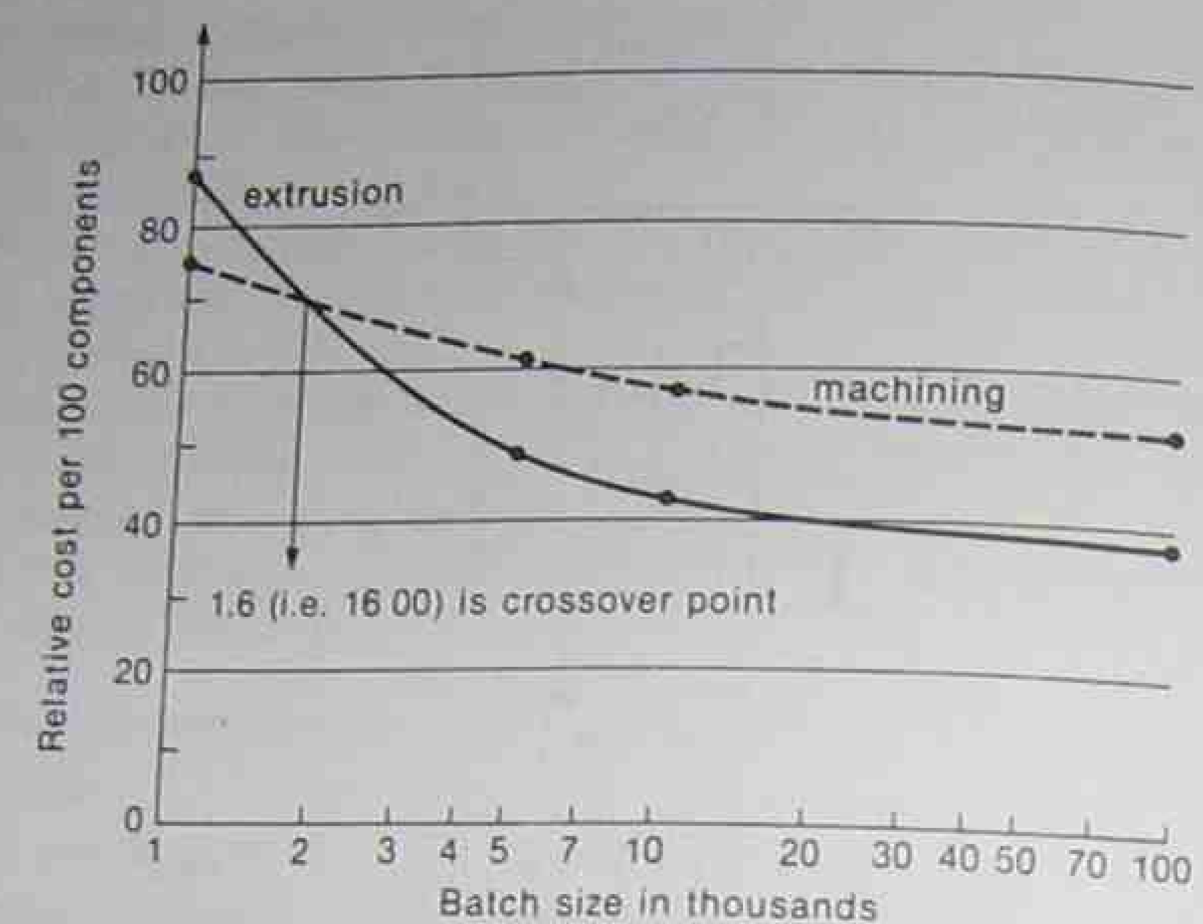


Figure 14–23 Comparison of costs versus batch size for machining or cold forming and machining the gear in Fig. 14–22.

1. Machining from an as-purchased hot-forged blank
2. Cold forging a purchased billet to a semi-finished condition followed by finish machining.

Using sequence (2) cut costs almost 50 percent compared with sequence (1).

TABLE 14–4

Component	Crossover Batch Quantity
Double can component	1400
Spur gear	1100
Synchronizing gear	1600
Annular component	5000
Flanged component	1100
Stepped can component	1300

TABLE 14–5

Component	Saving on Machining
Double can component	47%
Spur gear	63%
Synchronizing gear	29%
Annular component	8%
Flanged component	56%
Stepped can component	43%

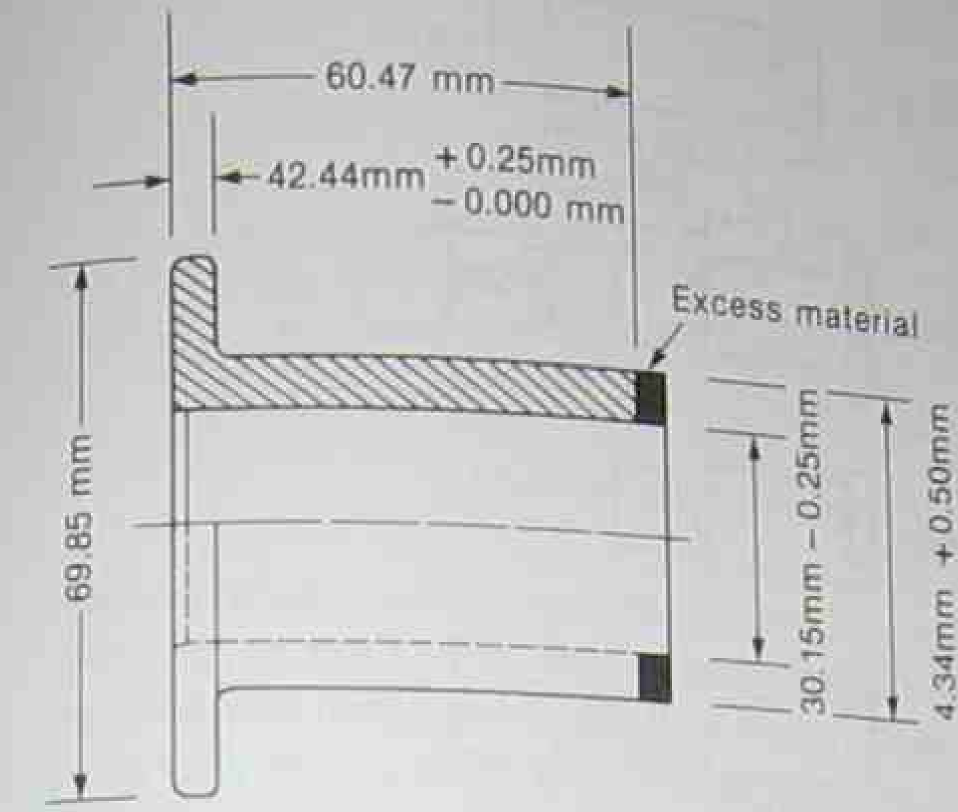


Figure 14–24 A bearing race sleeve.

Other interesting cases of choice of alternative production methods on the basis of quantity and cost are described by the ASM Committee on Cold Heading.* One example involves the case of the production of lawnmower wheel bolts shown in Fig. 14–25. These bolts were originally produced by (1) heading the slug, and simultaneously extruding the opposite end to 0.525-in. diameter; (2) coining and trimming the round head to hexagon shape; and (3) turning the bolt blank to 0.331-in. diameter in a secondary operation prior to rolling the thread. With the improved method shown in Fig. 14–25, the slug was first extruded to form two diameters on the shank end, then headed, coined, and trimmed. By this procedure the minor extruded diameter was ready for thread rolling, and no turning was required. The improved method not only reduced cost by 40 percent, by eliminating the secondary turning operation, but also produced a stronger part, because metallurgical flow lines were not interrupted at the shoulder. Owing to the turning operation, production by the original method was only 300 pieces per hour. By the improved method, 3000 pieces could be produced per hour. The ASM Committee quoted another example of a study at one plant which compared the costs of producing the pin shown in Fig. 14–26 in the same quantity (25,000), using the same material (AISI 8740 steel) by three different processes, machining, hot heading, and cold heading. Table 14–6 gives the results. Although tool and setup costs were greater for cold heading than for the two competitive methods, these higher initial costs were outweighed by the lower costs for material and production when cold heading was used. The lower production cost for cold heading resulted from the high production rate. For larger quantities, the cost differential would have increased in favor of cold heading; for smaller quantities, the cost advantage of cold heading would have decreased.

All these sorts of cost estimates from different sources lead to the following main conclusions:

* Source Book on Cold Forming (Metals Park, Ohio: American Society for Metals, 1975).

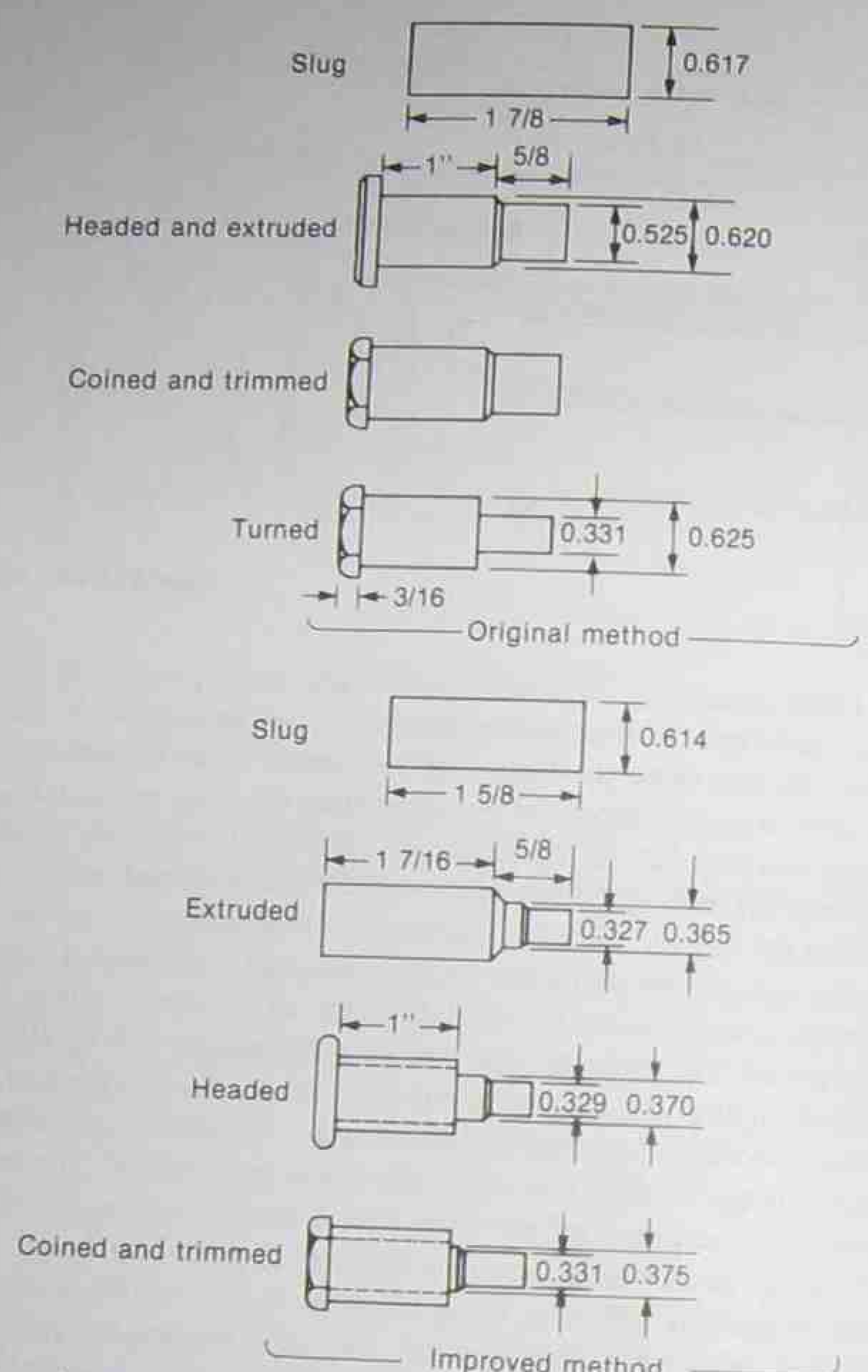


Figure 14-25 A lawnmower wheel bolt produced by two different methods.

1. Machining costs are usually less for cold than for hot forging, because cold forgings are nearer net shape.
2. Cold forging costs decrease with increasing batch size.
3. Cold forging requires less raw material.

A remarkable example of economic savings which can be achieved when bulk machining is replaced by cold forging concerns the manufacture of helical planetary

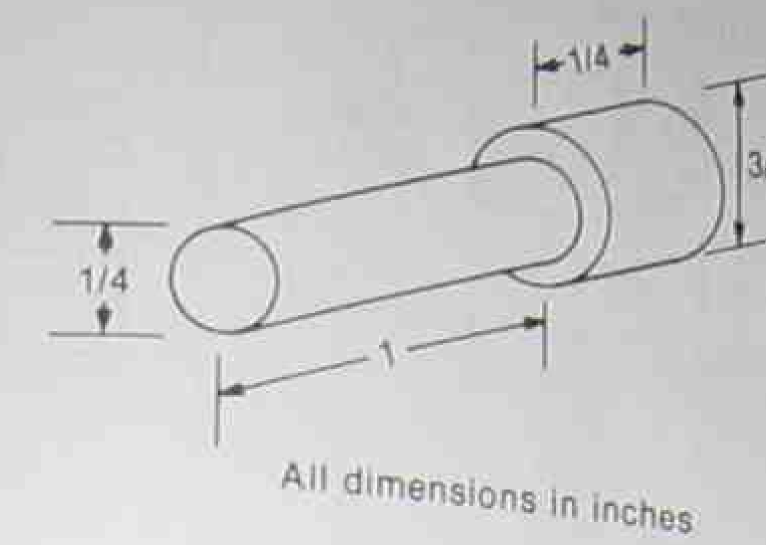


Figure 14-26 A pin produced by three different methods; see Table 14-7 for cost comparisons.

pinions of automobile transmissions. In 1976, about 60 million of these were being made per year in the United States, all by traditional gear cutting methods. An alternative method was developed at the Ford Motor Co. by which the gears can be cold extruded with fully formed teeth, have virtually no defects, and require no subsequent machining. Merely on the basis of using lower-cost feedstock with minimal scrap generation, the new process showed cost savings of between 30 and 50 percent, and savings are further increased when production levels and rates are increased. Much greater production rates are achievable by the forming route. Hobbing *one* gear takes between 1 and 2.7 minutes, whereas gears can be extruded at the rate of 10 per minute. A Ford report* indicates that the saving per car is around $7.

In place of machining versus forming of solid billets (which has been the basis of most examples so far), other manufacturing routes may be possible. Consider Fig. 14-27, which shows a spur gear that is traditionally made by cutting from a blank but could be made alternatively by using the techniques of powder metallurgy. Specifically, these involved powder compaction and sintering of a suitable preform, followed by forging and grinding to the precise finished dimensions. A detailed comparison of the various operations required for both the machining and the powder metallurgy/forging routes is given

* S. K. Samanta, *Proc. NAMRC-IV*, **199** (Battelle Labs., Columbus, Ohio, 1975).

TABLE 14-6 COST OF PRODUCING 25,000 STEEL PINS BY THREE DIFFERENT METHODS

Item	Machining	Hot Heading	Cold Heading
Material cost	$295.00	$212.50	$192.50
Tooling cost*	60.00	90.00	150.00
Setup cost	30.00	35.00	50.00
Production cost	382.50	432.50	65.00
Total cost, 25,000 pcs.	$767.50	$770.00	$457.50
Cost per piece	$0.0307	$0.0308	$0.0183
Production in pcs/hour	285	430	5000

*Amortization of machines not included.

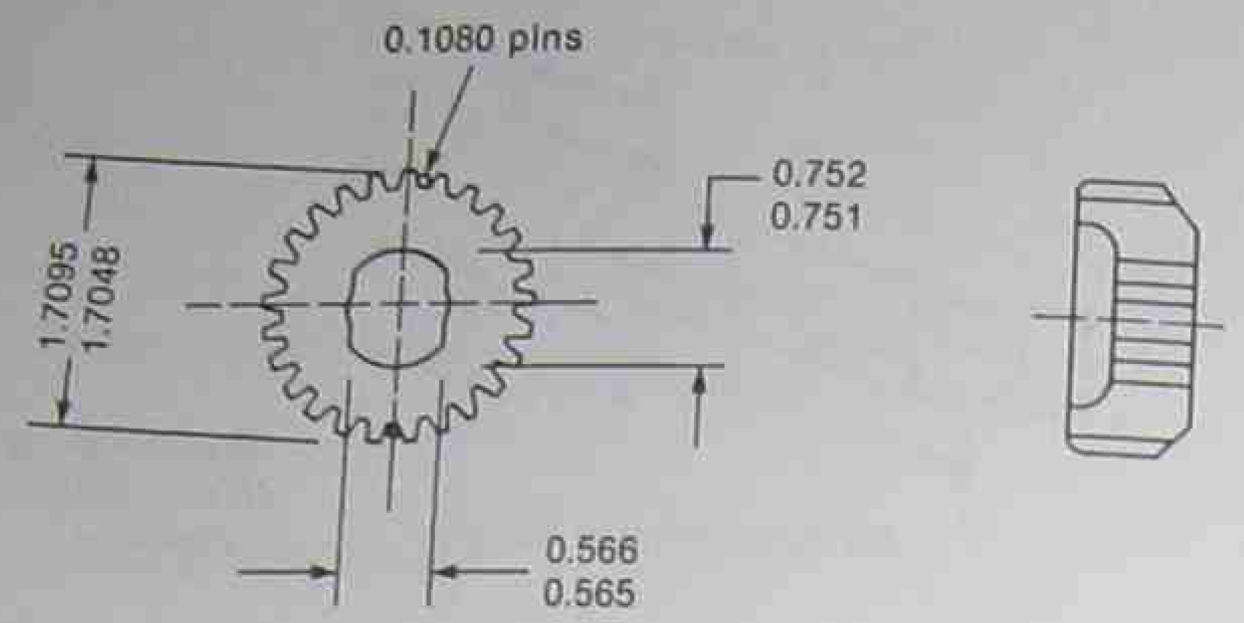


Figure 14–27 A typical spur gear. Comparative costs to produce by machining and by powder metallurgy are given in Table 14–18.

in Table 14–7. A cost advantage is shown for the latter method, mainly due to reduction in the quantity of raw material required.

When it is possible to completely *redesign* a component, or an assembly of components, there can be considerable savings in achieving the same end result, that is, a different design of component or assembly that still performs as appropriately as the original design. This is, of course, the essence of *design for production*. By way of illustration of the principles involved, Fig. 14–28 shows how a shaft can be redesigned on the basis of equivalent functional requirements and suitability for cold forming. The redesign on the right-hand side of the illustration satisfied both these criteria, and additionally saved the material shown crosshatched.

Another pertinent example is shown in Fig. 14–29. The part (a) is a simple shaft with a central hub that might later have gear teeth machined in. To turn it on a lathe (b) requires 18 cu in. of solid barstock, whereas the finished component is only 8 cu in. An alternative manufacturing route (c) would be to redesign the component so that a ma-

TABLE 14–7 COSTS TO PRODUCE A SPUR GEAR BY MACHINING AND POWDER METALLURGY FORMING

Machining		Powder Metallurgy Forming	
Material	26.00	Material	14.00
OD turning, boring, drilling, parting	0.34	Compaction	0.76
Gear cutting, broaching	4.83	Sintering	2.61
Case hardening	2.00	Forging	5.68
Grinding	3.50	Machining, grinding	4.54
Inspection	1.89	Inspection	1.89
Scrap	1.12	Scrap	3.78
Total cost*	39.68		33.26

*These are in cents/part, with labor and overhead excluded.

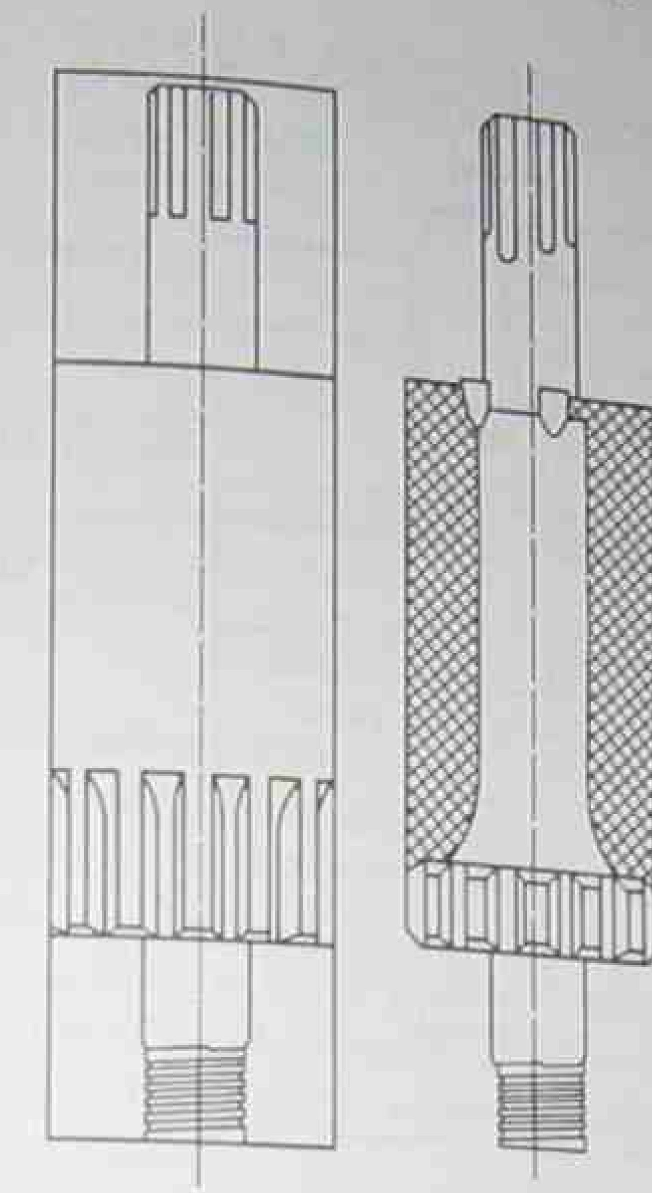


Figure 14-28 Two designs of a shaft made by machining (left) and cold forming (right). Crosshatched area on right indicates material savings.

chined sleeve is pressed over the shaft. But following the theme of earlier examples in this section, much less material would be wasted were the hub hot-upset-forged integrally with the shaft (d) or, better still, produced by cold extrusion/cold forging.

We conclude this section by repeating an observation made earlier in Sec. 13.2: When saying a certain process is the best to produce a component, we must recognize that the equipment appropriate for the task may not be available in a given organization. Even if suitable equipment is available, it may be fully committed to the manufacture of other products. If the projected costings relate to large production volumes, there may be financial justification in purchasing new equipment. However, for smaller quantities, it may be cheaper overall to use less optimum, but available, equipment.

14.8 ENERGY COSTS AND MANUFACTURING

The production of metals, polymers, and ceramics used in manufacturing processes requires energy, so this cost must influence the direct starting price of materials before any manufacturing operations are carried out. Table 14–8 shows the energy/ton required for overall metallurgical and materials extraction and processing, where some materials require proportionately more energy in the original making stage from the raw materials and less on subsequent processing, as seen in Table 14–9. Also shown in Table 14–8 is the energy content involved in providing equivalent stiffnesses and strengths in a body de-

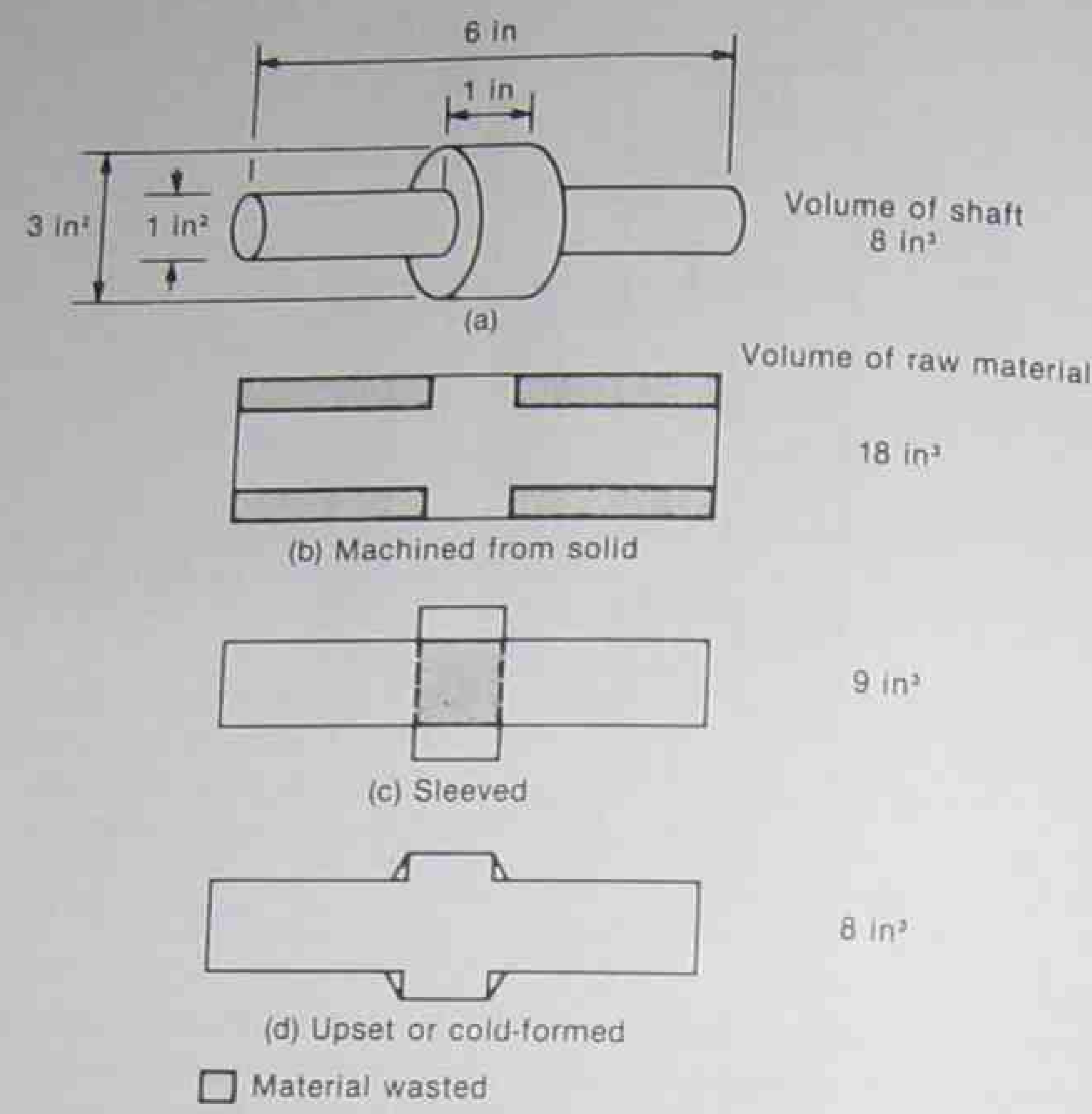


Figure 14-29 Various methods used to produce a stepped shaft to illustrate the interaction between design and manufacturing on material utilization.

TABLE 14-8 ENERGY CONSUMPTION IN THE MANUFACTURE OF VARIOUS MATERIALS

Material	Energy/Tonne-GJ	Energy Per Unit Stiffness	Energy Per Unit Strength
Aluminum	200	2.8	870
Cast iron	45	0.3	300
Low-carbon steel	50	0.24	240
Glass	20	0.30	333
Brick	6	0.10	30
Concrete	2	0.05	80
Wood	1	0.07	10
Polyethylene	45	90	1500
Carbon fiber-reinforced composite	4000	27	6000

TABLE 14-9 ENERGY REQUIRED OR AVAILABLE AT DIFFERENT STAGES IN THE MANUFACTURE OF STEEL AND ALUMINUM COMPONENTS

Stage	Carbon Steel	Aluminum
Manufacture of material	36	300
Primary working	3	40
Machining	1	1
Swarf	40	340
Energy saved by using scrap	10	295
Energy required to melt	1	1

signed to be manufactured from a given material. On this basis, the pursuit of materials based on exotic fibers is quite inefficient in terms of energy. Clearly, those materials requiring high energy for production are most vulnerable to changes in energy prices.

The oil crisis of 1973–1974 resulted in significant increases in the price of oil, and hence increases in the cost of all processes utilizing oil. This applied to processes involving oil both directly and indirectly. Oil-fired furnaces, electricity generated by steam turbines fed by oil-fired boilers, and the products of the petrochemicals industry all demanded higher costs.

It is clear that any increasing shortages, with the subsequent rising prices of energy, will necessarily lead to a shift to production methods that save energy. The saving will be on a global basis, although actual changes in methods of processing and manufacturing seen in different industries will vary with the size of the enterprise. Some who buy all of their starting raw materials will be concerned primarily with changes in traditional methods of manufacture to more energy-saving routes. However, their decisions will also be based on increased prices of the starting material, increases which may depend on the type of product. As a rough guide, 25 percent of the costs of common metal feedstock are energy-related costs.

We have already seen that all machining processes are wasteful in energy terms. It costs money to produce the feedstock in the first place; energy is consumed in the manufacturing operation, and much of the raw material becomes scrap, the value of which is far less (<20 percent) than that of the cost of starting material. It also takes energy to recycle the scrap back to good feedstock, and that is not always possible. Forming processes, on the other hand, are more energy-efficient in that there is far less scrap. Energy is required to conduct the forming operation, of course, and the capital cost of forming equipment is often greater than that of metal cutting equipment. The examples in Sec. 14.7 show that metal cutting has always been cost-effective for small batch sizes, but forming wins out for long production runs. There will be a breakeven point beyond which it will be cost-effective to alter the manufacturing method, as indicated in Fig. 14-23.

Given all these favorable indicators to change away from metal cutting methods, we may wonder why machining processes "hang on." One reason relates to the fact that many factories are already equipped with machine tools, which are still being paid for, and also that the labor force is more skilled and experienced at cutting than at forming.

Even so, economics must rule at the end of the day and cheaper manufacturing methods will predominate.

When anticipated increases in energy costs are considered, the breakeven point moves more in favor of forming processes. Of course, there are some operations which, owing to the geometries involved, are unlikely to be displaced by forming; drilling, tapping, slotting, and the generation of thin-walled sections are examples. Also finish-turning and grinding operations, which produce fine tolerances and fine surface finishes, will inevitably remain important.

Acknowledgements, continued

Figure 4–5, 4–6, 4–8, 4–9
From K. H. Moltrecht and R. M. Caddell, "How to Determine Production Tolerances–Part 2," *The Tool Engineer* (Nov. 1957), pp. 85–89. Reprinted courtesy of The Society of Manufacturing Engineers.

Figure 6–1, 6–2, 6–3, 6–4, 6–5, 6–6, 6–8, 6–9, 6–11, 8–1, 8–10, 8–11, 8–12, 8–13, 8–14, 8–15, 8–16, 8–17, 8–18, 8–21, 8–22, 8–23, 8–24, 8–25, 8–26, 8–27, 8–28, 8–29, 8–30, 8–31, 8–32, 8–33, 8–34, 8–35, 8–36, 8–37, 8–38, 8–39, 8–40, 8–41, 8–42, 8–43, 8–44, 8–45, 8–47, 8–48, 8–49, 8–50, 8–51, 8–52, 8–53, 8–54, 8–55, 8–56, 8–58, 8–59, 8–60
From William F. Hosford and Robert M. Caddell, *Metal Forming: Mechanics and Metallurgy* (1983). Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, N.J. pp. 2–86.

Figure 6–10
From R. M. Caddell and R. Sowerby, *Bulletin of Mechanical Engineers Education*, **8** (1969), pp. 31–43. Used with permission of the *International Journal of Mechanical Engineering Education* and Pergamon Press.

Figure 6–7, 6–10, 7–33, 7–34, 7–35, 7–36
From Robert M. Caddell, *Deformation and Fracture of Solids*, (1980). Reprinted with permission from Prentice-Hall, Inc. Englewood Cliffs, N.J.

Figure 6–13, 6–16
From *Materials Science and Engineering*, American Society for Metals (1975), pp. 213–20. Used with permission ASM.

Figure 6–13, 6–14, 6–15, 6–16, 7–4, 7–5, 7–10, 7–11, 7–24, 7–25
From David K. Felbeck and Anthony G. Atkins, *Strength and Fracture of Engineering Solids*, (1984), p. 45. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, N.J.

Figure 6–14, 6–15
Reprinted from R. Raghava, R. M. Caddell, L. Buege, and A. G. Atkins, *Journal of Macromolecular Science and Physics*, **B6**, 4 (1972), pp. 655–65, courtesy of Marcel Dekker, Inc., N.Y.

Figure 7–1, 8–57, 8–58
From *Metals Handbook*, Vol. **7**, 8th ed., American Society for Metals, (1972), p. 4–14. Used with permission ASM.

Figure 7–4, 7–5
From *Metals Handbook*, Vol. **8**, 8th ed., American Society for Metals, (1973), p. 275–76. Used with permission ASM.

Figure 8–11, 8–50
From D. J. Meuleman, Ph.D. thesis, The University of Michigan (1980).

Figure 8–12
From G. Dieter, *Mechanical Metallurgy*, 2nd ed., (New York: McGraw-Hill Book Co., 1976). Used with permission.

Figure 8–17
This was adapted from J. Wistreich, *Metals Review*, **3** (1958), pp. 97–142.

Figure 8–28, 8–29
From R. M. Caddell and A. G. Atkins, *Journal of Engineering for Industry Transactions of ASME*, Series B, No. 90 (1968). Used with permission of ASME.

Figure 8–30
From B. B. Hundy and A. R. E. Singer, *Journal of the Institute of Metals*, **83** (1954–55), p. 402. Used with permission of The Institute of Metals.

Figure 8–31
From H. C. Rogers, General Electric Report No. 69-C-260 (July, 1969).

Figure 8–32
Reprinted with permission from *Transactions of the Metallurgical Society*, Vol. **175** (1948), pp. 337–54. A publication of The Metallurgical Society, Warrendale, Pa.

Figure 8–33
From D. J. Blickwede, *Metals Progress*, **97** (May 1970), pp. 76–80. American Society for Metals. Used with permission ASM.

Figure 8–36
From M. G. Cockroft and D. J. Latham, *Journal of the Institute of Metals*, **96** (1968), pp. 33–39. Used with permission of The Institute of Metals.

Figure 8–37
From A. L. Hoffmanner, Interim Report, Air Force Contract F 33615-67-C-1466, TRW (1967). See also G. E. Dieter in *Ductility*, ASM (1968), pp. 1–30. Used with permission.

Figure 8–38
From H. C. Rogers and L. F. Coffin, Final Report, Contract NOW-66-0546-d, Naval Air Systems Command (June 1967). See also H. C. Rogers in *Ductility*, ASM (1968), p. 48. Used with permission.

Figure 8–39, 8–40
From H. A. Kuhn, "Formability Topics–Metallic Materials," ASTM STP 647 (1978), pp. 206–219. Used with permission of ASTM.

Figure 8–45
From C. T. Yang, *Metals Progress*, **98** (Nov. 1970), pp. 107–110. Used with permission of American Society for Metals. See also J. Datsko, *Materials Properties and Manufacturing Processes* (New York: John Wiley and Sons, 1966), p. 318

Figure 8–47
From W. F. Hosford and W. A. Backofen, "Strength and Plasticity of Textured Metals," *Fundamentals of Deformation Processing*, Proc. 9th Sagamore Army Materials Research Conference, eds. W. A. Backofen, et al. (Syracuse, N.Y.; Syracuse University Press, 1964). Used by permission of the publisher.

Figure 8–49
Reprinted with permission from *Formability: Analysis, Modeling and Experimentation*, ed. by W. F. Hosford, ed. Used with permission of The Metallurgical Society, Warrendale, Pa.

Figure 8–51, 8–53
From D. V. Wilson and R. D. Butler, *Journal of the Institute of Metals*, **90** (1961–62), pp. 473–83. Used with permission of The Institute of Metals.

Figure 8–54, 8–55
From S. S. Hecker, *Sheet Metal Industries*, 52 (1975), pp. 671–75. Used with permission of Sheet Metal Industries.

Figure 8–56
From A. K. Ghosh, *Journal of Engineering Materials and Technology*, ASME, H, 99 (1977), p. 269. Used with permission of ASME.

Figure 8–60
From C. E. Pearson, *Journal of the Institute of Metals*, **54** (1934), p. 111. Used with permission of The Institute of Metals.

Figure 8–63, 8–64, 8–65, 9–1, 9–14, 9–16, 9–17, 9–18, 9–19(a), 9–21, 9–24 (a), 9–36(a), 11–5, 11–6, 11–9, 11–10, 11–12, 11–13
From Lawrence E. Doyle, Carl A. Keyser, James L. Leach, George F. Schrader, and Morse B. Singer, *Manufacturing Processes and Materials for Engineers*, 3rd ed. (1985). Reprinted with permission from Prentice-Hall, Inc., Englewood Cliffs, N.J.

Figure 9–2, 9–10(a), 9–19(b), 9–25
From *Fundamentals of Tool Design*, ASTME (1962). Reprinted by courtesy of the Society of Manufacturing Engineers.

Figure 9–4(b), 9–5, 9–24(b)
From Amstead, Ostwald and Begeman, *Manufacturing Processes*, 7th ed. (1977). Used with permission of John Wiley and Sons, Inc.

Figure 9–20
Photo orginally supplied by the Carborundum Co.

Figure 9–22
Used with permission of Warner and Swasey.

Figure 9–40, 9–41
Adapted from *Manual on Cutting of Metals*, 2nd ed., ASME (1952), pp. 270–71.

Figure 11–1
Courtesy of Central Foundry Division, General Motors Corporation. Used with permission.

Figure 11–2, 11–16, 11–17
From *The Making, Shaping, and Treating of Steel*, 9th ed. U.S. Steel Corp. (1971). Used with permission of the U.S. Steel Corp.

Figure 11–18(a), 11–18(c)
From D. S. Clark and W. R. Varney, *Physical Metallurgy for Engineers*, 2nd ed. (D. Van Nostrand and Co. Inc., 1962). Reprinted by permission of PWS Publishers, Boston.

Figure 11–18(b), 11–19
From M. L. Begeman and B. H. Amstead, *Manufacturing Processes*, 6th ed. (New York: John Wiley and Sons, Inc., 1969). © John Wiley and Sons, Inc. Used with permission.

Figure 11–22
From F. R. Samp and M. E. Shank, *Aircraft Turofans: New Economical and Environmental Benefits*, (Sept., 1985), Mechanical Engineering, p. 47. Used with permission of Pratt and Whitney Engineering Division, United Technologies, CT. Photograph courtesy of D. M. E. Shank.

Figure 13–2, 13–3
From B.P. McMahon, "Welding and Metal Fabrication," (1970). Used with permission of Fuel and Metallurgical Journals Ltd., Industrial Press, Sutton, Surrey.

Figure 14–3, 14–4, 14–6, 14–7
Courtesy of the Fulmer Research Institute Ltd., England.

Figure 14–8, 14–10
R. G. Salter, FWP Journal (August 1983). Used with permission.

Figure 14–9
J. G. Hicks, "Machine Design and Control," 1970. Used with permission.

Figure 14–25
From "Source Book on Cold Forming", used with permission of The American Society for Metals, Metals Park, Ohio.

Figure 14–29
From George E. Dieter, *Engineering Design* (New York: McGraw-Hill Book Co., 1983), p. 210. © 1983 McGraw-Hill Book Co. Used with permission.

index

a

b

c

d

e

f

g

h

i

j

k

l

m

n

o

p

q

r

s

t

u

V

W

Y

MANUFACTURING ENGINEERING

Economics and Processes

Kenneth C Ludema Robert M. Caddell Anthony G. Atkins

This book addresses the design of manufacturing systems and the fundamental aspects of their individual processing methods. Information on numerical problem solving and all three types of units—SI, metric, and English—is also found throughout. The primary principles of manufacturing engineering are broken down into three sections:

- Philosophical aspects and the importance of economic considerations (with a concise coverage of surface [illegible] tolerances)
- Review of mechanical properties and engineering materials; a discussion of how these apply to both [illegible] nontraditional, or special, processing methods
- Case studies that [illegible] producing parts by different manufacturing methods [illegible] integrated design and manufacturing.

You'll also find information [illegible] automation and computers, engineering mat[illegible] of integrated manufacturing . . . and much more.

PRENTICE-HALL, INC., Englewood Cliffs, N.J. 07632

ISBN 0-13-555582-5

Management For Quality Improvement

THE 7 NEW QC TOOLS

Shigeru Mizuno, Editor

Management for Quality Improvement

THE SEVEN NEW QC TOOLS

Edited by

Shigeru Mizuno

Foreword by Norman Bodek, President
Productivity, Inc.

0120LF0207000

Productivity Press

Cambridge, Massachusetts and Norwalk, Connecticut

Original Japanese edition, *Kanrishi to Sutaffu no Shin-QC-nanatsu-dogu*, edited by Shigeru Mizuno, published by JUSE Press, Ltd., Tokyo, Japan,

English translation rights arranged with JUSE Press, Ltd. through Japan Foreign-Rights Centre

Productivity Press
P.O. Box 3007
Cambridge, MA 02140
(617) 497-5146

Library of Congress Catalog Card Number: 88-042625
ISBN: 0-915299-29-1

Cover design: Bill Stanton
Typeset by Publication Services, Boston, MA
Printed and bound by Maple-Vail Book Manufacturing Group
Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Management for quality improvement.

Translation of: Kanrisha to sutaffu no shin QC nanatsu-dōgu.
Includes index.
1. Production management — Quality control.
I. Mizuno, Shigeru, 1910- . II. Title.
TS156.K35413 1988 658.5'62 88-42625
ISBN 0-915299-29-1

93 10 9 8 7

Contents

Publisher's Foreword

Productivity Press is pleased to present *Management for Quality Improvement: The Seven New QC Tools.* This book introduces the latest advances in tools for quality management—tools that promote a new, more creative and effective approach to quality planning and project management. *The Seven New QC Tools* are used by top executives for strategic planning as well as at all levels of management for planning, goal-setting, and problem-solving.

Most of us are familiar with the original seven tools used in statistical quality control: the cause-and-effect diagram, pareto chart, histogram, check sheet, control chart, bar graph, and scatter diagram. These are used in data-gathering and analysis to solve specific QC problems while total quality control (TQC), as the name implies, involves problem-solving companywide. TQC marshals the skills, information, and efforts of many people—across different departments and over extended periods of time—to address quality problems that go beyond manufacturing into such areas as design, delivery, and service. Total quality control demands that we build quality into our products; the seven new QC tools were designed to help us build quality into every single management decision.

How can we assure quality in planning and management?

Lately, we have been hearing more and more about the hidden "software" of the Japanese – the management techniques that permit the most productive companies to plan and successfully implement wide-ranging and detailed TQC objectives. *The Seven New QC Tools* are at the heart of this management "program for success." Look at almost any Japanese book published in the last ten years on achieving manufacturing excellence: You don't have to understand Japanese to see that these new graphic QC management tools are essential to the Japanese quality improvement effort. Over and over, we see them used both to analyze information and to communicate it graphically and effectively.

For example, the affinity and relations diagrams introduced in this book are widely used during the planning stage to identify problems by organizing diverse forms of verbal data and clarifying complex causal relationships. Once problems are identified, the systematic diagram and the matrix diagram methods facilitate the search for appropriate solutions and organize the steps toward achievement of quality objectives. Finally, the arrow diagram (used in PERT) and the process decision program chart (PDPC) assist in planning and controlling actual implementation.

Our mission at Productivity Press is to make available the ideas and tools that have revolutionized manufacturing in Japan. We are especially proud to introduce the first book dealing with this important new approach to total quality management.

We extend our special thanks to Mrs. Haruko Mitsuaki, Managing Director of JUSE Press, Ltd. (Japanese Union of Scientists and Engineers) for her help and cooperation in bringing this book to America. Thanks also to Connie Dyer who supervised the project and to the staff of Editorial Services of New England, Inc., who helped produce the book.

Norman Bodek

Preface

The meetings of the Society of QC Technique Development, part of the quality control basic course sponsored by the Japanese Federation of Science and Technology, served as the birthplace of this book. The Society's first meeting was held on April 26, 1972. Now, nearly seven years later, it gives us great pleasure to observe the publication of this book. The authors would like to take this opportunity to briefly note the main points of *Seven New QC Tools*.

The general consensus remains that the main objective of company-wide promotion of quality control is achieving the company's business goals through basic reform in the following five areas:

1. *Distinguishing potential future development projects*. Merely finding out what sells and then trying to produce it less expensively than competitors is not enough. Instead, development of technology and systems capable of competing in the world market must be found. It is necessary to transform potential market needs into future development projects.

2. *Planning seriously for the future.* Simply "putting out the fire" once trouble has begun is not sufficient. Anticipating likely trouble spots before problems occur is much more important.
3. *Paying strict attention to processes.* Increased profit is not always an indicator that systems are functioning well. If that were the only criterion, recessions would indicate a poorly operating system. Processes need to be continually evaluated to facilitate needed improvements.
4. *Prioritizing and focusing attention on problems.* Efforts should be made to achieve business objectives within set resource and expense limits. From all the problems a company faces, being able to distinguish those which must be dealt with first is necessary in order to meet the corporate goals.
5. *Focusing attention on the corporate system.* Individuals working alone, even though they are doing their best, cannot compete with a company in which members cooperate in a closely interrelated system.

In order to promote these basic reforms, every employee of a business should be thoroughly familiar with the philosophy of "thinking quality management." The need for managers and staff to "think quality management" has increased dramatically in recent years as a result of the continually changing social and economic atmosphere.

The tools used in the past for quality management include the former seven tools, various statistical methods, experimental design methods, etc. These tools have been used effectively in all fields of company-wide quality control by managers and QC circle members. But now, to meet the demands that social change requires, managers and staff must supplement the traditional tools of quality control with new techniques. This served as the motivating factor when the Society for QC Technique Development was originally organized.

Managers and staff who promote company-wide quality control should not place heavy emphasis only on data collection and analysis. More appropriately, their duty lies in identifying problems, establishing plans, and supporting interdepartmental coordination. Managers and staff must assimilate diverse verbal as well as technical information and develop it into specific plans with innovative

flair. The tools they need are those which can be most useful now. Some of these new techniques have already been used in various ways by QC "pioneers" with considerable success. Hoping to promote these new methods, the Society for QC Technique Development presents *Seven New QC Tools.*

Since April of 1972, the Society for QC Technique Development has held monthly study meetings for the purpose of putting together the seven new tools. The Society's activities consisted of identifying and evaluating various management-control techniques used in areas such as operations research (OR) and value engineering (VE), diverse creativity techniques, and other company-wide QC techniques, looking for those which have proven most effective. Each technique was investigated through applications and outcome. In addition, each technique was also tested in companies not affiliated with Society members. *Seven New QC Tools* is based on the cumulative results of these investigations and other experimental trials.

The proposed seven new tools are as follows:

1. Relations diagram
2. KJ method (affinity diagram)
3. Systematic diagram
4. Matrix diagram
5. Matrix data analysis
6. Process decision program chart (PDPC)
7. Arrow diagram

The rationale for designating these techniques as the seven new tools is based on the collective experience of Society members that an outcome is assured and success is heightened when the techniques are applied to all aspects of company-wide quality control in a closely coordinated manner. It is further believed that these techniques in no way contradict or detract from earlier QC techniques; they actually complement each other, thus contributing to the promotion of total quality control.

The Society for QC Technique Development finished its research late in 1976. Starting in January of 1977, the Society devoted itself to the refinement and promotion of its ideas through lectures, conferences, seminars, and symposia. In 1978, the Society invited Mizuno Shigeru, professor at the Tokyo Industrial University and Kondo Yoshio, professor at Kyoto University to act as advisors. For nine months, monthly meetings were held under the name the Research Society for the Seven New QC Tools. These meetings contributed enormously to the development of a conceptual framework and techniques for the *Seven New QC Tools*. At the same time, a large number of applications of these principles began to pour in from various companies.

As our promotional activities progressed, we received many requests from different people in diverse fields asking us to teach more about the theory, use, and application of the seven new tools. In addition, many people whose opinions we value advised us to make this material on the seven new tools more widely available. In response to these demands and this advice, we have compiled the present book.

The book consists of two parts. Part I describes the background and rationale behind the seven new tools, provides an outline of each technique, examines the relationship of each tool to other QC techniques, and discusses the graphic and linguistic basis of the new tools. We also introduce examples of systematic application in policy management, hoping that the reader will be able to grasp fairly quickly not only the main features, but also a few of the different types of applications.

Part II provides a detailed discussion of the conception, construction, and use of each one of the seven new tools. A variety of examples to which these methods can be applied is given so readers will be able to modify and apply the tools to their respective companies and situations. We hope our readers will apply these techniques and inform us of their experiences in order for us to improve on any weak points that might surface.

We gratefully acknowledge the guidance and advice of Professors Mizuno and Kondo. In spite of his busy schedule, Dr. Mizuno agreed to take on the role of editor-in-chief and has written a gracious foreword. He has also been my lifelong mentor, for which I am extremely thankful.

We are indebted to Chizumi Shizuo, professor at Keio University, who provided guidance on relations diagrams to the committee

members of the Society, and to the numerous executives and managers of various companies for their encouragement and assistance. Unfortunately, space will not allow all of them to be mentioned here.

Messrs. Aima and Tanaka of the Osaka Office of the Japanese Federation of Science and Technology have made personal sacrifices in supervising the Society for QC Technique Development since its inception in 1977, and they have provided continuing assistance to us. We would like to express our extreme gratitude to them.

Finally, we are indebted to Mr. Tahara and his colleagues at JUSE Press, Ltd. for helping us with the graphs and other complicated tasks inherent in publishing this type of book. It would be impossible to imagine this book ever being completed without the assistance and guidance of these people. We are indeed fortunate to have had their assistance, and accordingly, we are very much indebted to them.

Nayatani Yoshinobu
President
Society for QC Technique Development
April 1979

Contributors (in alphabetical order)

Futami Yoshiji (Osaka Electro-Communication College)
Kato Shoichi (Nippon Paint Co.)
Kurabayashi Mikihiko (Mitsubishi Electric Co.)
Nayatani Yoshinobu (Osaka Electro-Communication College)
Sano Motohiko (Sekisui Chemical Co.)
Yagi Juichi (Mitsubishi Electric Co.)

Introduction

It probably goes without saying that the quality of Japanese products, such as automobiles, home appliances, and cameras, is far above the quality of similar foreign-made products. These products were originally designed and manufactured overseas, but the recent superior quality level of Japanese products has resulted in massive exports, giving rise to a growing chorus of objections to the further expansion of Japanese manufacturing, which supposedly deprives workers in foreign markets of their jobs. Japan's future options appear limited to either manufacturing at foreign plants that draw from the local work force, manufacturing products that do not compete with foreign products, or developing creative new products that meet the needs of foreign markets.

As Dr. Juran pointed out at the International Conference on Quality Control held in October of 1978 in Tokyo, the quality of Japanese products in the 1950s was so poor that it seemed that Japanese industry would not survive unless it could improve its products enough to be able to export to foreign markets. Since that time, Japanese industry's efforts for improved quality control, fueled by survival instincts, have brought about splendid results, and now Japan faces the new challenge of searching in new directions for the future. Even prior to Dr. Juran's comments, the Japanese business sector had been expanding the application of quality control (QC) from traditional areas such as manufacturing to planning, development, and design and even extending it to include postmanufacturing areas such as sales and service. In other words, the scope of quality control had been expanded companywide. As part of these promotion activities, efforts were made to improve the fundamental company culture, for example, through companywide participation in quality control activities from top management to the basic quality control circles.

There are seven traditional tools of quality control, including the well-known pareto chart, the cause-and-effect diagram, and the control chart. These tools have been used as an effective means of analysis and control, and they have contributed significantly to quality improvement. The seven tools have been the favorites of QC staff as well as of QC circle members, largely because the tools are easy to understand. No one doubts that these tools will continue to be used extensively. Quality control has entered a new era of development, however, and there is no room for complacency with the present tools. A new era demands new tools.

In this new era of quality control, the boundaries for QC involvement are limitless: Activities of managers and staff are expanding to include resolving major quality problems, developing products with new levels of quality, and setting up and managing the systems necessary for attaining goals such as these.

Management for Quality Improvement: The Seven New QC Tools is designed to meet the needs of this new QC era. This book is the result of the tireless efforts of Dr. Nayatani and his colleagues with the Kansai QC Leadership Group. Although designation as the "seven new tools" runs the risk of creating what may become a passing fad in facing the new era, and although the designation may impart the impression that the seven new tools are better than the original tools, the seven new tools are new techniques not intended to replace the original tools, but to aid in coping with the problems that the new era poses.

In varying degrees, these tools have been used in other fields; however, they have almost never been used for quality control. It is significant that recognizing the need for these tools in quality control, the Kansai group has named them the seven *new* tools.

Although an editor-in-chief might typically be involved in all the details of a book's production, my role was limited to offering suggestions I felt might be useful based on my experience. I wish to emphasize the limited nature of my contribution, lest I be given any undue credit simply because of the title editor-in-chief.

Management for Quality Improvement presents tools that lend themselves to an approach that is forward-looking. This is worthy of special note. In this sense the new tools may provide momentum for new directions in this new era. The practical examples enhance this potential. The many examples reflect actual applications that were of major importance to the companies involved and which, in fact, were candidates for the respective company's President's award. For reasons of privacy and to prevent premature disclosure, some information and a few names had to be changed. Nonetheless, these tools have great promise and potential for immediate results.

I sincerely hope that the seven new tools will be adopted by managers and staff and will contribute to the establishment of a new era for quality control.

Mizuno Shigeru

A Note on Japanese Names

In Japan, the family name appears first. Thus, the famed inventor of the Toyota production system is known in Japan as Ohno Taiichi, and not Taiichi Ohno as usually written in the West.

In Productivity Press books we try to follow the Japanese practice of placing the surname first, in part, to make the representation of Japanese names uniform but primarily out of common courtesy. The reader therefore will find contributors and authors referred to surname first in the text and notes.

Management for Quality Improvement

The Seven New QC Tools

PART
ONE

An Overview of the Seven New Tools for Quality Control

1

Total Quality Control and the Seven New QC Tools

Expanding the promotion of total quality control

The basic objective of total quality control (TQC) consists of bringing about company reforms in the following areas: (1) distinguishing potential future development projects, (2) planning seriously for the future, (3) paying strict attention to processes, (4) prioritizing and focusing attention on problems, and (5) focusing attention on the corporate system. Concerted efforts to improve in these areas make it possible to expect progress in fulfilling corporate duties to society as well as to develop better management systems.

In order to promote and fulfill the basic objectives of total quality control, the following four elements are extremely important:

1. Ideals and viewpoints concerning quality control. What are the company's management goals? What are the company's viewpoints concerning the promotion of total quality control? What is the company's long-range plan for the promotion of total quality control? What are the social responsibilities connected with the business?

2. Specifying and attaining policy directives. What needs to be done to attain corporate ideals? Are the goals stated clearly? Are expected levels of achievement clearly defined?

3. Establishing management systems. Are the quality assurance system, cost-control system, and other necessary systems in place? Are the systems functioning effectively?

4. Tools for quality control. Are the seven tools for quality control and other statistical methods being utilized? Are new ideals, goals, policies, systems, and QC tools being researched?

Because these four things are interrelated and complement each other, if a company is willing to attempt to implement them effectively, then that company can expect the practical results from their application to soar. Essentially, then, the true purpose in desiring reform in the five areas listed in the first paragraph is for all employees of a company, including managers and staff, to develop the talent of "better thinking." This is exactly the expectation of every business manager and the goal of TQC human-resource development. A wise manager once commented that he wished his employees would bring their minds to work, not just their bodies. From this perspective, the adoption of "thinking total quality control" and "thinking quality management" is necessary for expanding the promotion of total quality control.

The seven new QC tools are presented with the conviction that they are the QC techniques that offer the best methods for the stimulation of thinking.

Background behind "thinking total quality control"

In the past several years, dramatic changes in TQC thinking and promotion have taken place. The major social and economic factors that have helped bring about these changes can be summarized as follows:

Continuation of stable economic growth

Even though an era of stable economic growth started in Japan at the end of the 1973 oil crisis, it is unlikely that a growth rate equivalent to the Jinmu-Iwato boom* will ever be experienced again. It is now necessary to reinforce the business systems that must be adopted in order to maintain stable economic growth.

Shift to multiple economic indicators

In the past, countries were concerned merely with their own economic welfare, but recently, the interrelationships among countries in terms of the world economy have become increasingly evident. Countries have realized that GNP alone does not suffice as an adequate economic indicator. A comprehensive evaluation of industrial activities and economic prosperity must now be made on the basis of a variety of indicators such as business trends within other countries, financial market trends, relative trade balances among countries, and the trade of specific products. Many nations need to reevaluate policies designed to simply boost their own GNP.

When only one indicator was used, the obvious goal was to maximize that indicator's value. Presently, however, with the use of multiple indicators to measure an economy, it is neither feasible nor permissible to obtain the highest values for every indicator. A careful, balanced selection of individual indicator levels must be made in terms of overall merits and demerits. This is an era of searching.

Conservation of energy and resources

After surviving the confusion of the oil crisis, the worldwide availability of energy has stabilized. However, many observers forecast a shortage of oil resources again in the 1980s. Although alternate energy plans, such as the "sunshine plan," the "moonlight plan," and nuclear fusion research geared toward the twenty-first century have

*The Jinmu-Iwato boom was a post-World War II period of extreme economic growth for Japan in the 1950s.

been promoted, the prospects for a practical substitute for petroleum are not bright.

While exhaustion of the world oil supply during this generation may not be realistic, our obligation to future generations to pass on all the natural resources possible remains. These considerations highlight the task of businesses to conserve resources in product design and production.

Environmental and public hazards

Viewed from the standpoint of improving the quality of life and respecting human dignity, protecting the environment and avoiding situations that might prove hazardous to the public become increasingly important. Systems for the production, delivery, circulation, use, and disposal of products must be evaluated to confirm that they have no adverse effects on the environment. These are important constraints that relate to all business activities.

Product liability

Judging from recent trends in the United States and the rising awareness of human rights in Japan, the importance of product liability will continue to increase. This also will act as a large industrial constraint, along with the environmental and public concerns just mentioned.

Awareness of customer needs

Despite the various constraints under which businesses must operate, consumers are demanding increasingly sophisticated and advanced products that complement their diverse lifestyles and values. Businesses that want to remain competitive both nationally and internationally, must research and develop products that will maintain a certain level of exportability despite the reduced export competitiveness of Japan caused by the present yen-dollar exchange rate.

Reduced prospects for technology import

Postwar economic development of Japan was based on the import of new technologies from foreign countries. The present situation is drastically different from the 1940s: New and radical technologies are emerging constantly. Fewer technological "seeds" can be developed into giant industries, the foreign ideas that might become these "seeds" are next to impossible to import. The advancement of developing countries has narrowed the technological gap in a number of industries and products. These technological constraints further emphasize the need to promote total quality control for the development and refinement of new products.

A new era for quality

The "new era for quality"[1] proposed by Professor Mizuno could have been anticipated if thoughtful consideration had been given to the economic and social factors discussed above. The first requirement of the new era for quality is the creation of an "added value" over and above consumer needs. It is necessary to first uncover latent customer needs and then, in response, not only meet those needs, but also to discover an added value that will surpass them. The new era for quality expects the generation of new ideas.

The second requirement inherent in the new era for quality is the ability to cope with varying limitations, hopefully without missing any necessary items: In other words, the key is to prevent failure in meeting customer needs. The constraints businesses must work around are numerous: environmental pollution, product liability, efficient use of resources, cost, and productivity, just to mention a few. In the future, business activities will undoubtedly be subject to additional constraints and limitations. In this sense, as Professor Kigure proposed, the major task facing businesses is to shift "from defensive QC to offensive QC."[2] It is important to solidify defenses against constraining factors while at the same time emphasizing an aggressive posture toward new product development.

The book *Quality Deployment*, edited by Professors Mizuno and Akao, represents an outstanding achievement as a quality system for new product development, responding to the needs of the new era.[3] Hopefully, a wide adoption of this system in industry will help businesses to march through this new era for quality.

This is the true objective of the *Seven New QC Tools*.

Highlights of "thinking total quality control"

Based on the background presented earlier, when the four elements of total quality control are applied company-wide in a comprehensive manner, the roles of managers and staff increase dramatically in company reform. In order to promote "thinking total quality control," seven items that highlight this way of thinking are listed and explained below.[4]

Conducting multidimensional evaluations

Managers and staff should always keep multidimensional characteristics and their evaluation in mind, even when pursuing a single objective. In the early stages of importing total quality control, concern frequently centered on reducing the rate of defective products or reducing costs. Now, however, desired outcomes would be difficult to obtain if quality is pursued as a single issue. Disruptions in other related functions often accompany any improvement in quality. Simply put, improvement in one function must be carried out while at the same time considering the constraints it might place on other functions and characteristics.

Thinking that "cost is cost and quality is quality" is no longer sufficient; these concepts are not exclusive of each other. An awareness of functional interrelationships is necessary: "Does the lowered cost bring about irregularities in quality?" "Is there a firm basis to back up the guarantee for product longevity?" Unless one is fully prepared to answer related questions, there will be unexpected claims later. The pressures associated with lowering costs have often

created unexpected accidents after shipping that ultimately brought near disasters to businesses. Thus multidimensional evaluation refers to looking at a problem in its context and totality.

Eliminating the phrase "recurrence prevention"

An important point in the current promotion of total quality control is forbidding the use of the phrase "recurrence prevention." Promotion should proceed with the firm conviction that failure is unacceptable from the beginning. Previously, the phrase "recurrence prevention" was used frequently, and it was sometimes said that "in quality control, the first mistake is acceptable, although its recurrence is not." However, this is an era in which even an initial failure or mistake in the design or development of a new product is simply not permissible. It is crucial to deliver a product to the market as planned and therefore to ring up sales as planned.

Such failures as lower profit because of higher production costs and unanticipated claims after shipment cannot be afforded. When faced with the problem of environmental pollution or product liability, it is totally inconceivable for anyone to maintain the naive posture that "It was a design mistake. Quality control now only has to prevent it from recurring."

In this era, failure is unacceptable from the very beginning. Merely "putting out the fire once it has started" is not sufficient. In the promotion of total quality control, preventing mistakes is necessary. Moreover, to prevent mistakes and errors during the promotion of total quality control, it is important to list all the correlated items.

Consider this issue from the standpoint of problem solving. A current problem causing a malfunction must be corrected as soon as possible. A secondary problem, however, is a situation which, left unattended, has the potential to create malfunctions in the future: It is necessary to be able to predict such malfunctions and introduce corrective measures now. The former type of problem is sometimes called an "emergent" problem; the latter is known as a "prognostic/predictive" problem.[5] Stating that recurrence prevention is not a permissible approach to quality control reflects an emphasis on the predictive type of problem.

Specifying a desirable condition

Understanding the statement "Specify a desirable condition and move toward achieving it" is easier than actually dealing with a situation within the framework of a predicted problem. There are several reasons for this. First, after talking with someone, a person may admit that there is a problem and strive toward a desirable condition. Second, in improving predictive ability, experience shows that thinking in terms of striving for a more desirable condition is more likely to produce interesting and creative ideas than talking only in terms of the problem.

In striving for a "desirable condition," it is important to note that desirable conditions, products, and systems all vary from one type of business to another. Even within a specific type of business, desirable conditions may differ depending on the size and scope of the business. To illustrate this point, consider that some taxis in metropolitan areas are painted yellow or orange, while others have sober colors such as black. Both companies believe they have the right color.

Further, an interview was held with the executive officer of a small taxi company that uses a bright color for its taxis. This company prefers a bright color because it is easier to detect from a distance. Since it changed from black to a brighter color, revenues have increased. Because there is a greater ability to identify the company, the drivers are less likely to refuse customers, and this has resulted in a reduction of complaints. As secondary effects, job stability among drivers is up and their attitudes toward customers have improved as well. The executive concluded that choosing the bright color has been very beneficial. There are some drawbacks, however. For example, the bright-colored taxis cannot be used for funeral services.

Another larger company prefers black because its cars are often hired by large corporate customers and the sober color is more fitting to their tastes and status. Owing to the large number of cars in operation, there is a high risk of minor accidents. In cases of minor accidents, the sober color makes the taxis indistinguishable from private automobiles, which is to the company's advantage. However, the sober color shows dust and dirt easily. For these reasons, these cars are probably not as suitable for operating as taxis.

Although the preceding example is mundane, it illustrates the importance of "thinking total quality control." In other words, this is an era in which simply copying other companies' quality control will not suffice. Each business must establish its own "desirable condition."

As is clear from the preceding example, a "desirable condition" may not be the best one for every evaluative dimension. Both advantages (merits) and disadvantages (demerits) must be considered, various means of achieving the objectives must be compared, and the solution that maximizes merits over demerits must be selected. This is similar to the importance of multidimensional evaluation explained earlier.

Making a truly prioritized effort

As described earlier, total quality control entails the task of prioritizing efforts. Probably foremost in prioritizing efforts is the allocation of resources. This refers to allocating the limited resources of a nation or a business (i.e., human, material, and financial resources, as well as facilities) to those objectives with the most merits. It is particularly important to place emphasis on investment in research and development.

Although the phrase "prioritized effort" has long been emphasized in the field of quality control, the reality of TQC promotion in the business world reveals that the principle of overall harmony, a typical Japanese trait, prevails. In order to determine the relative importance of suggestions made by the subordinates, it is critical for the manager to be decisive and able to distinguish the important from the unimportant.

The second task of the prioritized effort is to assign relative importance to the various steps and processes in promoting reform. Generally speaking, a single project is comprised of many tasks that have to be implemented. In evaluating a project after the passage of a certain amount of time, to cite an extreme example, one may find that 95 percent of the tasks have been completed, but that the overall result is nil. Upon closer examination of such a case, the remaining 5 percent of tasks frequently are discovered to be technically quite difficult. There is a natural human tendency to start with easy tasks and postpone difficult ones. Although the number of tasks to be implemented may be large, usually it is only a small portion that poses any real difficulty. Yet without executing the difficult tasks, the desired objective can never be obtained. In executing long-range plans, an efficient manager identifies the difficult tasks at the beginning and works gradually toward completing them.

For this reason, it is necessary at the inception of a project to list all items or tasks that need execution and to assess the level of ease or difficulty of each by attaching as much technical detail as possible.

Encouraging system-wide promotion

The system-wide promotion of total quality control requires that every member of a company or team cooperate fully. A system functions like a human body. The entire organization of a business should act in organic harmony. For this purpose, it is essential that the function of information transmission, like the brain function in the human body, be performed faithfully. Professor Oba proposed a method called the "theory of QC hearing"[6] which articulates the need for each project director to develop and maintain a close relationship with other project directors or counterparts. It is an excellent idea that attempts to get everyone who is involved in a project to the same level of awareness and knowledge.

In addition to transmitting simple information, it is necessary to study information pertaining to other fields or projects, including their complex language, so that information can be transmitted accurately. For example, suppose that a business planning department's top directive is to "secure *x* new users." Under such circumstances, those in charge of field operations would naturally endeavor to secure the designated number of new users. In the process of securing new users, however, it often happens that old users are lost to competitors. Unknown to the directive's audience, though, the idea of securing a designated number of users presumed certain unstated preconditions. Among these, for example, may be that the existing users be retained or that limits be set on the types of orders received, the suggested price, or the quantity per customer. Directives usually do not spell out such details, but managers and staff are expected to grasp the underlying presumptions, as well as

how they relate to other departments or projects, rather than simply to carry out the directives blindly and literally.

In the example just described, the desired result can be obtained only when there is system-wide cooperation among the planning department, the plant, and the field office. In gaining the cooperation of other departments, it is not sufficient to resort only to official channels. It is nothing more than an excuse if one has to say, "We have talked to them, but they just wouldn't do it." It is clear that unless the related sectors are mobilized, any system-wide reform is hard to achieve. What is important here is to think of ways that encourage one's counterparts to *want* to work. Helpful in this connection are ideas that motivate one's counterparts and provide incentives.

The education of subordinates emerges as an important problem in executing system-wide management. Managers and staff need to meet with subordinates at least once a month to share thoughts, perspectives, and progress. In other words, it is necessary to lead one's subordinates with words and theory and at the same time let them possess the same information. Here the problem of transmitting complex information in a way that is simple to understand is encountered.

Actively making changes

One of the important ideas permeating the present concept of total quality control is that of change. In order to create products with "added value" and to avoid not meeting consumers' needs, previous methods will undoubtedly have to be changed somewhat. Basically, change is the essence of reform. No one could reasonably suggest that drastic reductions in a product's fraction defective rate or a significant lowering of cost could occur if, as in the past, the same employees in the same organization use the same system with the same equipment and the same methods. At any rate, change is critical. It would not be overstating the point to say that managers and staff need only to think of "What should we change?"

Change is important not only in terms of results, but it is also necessary in terms of avoiding becoming entrenched in routines: Through change, people are able to avoid job stagnation. People often tend to become invigorated when they join a new company, change job locations, or get promoted. Managers and staff should constantly introduce changes in their TQC activities to prevent the organization from becoming lifeless and stale.

Looking at the other side, however, failures also tend to occur at the time of change. Trouble might occur when systems change or operators change as a result of personnel transfers, yet without constant change, improvement and reform are hopeless. This is a time when companies cannot afford not to change. In this new era for quality, one of the prime responsibilities of managers and staff is to actively promote change based on a balanced grasp of the various merits and demerits entailed in that change.

Anticipating and predicting the future

As an idea about adjusting to the new era for quality, one last comment is necessary: A smooth and quick rotation through the PDCA (plan, do, check, act) cycle is invaluable. Stated another way, it is better not to waste time rotating through PDCA cycles that will be useless. Therefore, it is important to try to anticipate and predict the expected outcomes. In the future promotion of total quality control, managers and staff will be required to be able to predict and be prepared for future events.

Concerning the problem of prediction, even QC specialists have voiced their concern about whether it is possible to "predict what has not yet happened." However, such predictions are made unconsciously every day. When people play "Go" or chess, they anticipate what their opponents will do during the next two or three moves. Players will at least predict what their opponents will do in response to this move or that move. Professional players can think through various contingency outcomes for a number of moves ahead. QC personnel are professionals in design, production, research, development, business operations, and sales. Unlike the situation when amateurs play "Go" or chess, QC personnel should be able to consider a multitude of factors and contingencies and take action with a certain degree of seasoned prediction. This opinion probably results from the belief that predictive potential has not been fully utilized.

In order to be able to predict, one must be able to assimilate the relationships between potentials, such as verbal information and likely outcome.

QC techniques designed to encourage thinking

As discussed in the preceding section, riding through the new era for quality requires the promotion of total quality control with a new frame of mind and a new perspective. Therefore, it is necessary to introduce new QC techniques which are appropriate to the demands of the new era.

The seven well-known tools of quality control[7] are as follows:

- Cause-and-effect diagram
- Pareto chart
- Checksheet
- Histogram
- Scatter diagram
- Control chart
- Various graphs

These constitute fields of company-wide QC activities. Results from using these tools have proven to be most effective when they are used comprehensively throughout companies, from managers to TQC circle members. However, managers and staff need special techniques in order to utilize the tools mentioned here and elsewhere.

In addition to the seven tools listed above, the tools used in quality control include a number of statistical methods, experimental design methods, and multivariate analysis methods. These tools, however, are used for obtaining data and analyzing available data after the objectives of the investigation have been decided. Although statistical quality control is based on these methods, as a practical matter, not all managers and staff can become proficient enough with statistics to use these tools.

Hopefully, managers and staff will consider the complex relationships among technical details and between departments and then organize and systematize this information as they initiate the phases of their QC activities. In order to accomplish this, managers and staff need new tools to help sort the confusing elements, uncover the underlying problems, and devise a measure of the extent of implementation. This need has long been recognized by authorities in the field of quality control, and the development of a partial solution has been advocated for some time now (see the first section of Chap. 3). It is important to think about how to treat previous problems rather than to simply gather more data. Of the original seven tools, only the cause-and-effect diagram appears to accommodate this need. However, expectations for the seven new tools run high.

Basically, there are seven desirable prerequisites for any tools or methods designed to be of use in the new era. The following subsections illustrate these prerequisites.

The ability to process verbal information

In general, the problems confronting managers and staff involve more verbal data. This is usually data involving both in-house and out-of-house matters, i.e., dealing with technological as well as market information. Managers and staff have a high-level ability to make use of this information. In the promotion of total quality control, it is important to transform language data into either graphs or some other quantitative form so that everyone in the company has equal access to it.

In other words it is hoped that any new techniques will be useful in helping managers and staff consolidate complex and varied verbal information. Such new techniques do not necessarily have to be quantitative or computational. In fact, if possible, computational techniques should be avoided. Rather, techniques that develop and express complex phenomena in terms of graphs and diagrams and uncover underlying problems clearly are sought. An appropriate new technique would be able to identify and adjust problematic elements when quantitative data do not exist, and it would specify the types of quantitative data needed for future analyses.

The ability to generate ideas

In the future, no one could contest the need to consistently promote total quality control using new ideas (see preceding section). New techniques should generate new ideas. Therefore, with this in mind, two needs surface.

The first is that any new technique should make managers and staff feel that they are utilizing their intellectual capabilities. The generation of ideas starts with the use of brainpower. In addition to this "thinking" process, thoughts and ideas need to be expressed in clear statements and diagrams. Through repetition of this process, ideas are transformed into sentences and diagrams that can serve as the starting point for ensuing rounds in the thinking process.

Methods of generating ideas also vary greatly from person to person. New techniques also should be capable of yielding results in proportion to the user's ability. In addition, new techniques should have some positive effects by virtue of having been utilized. For example, anyone can construct a cause-and-effect diagram, brainstorming and listing ideas to be sorted according to their relative importance. However, frequently it is merely the drawing of this cause-and-effect diagram that contributes to a better articulation of problems.

The second need for new techniques in terms of the generation of ideas is that there should be an orderly procedure for the construction and use of cause-and-effect diagrams, while at the same time leaving room for improvement. Because there are numerous ways of drawing these diagrams, any method should be flexible enough to accommodate new ways to utilize these diagrams.

The ability to complete tasks

When a directive is received to proceed with or complete a certain task, one should take into consideration the directive's relationship with other tasks. Any new technique brought to bear should incorporate ways to dissect tasks and assist in forming a step-by-step work schedule.

One of the important aspects of completing a task closely relates to the problem of prediction (see the preceding section). For example, QC activities in research and development aim at completing constantly changing themes and objectives. An important element here is the establishment of a hypothesis designed for problem resolution based on various kinds of verbal and technical information. QC methods in research and development quickly and thoroughly test such hypotheses. Such hypotheses are evaluated from a theoretical perspective and examined for potential value; then a detailed test is designed based on the resulting predictions. It is commonly believed that the quality of plans cannot be evaluated. In specific instances, this is probably due to the fact that there has not been sufficient appreciation of either the problem-solving hypothesis or its predicted outcome.

This situation is not restricted to research and development applications. Since the new era for quality aims to improve conditions across the board in industry, this concept deserves serious consideration in all departments, including the QC activities for high-level products. Hopefully, new techniques will be responsive to these requirements.

The ability to eliminate failure

When quality improvement is suggested in a department and, accordingly, reform plans are implemented, complex problems surface relating to cost, payment period, and facilities. Reform plans can rarely be implemented at the upper levels of management alone; they have to be coordinated with the production process, outside clients, supervisors, and subordinates. It is necessary to promote improvement and consolidate efforts on all fronts while at the same time working to prevent failures or eliminate slippage that might occur during the transition period. New techniques should be useful here as well.

The ability to assist in the exchange of information

As explained earlier, research, design, and development in total quality control, along with QC development in such new fields as atomic power plants, shipbuilding, and building construction, require a higher degree of information sharing between the departments

concerned. In order to achieve this, several things are required: Information should be made accessible to everyone involved. Departmental relationships should be well-defined. Each manager or supervisor should list all the tasks for which they are responsible, as well as the specific tasks inherent to their project. Efforts should be made to merge company-wide technical forces. Presumably, new techniques should be responsive to these needs.

The ability to disseminate information to concerned parties

Since TQC activities are aimed at merging company-wide intellectual power, new techniques should provide a process whereby individuals' ideas can be clearly communicated to others who could benefit from that information. As long as reform is a group activity, ideas and thoughts need to be expressed in an easy to understand manner. The essential aim of this prerequisite, then, is for all the information contained in the group to become the possession of the entire group, and this, in turn, leads to the generation of new ideas.

The ability to use "unfiltered expression"

Quality control materials often suggest "resorting to unfiltered expression," in other words, saying things as they are. This is the equivalent of encouraging an uncensored, unfiltered, lively expression of successful results, the troubles endured during the process of some improvement, or the birth or death of a new idea.

Essentially, "unfiltered expression" correlates closely with another well-known saying in quality control: "Evaluation of the process is more important than the outcome." When one listens to the history of quality improvement in another department in one's company, it is very helpful if the changes in deficiency rate and values of characteristics have been recorded over the years. However, people rarely go so far as to record implementation details over the course of several years of reform. Practically speaking, however, it is hopeless to expect to rouse the interest of the younger generation of employees by sharing with them only tales of trials and hardships.

Preferably, new techniques should encourage an unfiltered, lively, and truthful expression of the promotion process, as well as presenting the information obtained in a manner that is accessible to succeeding generations.

As an attempt to fulfill these prerequisites, seven new QC tools are advocated and promoted here. Details of these tools are provided generally in Chapter 2 and more specifically in Part II. At this point, let us just list the seven tools briefly:

1. Relations diagram method
2. KJ method (affinity diagram method)
3. Systematic diagram method
4. Matrix diagram method
5. Matrix data-analysis method
6. PDPC (process decision program chart) method
7. Arrow diagram method

These methods have all been used in other fields, but an attempt has been made to avoid simply importing these techniques. Each method was assessed in terms of its potential effectiveness in quality control or in terms of previous experience with the techniques in QC applications, especially in areas that pose QC problems. These new methods do not have to be newly invented ideas; if they were techniques that had not previously existed, their utility would be rather suspect.

These seven new tools are being promoted in the belief that they can be used most efficiently when they are combined in an interrelated manner. It also needs to be made clear that the seven new QC tools do in no way contradict or replace the existing tools; the new tools are simply meant to supplement and complement the previous tools (see the second section of Chap. 2).

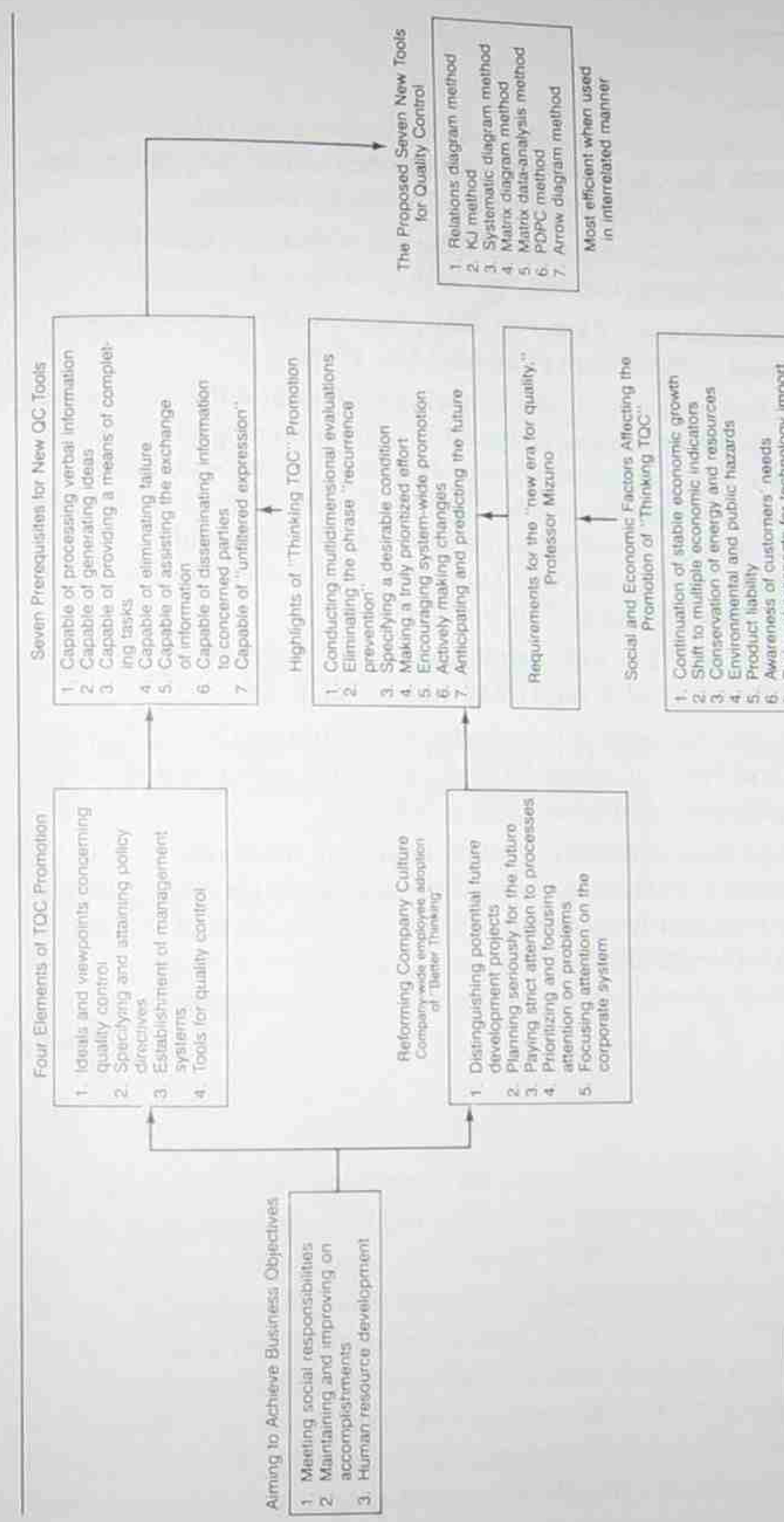


FIGURE 1.1
Relationship between the seven new tools for quality control and the promotion of company-wide quality control

Notes

1. Mizuno Shigeru and Akao Yoji, *Hinshitsu Kino Tenkai* (JUSE Press, Ltd., 1978). See Chapter 1 by Dr. Mizuno; he coined the phrase "new era for quality" at the 1972 Quality Control Convention.
2. Kigure Masao, "Quality Control in an Era of Reduced Quantity," *Hinshitsu Kanri (Quality Control)*, vol. 29 (July 1978), p. 44.
3. Mizuno Shigeru and Akao Yoji (Eds.), *Hinshitsu Kino Tenkai (Quality Deployment)* (Nikka Giren Publishing Co., 1978).
4. This section benefits from the following sources, and the writer gratefully acknowledges them: Asayoshi T., "Strengthening of Company Culture in TQC," *Standardization and Quality Control* (Sept. 1978); Asayoshi T., "Suggestions to Managers – Coping with a Low-Growth Stable Economy," *Monthly Quality Text*, vol. 78 (1975); and Ishikawa T., "Riding Through Turbulent Periods with Quality Control," *Monthly Quality Text*, vol. 71 (1974).
5. Sato Mitsuichi, "How to Structure and Solve a Problem," *Diamond Harvard Business*, vol. 3, no. 3 (May–June 1978), p. 49.
6. Oba Koichi, "Problems in Introducing and Planning QC – A Few Examples and Practical Advice," *Hinshitsu Kanri, (Quality Control)*, vol. 20, Supplement (November 1969), p. 5–8.
7. There is no fixed or definitive theory regarding the seven tools of quality control. This section relies on *Techniques for On-Site Improvements* by Imazumi. In other lists of the seven tools, control charts are replaced by stratification.

2

Seven New QC Tools

Overview of the seven methods

This section will introduce only the fundamentals of each method; detailed explanations will be presented in Part II.

The relations diagram method

This is a technique developed to clarify intertwined causal relationships in a complex problem or situation in order to find an appropriate solution.

The method

In order to analyze problems with a complex network of cause-and-effect relationships, a relations diagram is constructed by indicating the logical relationships that exist between the causal factors (Fig. 2.1). Such a diagram facilitates solutions to problems by allowing the whole problem to be viewed from a broad perspective.

To solve problems using the relations diagram method, a team composed of as many members as necessary should draft diagrams several times. By constructing diagrams in this way, the team generates new ideas that may lead to an effective solution.

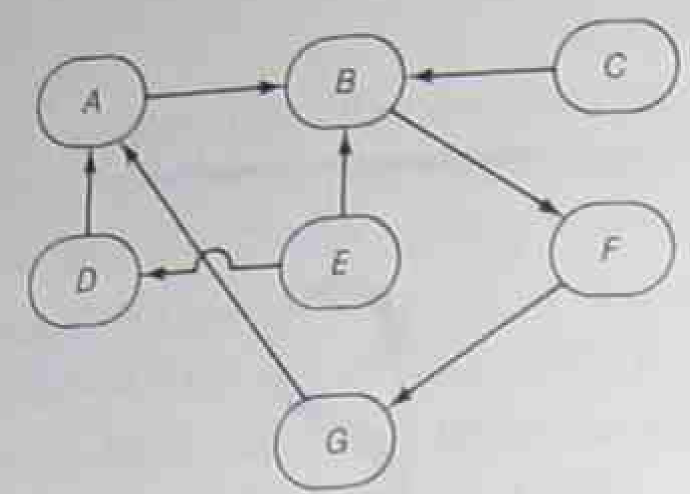


FIGURE 2.1
Abstract relations diagram

Uses

The relations diagram method can be used to

- Determine and develop quality assurance policies
- Establish promotional plans for TQC introduction
- Design steps to counter market complaints
- Improve quality in the manufacturing process (especially in planning to eliminate latent defects)
- Promote quality control in purchased or ordered items
- Provide measures against troubles related to payment and process control
- Effectively promote small group activities
- Reform administrative and business departments

Applications

Example 1. At the first symposium on quality control, a survey was conducted on the major items necessary for the introduction

and promotion of total quality control. The items obtained from this study are presented in the relational diagram pictured in Fig. 2.2. This diagram clarifies the important items that companies might consider in their promotion of total quality control.

Example 2. Company U investigated the causes of a chronic deficiency in its assembly line of a certain product by using a relational diagram (Fig. 2.3). As a result of drawing the relational diagram, the staff's preconceptions regarding the causes of the deficiency were corrected and countermeasures were taken that result in a drastic reduction in the defective rate.

The KJ method*

This technique clarifies important but unresolved problems by collecting verbal data from disordered and confused situations and analyzing that data by mutual affinity.

The method

The KJ method attempts to clarify the nature, shape, and extent of problems that affect the near and distant future in fields where there is little or no prior knowledge and/or experience. This technique consists of gathering ideas and opinions in the form of verbal data and drawing a complete diagram based on the common relationships and similarities found among the data.

The KJ method is an organizational technique based on "participatory group formation." Problems are solved through the creation of teams that gather the opinions, ideas, and experiences of diverse people and then coordinate and organize those data in terms of mutual affinity.

The KJ method was originally conceived, developed, and promoted by Kawakita Jiro. Mr. Kawakita attempts to solve all problems by cumulatively using the affinity diagram, which is further explained in Chap. 5. Contrasting the KJ method with statistical techniques (Table 2.1 and Fig. 2.4) highlights its effectiveness as one of

*The KJ method is a trademark registered by the founder of the Kawayoshida Research Center. The trademark is still held by the center. We gratefully acknowledge the center's permission to use its materials in this chapter.

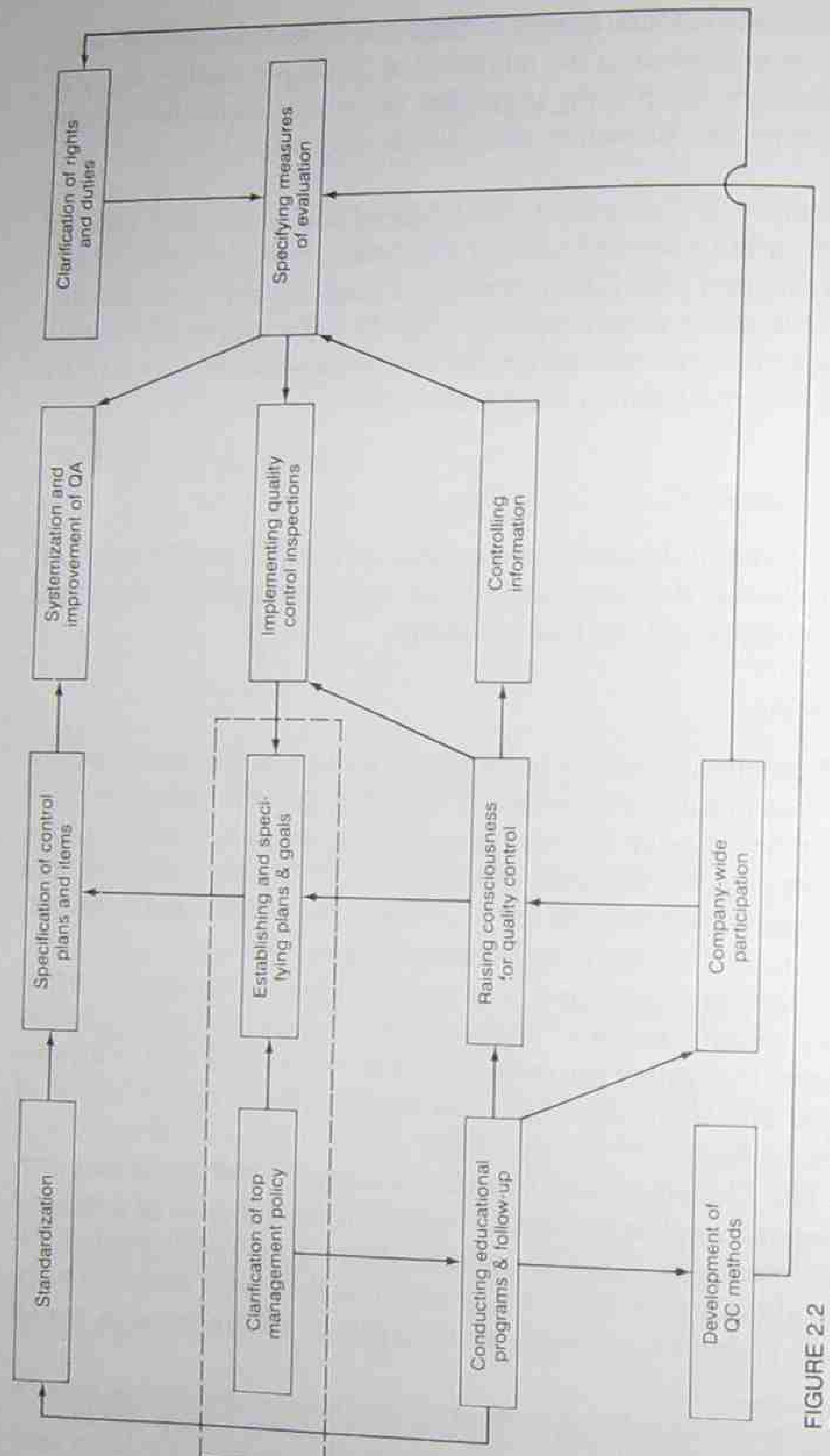


FIGURE 2.2
Relations diagram depicting the group of items important in the promotion of total quality control

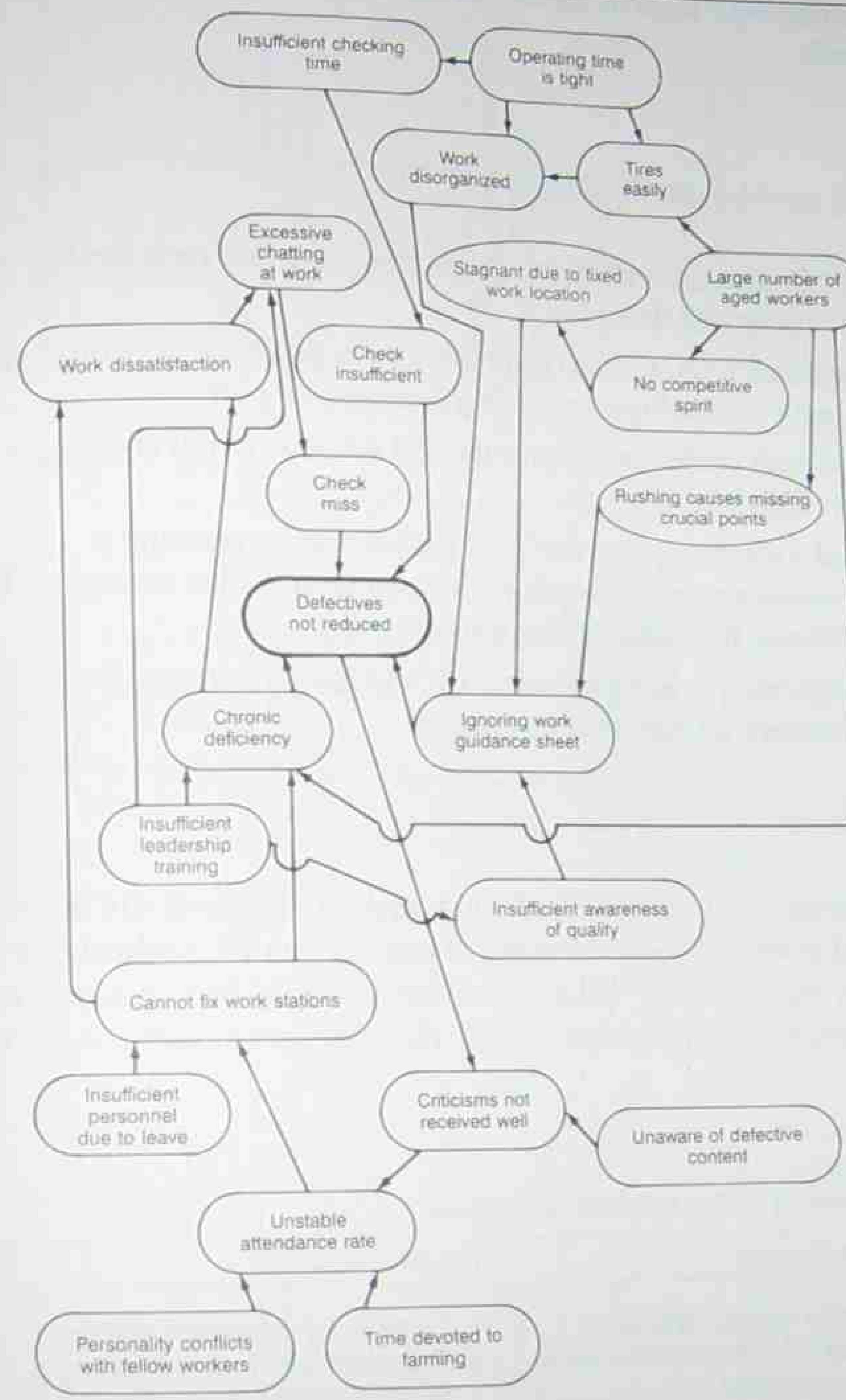


FIGURE 2.3
Relations diagram examining the causes of a deficiency in assembly-line production

the seven new tools in solving problems in conjunction with other methods.

Uses

The KJ method can be used to

- Establish a QC policy for a new company or a new factory and to implement that plan
- Establish a QC policy concerning new projects, new products, or new technology and to implement that plan
- Conduct quality assurance market surveys when entering a new untested market
- Find a starting point for TQC promotion by creating a consensus among people with varying opinions regarding the problems that arise within each department
- Invigorate project teams and QC circles and promote teamwork within various groups.

Applications

Example 1. Shown in Fig. 2.5 is a portion of the affinity diagram obtained from the second round of cumulative KJ method discussions on the topic of "Where and how should our research and development proceed from here?" This discussion was conducted

TABLE 2.1
KJ method versus statistical techniques

Statistical techniques	The KJ method
1. Oriented for testing hypotheses	1. Oriented for discovering problems
2. Quantifies and transforms an event into numerical data	2. Expresses data in language and symbols without quantifying
3. Capable of analytic understanding; ability to stratify	3. Overall understanding possible; harmonizes heterogeneous elements
4. Can grasp by reasoning	4. Can grasp through feeling
5. Western way of thinking introduced through translation	5. Thinking based on Japanese language (said not to be amenable with language written horizontally)

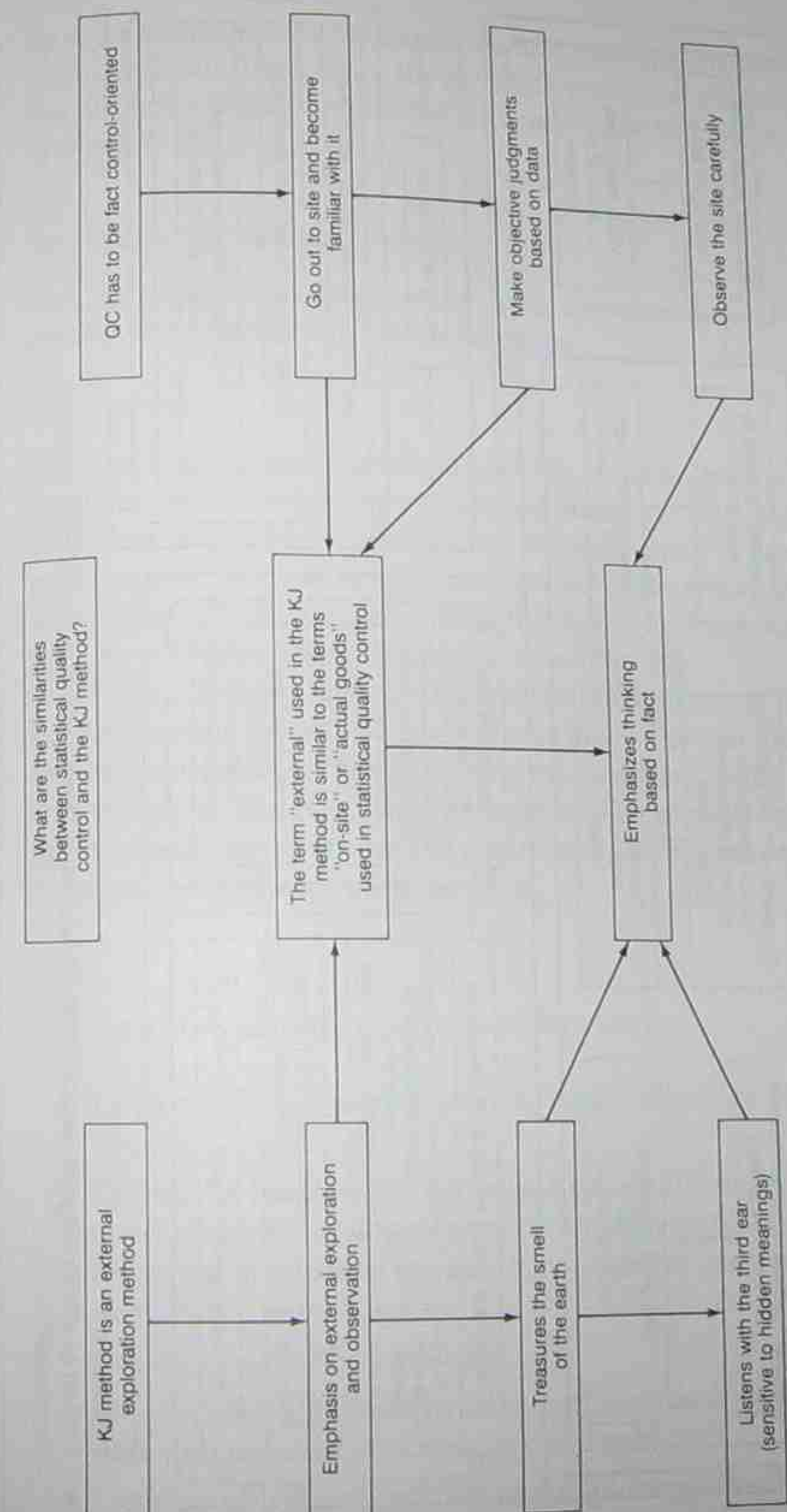


FIGURE 2.4
Similarities between statistical quality control and the KJ method

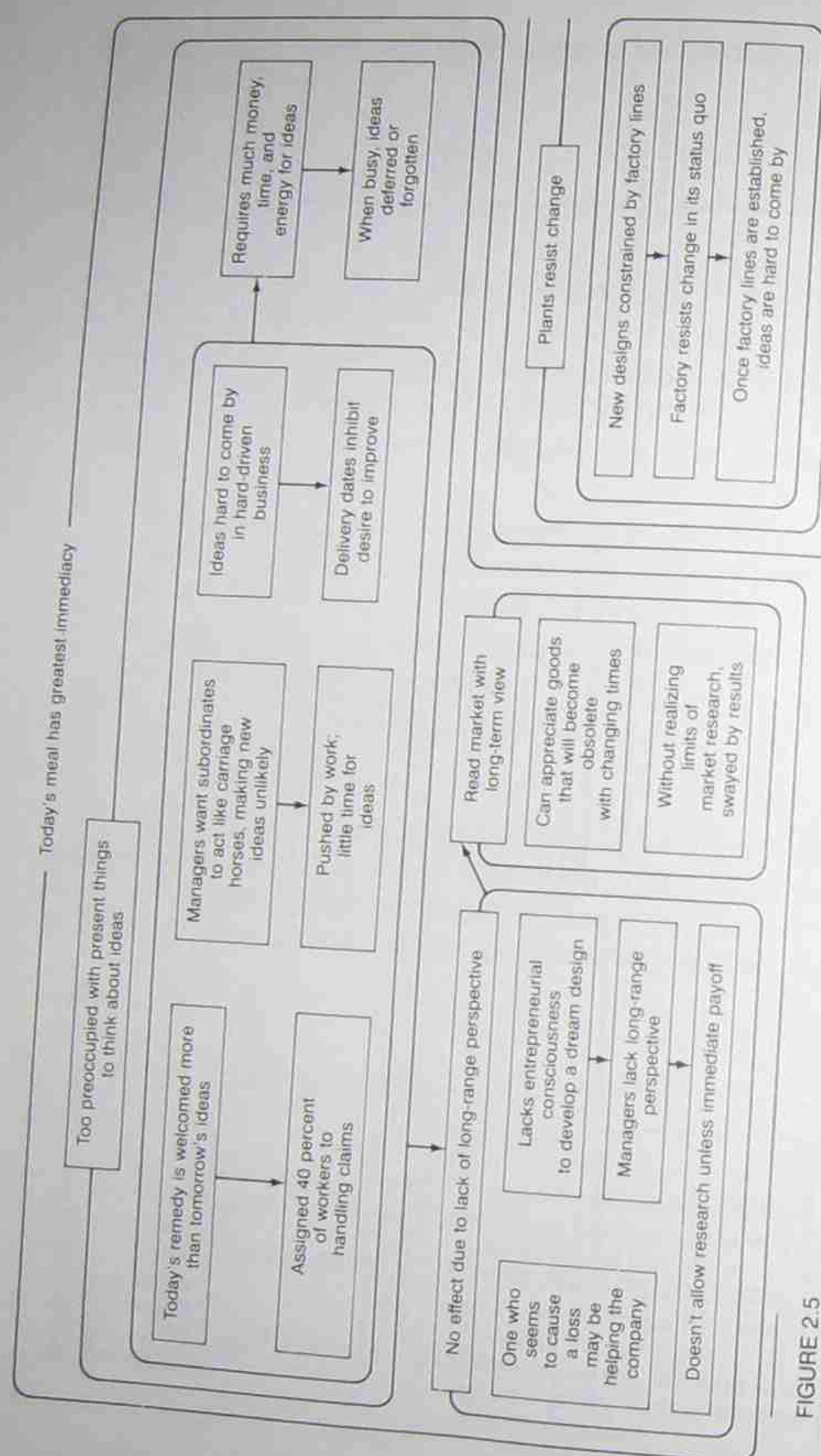


FIGURE 2.5
Portion of affinity diagram depicting the discussion of "where and how should R&D proceed from here?"

by a 10-member team of engineers and research managers from heterogeneous fields. Since the members were from different fields, the synthesis of their opinions led to a conclusion that will have wide applicability.

The systematic diagram method

This technique searches for the most appropriate and effective means of accomplishing given objectives.

The method

The systematic diagram method searches for techniques that will be most suitable for attaining established objectives by systematically clarifying important aspects of the problem. Such systematic diagrams enable workers to have an overview of the whole situation at one glance, effectively delineating the means and measures necessary for achieving the desired objectives (Fig. 2.6).

Systematic diagrams can be divided into two types: The *constituent-component-analysis* diagram breaks down the main subject into its basic elements and depicts their relationships to the objectives and means of attaining those objectives. The *plan-development* diagram systematically shows the means and procedures necessary to successfully implement a given plan.

Uses

The systematic diagram method can be used to

- Deploy a design-quality plan in the development of a new product
- Depict the relationship between a QC production process chart and the development of certified levels of quality designed to improve the accuracy of quality assurance activities
- Create a cause-effect diagram
- Develop ideas in order to solve problems dealing with quality, cost, and delivery that arise in new businesses
- Develop objectives, policies, and implementation steps

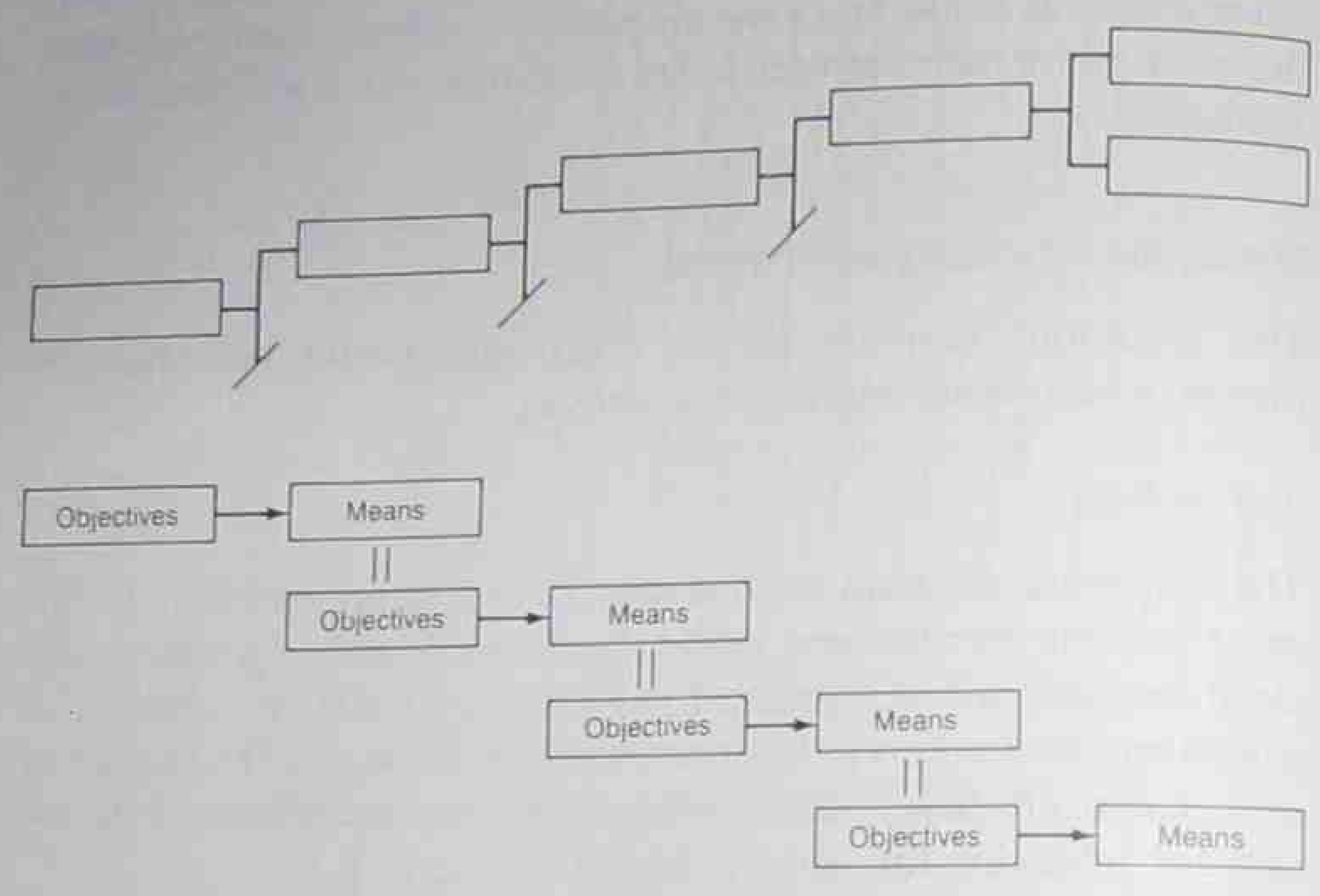


FIGURE 2.6
Conceptual systematic diagram

- Pursue the specification of increased efficiency in parts and control functions

Applications

Example 1. A systematic diagram of the plan-development type is shown in Fig. 2.7. This diagram illustrates the development of design-quality and implementation steps necessary for production of a UHF tuner for a television.

The matrix diagram method

This technique clarifies problematic spots through multidimensional thinking.

The method

The matrix diagram method identifies corresponding elements involved in a problem situation or event. These elements are arranged in rows and columns on a chart (Fig. 2.8) that shows the presence or absence of relationships among collected pairs of elements. Hopefully, this method will assist in specifying (with a two-way layout) the nature and/or location of problems, enabling idea conception on the basis of two-dimensional relationships. Effective problem solving is facilitated at the intersection points, also referred to as "idea conception points."

Matrix diagrams are classified on the basis of their pattern into five different groups: (1) the L-type matrix, (2) the T-type matrix, (3) the Y-type matrix, (4) the X-type matrix, and (5) the C-type matrix.

Uses

Matrix diagrams can be used to

- Establish idea conception points for the development and improvement of system products
- Achieve quality deployment in product materials
- Establish and strengthen the quality assurance system by linking certified levels of quality with the various control functions
- Reinforce and improve the efficiency of the quality evaluation system
- Pursue the causes of nonconformities in the manufacturing process
- Establish strategies about the mix of products to send to market by evaluating the relationships between the products and market situations

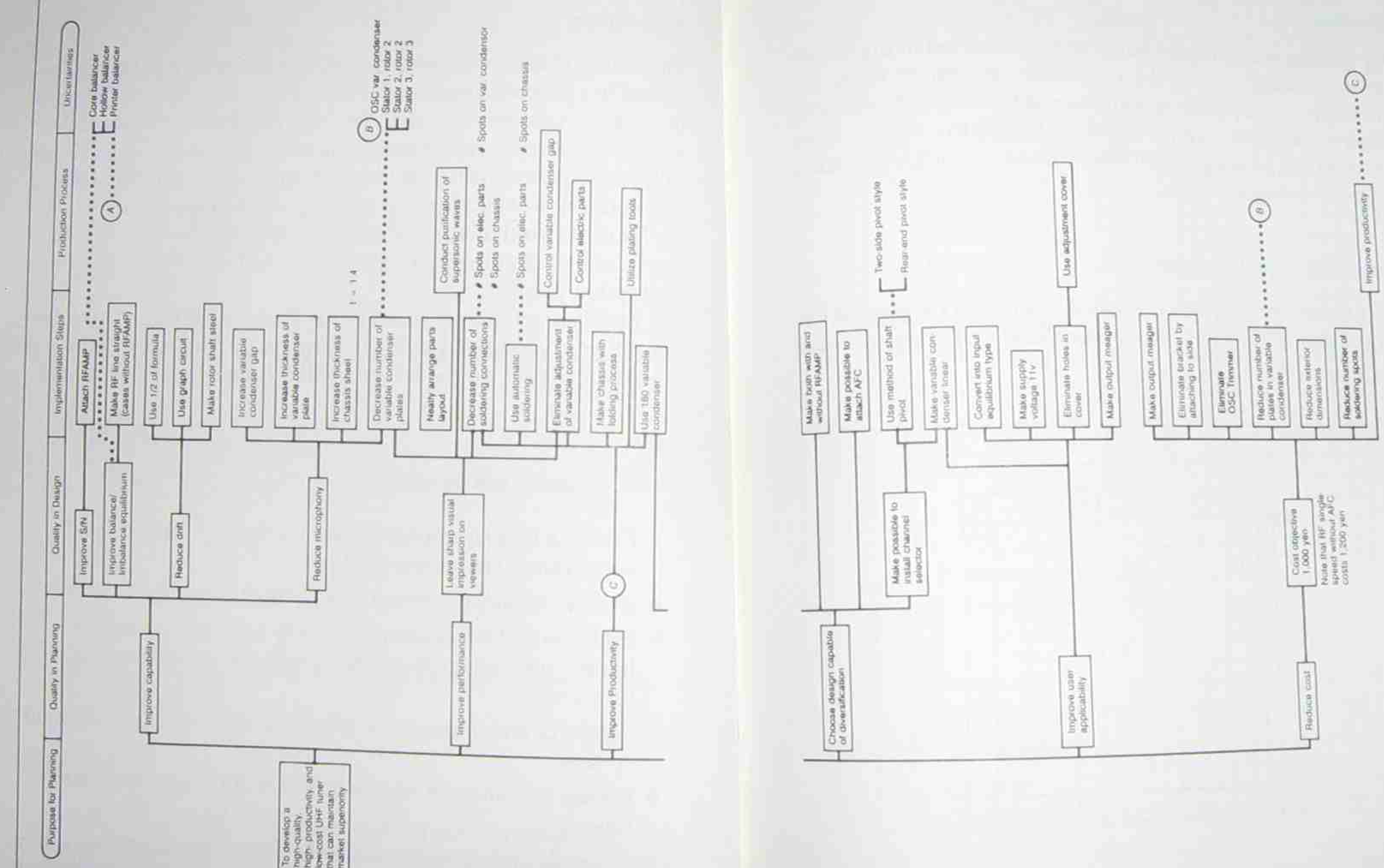


FIGURE 2.7
Systematic diagram for design quality for a television UHF tuner

- Clarify the technical relationships among several projects
- Explore the application potential of currently available technology and raw materials

Applications

As Fig. 2.9 indicates, a T-type matrix was constructed in the investigation of smearing during the production of printed cloth. The matrix helped to clarify the relationships between nonconformities and their causes. Based on the results of this matrix, a list of countermeasures was produced (see Table 7.1), and their implementation considerably reduced smears.

The matrix data-analysis method

This technique arranges data presented in a matrix diagram so that the large array of numbers can be visualized and comprehended easily.

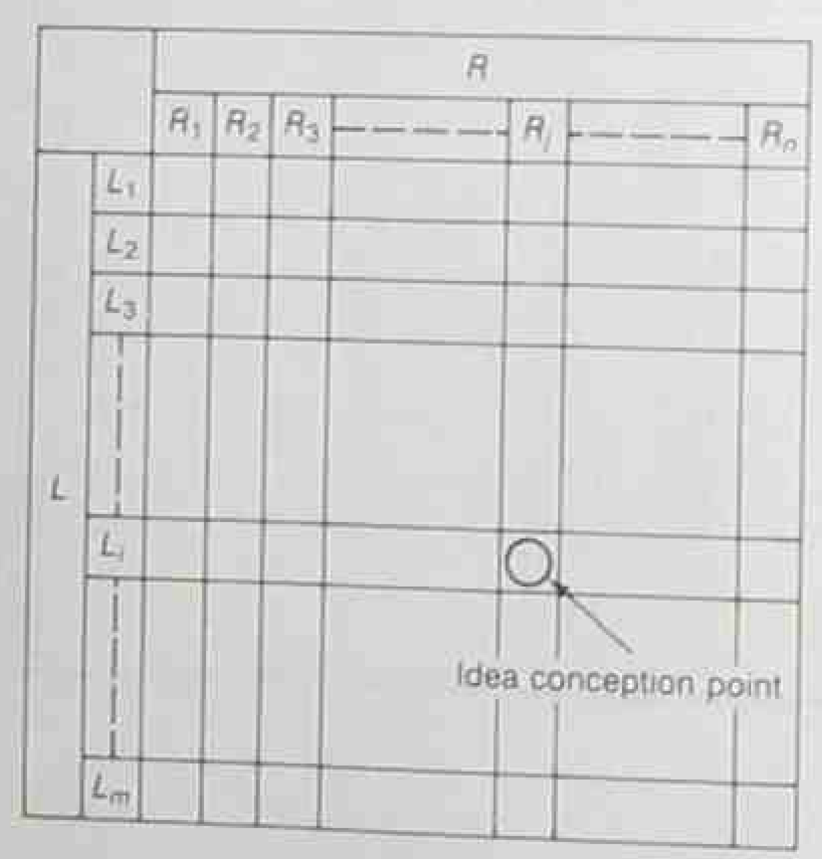


FIGURE 2.8
Conceptual matrix diagram method

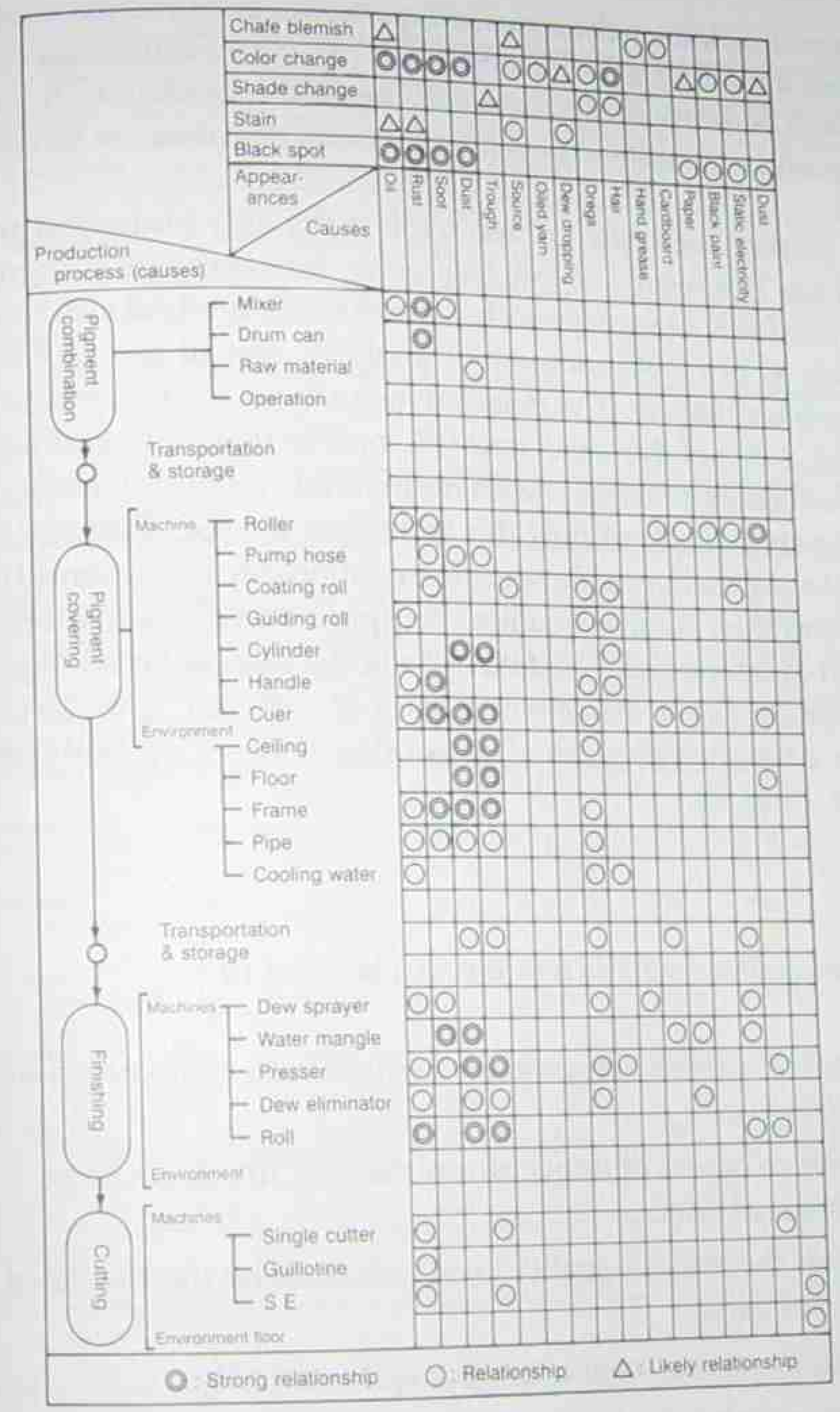


FIGURE 2.9
T-type matrix searching for causes of smears in printing of cloth

The method

The matrix data-analysis method quantifies and arranges matrix diagram data so that the information is easy to visualize and comprehend. The relationships between the elements shown in a matrix diagram are quantified by obtaining numerical data for intersection cells.

Of the seven new QC tools, this is the only numerical analysis method. The results of this technique, however, are presented in diagram form. One major technique that this method also utilizes is known as principal-component analysis, one of the multivariate analysis techniques. The matrix data-analysis method has been included as one of the seven new QC tools so that managers and staff can become more familiar with multivariate analysis techniques.

As an example of how this technique is used, Table 2.2 is presented as a data matrix showing the relationships between 40 uses of cloth and their desired qualities. Suppose that a new material, material *A*, is developed. The data in Table 2.2 does not readily provide information about the potential uses of material *A*; however, the matrix data-analysis method does offer answers by analyzing the data in Table 2.2.

Uses

The matrix data-analysis method can be used to

- Analyze production processes where factors are complexly intertwined
- Analyze causes of nonconformities that involve a large volume of data
- Grasp the desired quality level indicated by the results of a market survey
- Classify sensory characteristics systematically
- Accomplish complex quality evaluations
- Analyze curvilinear data

TABLE 2.2
Product uses and their desired qualities

Uses \ *Desirable Qualities*	1 Resists Fading	2 Washable	3 Resists perspiration	…	23 Flame retardant	24 Chemical resistant	25 Non-irritating to skin
1. Mens' summer suits	$x_{1\text{-}1}$	$x_{1\text{-}2}$	$x_{1\text{-}3}$	…	$x_{1\text{-}23}$	$x_{1\text{-}24}$	$x_{1\text{-}25}$
2. Mens' all-season suits	$x_{2\text{-}1}$	$x_{2\text{-}2}$	$x_{2\text{-}3}$	…	$x_{2\text{-}23}$	$x_{2\text{-}24}$	$x_{2\text{-}25}$
3. Ladies' summer dresses	$x_{3\text{-}1}$	$x_{3\text{-}2}$	$x_{3\text{-}3}$	…	$x_{3\text{-}23}$	$x_{3\text{-}24}$	$x_{3\text{-}25}$
4. Ladies' all-season dresses	$x_{4\text{-}1}$	$x_{4\text{-}2}$	$x_{4\text{-}3}$	…	$x_{4\text{-}23}$	$x_{4\text{-}24}$	$x_{4\text{-}25}$
5. Skirts	$x_{5\text{-}1}$	$x_{5\text{-}2}$	$x_{5\text{-}3}$	…	$x_{5\text{-}23}$	$x_{5\text{-}24}$	$x_{5\text{-}25}$
6. Trousers	$x_{6\text{-}1}$	$x_{6\text{-}2}$	$x_{6\text{-}3}$	…	$x_{6\text{-}23}$	$x_{6\text{-}24}$	$x_{6\text{-}25}$
7. Overcoats	$x_{7\text{-}1}$	$x_{7\text{-}2}$	$x_{7\text{-}3}$	…	$x_{7\text{-}23}$	$x_{7\text{-}24}$	$x_{7\text{-}25}$
8. Raincoats	$x_{8\text{-}1}$	$x_{8\text{-}2}$	$x_{8\text{-}3}$	…	$x_{8\text{-}23}$	$x_{8\text{-}24}$	$x_{8\text{-}25}$
9. Office wear	$x_{9\text{-}1}$	$x_{9\text{-}2}$	$x_{9\text{-}3}$	…	$x_{9\text{-}23}$	$x_{9\text{-}24}$	$x_{9\text{-}25}$
10. Work clothes	$x_{10\text{-}1}$	$x_{10\text{-}2}$	$x_{10\text{-}3}$	…	$x_{10\text{-}23}$	$x_{10\text{-}24}$	$x_{10\text{-}25}$
11. Sports wear	$x_{11\text{-}1}$	$x_{11\text{-}2}$	$x_{11\text{-}3}$	…	$x_{11\text{-}23}$	$x_{11\text{-}24}$	$x_{11\text{-}25}$
12. Student wear	$x_{12\text{-}1}$	$x_{12\text{-}2}$	$x_{12\text{-}3}$	…	$x_{12\text{-}23}$	$x_{12\text{-}24}$	$x_{12\text{-}25}$
13. Home wear	$x_{13\text{-}1}$	$x_{13\text{-}2}$	$x_{13\text{-}3}$	…	$x_{13\text{-}23}$	$x_{13\text{-}24}$	$x_{13\text{-}25}$
14. Baby wear	$x_{14\text{-}1}$	$x_{14\text{-}2}$	$x_{14\text{-}3}$	…	$x_{14\text{-}23}$	$x_{14\text{-}24}$	$x_{14\text{-}25}$
15. Dress shirts	$x_{15\text{-}1}$	$x_{15\text{-}2}$	$x_{15\text{-}3}$	…	$x_{15\text{-}23}$	$x_{15\text{-}24}$	$x_{15\text{-}25}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
40. Foot warmer blankets	$x_{40\text{-}1}$	$x_{40\text{-}2}$	$x_{40\text{-}3}$	…	$x_{40\text{-}23}$	$x_{40\text{-}24}$	$x_{40\text{-}25}$
Material *A*	x_1	x_2	x_3	…	x_{23}	x_{24}	x_{25}

Applications

A principal-component analysis is performed on the matrix data provided in Table 2.2. The resulting first and second principal components are shown graphically in Fig. 2.10. This figure reveals that the new material *A* would probably be better suited for skirts and trousers than for sports wear, work uniforms, or gloves.

The PDPC method

This technique helps determine which processes to use to obtain desired results by evaluating the progress of events and the variety of conceivable outcomes.

The method

Implementation plans do not always progress as anticipated. When problems, technical or otherwise, arise, solutions are frequently not apparent.

The PDPC (process decision program chart) method, in response to these kinds of problems, anticipates possible outcomes and prepares countermeasures that will lead to the best possible solutions. By anticipating potential outcomes of events, this technique allows process adjustments in light of actual progress.

If an unanticipated event occurs, then it becomes necessary to rewrite the process decision program chart (PDPC) at once so that adjustive countermeasures can be taken.

The PDPC method[1] is borrowed from the operations research field for use in quality control.

Uses

The PDPC method can be used to

- Establish an implementation plan for management by objectives
- Establish an implementation plan for technology-development themes
- Establish a policy of forecasting and responding in advance to major events predicted in the system

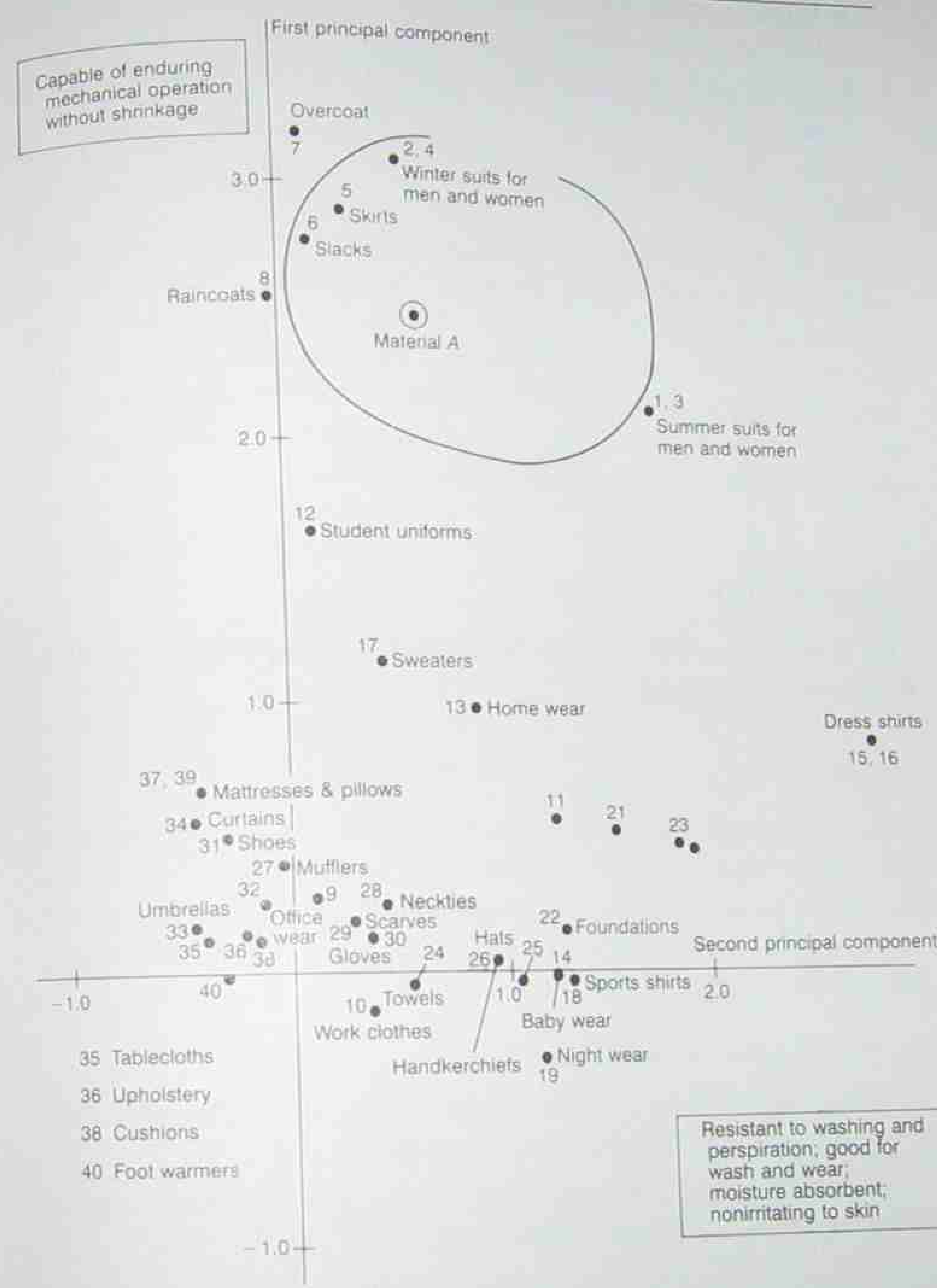


FIGURE 2.10
Searching for uses of new material *A*

- Implement countermeasures to minimize nonconformities in the manufacturing process
- Set up and select adjustment measures for the negotiating process

Applications

Suppose that a company's objective is to deliver a fragile item to an addressee in a developing country. The company must anticipate various contingencies from shipment time to delivery, and address all the problems that might arise as a result of transportation and landing. The company must then develop countermeasures to avoid possible mishaps. The PDPC method approach to this hypothetical example and the results are shown in Fig. 2.11.

The arrow diagram method

This technique establishes the most suitable daily plan and monitors its progress efficiently.

The method

The arrow diagram method, utilized by PERT and CPM, is a network diagram for daily plans. It illustrates the network of lines that connect all the elements related to plan execution, as shown in Fig. 2.12.

Use of the arrow diagram method in advancing and monitoring daily plans has the following advantages:

1. It establishes a finely tuned plan.
2. It establishes the most suitable daily plan, since changes can be made easily during the early planning stages.
3. It allows one to cope easily with changes that occur in a given situation or during plan execution.
4. It expedites necessary action by quickly providing information on the impact delays in certain subparts will have on the operation as a whole.
5. It is increasingly useful in proportion to the size of the plans.
6. It controls the process efficiently because the progress highlights are easily discernible.

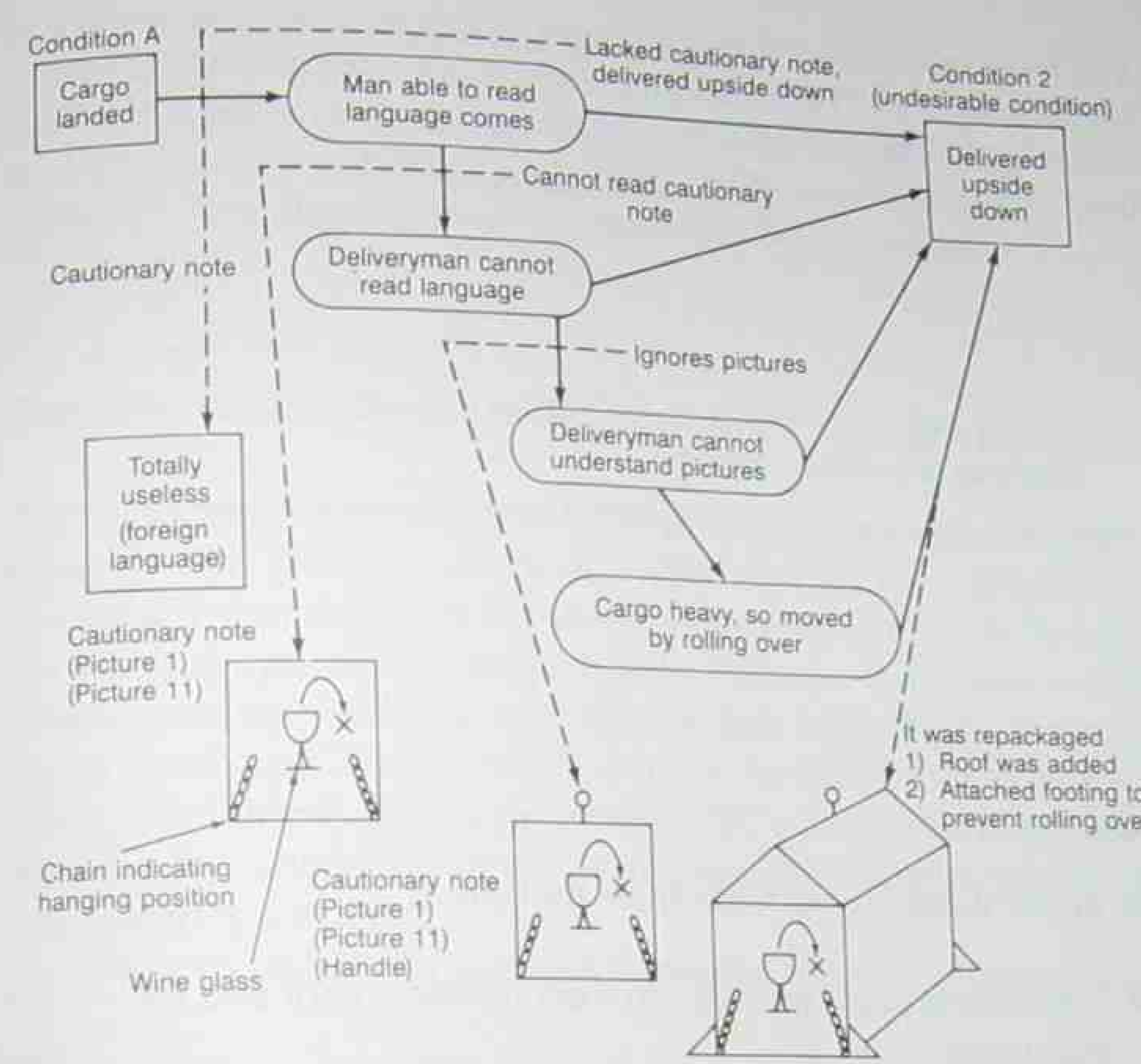


FIGURE 2.11
Delivery of fragile item (pattern II)

The control of daily plans is extremely important in the promotion of QC activities. An efficient method of constructing and utilizing arrow diagrams that employs cards will be introduced mainly to assist staff members.

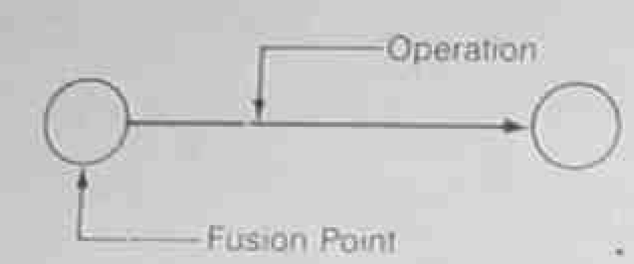


FIGURE 2.12
Elements of an arrow diagram

Uses

The arrow diagram method can be used to

- Implement plans for new product development and its follow-up
- Develop product-improvement plans and follow-up activities
- Establish daily plans for experimental trials and follow-up activities
- Establish daily plans for increases in production and their follow-up
- Synchronize the preceding plans with QC activities
- Develop plans for a facility move and for monitoring follow-up
- Implement a periodic facility maintenance plan and its follow-up
- Analyze a manufacturing process and draw up plans for improved efficiency
- Plan and follow up QC inspections and diagnostic tests
- Plan and follow up QC conferences and QC circle conferences

Applications

In a low-cost project to produce special resistor electrodes used in a starter for an electric motor, value engineering (VE) experimental trials and quality confirmation tests were implemented based on an arrow diagram (Fig. 2.13) of daily plans constructed using the "card method." This diagram made the trials and tests possible.

The role of the seven new QC tools in quality control

The seven new tools proposed fulfill the planning steps often mentioned in the "plan, do, check, act" (PDCA) TQC cycle. Figure 2.14 graphically demonstrates the placement of various QC techniques applied in the *plan* and *do* stages to solve an important problem. If an adequate amount of past quantitative data is available, the traditional seven tools would probably suffice; however, as mentioned in Chap. 1, this is not always the case in TQC problem solving. Therefore, Fig. 2.14 should be viewed as relating to a situation where data are relatively scarce.

The plan stage can be divided into the following three phases:

Plan 1: This phase reviews a confusing event and arranges the information so to clarify the underlying nature of the problem.

Plan 2: This phase searches for various means that might be employed to solve the problem and identifies the relationships between the objectives and the means.

Plan 3: This phase establishes an implementation strategy in a time-order sequence in order to increase the chance for success.

During the plan 1 phase, the relations diagram and the KJ methods are used. During plan 2, the systematic diagram and matrix diagram methods, as well as the cause-and-effect diagram and matrix data-analysis methods are used. The plan 3 phase relies on the arrow diagram and the PDPC methods. The arrow diagram method is most often used when the sequence of steps is relatively fixed and

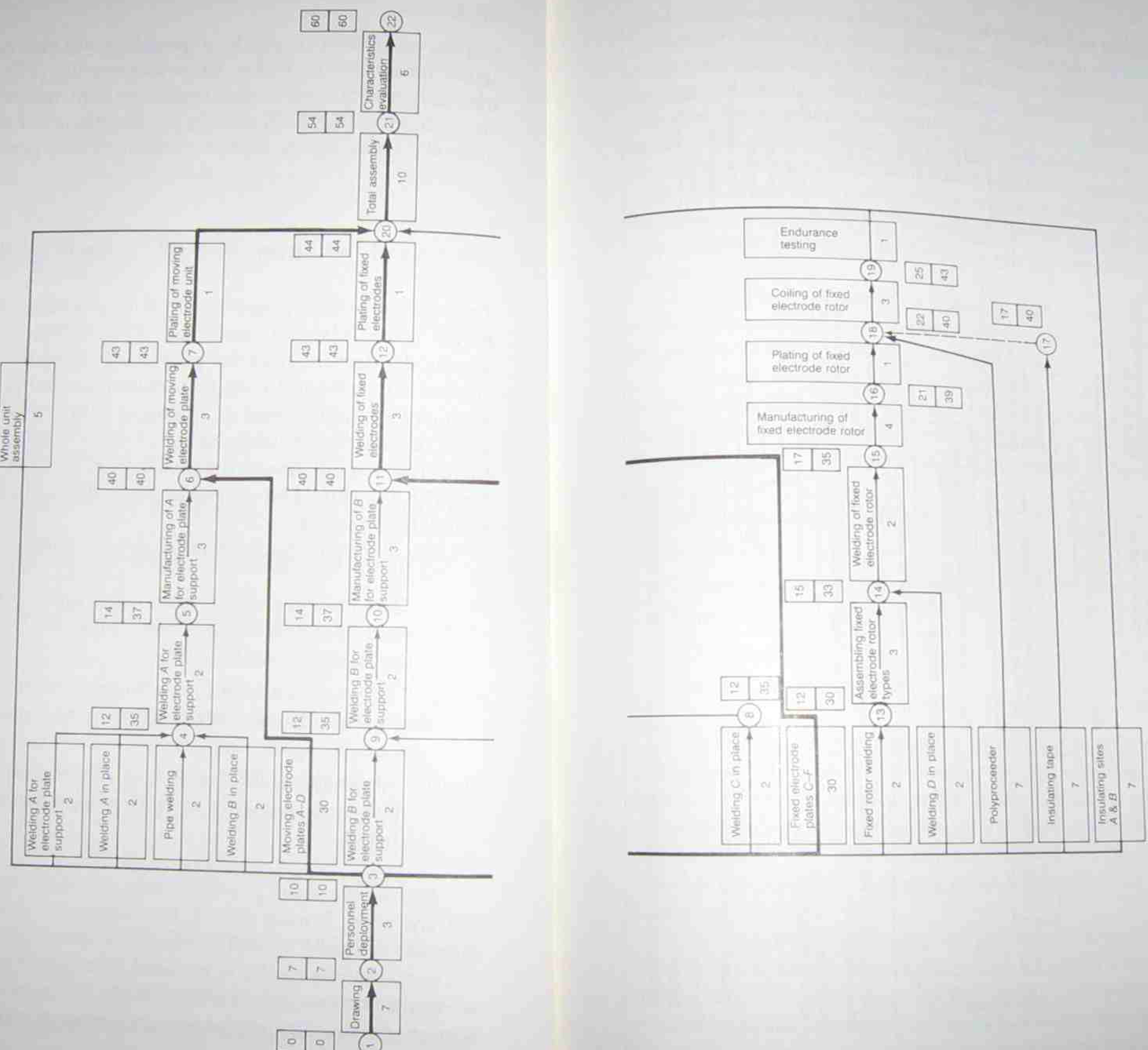


FIGURE 2.13
Experimental trials and quality confirmation plan for VE improvement of special resistor electrodes

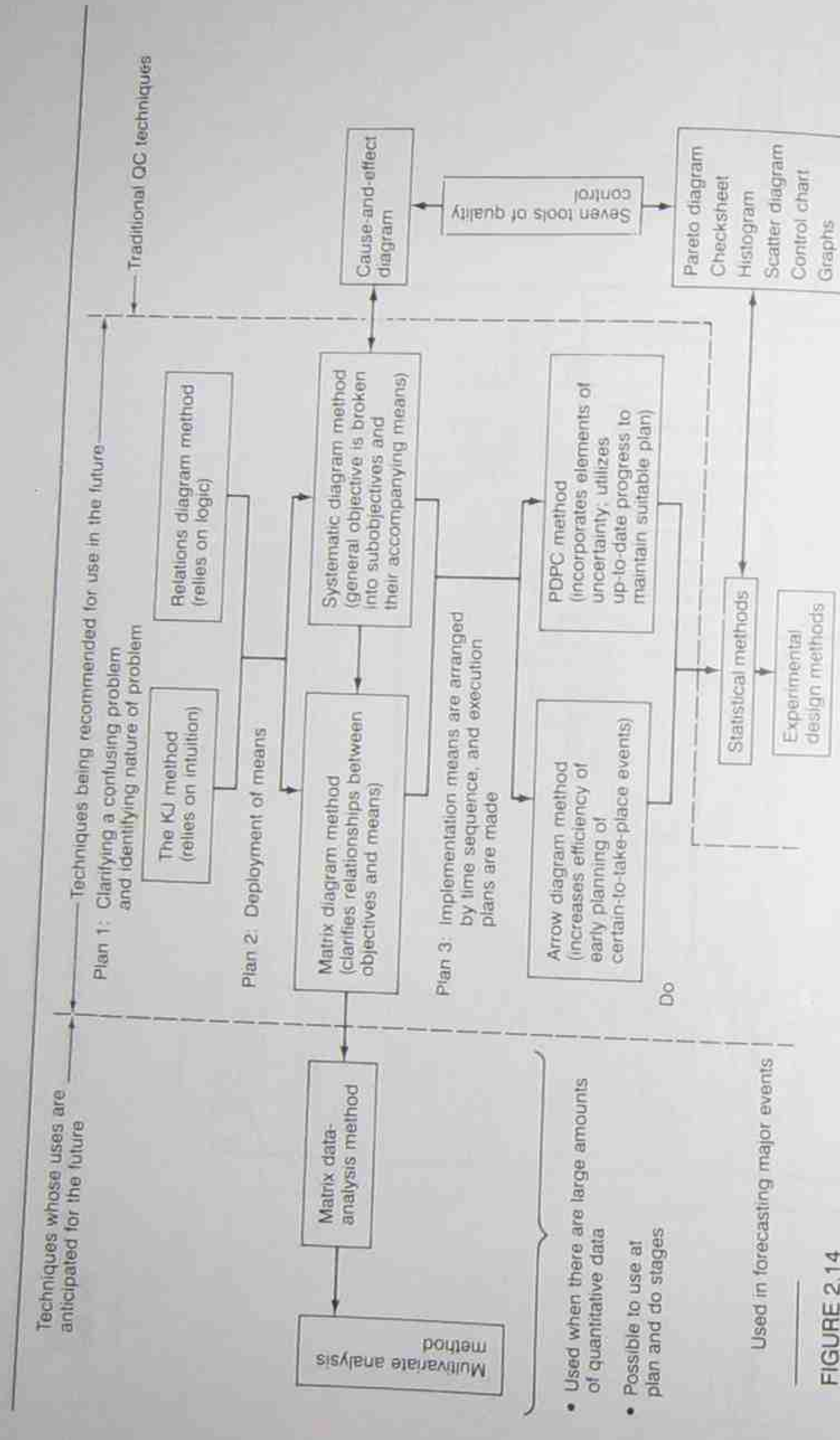


FIGURE 2.14
Relationship of seven new QCtools to traditional QC techniques

predetermined, as in the construction of a building or in shortening the delivery time of a product. However, in QC-related activities, the original readings or plans frequently must be modified because of an unexpected development or occurrence. Whenever there is an unexpected development, information collected up to that point must be analyzed, underlying causes must be understood, and appropriate changes in plans must be made in order to achieve the overall objectives. The PDPC method is an extremely useful technique in such situations, and the PDPC method also can be used to forecast major accidents in such related areas as environmental hazards and product liability.

During the *do* stage, other statistical and experimental design methods are selected from the traditional seven tools depending on the particular circumstances of implementation.

Although the *check* and *act* stages of the PDCA cycle are not shown in Fig. 2.14, they proceed from the outcome of the *do* stage. But in Fig. 2.14 it is clear that the following six of the seven new tools should be used in every stage of TQC promotion in the future:

- Relations diagram method
- KJ method
- Systematic diagram method
- Matrix diagram method
- PDPC method
- Arrow diagram method

The matrix diagram method provides a readily comprehensible graphic representation of complex data when each cell in the matrix is given a numerical value. It corresponds to the principal-component analysis method, one of the multivariate analysis methods. Hopefully, the matrix diagram method will be accepted as a tool that can simplify and bring order to numerical or quantitative confusion.

During product planning and process-improvement activities, situations are frequently encountered in which numerous intertwined causes must be untangled. Keeping pace with the development of computers, many businesses have aggressively introduced this method with good results. Even though this technique should be brought to the attention of managers and staff, it should not necessarily be overused.

To prevent misunderstandings about the origin of the seven new QC tools, five points should be highlighted:

1. The seven methods described here do not necessarily exhaust the list. Selection of these seven tools came after much hesitation and thought. Actually, it is hoped that these techniques enrich QC techniques even further. Even people who use the traditional seven tools use them differently over time. Therefore, some modifications and revisions of the seven new tools in response to future developments are necessary. There is also nothing sacred about the number seven. What is essential is that one's toolbox contain all the useful and necessary tools.
2. Most of the techniques described above are already known and were not recently created as QC methods. This also applies to the traditional seven tools, such as the histogram, pareto chart, scatter diagram, and other statistical methods. Some of these even appear in elementary school textbooks. However, because of their promotion, along with the cause-and-effect diagram, as part of the seven tools for quality control, they have come to be used extensively in QC activities. By christening the preceding methods as the seven new QC tools, it is hoped that they too will receive the attention necessary to allow the development of diverse applications in fields concerned with quality control.
3. There are two important reasons for using the seven new tools as a group. One is that, as previously mentioned, the seven tools act like an organically integrated set; it would be hard to expect excellent results if they were used independently. The seven new tools should be used in combination, as shown in the second section of Chap. 3, for best results in solving problems. The second reason is that the seven new tools are best used when promoted in all facets of operation under the direction of the company or department head. Effects are limited if the tools are used sporadically in isolated divisions or units.
4. The pioneers who promoted quality control deserve our thanks for having introduced various techniques and for having successfully completed the original seven QC tools, the statistical and experimental design methods for quality control. Unfortunately, however, over the course of the past decade or so, few innovative proposals similar to the ones promoted by the first-generation pioneers have been set forth. As the succeeding generation, though, it is important for us to gather all the seedlings and fruit of the preceding generation, develop them as much as possible in a systematic fashion, and pass them on to the next generation.
5. All the techniques described here have been used previously in various fields and have demonstrated some level of effectiveness. Although not unique to quality control, the requirements for using these tools include a keen awareness of the problem at hand, an incessant desire for improvement, and an enhanced spirit and thought process. Without these kinds of mental preparation, the tools cannot be used effectively. Anyone who expects a tool to do all the work cannot expect good results and cannot really be considered a "user."

The seven new QC tools and the basics of graphics-language theory[2]

As discussed earlier, the seven new tools make considerable use of graphics. Table 2.3 presents a classification system developed by T. Kahn[3] that shows the differences between language and graphics as they apply to manner of recognition and relative ease of understanding. When viewing graphics, we first comprehend the overall structure (pattern, balance, and trends of dots and lines) and then reach out for that which is interesting.

Regardless of nationality and race, any person can understand pictures. Language, however, presupposes an understanding of pre-existing rules, without which the language is totally incomprehensible, for example, when a non-Japanese speaking person tries to make sense of Japanese characters. The understanding of graphics and language is analogous to the human developmental process. Infants start out as "contact beings," whose exchanges of information with others, including their mother, occur through physical

TABLE 2.3
Differences between graphics and language

	Graphics	Language
Manner of recognition:	First, the whole is grasped. Next, elements are analyzed.	First, elements are recognized. Next, whole is constructed.
Ease of understanding:	Understood by almost anybody immediately (pictures).	If rules are not understood, then it is incomprehensible (foreign languages).

contact. These contact beings later grow into "picture beings," who are able to understand information based on pictures, drawings, and graphics of various sorts. Finally, these picture beings turn into "character beings," who are capable of information transmission through written characters. The recent popularity of commercial drawings and comic books may be viewed as an extension of the picture-being stage. In other words, the human capability to understand graphic forms easily seems to be a developmental characteristic. It should be evident that tools anchored in graphics will emerge as powerful techniques in the promotion of company-wide total quality control because of their ready comprehension by all concerned.

Computer graphics theorists distinguish pictures that contain characters, such as the seven new tools, from graphics that contain only drawings by calling them "graphics language." The graphics-language group is further divided into the following four types:

1. Relational system
2. Network system
3. Column-row system
4. Coordinate system

A further explanation of these system types is contained in *Computer Graphics Theory*. However, their names alone suggest their meanings. Of the four types listed, the coordinate system has been in use the longest. Table 2.4 shows the seven new and seven "old" tools classified into the different graphics-language types. Notice that the "old" tools are primarily coordinate-based, while the new tools rely,

TABLE 2.4
The seven new tools and the traditional seven tools classified by graphics-language systems

Graphics-language system	Seven new tools	Traditional seven tools
Relational system	KJ method	–
Network system	Relations diagram method Systematic diagram method Arrow diagram method PDPC method	Cause-and-effect diagram
Column-row system	Matrix diagram method	Checksheet
Coordinate system	Matrix data-analysis method	Pareto chart Histogram Scatter diagram Control chart Graphs

primarily on the relational or network systems.

Experience has shown that the ease or difficulty involved in constructing graphics for the seven new tools varies from one person to another. The graphics tools used can be divided into two groups:

Soft graphic tools: The KJ method, the relations diagram method, and the PDPC method offer a relatively greater degree of freedom in graphics construction.

Hard graphic tools: The systematic diagram method, the matrix diagram method, and the arrow diagram method have considerably less freedom in graphics construction.

Even though the former may appear to be the easier to people just starting to use the new tools, they are actually more difficult because they allow a greater degree of freedom. Nevertheless, some beginners have produced outstanding graphics.

A necessary step in company-wide QC activities is for all employees to become thoroughly acquainted with pictorial or graphic thinking. This is accomplished by practicing the construction of various graphics and through the cooperation of everyone involved. All employees should be exposed to a wide variety of graphics so they may identify their own weak areas and learn to construct the diagrams properly. Only this kind of total exposure will make the combined use of all the seven new tools feasible.

Notes

1. Kondo Jiro, *Operations Research* (Tokyo: JUSE Press, Ltd., 1973), pp. 128–136.
2. This section relies heavily on information in Chaps. 2 and 3 of *Computer Graphics Theory*, by Yoshikawa Hiroyuki, published by JUSE Press in 1977. Grateful acknowledgment is due.
3. Yoshikawa Hiroyuki, *Computer Graphics Theory* (Tokyo: JUSE Press, Ltd., 1977).

3

Applying the Seven New QC Tools

Fields of application for the seven new QC tools

Each one of the seven new QC tools has been in use for some time now in the field of quality control. A review of the proceedings of quality control conferences held between 1966 and 1978 shows that use of the techniques included in the seven new tools has increased dramatically since the seven new tools were announced in 1977. (See Fig. 3.1.) This increase should continue in the future.

Application of the seven new tools began with the analysis and consolidation of verbal data and is continuing to expand. Diverse applications of the new tools in different fields and for varied purposes have been widely publicized in Japan, which has contributed to a healthy momentum for further promotion. We can expect to see applications in unexpected, heterogeneous fields.

Table 3.1 shows both the publicized and unpublicized applications, indicating the fields where we may expect future use. Although there may have been some unrefined usage, the table will be helpful as a guide to applications of the seven new QC tools in total quality control.

Examples introduced in individual chapters are also listed in Table 3.2 (by technique), so this table serves as a guide to various examples, as well as a way of finding fields of future application.

TABLE 3.1
Listing of fields where applications are expected

Step	Primary implementation	Secondary implementation	Relations Diagram Method	KJ Method	Systematic Diagram Method	Matrix Diagram Method	Matrix Data-Analysis Method	PDPC Method	Arrow Diagram Method
Overall	Policy control	Priority imposed on many implementation items	●	○	○	○			
		Policy transmitted from top down	●		◎				
		Clarify responsibilities for implementation items	○		◎	○			
		Develop a plan with high probability of success for objectives			◎				
	Education and training	Training items are prioritized						◎	◎
		Relate training items with the expected standards	○	○		○			
	Labor and personnel relations	Clarify lines of responsibility by job title	○		○	○			
		Smooth out interpersonal conflicts			○	◎			
	Financial matters and accounting	Comparative diagnosis of one's business culture	○					◎	
	Promotion of total quality controls	Smooth out interdepartmental relationships				○	◎		
	QC circle activities	Invigorate QC circle activities	○	○	○	○			○
Develop-ment of plans	Market analysis	Monitor and classify needs	○	●	○				
		Perform demand forecasts		◎	○		●		
		Analyze competitors			○		●		
		Explore uses for new products				◎	●		
		Survey circulation routes	○	○		○	●		
		Conduct search for themes			○	○			
	Quality design	Understand quality characteristics	○	○		○			
	Quality assurance	Transform needs into alternative characteristics			●	○	●		
		Relate soft and hard functions		●	●	●			
		Relate alternative characteristics to process-control items	○	○		●			
		Develop planning policy into design quality			●	●			
		Clarify factors that enhance reliability		●	●	○			
		Attempt to complete PLP activities	○	○	●		◎	○	

Step	Primary implementation	Secondary implementation	Relations Diagram Method	KJ Method	Systematic Diagram Method	Matrix Diagram Method	Matrix Data-Analysis Method	PDPC Method	Arrow Diagram Method
Develop-ment of plans	Evaluation and development of test trials	Develop production techniques			○		◎	●	
		Clarify relationship between functions and cost			◎	○			
		Examine status of research and development		◎					
Production	Stocking and purchasing	Conduct price negotiations				○		○	
		Clarify the reasons for shipment delay	◎		○				○
		Reduce nonconforming items in received shipments	◎			○			
	Inventory control	Understand conditions of cargo movement by item	◎			○			
	Quality control and improvement	Analyze reasons for nonconforming items and improve production process	●		●	●	●		
		Relate experimental characteristics and measurement items			○	●			
		Analyze and solve claims and improve quality	●		○	○	◎		
	Productivity control and improvement	Clarify problems and make improvement; increase rate of operation	○			○		●	
		Analyze obstacles to timely delivery and meet deadlines	●		○				●
		Conduct high-precision quality control	○						◎
	Cost control and reduction	Reduce cost by VE	○		◎				
	Facilities control and improvement	Execute investment in facilities as planned			○				◎
		Adopt measures that require small investment but yield big results						●	
	Safety control	Examine reasons for worker injuries and minor accidents	○	○		○			
		Examine safety measures against major accidents and disasters	○	○				●	
Sales	Sales policy and strategy	Adjust relationship between needs and product types	○	○		◎	●		
		Critically compare one's own company against competitors' companies regarding needs			○	◎	●		
		Understand regional characteristics of high-volume sales items				○			
		Develop sales-promotion measures	◎					○	
	Sales record and profit control	Understand sales by retail stores				○			
		Understand sales by product items				○			
	Before service and after service	Acquire more accurate user needs	○		○	○	◎		
		Feed information obtained back to design			○	○			

Note: ● Listed in this text as examples ◎ Existing examples but not publicly announced ○ Fields where uses are conceivable.

TABLE 3.2
Applications of the seven new QC tools (mentioned in text)

Techniques	Applications	Titles	Contents	Page Nos.
Relations diagram method	TQC promotion	Relations diagram for major item clusters in promoting the introduction of TQC	Clarify items emphasized by own company	20
		Relations diagram for substance of TQC implementation	Clarify individual responsibilities for promotion and introduction of total quality control	82
		Relations diagram for administrative improvement in administrative department	Determine priority items for operational improvement in administrative department	82
	Policy control	Simple examples of relations diagrams made for selection of implementations items	Consider priority of counter measures to develop in light of economic environment, cost and quality assurance	55
		Relations diagram for policy control	Clarify relations between development of overall plant policy and implementation priorities of division	81
	Quality control and improvement (analysis of the production process)	Relations diagram designed to help identify causes of nonconforming items in assembly line	By constructing a relational diagram for chronic nonconforming items, causes previously overlooked were detected and successful countermeasures were adopted	21
		Relations diagram for market claims	A relations diagram was used in investigating the claims against missing labels often caused by human error	85
		Relations diagram for dispersion of characteristic values	To reduce the dispersion of hardness in product A, a cause analysis was conducted and the relationship among causes clarified	85
		Reduction of nonconforming items among items received	After a search for causes of nonconforming items in shipment received, an effective measure was adopted based on the interrelated causes that emerged	84
	Production method	Relations diagram for introduction of new production method	A storyboard method exposed the relationships among major items	83
	Production control and improvement	Relations diagram for delays in assembly process	Causes of delay in electric motor assembly identified and improved	86
KJ method	Development	Direction of future research and development (first and second round)	Examined various means of research and development and identified problems	106 108
	TQC promotion and QC circle	How to promote QC circle smoothly	Adjusted problems for smooth operation of QC circles	110 111
	Quality assurance	What quality assurance will be	Examined various means of quality assurance and clarified means of improvement	112 114
Systematic diagram method	Development	Systematic diagram for design quality of television UHF tuner	In response to increased demand for an all-channel tuner and the need to develop an advanced quality UHF tuner, a systematic diagram was used to develop the design quality needed and efficiency was increased	25
	Quality assurance	Systematic diagram for quality guarantee of a television VHF tuner	The quality required of the VHF tuner was selected from the design qualities used by television set makers to respond to demands of television users; quality assurance activities were simplified by using control points in the manufacturing process and by developing control quality characteristics and standards as well as design check items based on the desired quality for tuner design	126
		Systematic diagram for quality guarantee of automobile brakes	Required properties of automobile brakes were translated into design quality and control characteristics; by relating these control characteristics to specific aspects, quality could be better ensured	129
	Quality improvement	Systematic diagram for identification of causes regarding dispersion of wall panel thickness	Systematic diagram for dispersion of wall panel thickness was constructed and related to work standards and control materials as a means to detect major causes for dispersion; appropiate measures were taken, dispersion was reduced, and production capacity was increased	131
		Systematic diagram for reducing loss due to trunk nonconformities	To reduce the loss due to trunk nonconformities, a systematic diagram of possible remedies was constructed; after discussion, appropriate action resulted in improvement	133

TABLE 3.2 continued

Techniques	Applications	Titles
Matrix diagram method	Development	Functional family matrix regarding the weight and combination of granules
	Quality assurance	Substitute characteristics matrix for required properties of water pipe couplings
		Matrix of process-control items for substitute characteristics of water pipe couplings
		Matrix related to guaranteed properties, test items, and test machines for automotive brakes
	Quality improvement	Matrix designed to investigate causes for blemishes in printed cloth
Matrix data-analysis method	Planning and development	Classification of preferences for a variety of food products
		Exploration of the uses for new product, cloth A
		Spectrum analysis of the pleasing fluorescent lamp
		Forecast of the fashion cycle
		Expected automotive style
	Analysis of the manufacturing process	Analysis of nonconformities in automotive parts caused during pressing
		Search for the causes of scars on metal surfaces
		Analysis of the change in the dyeing outcome in the photo industry
PDPC method	Planning and quality assurance	PDPC of useless delivery
	Safety control	PDPC of the Gemini Plan
		Applicability to inspection of PDPC system safety
	Test evaluation (development)	PDPC for research and development

Contents	Page Nos.
In developing a weight-distribution system for a new product, functions were divided into three categories: soft, hard, and layout; the use of a matrix diagram of their functions made functional design smooth and easy	147
A systematic diagram was constructed for the required properties and substitute characteristics of water pipe couplings; by expressing terminal elements in a matrix diagram format, the importance of substitute characteristics was established, which in turn brought about a higher level of quality assurance	150
By placing the substitute characteristics described above in correspondence with the process-control items in a matrix diagram format, it was possible to reevaluate the importance of process-control items; reorganizing the process-control system then succeeded in increasing control efficiency and reducing nonconforming items	151
By relating the test and measurement items of the guaranteed characteristics of automotive brakes and by constructing a matrix diagram for test and measurement items and machines, needed improvements in the evaluation system were clarified and efficiency increased	152
By relating the blemishes in printed cloth to their causes and by further relating the causes to the sources in the manufacturing process, ideas were generated for reducing the sources of nonconformities	27
Preferences are classified by generic preferences, age, and gender; food products are also classified in a similar manner	166
Characteristics of the newly developed cloth are compared to various areas of use, facilitating identification of the most appropriate use	30
Extract spectral components with desirable color reproduction based on the results of the spectrum analysis of various fluorescent lights	169
Predict next year's fashion trends from the surveys of the past year's fashion designs	171
Analyze the dimensions of each automotive part and relate it to user preference	171
Search for the causes of creasing during pressing and adopt appropriate manufacturing process control	172
Matrix data analysis method was useful in investigating causes of scars on metal surfaces	172
Based on the observation regarding spectral absorption curves during a chemical reaction, distinguish a mixture of foreign substances from a change in the amount of ingredients	173
Examination of measures to prevent damage at time of cargo delivery	32
Examination of measures to retrieve astronauts safely from a space flight	177
Examine the system of prevention of street car derailment and change portions of track	195
Faced with troubles in the development of new products, product-devlopment and manufacturing departments employ PDPC and solve their problems	187

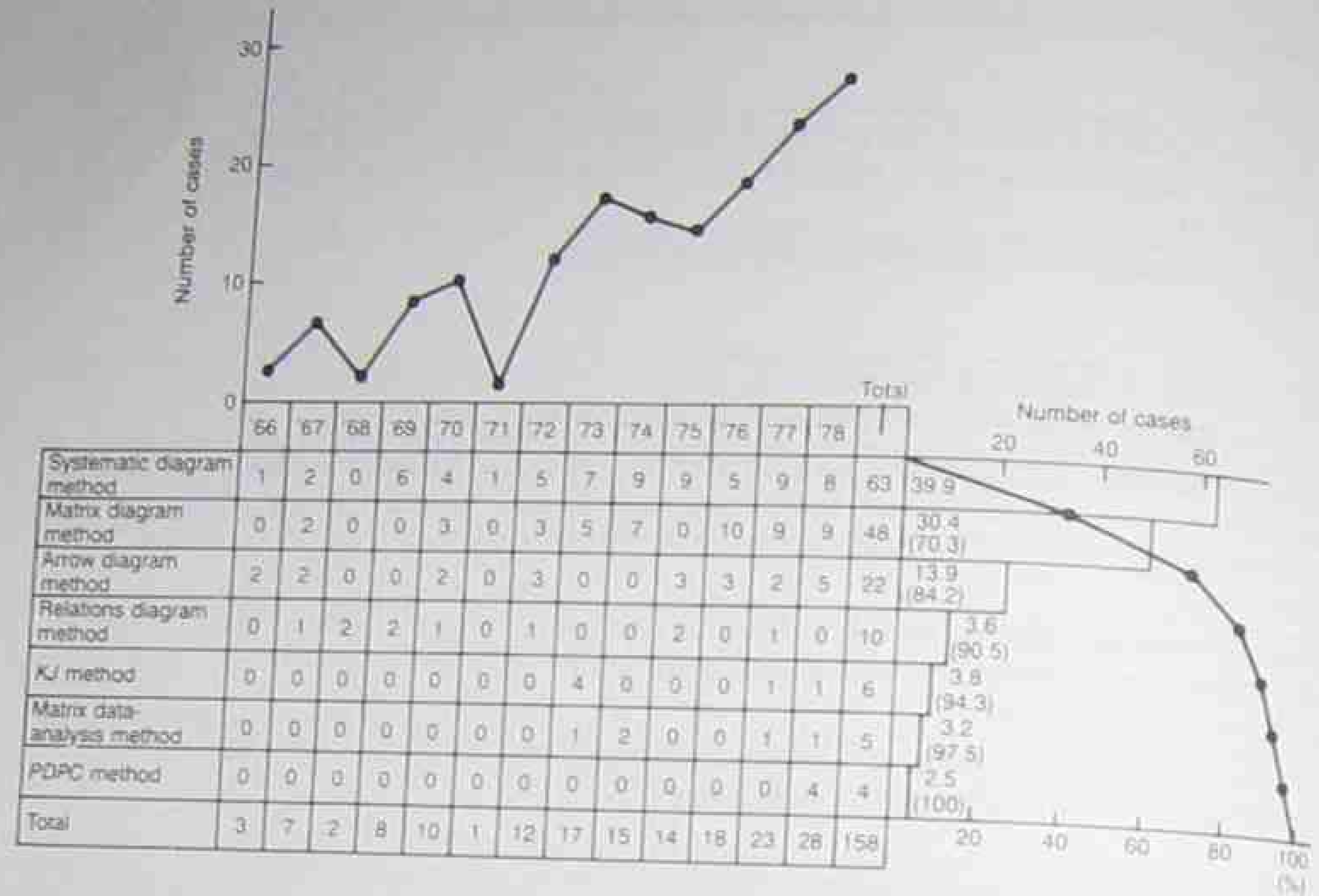
TABLE 3.2 continued

Techniques	Applications	Titles	Contents	Page Nos.
PDPC method	Test evaluation (development)	Examination of techniques and PDPC	In order to solve a problem in the manufacturing technique, mobilize all knowledge and information to generate possible solutions; write and rewrite PDPC three times to arrive at a solution	192
	Control and improvement of production volume	PDPC for increased productivity	Devise a measure to increase productivity by 20 percent	190
	Management and improvement of facilities	Ways of reducing NO_2 using PDPC	There are five measures for reducing NO_2; examine alternatives and select among them so that the result may be obtained with minimal investment	191
		Example where solution was found by writing the process of examination in terms of PDPC	Since an examination of accident-prevention measures was too complicated, a PDPC was written; as a result, what was considered unfeasible in the past analysis emerged as a promising solution; the whole situation was reexamined and appropriate measures were taken	195
Arrow diagram method	Quality design	Experimental run of the VE improvement plan for special resistor electrodes and quality confirmation plan	Both experimental runs were based on the VE improvement plan for special resistor electrodes used in electric motors, and daily plans for quality confirmation were constructed using cards	34
	Development	Development plan for electronic device model 52 (plan for initial stages; plan for shortening production schedule)	In making development plans for electronic device, schedule was shortened by reducing number of days required on critical paths	213 215
		Development plan for electronic device model 60 (plan for initial stages; plan for shortening production schedule)	Schedule was shortened through a modification of arrow diagram introduced by an addition of work	216 217
		Promotion plan for quality production of model 62	The daily plan for volume production of a new product was expressed in a time-scale arrow diagram for each participating department	220
	Quality improvement	Improvement plan for model T-827	The quality-improvement plan was expressed in a time-scale arrow diagram for each participating department	219

Applying the seven new QC tools to policy control

At the first QC symposium, entitled Quality Control: Introduction, Promotion, and Establishment (sponsored by the Japanese Federation of Science and Technology (JUSE), participants completed a questionnaire regarding important items of common interest in introducing and promoting total quality control.[1] These items, shown in Table 3.3, range from setting and clarifying policies and objectives to developing QC techniques. The "Rank" column indicates the ranking given to each item by the participants. The "Frequency" column refers to the number of participants who chose a particular item, and the "Points" column is the sum of the value of each rank. (*Note:* Assign a value of 6 to rank 1, a value of 5 to rank 2, etc., and a value of 1 to rank 6. Multiply the frequency by the value and add up the totals; this results in the sum total of points.) Table 3.3 lists items in order of total points.

According to this table, the establishment (and clarification) of policies, objectives, and plans, which constitutes the theme of this section, occupies the highest rank; it is the most important item in the introduction and promotion of quality control. We frequently consider the question of what steps have the highest priority in the implementation of total quality control. Other related questions we are often asked are how far on the list one should go in selecting implementation measures, whether starting with standardization is a detour, and so on. These questions arise from a misconception that the items listed in Table 3.3 exist in functional isolation from each other. Although Table 3.3 shows the relative importance of the listed



	'66	'67	'68	'69	'70	'71	'72	'73	'74	'75	'76	'77	'78	Total	
Systematic diagram method	1	2	0	6	4	1	5	7	9	9	5	9	8	63	39.9
Matrix diagram method	0	2	0	0	3	0	3	5	7	0	10	9	9	48	30.4 (70.3)
Arrow diagram method	2	2	0	0	2	0	3	0	0	3	3	2	5	22	13.9 (84.2)
Relations diagram method	0	1	2	2	1	0	1	0	0	2	0	1	0	10	3.6 (90.5)
KJ method	0	0	0	0	0	0	0	4	0	0	0	1	1	6	3.8 (94.3)
Matrix data-analysis method	0	0	0	0	0	0	0	1	2	0	0	1	1	5	3.2 (97.5)
PDPC method	0	0	0	0	0	0	0	0	0	0	0	0	4	4	2.5 (100)
Total	3	7	2	8	10	1	12	17	15	14	18	23	28	158	

FIGURE 3.1
Trends in the applications of the seven new QC tools (based on the Proceedings of the quality control conferences)

items, as indicated by the participants, it does not show that the items are interdependent and related to each other, as shown in the relations diagram in Fig. 2.2.

The establishment and clarification of policies, objectives, and plans are important, but they presuppose a clearly established policy. Implementation would be troublesome without thorough control and quality consciousness. Furthermore, without effective quality and QC inspection, it would not be possible to anticipate favorable results either in setting and clarifying policies, objectives, and plans or in the PDCA (plan-do-check-action) cycle. In other words, the items listed in Table 3.3 are all interrelated, and we may start with any one we choose. The shortcomings of the items listed in Fig. 2.2 will be corrected regardless of where one starts, as long as implementation follows and the PDCA cycle is repeated, with the utmost effort, toward company-wide improvement. Of the items listed in Table 3.3, we should first introduce quality control as a policy issuing from the highest level and then aim for overall improvement in the company's position and product quality.[2]

TABLE 3.3
Rank ordering of major items in the introduction and promotion of quality control

Ranking	Major items	Order or selection 1	2	3	4	5	6–10	Sum of frequencies	Sum of points
1	Set and clarify policies, objectives, and plans	𝍸 ////	𝍸 //	/	/		//	20	98
2	Thoroughly develop awareness for control, management, and quality	𝍸 ///	//	///		/		15	76
3	Conduct training programs and follow up	///	/	𝍸 /	////	//	///	19	66
4	Conduct quality and QC inspection			///	𝍸 /	𝍸 /	𝍸 //	22	49
5	Clarify control plans and control items		/	////	///		//	11	36
6	Clarify priority policies	///	//				/	6	29
7	Systematize and improve QC system	/		//	//			5	20
8	Standardization		/		/	//	////	8	16
9	Clarify measures of evaluation		/		/	/	///	6	13
10	Clarify duties and rights		//				///	5	13
11	Information control and management		/			/	////	6	11
12	Company-wide participation		/		/	/		3	10
13	Development of QC techniques				/	/	////	6	9

The authors have been fortunate in having participated in numerous discussions regarding the promotion of total quality control in different companies and examining in detail the problems of policy management. This section is based on experience gained from these discussions through the cooperation of various companies and is intended to serve as a guide for policy management. Out of this participation, the seven new QC tools emerge as highly effective QC techniques for policy setting and implementation.

This section provides a brief description of the use of each technique at each step of policy management and notes where to use caution. Policy management concerns effective implementation of decisions made by top management; this procedure is not directly applicable to the decision-making process involved in entering a new field of industry and embarking on a new overseas market, which are more properly the domain of top management. However, the procedure described in this section will be useful for gathering information to be used for such decision making.

Promoting policy management

Figure 3.2 outlines the procedure for policy management. Listed at the top of the figure are points of caution and results that may be anticipated for each step.

It is clear that among the steps involved in the control cycle, policy management requires a most careful and serious planning effort. In various fields, the psychological and organizational climate of Japanese companies harbor elements that hinder planned policy management.[3] For this reason, successful promotion of TQC through effective policy management depends on dispositional reform of companies.

Step 1: Determine policy and direction

Be certain of the needs for policy and direction. Identify problems of quality, quantity, and cost. Keep in mind that the ultimate goal is to upgrade achievement levels and to train staff resources.

a. Clarify the policy, implementation items, implementation process, levels of achievement and criticisms for the previous year or preceding period. In doing this, we should not be concerned exclusively with the levels achieved, but rather should focus on evaluation of processes involved.

b. For the current year or period, list fully business conditions, problems that may arise in connection with planned projects, and problems that may be encountered in one's own department.

c. Whether a problem is one of quality, cost, or timely production quantity, it ultimately becomes a quality problem. For example, improving the rate of operation may appear to involve problems related to improving the timing or quantity of production, but the underlying problem really may be that of quality, such as insufficient materials and inadequate facilities.

d. As mentioned in the preceding item, various functions, including quality, cost, and production volume, i.e., delivery (QCD), are interrelated in a complex way. Therefore, it is important at this

stage to construct a relational diagram for the implementation items of each function. Figure 3.3 is a hypothetical simplified relations diagram of policy and implementation items at the division level.[4] From this figure we learn that an examination of suppliers' products is necessary to maintain acceptable levels. To improve the quality of outside products, we may provide guidance to the supplier or shift production from outside to inside the company if the supplier is not amenable to quality improvement. In this situation, we would need to improve efficiency further and embark early on the training of personnel and the adjustment of specification instructions and facilities. Although Fig. 3.3 is incomplete, it helps to clarify the relationships between departmental policies and implementation items.

e. The format of a relations diagram is not restricted to that shown in Fig. 3.3. We may place problems, business environment, and projects at the center, write around these entries the policies and implementation items of top management, and note the lower-level management's implementation items outside those of top management. In this way mutual relations are easier to discern.

f. For each position, the implementation items can be divided into three categories: items to be implemented personally, those to be delegated to subordinates, and those to be implemented by other departments. We underscore items to be implemented personally because regardless of one's position, whether one is a division chief or a departmental chief, one must execute certain items if top management's policies are to be fulfilled successfully. Delegation of tasks to lower-level personnel is never sufficient. Managers must think about what they themselves ought to do. A systematic diagram is often used for developing the implementation items for each position.

g. List the needs ($g_1, g_2, \ldots, g_l$) determined by business condition, project, and problems of each department, and through the KJ method or brainstorming, generate as many small implementation items ($d_1, d_2, \ldots, d_m$) as possible that respond to these needs. Group them to obtain large implementation projects ($D_1, D_2, \ldots, D_n$) and remedies ($P_1, P_2, \ldots, P_r$). Be careful, in this process, that no implementation items are omitted for any position.

h. Since experience has accumulated on quality, cost, and delivery over many years, it is customary to revise improvement plans for the current year on the basis of cumulative experience. By analyzing

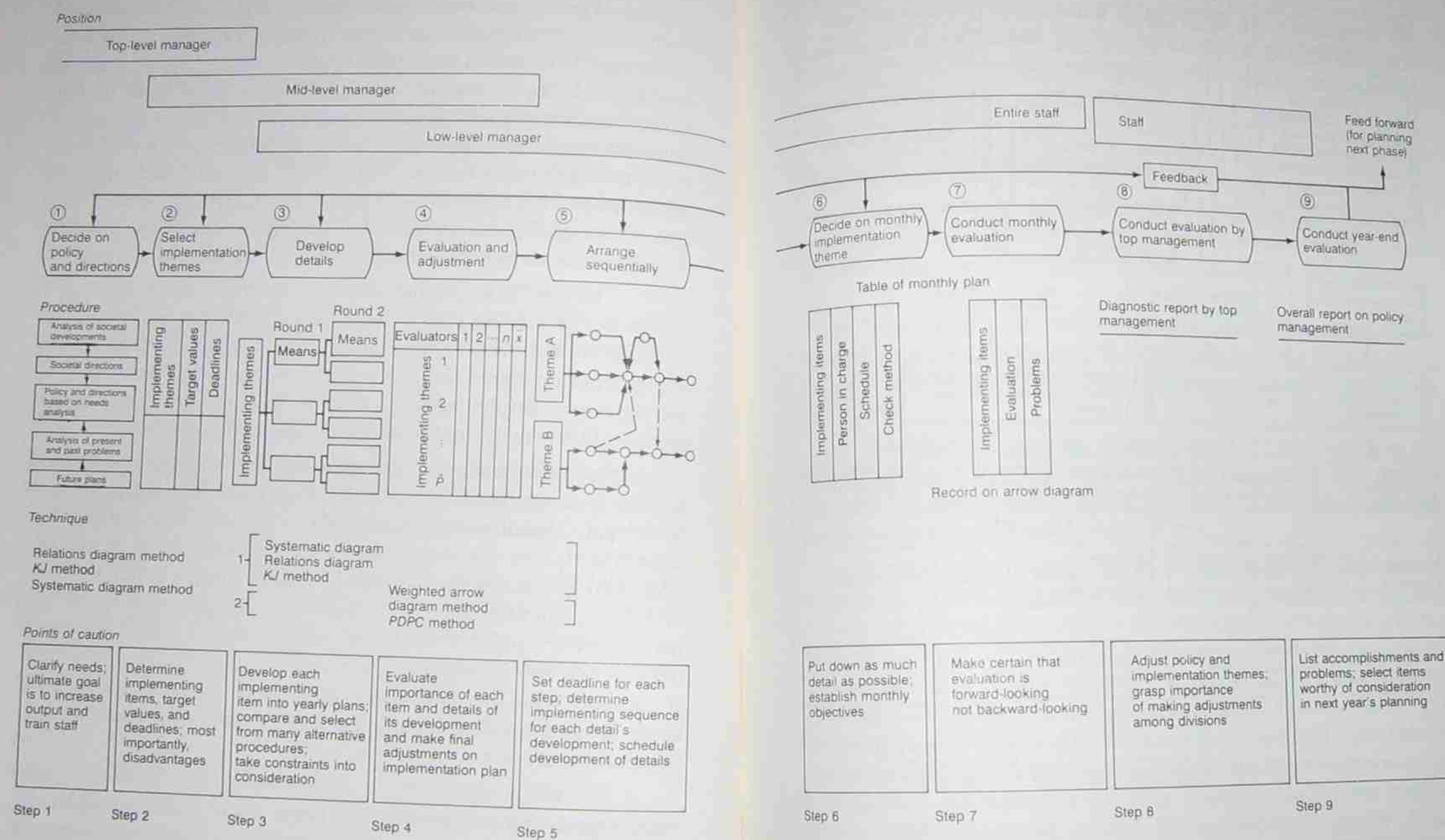


FIGURE 3.2
Outline of the policy control procedure

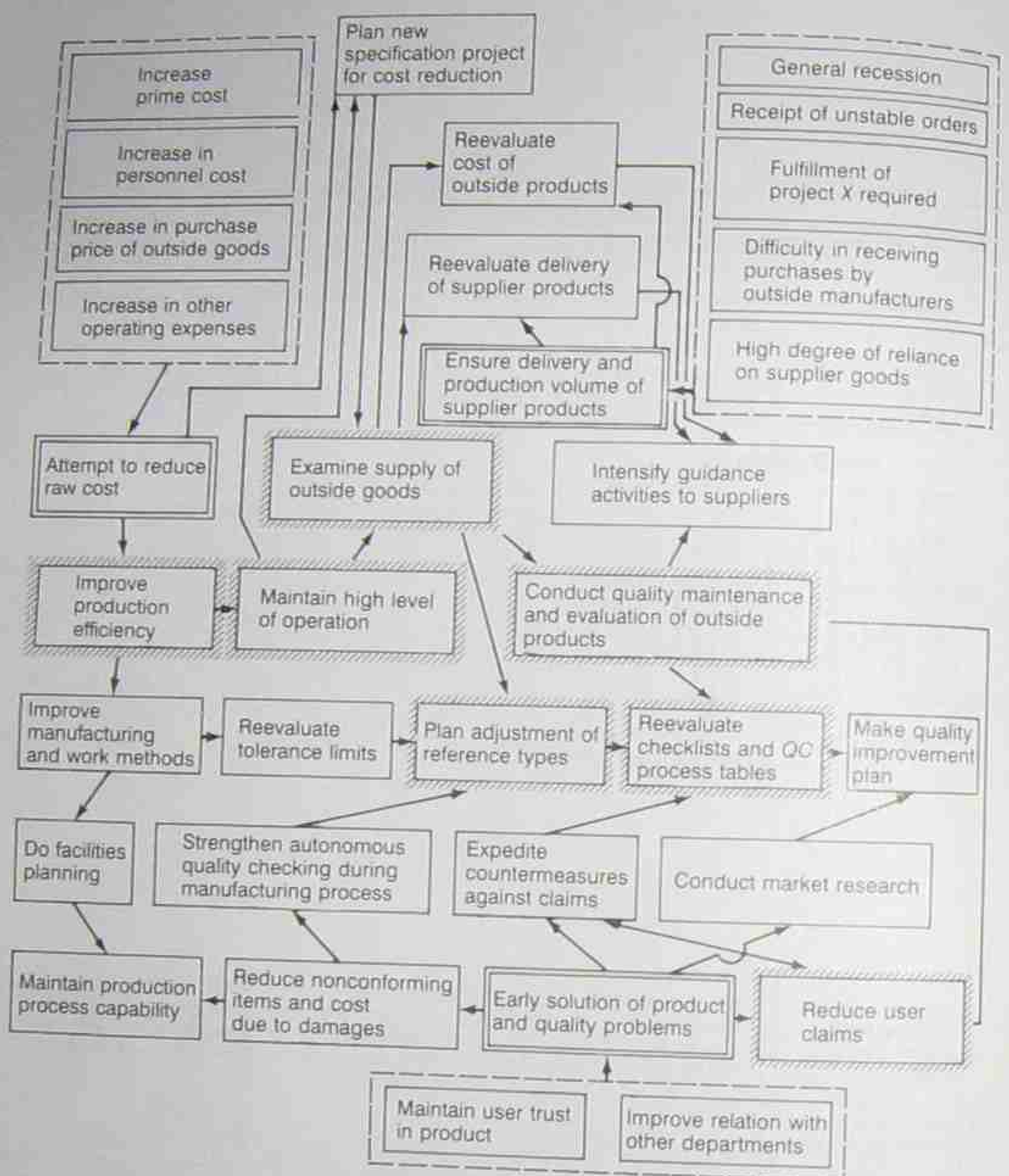


FIGURE 3.3
Simplified relations diagram for selection of implementing items

past experience with the PDPC method, we find that it is frequently possible to generate new ideas for implementing this year's objectives.

i. The implementation items mentioned here are those that require sustained implementation for the current year or period for

a given position; developing detailed steps is described in step 3 and thereafter.

j. The policies determined by the steps described up to this point can be consolidated. It is helpful to capture in a few sentences the background or need for such policies and their relation to implementation items.

Step 2: Select implementation activity items

Select activity items and determine their projected value and delivery dates. Weighing the costs and benefits of an activity is most important here.

a. Build on the results of the first step by selecting activity items for each grouping. In doing this, determine not only the target and delivery date, but also the required resources (human, technical, and monetary).

b. Establish whether the targets assigned to lower-ranking positions are related to accomplishment of the overall objective assigned to the department. Through this process activities at different levels can be brought into harmony.

c. Consider any constraints. In discussing quality (*Q*) improvement as an implementation activity item, we should ask whether it raises problems of cost (*C*), production volume (*D*), safety (*S*), the environment (*E*), or morale (*M*). If there are grounds for concern, we should naturally come up with activities that counteract this constraint. Similarly, it is necessary to give full attention to the potential of negative effect of cost reduction (*C*) on quality, production volume, safety, the environment, and morale. It is a common situation for cost reductions to result in claims of inadequate quality, which nullify the profit gained. Since product longevity is not readily controllable by inspection, this requires much more attention and consideration.

d. Another aspect of the cost-benefit analysis deserves comment. One should ask whether an improvement in one's department or segment of the production process may cause problems in the preceding or subsequent portions of the process. An improvement from one's standpoint may hinder or adversely affect the next phase of the production process.

e. An activity may be consistent with today's purpose, but we should ask whether the situation may not change if production volume or equipment facilities change in the future. Our thinking should be oriented toward the future.

f. The target level should be a value that can be achieved only with considerable effort, and it should be expressed in specific numerical terms. It is meaningless to set a goal that is beyond reach. There is a general tendency, however, not to set high targets. Since the spirit of policy management is to evaluate not only the results, but also the process, it is important that personnel realize that projected values should be set high. A reasonably high target poses a challenge and inspires people toward achieving it.

Step 3: Detailed development of activity items

Consider all options for proceeding, then select a suitable goal and make certain that constraints are taken into consideration.

a. When activity items for each position are established using the steps described above, it is necessary to develop those items further for the current year or period. For example, assume that receiving outside supplier parts shown in Fig. 3.3 has been selected as an activity item. A delivery date and quantifiable objectives are fixed. We note that there are many items to investigate. For example, we may need to ask what type of products should be obtained, from where, when, and what coordination should be made with other departments. In addition, we need to determine how and when to coordinate with other departments. Handled in this manner, each activity item develops into several implementation items for the current year or period. As shown in step 3 of Fig. 3.2, we mainly use the systematic diagram method to develop detailed plans. Figure 3.4 illustrates the development of aportion of a problem into a systematic diagram.

b. Although a delivery date and objectives may have been decided on for an activity item, there is more than one way to achieve our goal. Presented with one yearly plan, we should always ask if it has been selected after considering all the alternatives. Often there are obvious ways of proceeding that even outsiders are able to list. Simple plans, evaluations, and status surveys do not have much

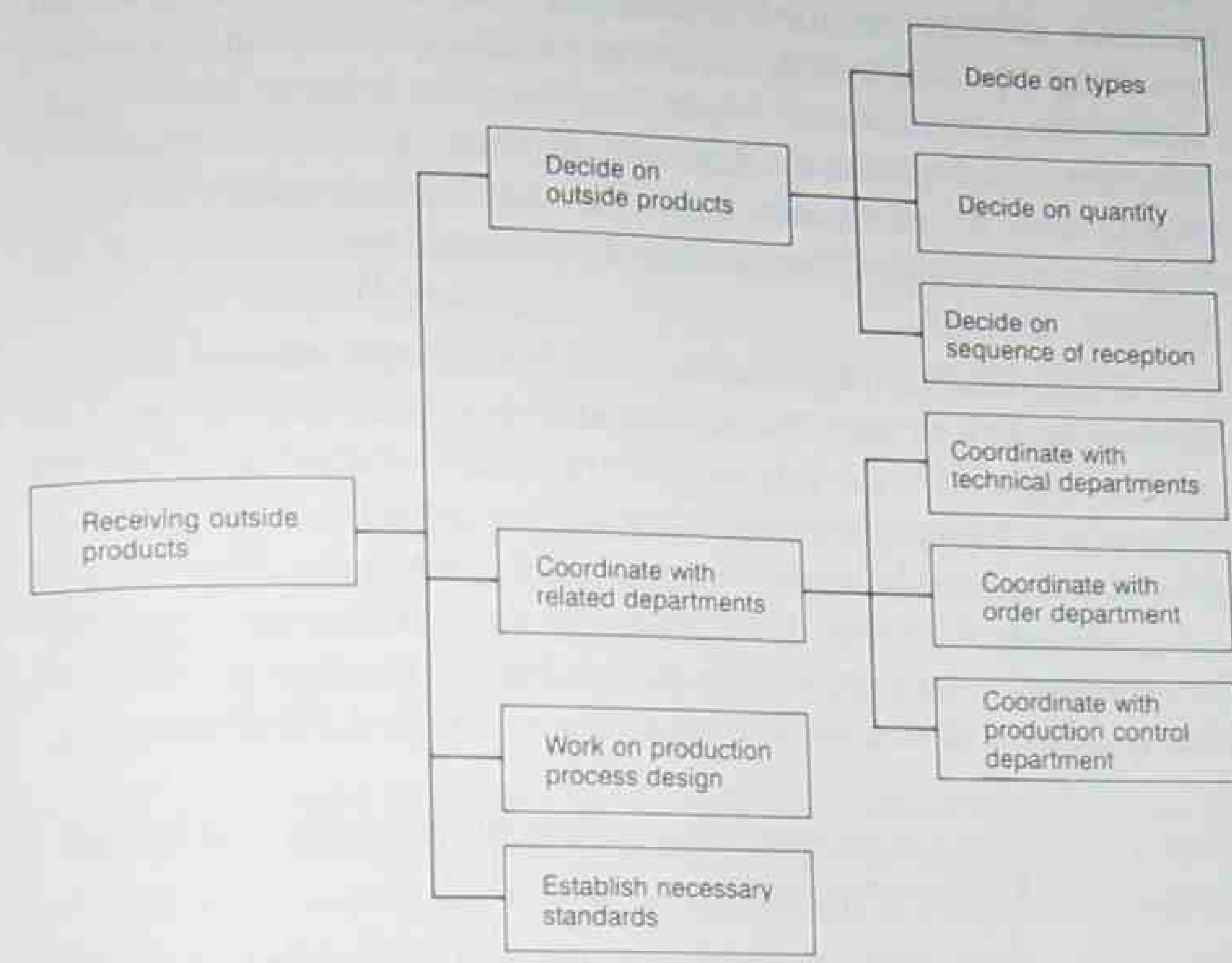


FIGURE 3.4
Detailed development of an activity item

meaning unless they are accompanied by the thoughtful evaluation of alternatives. Even for a matter as simple as training, we need to think about such things as who is best qualified to be trained, how a trainee should be selected, how to follow after training, how to train, and what training materials should be used.

c. In the preceding item, we underscored the selection of one method from a set of alternatives. What should guide the selection is the economic effect on resources. *Resources* here refer to budgetary resources, human resources, and technical resources. All these resources may boil down to the question of expense. If the solution to a problem requires the renovation of existing facilities and the building of new facilities, the resulting expense may turn out to be prohibitive and the solution impractical. However, constraints can sometimes be resolved by means of positions. The cooperation required for a solution may not be feasible at lower levels in the organization, although the same cooperation may be easy to obtain at a higher level. For this reason, upper management should be involved in assessing whether the constraints identified genuinely pose significant obstacles.

d. If a disadvantage is anticipated for an implementation activity item (part c in step 2), this fact also should be kept in mind in developing details for this step. If the creation of an activity item depends on the cooperation of another department, detail development must occur between the related departments. The detailed tasks developed for each implementation item are the keys to successful achievement of company objectives. For this reason, there must be clearly defined ideal criteria to guide our selection of objectives.

Step 4: Evaluate the importance of individual activity items and then finalize plan for implementation

Prioritize activity items through evaluation. Consensus should be a guiding principle in modifying the plan.

a. The completion of steps 2 and 3 result in an increased number of activity items and their details. Since it is often not possible to implement all these items, it almost inevitably becomes necessary to evaluate the importance of activity items and select a manageable number for implementation. Rarely will all items need to be implemented in a given year or period. It is appropriate to defer some items of lesser priority. Indeed, certain items may increase in importance at a later stage.

b. Evaluation can take place immediately after either step 2 or step 3. Use the number of items to be evaluated as a guide in determining the timing.

c. Several methods of evaluation are available.[5] If P items are to be evaluated by N persons, a simple method is to use a 5-point scale for each item, where a score of 5 indicates the greatest importance and a score of 1 stands for the least importance. The evaluation

score of an individual item is reached by the mean X of evaluation scores; the dispersion of evaluations among evaluators is expressed by the range R of evaluation scores. If $R = 4$ for an item, we know that one judge has given a score of 5, while another judge has assigned a score of 1. One judge thought it was the most important, while another judge thought it was the least important, reflecting diametrically opposed views. In such cases, we might discuss the evaluation criteria to reduce the R value. (Evaluation can also be done on a 10-point scale.)

d. Activity items and their detailed development are to be brought up for policy management in the order of evaluation ranks found in item c above.

e. In assessing the importance of items, as many people familiar with the evaluation items as possible (drawn from many fields and positions) should serve as judges. This objective can be accomplished by making a file of pertinent information available to the potential judges.

f. Prioritization is not the only goal of evaluation. It is better to develop a common perspective and consensus among those participating in the evaluation, even if misleading or incorrect values are sometimes obtained.

g. It is meaningless, even dangerous, for a majority of lower-level personnel to evaluate an item that is properly a decision for top management. Although the results of such evaluations may be considered by top management, the decision ultimately should be made by top management. It is, in fact, the duty of top management to make heart-wrenching, difficult decisions.

Step 5: Schedule implementation activity items

Determine how much time is needed to carry out the details of each step and develop a schedule. Identify the critical step upon which completion of an objective depends. Arrange the necessary steps in a chronological time sequence using an arrow diagram.

a. Through step 4, we have clarified the selection of items and their detailed development. Our next task is to determine the time to be allocated for implementing each step.

b. For each step in an implementation plan, there is a natural sequence to be followed. Implementation of one step might depend upon completion of several preceding steps. If the completion of a step depends on cooperation or input from other departments, fulfillment of their obligations by other departments is a prerequisite, and enough time should be allowed for coordination.

c. By going through this process, we determine the implementation sequence and deadlines for detailed development of each step. An arrow diagram is a suitable way to express this. We need not consider it a difficult task; rather, we should simply try to specify the connections among the detailed steps of each activity item and their delivery dates. For items that rely on another department's participation, mark the necessary delivery date clearly and make certain that the participating department agrees on the date. Furthermore, for activity items that depend on another department's input, make sure that the item is clearly incorporated into the work plan of the participating department and that the department's completion date matches the required date.

d. Clarify which steps are critical to final completion and fulfillment of the objective. Does an item that must be executed pose a deadline problem? Is there an item on which a technical or resource problem can be expected? When these problems are analyzed in greater detail, what emerges as the critical link? If a problem is expected, would additional personnel help? Would allocating more funds be of help? Would assistance from another department be of any help? It is necessary to think through possible solutions like these to enhance the probability of successful implementation.

e. In general, evaluation of situations described in item *d* should precede actual implementation. Frequently, only easy items are executed, deferring difficult ones to a later time. This makes it difficult to accomplish all activity items, however, and inevitably disrupts the following year's plan.

f. Typically, we proceed from planning to execution; we then discover problems, adjust our plan, return to execution and discover more problems. Despite the fact that the PDCA cycle appears to progress forward, achieving objectives sometimes turns out to be very time consuming. In such a situation, it is necessary to ask whether the person in charge could have anticipated the problem at an earlier stage. It is not uncommon to find in various companies that

certain problems could have been anticipated at the planning stage and thus prevented. For purposes of early identification and prevention, it is necessary to have a contingency plan for each step in the detailed development of an activity item—a plan of countermeasures and alternatives in case the primary plan fails to proceed as expected. Although a plan is usually based on the presumption that everything will proceed normally, it is helpful to have a contingency plan for areas where difficulty is conceivable.

A way to promote steps 3 through 5: the PDPC method

Steps 3 through 5 are sufficient if the execution sequence and the details of each activity item can be specified fairly clearly, such as with construction or assembly. With a development project or a complex quality-improvement task, however, the detailed execution plan laid out at the early planning stage can hardly be complete. During execution, we need to revise and develop the detailed plan on the basis of the results of implementation or research up to that point. In recent times, the PDPC method has been used to determine the process that can best bring about the desired outcome despite conceivable complications.

Consider the situation where parallel efforts on each activity item by all participating departments are needed to bring about the desired state B_p from the initial state A_0 (Fig. 3.5). A_1, A_2, and A_3 refer to the series of activity items assigned to each department. Discussion among the participating departments midway through the execution period increases the tightness and cohesiveness of the subsequent portion of the plan. Sometimes we need to revise the policy or activity items during an execution period, but use of the PDPC method will automatically do this.

In interim discussions each participating department comes to recognize the role it plays in the expanded detail development and to better appreciate what is required of it. Consensus is reached and the departments become better motivated to fulfill their obligations regarding the delivery target.

Series A_1, A_2, and A_3 of activity items shown in Fig. 3.5 represent independent paths from A_0 to B_p. When there are discrepancies in execution among the series, the PDPC method can be used to remedy the situation. At one company, in response to a request to reduce NO_x, five possible measures were listed in decreasing order of cost

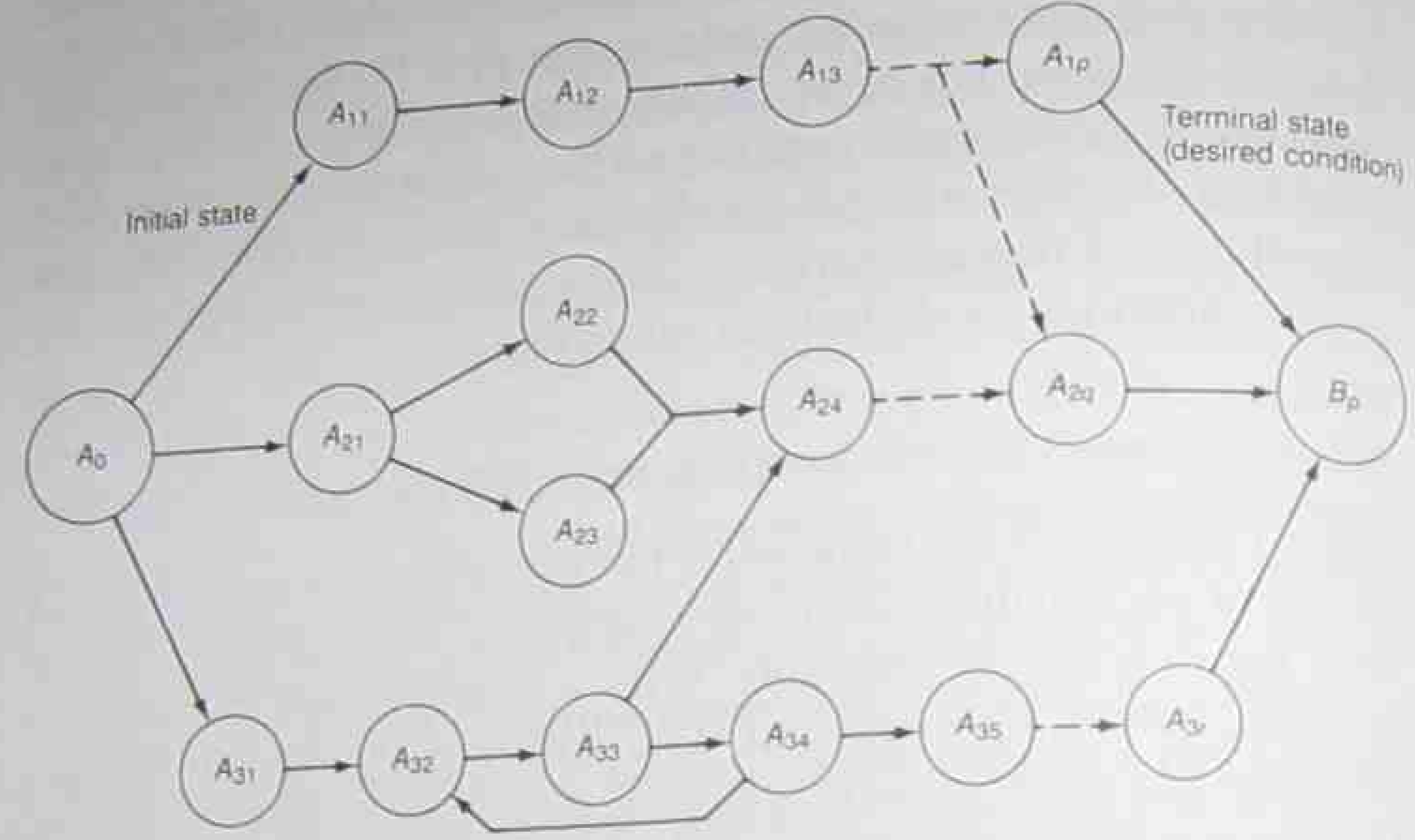


FIGURE 3.5
PDPC diagram to arrive at desired state

while always keeping the target date in mind. By this means, the company was able to achieve its goal with the least investment.[6]

Use of the PDPC method for policy-management is still under development. Depending on the situation, we can develop activity items using the systematic diagram in step 3 and further develop them into a chronologic sequence through the PDPC method. This point will be discussed further in the PDPC method section in Chap. 9.

Step 6: Decide on monthly activity items

Work out the plan in as much detail as possible. Set monthly goals for activity items.

a. Each month, set activity-item implementation goals for the month. At the early stage of policy management it may be difficult to

set up a detailed plan for the year. Even in such a situation, however, it is possible to work out activity items for the current month to at least aim toward the desired year-end result.

b. It is advantageous to specify as much as possible who is in charge, the length of the execution period, and the promotional measures for each activity item. In general, daily routines and fixed schedules keep people very busy, which often results in cramming most of the monthly improvement activities in toward the end or completing schedules in less than a fully satisfactory manner. It thus becomes necessary to determine what activities are scheduled for each week.

c. Establish clear objectives for each month. This is an important step in terms of the following month's evaluation. Within the context of detail implementation for a given month, it may not be possible to link monthly activity goals with the ultimate objectives of quality, cost, and delivery. In this case it is necessary to decide in advance how evaluation is to be conducted.

Step 7: Conduct monthly evaluations

Evaluation should focus on what lies ahead; an evaluation process should not reflect the past.

a. Evaluation is undertaken to ascertain if the proposed activity has taken place and to determine the level of achievement. Concerning the problem of standardization, for instance, we can ascertain the level by asking how the discussion progressed, what was discussed, and what was the result of trial testing.

b. An evaluation should not center on shortcomings in results and blame. What is important is to gain information and knowledge regarding how one may guide subordinates, what should be implemented in the following month, and how it should be implemented. Such an evaluation is what we call a forward-looking, action-oriented evaluation.

c. Steps 6 and 7 should take place simultaneously each month during the execution period. The main point of discussion, however, is to determine what must be achieved in the following month.

Step 8: Conduct a diagnosis by top management

Carry out adjustments regarding policy and activity items. Grasp the importance of interdepartmental coordination.

a. During the current year or period, it is necessary for the president or other top managers to conduct a diagnosis of the success of implementation activity related to the basic policy of each department or division.

b. Departmental policies and activity items must remain flexible so that they can be revised and supplemented in response to changing market conditions, problems of new standards, and progress status within each implementation year or period. One of the purposes of the diagnosis by top managers is to verify that such revisions and additions are instituted voluntarily by rank-and-file workers.

c. At the early stage of policy implementation, it is common for interdepartmental cooperation to be not so smooth, even though such cooperation may have been clearly specified. Another purpose of the top-management diagnosis is to evaluate the level of existing cooperation and to generate ways to bring about the needed cooperation if it is deficient.

d. Top management diagnosis is important from the standpoint of assessing progress and learning about the underlying process. It also contributes to an understanding and furthering of quality control at the top-management level itself.

Step 9: Conduct an overall year-end evaluation

List accomplishments, and assemble a list of problems. Identify items necessary for incorporation in next year's policy.

a. At the end of the year, list accomplishments for each activity item derived from the underlying policy. Rather than worrying about the results, pay attention to the process. The PDPC method may be useful for this purpose.

b. Problems listed and analyzed from the current year can be used as basic data for next year's policy management.

c. Acting out our commitment to policy management, we learn how to improve the quality of the policy-management system itself, including policy setting and daily schedule planning.

d. An analysis of problems that arise in the functional management of quality assurance, production volume and delivery, and cost may serve as a useful reference point for setting policies for the following year.

Conclusion

Regarding the seven new QC tools and policy management, further research and development are needed in the following areas:

Establishment and promotion of company-wide policies

It is hoped that efforts will be made to apply the technique of policy management discussed in this section to the establishment and promotion of company-wide policies. We look forward to the use of the seven new QC tools within the policy management system in order to harmonize top-management decisions with staff support activities and to advance company-wide TQC activity in an appropriate direction.

Coordination with a profit-control system

The preceding section focused primarily on the establishment and promotion of policies. In the policy management of individual companies, however, securing profit is important. This relationship between policy management and the profit-control system needs further study.

Standardization of the policy-control system

For the policy-management system discussed in this section to be introduced, established, and developed in individual companies, it will be necessary for such companies to announce the standardization of the system and to issue various processing forms as a systematic company policy.

Accumulation of experience on the use of each technique for policy management

Typically, in working toward a solution to a plant problem, a plant manager may meet with group leaders and collectively construct a relations diagram. Each group leader carries the group's tasks back to the group, which, in turn, constructs its own relations diagram and works toward the common goal of solving the problem. Also in some cases the promotion of company-wide policy management is expressed in an arrow diagram format and the solution of individual problems relies on the PDPC method. The former aims at a wider dissemination of knowledge, whereas the latter promotes the goal in terms of specific details.

Figure 3.2 is an example of such an application. In applying our techniques at each step, we hope our readers' cumulative experience will lead to many suggestions for the solution of existing problems as well as to the articulation of new problems.

We have discussed several points of general caution as well as the application of the seven new QC tools to promote policy management. In our experience the system described in this section helped to increase the likelihood of success at the stage of policy establishment. The greatest advantage reported by managers is the ability to obtain a comprehensive overview of the progress of improvement. We hope to see continued improvement in applications based on feedback from various fields.[7]

Notes

1. Mizuno, Asaka, and Ishikawa (eds.), *Dai Ikkai Hinshitju Kanri* (*The first quality control symposium*) (Tokyo: JUSE Press, Ltd., 1965).
2. A number of Japanese authors have considered the important issue of promoting policy control, for example, Ishihara Katsuyoshi, "Procedures for Establishing Quality Control Policy," *Hinshitsu Kanri* (*Quality Control*), vol. 27 (October 1976), pp. 8–12; Ikazawa Tachuo, "Managing Quality Control Policy," *ibid.*, pp. 13–15; Tamura Shoji, "Promoting Management Policy," *ibid.*, pp. 16–19.
   Mizuno Shigeru has also addressed the related problems involved in planning and programming in "Planning and Programming in Business," *Quality Control*, vol. 20 (November 1969), supplement pp. 1–4.

   Koura Kozo has studied the relationship between objectives control and the QC team in "Examples of QC Teams," *Quality Control*, vol. 19 (October 1968), pp. 32–37.
   Finally, a number of studies have been done on policy setting, implementation and promotion at different companies, for example, Ishihara Katsukishi, et al., "Staff's Role in Developing QC Policy and Plans (Nos. 1–4)," *Quality Control*, vol. 21 (November 1970), supplement, pp. 68–86; Yamamoto Mitsuo, "Examples of Policy Management in 1976," *Quality Control*, vol. 28 (June 1977), supplement, pp. 9–13.
3. Matsuda Takehiko, *Keikaku to Joho* (*Planning and Information*) (Tokyo: Nippon Hoso Kyokai Publishing Co., 1969).
4. In the field of quality control, the relations diagram was first introduced in Senmochi and Mizuno, *Hinshitsu Kanri no tameno Keizaikeisan* (Economic calculations for quality control) (Tokyo: JUSE Press, Ltd., 1971).
5. For more on evaluation *see* Shigeru Mizuno and Yoji Akao (eds.), *Hinshitsu Kino Tenkai* (Development of quality and function) (Tokyo: JUSE Press, Ltd., 1978).
6. Eguchi, Kagoyama, and Kishimoto, "On Reducing Nitrogen Oxide in LPG Boiler," in *Hin QC Nanatsu Togu Katsuyo irei Happyokai Jireihsyu* (Applications of the seven new QC tools) (Tokyo: JUSE Press, Ltd., 1978).
7. This chapter is based on research reported in Nayatani Yoshinobu, "Applying QC Techniques in Policy Management: *Hinshitsu* (*Quality*), vol. 8, no. 4 (1978), pp. 25–31.

PART
TWO

Details of the Seven New QC Tools

4

The Relations Diagram

The relations diagram clarifies intertwined causal relationships in complex problems or situations in order to find appropriate solutions.

The method

Many problems confront modern businesses, problems that pervade such areas as

- Quality assurance
- Cost control
- Delivery schedules
- Resource economy
- Product liability prevention
- Environmental protective measures

- Reliability
- Automation
- Personnel economy
- Energy economy

Moreover, these problems involve complexly interrelated causal factors. Even if a problem were limited simply to quality control, there would still be many intertwined causal factors. Such problems cannot possibly be solved through conventional problem-solving methods, which rely on the efforts of a single staff member eliminating the causal factors one by one. Today, teamwork must be widely and effectively used. The relations diagram method is suitable for such complicated modern problems.

The relations diagram shown in Fig. 4.1 employs arrows to show the cause-and-effect relationships among a number of problems and factors that influence them. The relations diagram method utilizes this type of diagram as a technique for problem resolution. The relations diagram method could further be defined as a technique used to solve problems that have complex cause-and-effect and objectives-means relationships through

1. Isolating all factors related to the issue
2. Expressing these factors freely and concisely
3. Identifying logically the cause-and-effect relationships and depicting them using arrows in a relations diagram
4. Producing a complete picture
5. Extracting the key factors

When this technique is used, diagrams should be drawn up and revised a number of times by several people. During this process, issues can be identified clearly, a consensus can be obtained, and ideas can be developed. In other words, the relations diagram method is an effective technique for reaching the root of a problem and devising solutions.

The relations diagram method was developed into a problem-solving method from diagrams used in management indicator relational analysis.[1] During an attempt to apply relational analysis, after

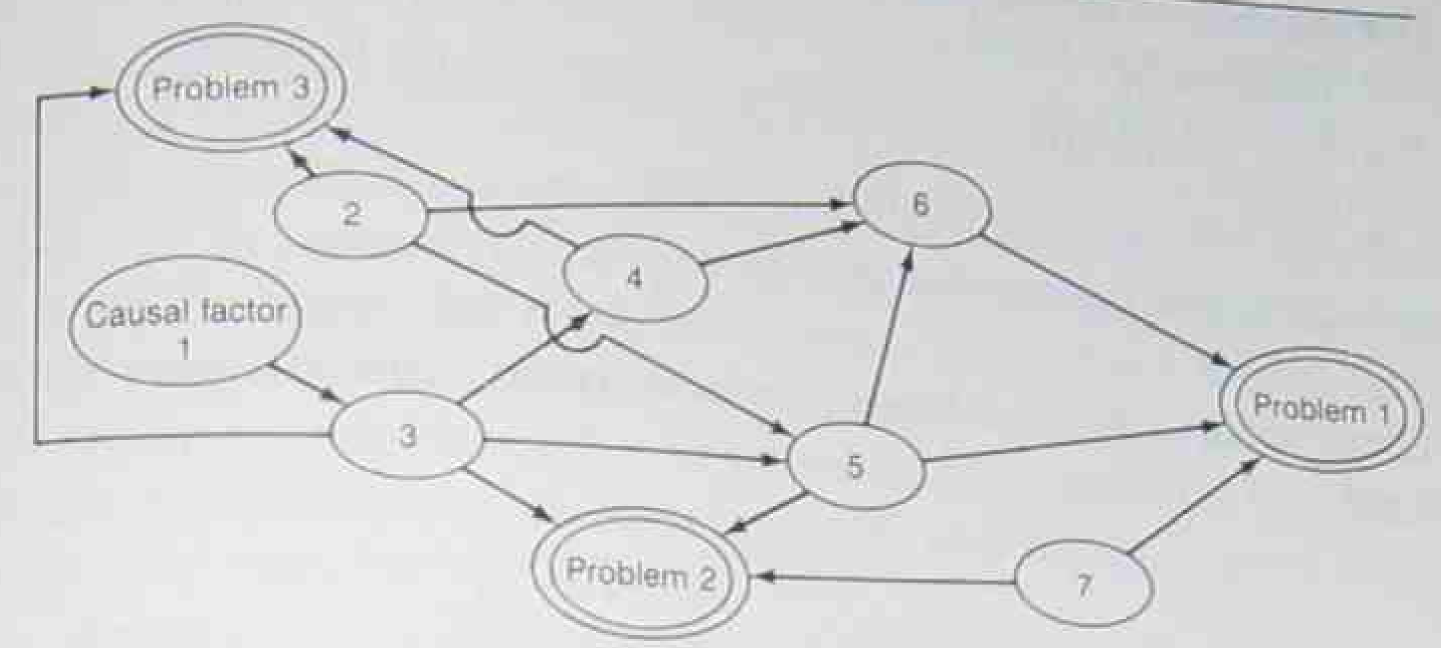


FIGURE 4.1
Sample relations diagram

the cause-and-effect relationships had been reorganized a number of times using relations diagrams, it suddenly became clear that the greater part of the problem had been solved. With this starting point, the use of relations diagrams has been expanded to other areas. Having been confirmed as a useful tool, the relations diagram has come to be viewed as one of the seven new QC tools.

Application fields

The relations diagram is used not only for problem resolution in business activities, but also for the analysis of social phenomena in other fields. In QC operations, this technique has been effective in

- TQC advancement operations
- Quality assurance (QA) and QC policy development

- Measures developed to counter manufacturing-process deficiencies
- The promotion and development of small group activities
- Market claim-prevention policies
- Operations improvement

There are many other fields in which the relations diagram can be used.

Special characteristics

The relations diagram can be used to organize information to highlight cause-and-effect relationships. The key characteristics of this technique are as follows:

1. It is a useful means for organizing problems that have complex factor relationships.
2. From the planning stages on, it provides a means for examining problems from a broad perspective.
3. It allows major factors to be accurately identified.
4. Its use facilitates consensus-building among participants.
5. Because it allows opinions to be expressed freely and does not restrict them with forms or models, problems and causes can be readily identified.
6. Because the diagram itself is not restricted to any specific framework, the conception and development of ideas are facilitated.
7. It helps in eliminating preconceptions.

Those who have successfully used the relations diagram have agreed that it has merits not found in characteristic cause-and-effect diagrams. This is so because, as stated earlier, a complete relationship can be mapped out without restricting expression or dictating the use of a specific framework; this, in turn, facilitates the conception and development of improvement ideas.

When the relations diagram method is introduced to study groups, many ask about the differences between this method and the KJ method. Both processes involve teamwork, the use of brainstorming sessions to develop ideas, and the use of a diagram to organize factors. Ideas are developed, and a consensus is reached. However, in contrast to the KJ method, which uses subjectivity in creating diagrams, the relations diagram method is a logical technique which connects the cause-and-effect relationships among factors by arrows. Moreover, it is a method that incorporates important factors and items from broad perspective, and in this respect, it varies greatly from the KJ Method.

Based on a number of case studies, the advantages and disadvantages of the relations diagram method are presented below.

Advantages

- Problems can be simplified into several major points, and this assists the development of improvement measures.
- The relationships among several departments can be clarified, and problem solving becomes easier when diverse groups are able to work in concert.
- Comments can be entered as they are stated, without restrictions.
- After a number of revisions, the key issues and substantiating points in problem resolution become apparent.
- When further information is added, it is easy to organize connections between factors, making forecasts possible.
- The relations diagram simplifies explanation of a complex problem to others (especially upper management).

Disadvantages and comments

- Since the format is unrestricted, the resulting diagrams differ from team to team, even if the same problem is being tackled (however, the conclusions ordinarily are similar).

- If factors are expressed too simply, arrows may point in misleading directions.
- If the diagram is too complicated, it becomes difficult to understand. When this occurs, important factors may be overlooked in drawing conclusions.
- Despite appearances, constructing an adequate diagram is a surprisingly difficult process.
- It is necessary to redraw the diagram in response to changing situations, a process that can be time-consuming.

Major uses of the relations diagram method

There are two major uses of the relations diagram method: single- and multiple-objective problem solving.

Multiple-objective problem solving (multiple-objective model)

Here the relations diagram is used in cases where there are two or more issues to be considered, such as TQC advancement, policy management, or operations improvement. The relations diagram is effective when applied to multiple-objective problem solving and demonstrates special characteristics not found in other techniques.

In order for a number of departments to cooperate in the development of a plan of action, the items essential in implementing these actions must be enumerated. After grouping all items to express their functional relations in a relations diagram, the major items are extracted. Generally, in such cases, it is also necessary to consider such objectives as quality, volume and delivery, cost, and safety, since the issue of quality control involves all these factors. For example, in order to raise productivity, it may be necessary to increase the speed of the machines used. However, such a step may adversely affect quality. For this reason, the process also may involve modifying material distribution so that when the machine speed is increased, quality products will still be produced.

Single-objective problem solving (single-objective model)

Here the relations diagram is used to solve problems having a single objective, such as decreasing product nonconformities, devising claims-prevention policies, or reducing schedules. Independent measures taken by individuals on the basis of their own view of a problem may result in a loss or no change in conditions. In most cases, it will be difficult to achieve beneficial results. It is therefore necessary to bring together the responsible personnel as a team to allow them to present their opinions. By constructing a relations diagram, a consensus can be obtained and an effective means to solve the problem can be devised.

Basic structure of a relations diagram

In a relations diagram, short sentences or phrases expressing factors or problem points are enclosed in rectangles or ovals, and cause-and-effect relationships are indicated with arrows. The goal to be achieved or the problem point to be solved is enclosed in a rectangle or oval, and important items or factors are shaded in so that they can be more readily identified. As a rule, the arrow in a cause-and-effect format points from the cause to the effect. Likewise, in an objectives-means format, the arrow points from the means or measure taken to the objective. However, where measure *B* is required in order to achieve objective *A*, it is sometimes better understood if the arrow points from *A* to *B*. Depending on the issue, however, an arrow pointing in the opposite direction may be more effective. It is necessary for the group in charge to determine beforehand the direction the arrow will point and its significance when creating a relations diagram.

Whether phrases or sentences are used is decided by the members of the group creating the diagram; however, it is best if these are kept short and easy to understand. In most cases it is best if at least a noun and a verb are used. (Expressions involving only the use of nouns often are not clear enough, and the quality of the diagram suffers as a result.)

Format of the relations diagram

A special characteristic of the relations diagram is its unrestricted format. However, general format considerations are shown for a centrally converging diagram in Fig. 4.2, a directionally intensive diagram in Fig. 4.3, and a relationship indicating diagram in Fig. 4.4.

Centrally converging relations diagram

The major item or problem to be solved is located in the center, and the related factors are arranged around the item or problem in such a way as to indicate close relationships (Fig. 4.2).

Directionally intensive relations diagram

The major item or problem to be solved is located on one side of the diagram, and the various factors arranged in accordance with the flow of their major cause-and-effect relationships on the other side (Fig. 4.3).

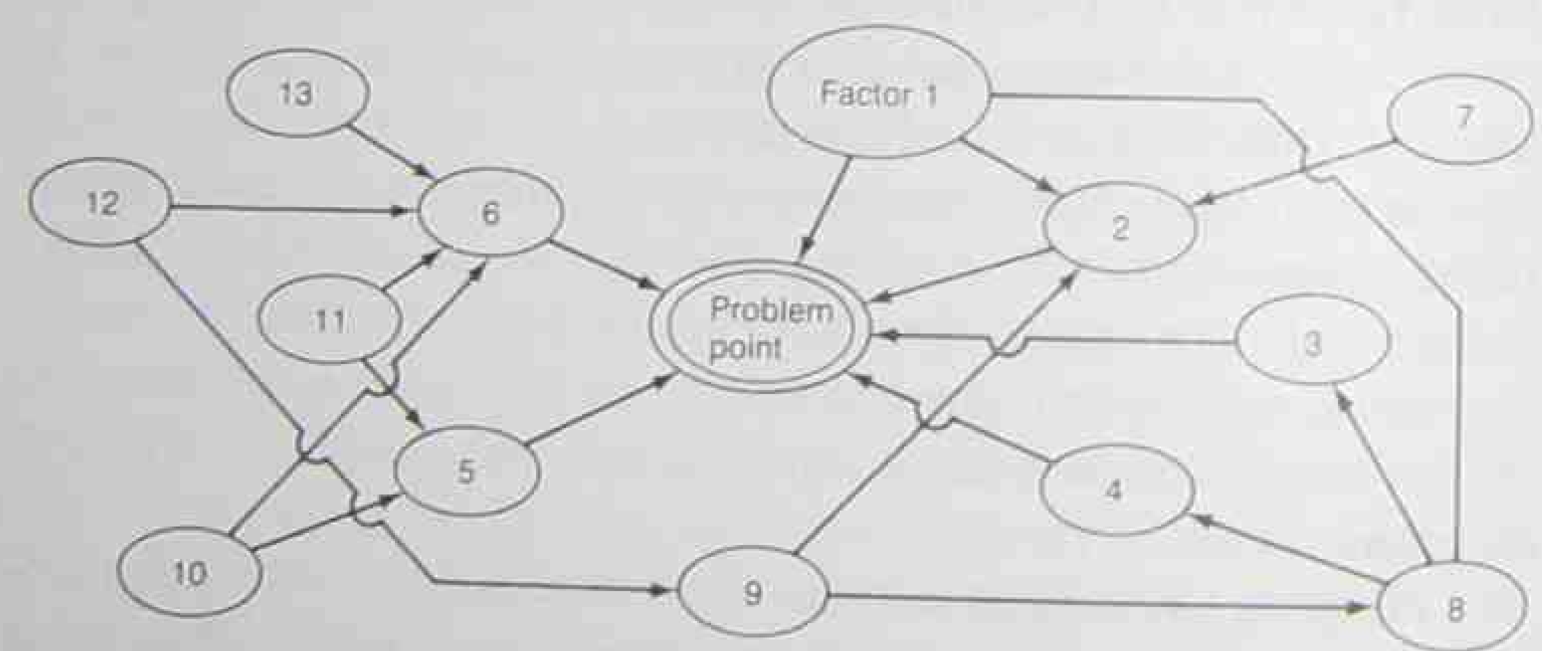


FIGURE 4.2
Centrally converging relationships

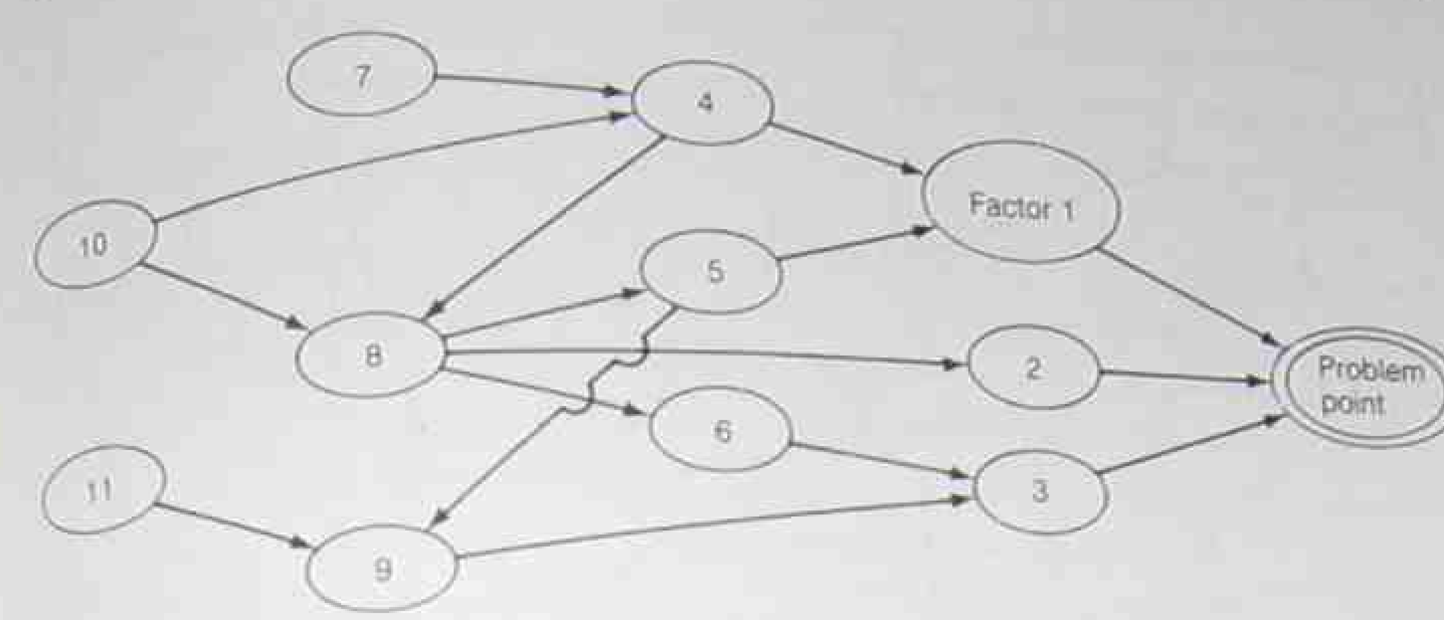


FIGURE 4.3
Directionally intensive relationships

Relationship indication relations diagram

There are no restrictions on this format because the main point is to arrange the cause-and-effect relationships of the application items or factors so that they are expressed in a straightforward manner in a diagram (Fig.4.4).

Application-format relations diagrams

Application-format relations diagrams are based on the three basic types of diagrams discussed above. This format includes diagrams in which the structure of the diagram is based on organizational structure, processes, the 5M's (man, machine, material, method, measurement), and other such items (Fig. 4.5). Other applied formats include diagrams in which a number of factors are collected, as in the KJ method (Fig. 4.6), and diagrams in which the interrelationships are entered after systematic development (Fig. 4.7).

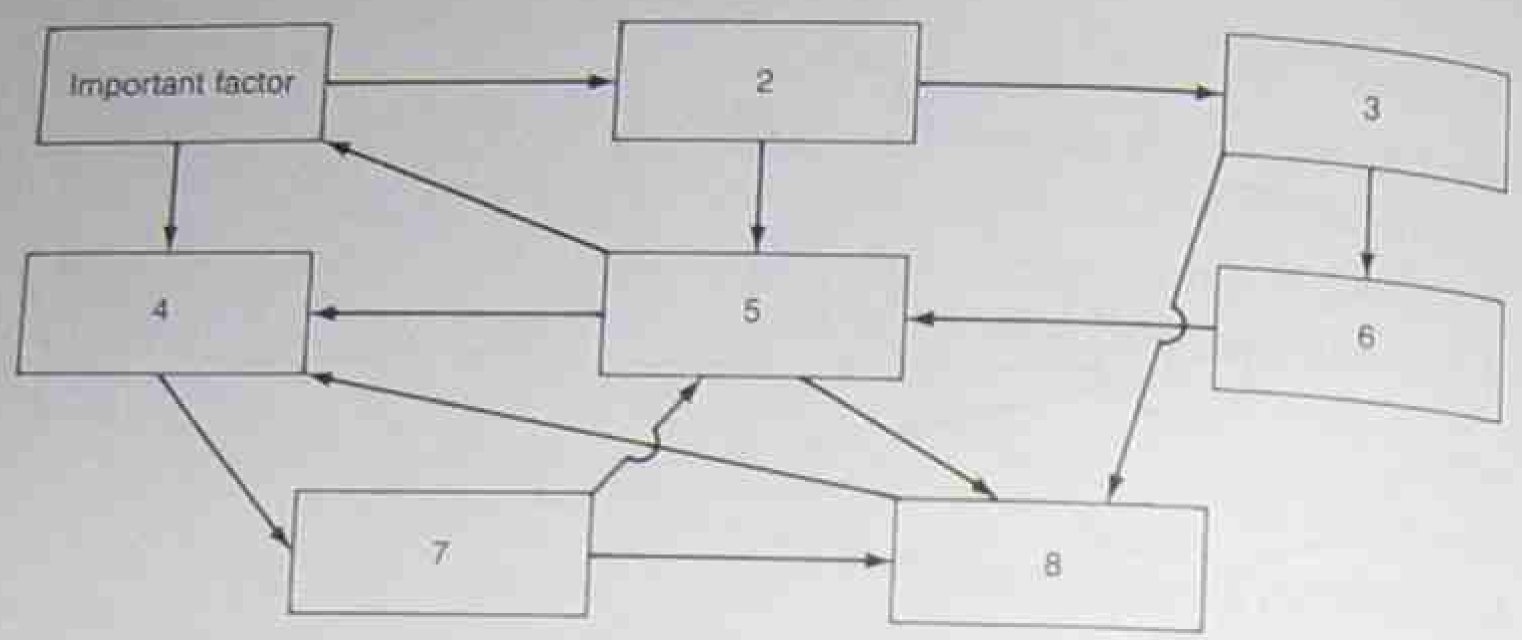


FIGURE 4.4
Relationship indication diagram

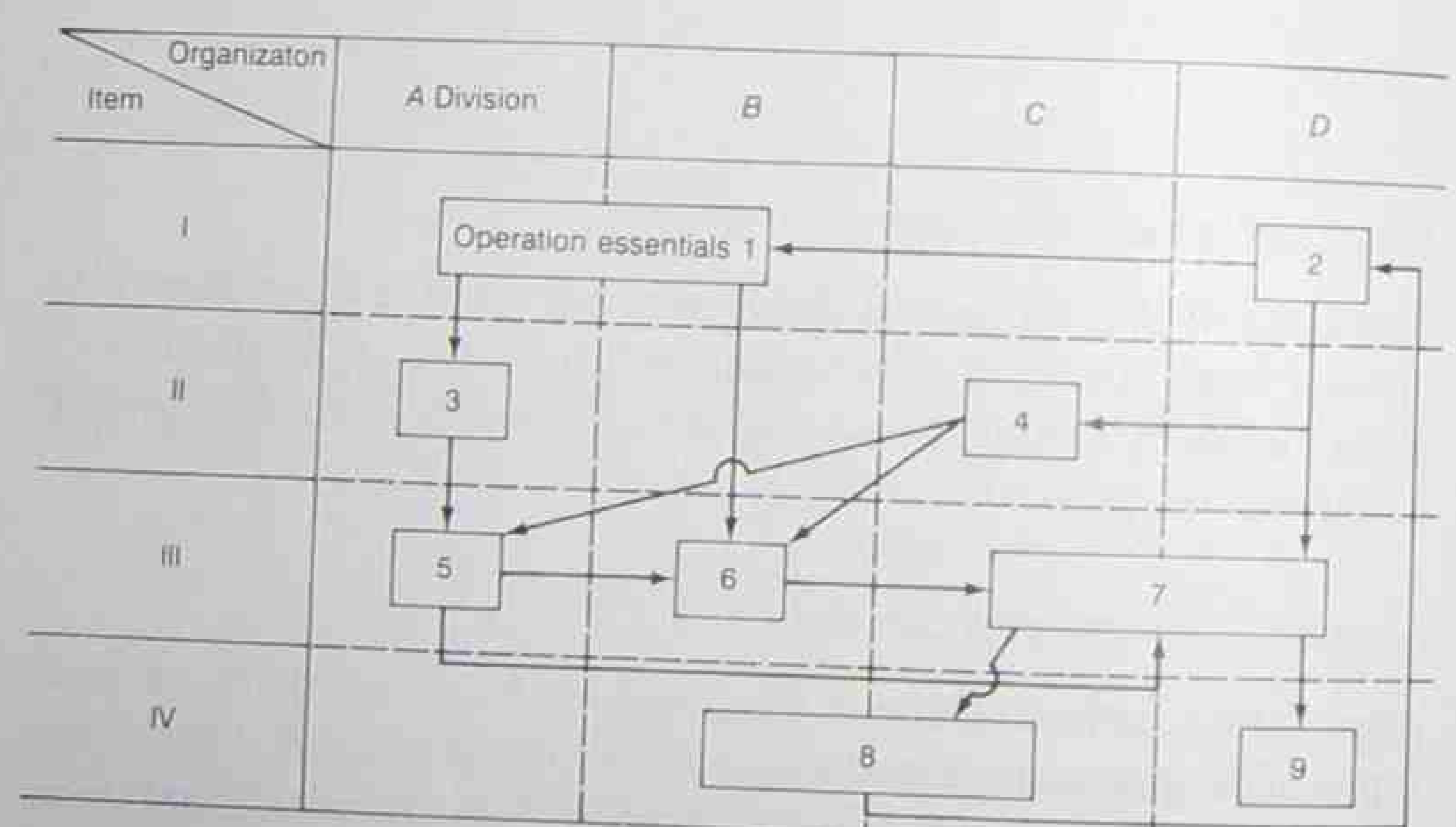


FIGURE 4.5
Applied relations diagram

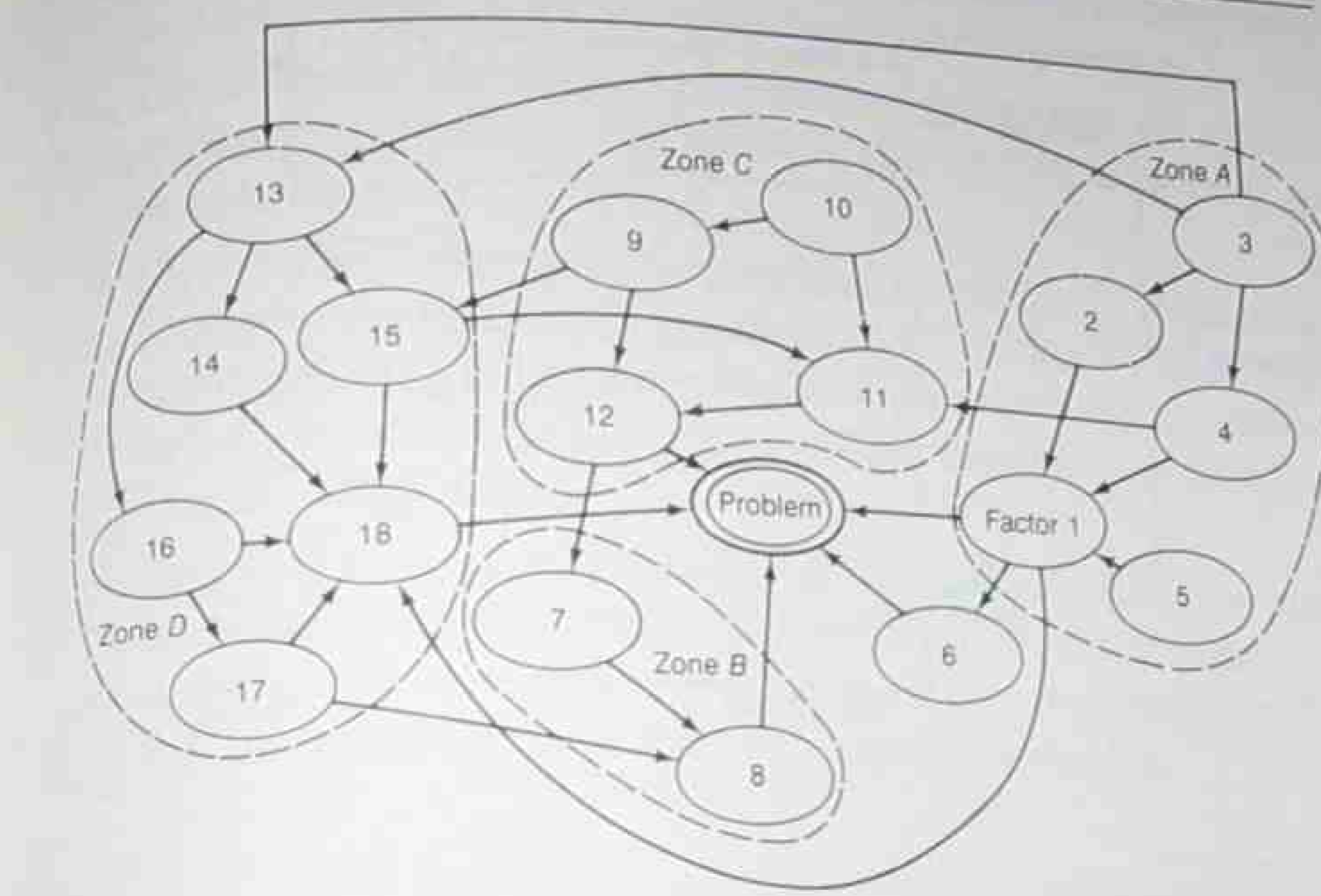


FIGURE 4.6
Applied relations diagram with collected factors

Of these diagrams, centrally converging, directionally intensive, and applied relations diagrams with systematic development are more often used in single-objective formats, whereas relationship indication and applied relations diagrams, both with organizational structure and with collected factors, are directed toward multiple-objective formats.

Relations diagram application process

As stated in the first section of this chapter, the relations diagram is a problem-solving technique. A problem is not solved simply by constructing a relations diagram, but the skillful use of such diagrams is an integral aspect of problem solving.

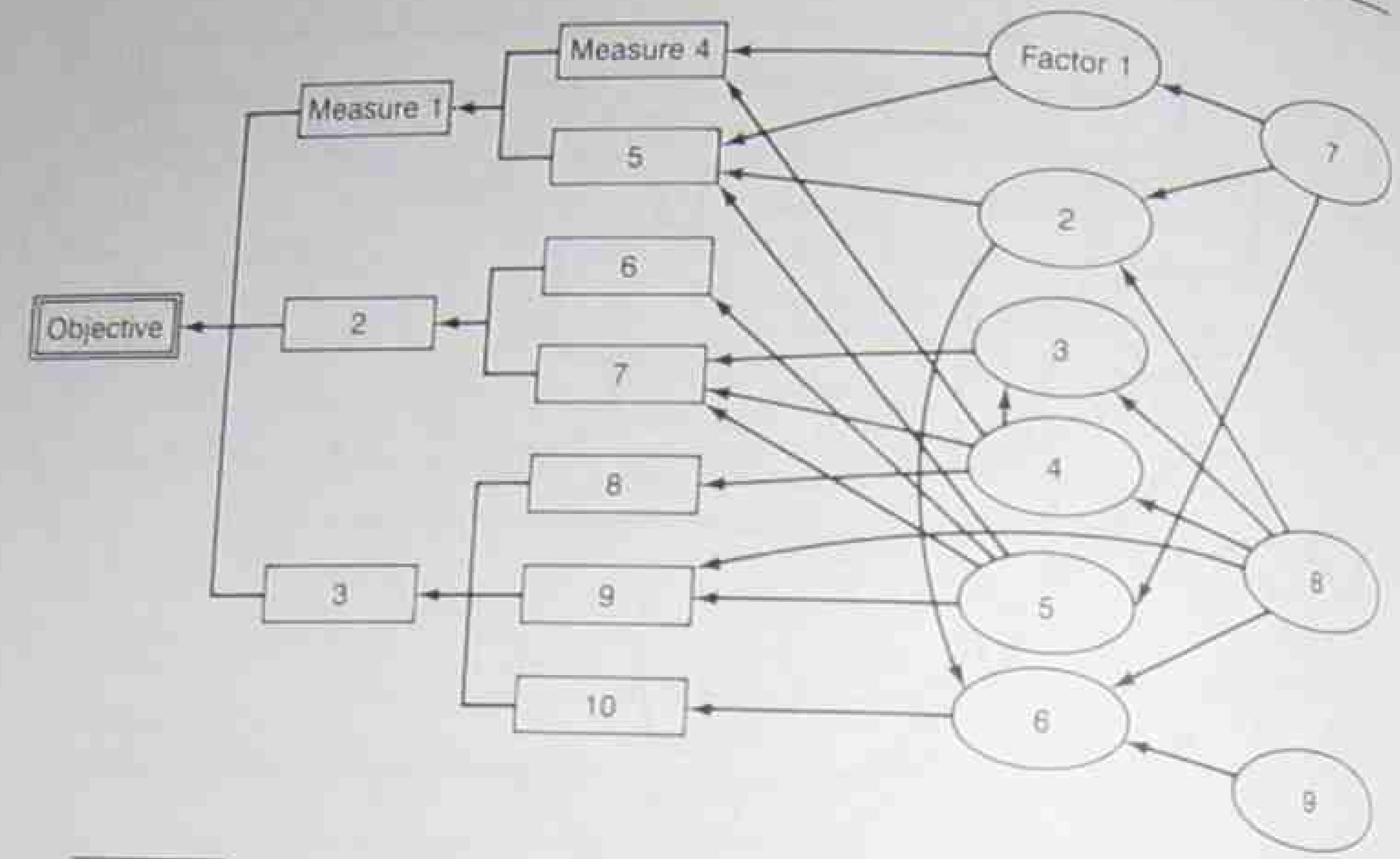


FIGURE 4.7
Applied relations diagram with systematic development

Multiple-objective format

Although methods of application vary significantly depending on the issue involved, multiple-objective models are generally applied in the following step-by-step manner:

1. *Formation of teams in response to problems.* In the case of policy development, for example, a team may be composed of department managers and their staffs.
2. *Close examination of factors and major items through team meetings.* Brainstorming or similar processes are used to allow team members to express their opinions on the measures and policies required to achieve an objective.These, in turn, are concisely entered as presented on a blackboard or on paper (cards also may be used).
3. *Creating the relations diagram.* After the ideas have been exhausted, arrows are drawn in the diagram to indicate the cause-and-effect relationships among the various items. (It is best to combine similar items and organize the diagram for clarity.)

4. *Diagram revision.* The diagram is comprehensively reviewed, and additions and revisions are made where necessary.
5. *Extracting the key items.* Key items in the process are clearly marked by shading or by marking the borders with thick or double lines.
6. *Planning concrete actions concerning key items.* If authorization for action on the key items (step 5) is given, the respective department managers begin preparing their plans for implementation of these actions. (Further stages are best developed using a systematic diagram.)
7. *Review of the relations diagram.* It is very important for the relations diagram to be reviewed in response to any changes. Additions, revisions, and rewritings must be completed in order for this technique to be used to its optimal potential.

Single-objective format

Single-objective relations diagrams are applied in such cases as problems involving decreasing product nonconformities, devising claims-prevention policies, and reducing schedules, among others. This type of relations diagram is usually drawn up either by a team or by a single staff member. As a rule, teamwork is preferable. Even in the case of individual development, the single staff member will not arbitrarily create a diagram, but rather will gather opinions from concerned personnel and organize these in the relations diagram, keeping the expressions as close to the original as possible. Then the staff member determines the actions to be taken. Since the process involved is basically the same as that for a team, the step-by-step application procedure discussed here will be that of a team approach:

1. *Team formation.* The best team is composed of four to five persons concerned with the problem (for example, the line manager, section chief, supervisor, and staff members).
2. *Preparatory meetings.* In a preparatory meeting, members decide what the problem is, how the relations diagram is to be applied, and how to conduct the approach. This step can be omitted if it is not necessary.

3. *First meeting.* Meetings should be held once or twice a week for approximately two hours. At the first meeting

   1. Define and clarify the problem.
   2. Let everyone talk freely about possible causes of the problem. Even excuses and factors based on guesswork should be welcomed.
   3. On the blackboard or a large sheet of paper write down every factor mentioned by the members and circle them. Use a bold circle for the problem definition.
   4. Study the relations between the problem and the factors, asking why over and over to reach root causes. Use arrows to connect factors with the problem. Add other factors as they come up along the way.
   5. When the chart is completed, decide what everyone has to do before the next meeting before closing.

      a. The group leader checks the chart, makes copies and distributes them to the members.
      b. Every member should review the chart.
      c. Follow up on guessed factors to see if they are realistic.
      d. Use data to confirm.
      e. Investigate the points members noticed during the meeting, and if you come across possible solutions, test them and bring the results to the next meeting.

4. *Second meeting.* Everyone brings their assignment from the first meeting. At this meeting make necessary corrections or additions to the chart and rewrite it. When it is completed, from this chart, identify: (a) those items for which data are available and those for which data are not available; (b) those items for which data are required; (c) those items on which action should be taken and those items on which action is not required; (d) those items on which action *must* be taken; and (e) those factors which do not require further consideration. Such factors should be so marked (e.g., by using: o = most important; Δ = less important; x = unimportant). At this stage, a consensus should be reached among the members. Each member should be encouraged to disregard any preconceptions in order to discern the key patterns from either ideas presented or the total picture.

Each should attempt to grasp the specifics necessary for problem solution. Then preparations for the third meeting are the same as for the first meeting.

5. *Third and subsequent meetings.* In the third and subsequent meetings, the following points should be covered: (a) the results of actions based on the relations diagram drawn up during the second meeting should be examined; (b) the relations diagram should be revised to reflect changes in the situation or environment; (c) the key factors brought out during the second meeting should be analyzed further using the relations diagram. Further meetings should be held as necessary, and the results of all actions should be reviewed.

Points on creating and using relations diagrams

The following principles have been compiled from the numerous case studies: First, collect information from a wide range of sources. In this process it is important to talk directly with operators at the site and to speak freely during brainstorming sessions. Second, in constructing the diagram keep written comments as close to the original as possible. Concise phrases or sentences are preferable to isolated words. Third, draw up diagrams only after a consensus has been reached among group members. Someone outside the group may not fully understand the process, and problems may arise should differences of opinion be apparent in the resulting diagram. Fourth, make an effort to rewrite diagrams two or three times. It is essential that this revision process be carried out so that key items can be isolated. And finally, don't be distracted by intermediate factors, use them only as a means to track down fundamental factors.

When working with a single-objective model, the following principles are useful: First, using the question "Why?" as the key to establishing cause-and-effect relationships, makes construction easier. Second, using the question "Why?" is also effective should the process slow down. Third, follow up on any actions to be taken and examine key factors that come to light during the creation or revision process promptly. Lastly, determine whether or not factual relationships that appear to return to the same point can be cut at some point in the cycle.

Relations diagram method case studies

Multiple-objective format

Thorough policy management

A relations diagram was used at the S Company to develop the main items to be implemented by each division in accordance with the plant manager's plan. In the past at the S Company, each division established its main objectives on the basis of its interpretation of the plant manager's plans. However, it was difficult to confirm whether these objectives were appropriate or sufficiently developed. This time, however, through the use of a relations diagram, as shown in Fig. 4.8, related items were organized, key points were extracted, and a much more accurate plan was implemented.

Figure 4.9 shows part of the relations diagram used during this process. The main points extracted were the creation of a quality chart and the integration of products. A structure that was able to respond to new user demands was created.

TQC advancement

In promoting total quality control, personnel at all levels within the company must work together and this policy must be given top priority. Although obvious, quality, cost, personnel, and other factors are all related in an extremely complex manner. Because they are not restrained in terms of format, relations diagram use by various businesses, such as the one presented in Fig. 4.10, will probably be beneficial to TQC advancement.

Administrative operations improvement

Administrative operations involve more complicated social factors and human relationships than do technical operations. Accordingly,

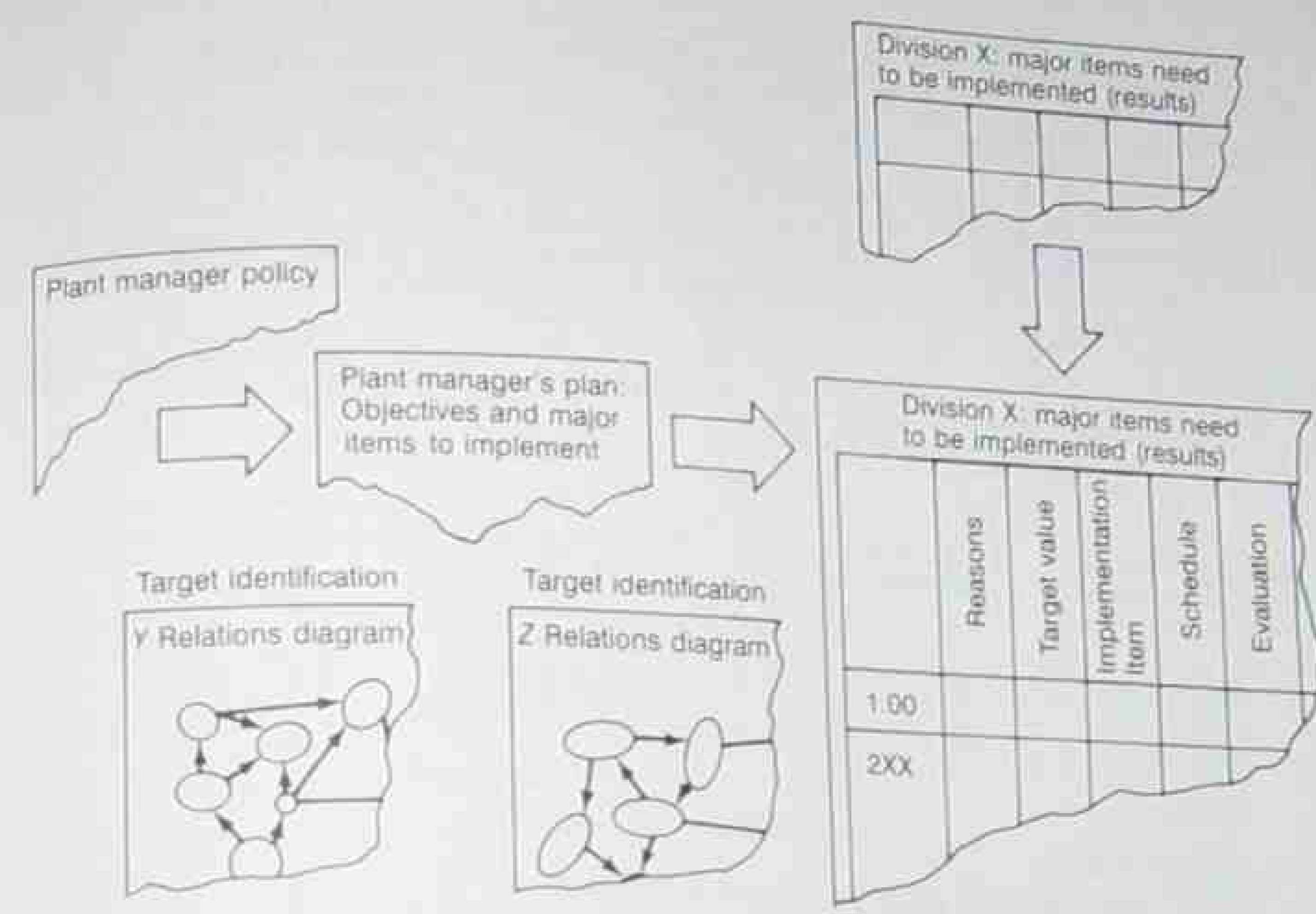


FIGURE 4.8
Policy management development and a relations diagram (partial)
Source: From *Shin QC Nanatsu Dogu Katsuyo Jirei Happyokai Tekisuto* (Case Studies on the Application of the Seven New QC Tools) (Tokyo: JUSE Press, Ltd., 1978).

it is important that the relations diagram be used to identify problems and extract key factors. In the relations diagram in Figure 4.11 the systematic connections can be readily identified.

Introduction of new production methods

In introducing new production techniques, staff and related departments, working with the support of the upper management, should prepare a careful implementation plan and seek uniform cooperation. Figure 4.12 provides a relations diagram used in the analysis subsequent to introduction of the just-in-time method, which is part of the Toyota production system. A superficial introduction of the

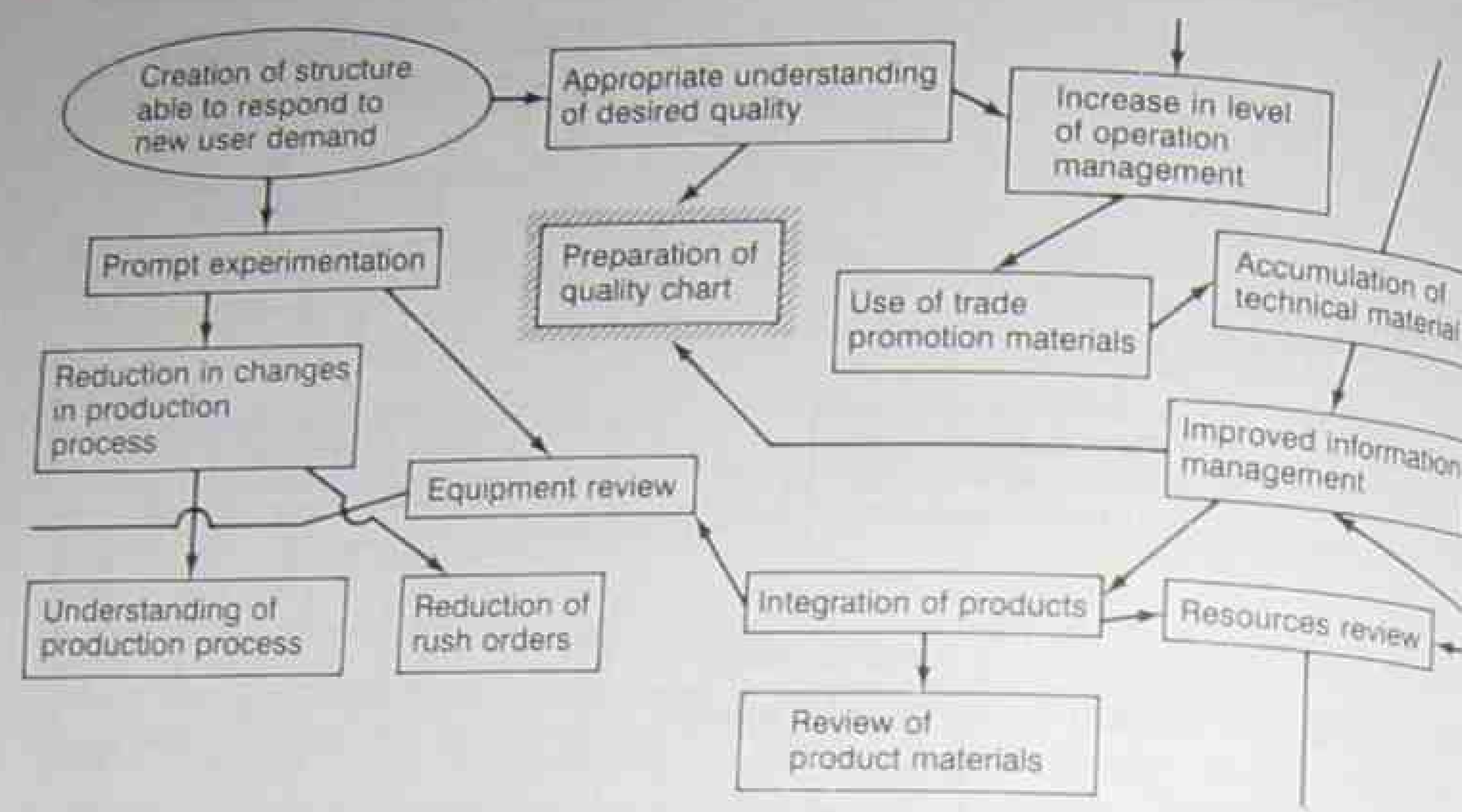


FIGURE 4.9
Policy management relations diagram (partial)

Source: From *Shin QC Nanatsu Dogu Katsuyo Jirei Happyokai Tekisuto (Case Studies on the Application of the Seven New QC Tools)* (Tokyo: JUSE Press, Ltd., 1978).

"just-in-time" method will not be effective. The analysis shows that in order to implement this plan, such fundamental aspects as thorough understanding of the concept throughout the company, standardization of production, and the creation of a supplier-management system must be addressed.

Single-objective format

Claim-prevention measures

Unexpected errors are often the cause of liability claims. Errors in labeling represent one form of such claims. Care must be taken, especially in terms of the human factor, to avoid errors in operating

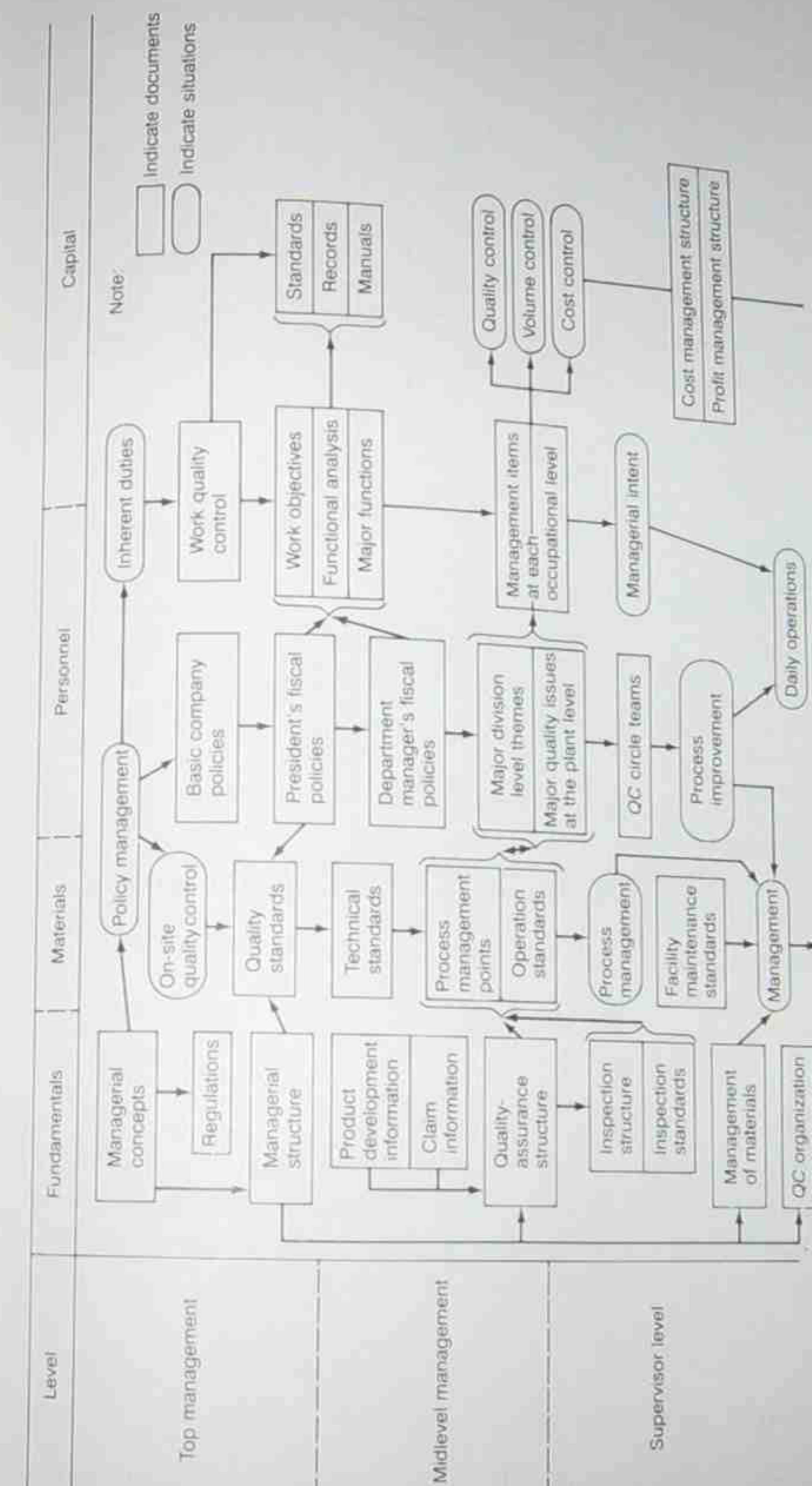


FIGURE 4.10
TQC implementation of the relations diagram (partial)

Source: From Nagatani Shotaro and Shinohara Ko, "Zenshateki Hinshitsu Kanri no Donyu to Suishin (Introduction and Development of Total Company Quality Control)," *Quality Control*, vol. 28 (June 1977).

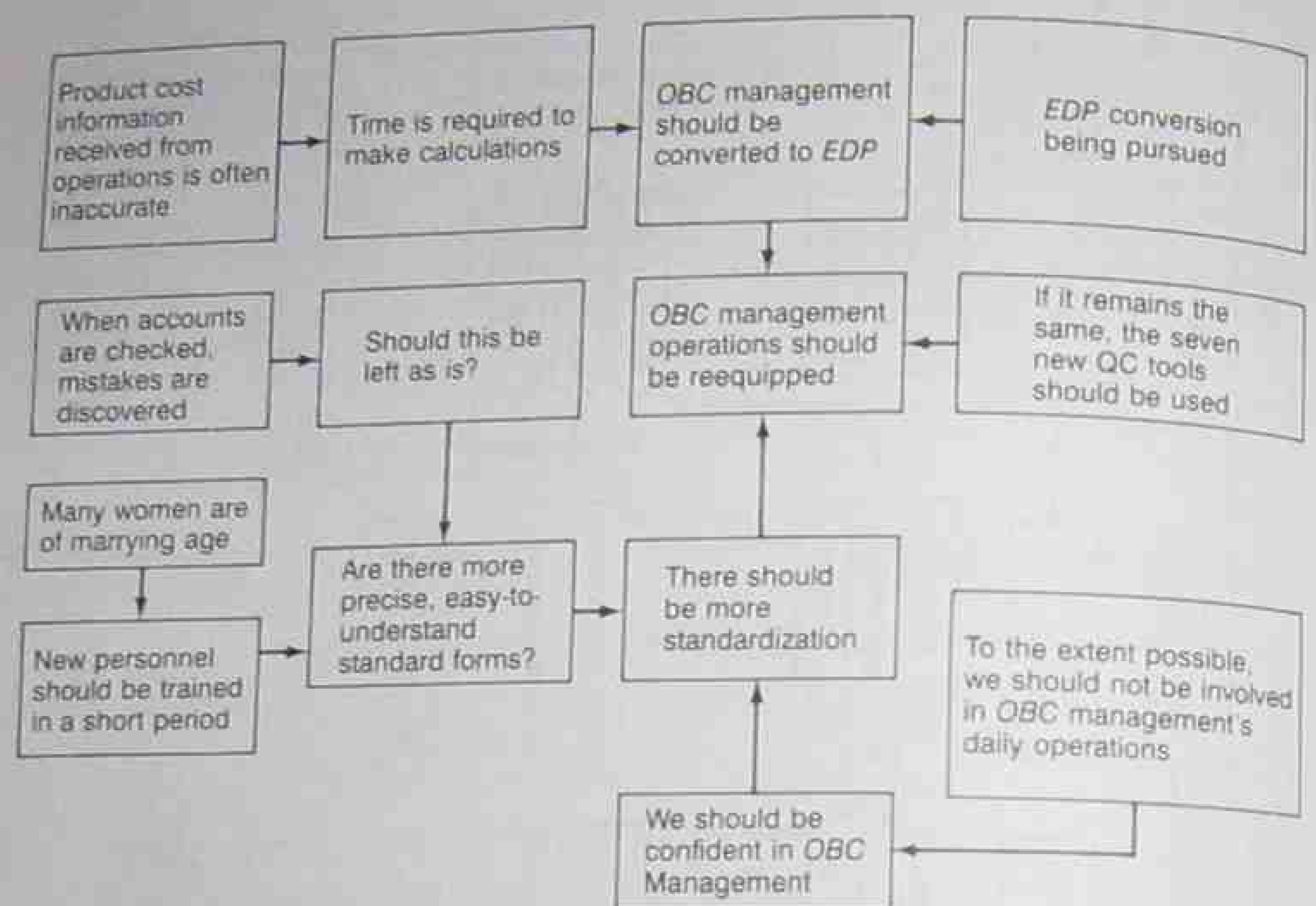


FIGURE 4.11
A relations diagram used in improving operations within the administrative department

Source: From *Shin QC Nanatsu Dogu Katsuyo Jirei Happyokai Tekisuto (Case Studies on the Applications of The Seven New QC Tools)* (Tokyo: JUSE Press, Ltd., 1978).

instructions as well as errors in judgment on the part of operators. In such cases, relations diagrams can be extremely useful in establishing clear policy, as shown in Fig. 4.13. Such diagrams also have the advantage of listing the key items in a short note in the bottom right-hand corner.

Decreasing production-line nonconformities

At Company U, the seven tools for quality control had been used in an attempt to solve the problem of chronic nonconformities in the

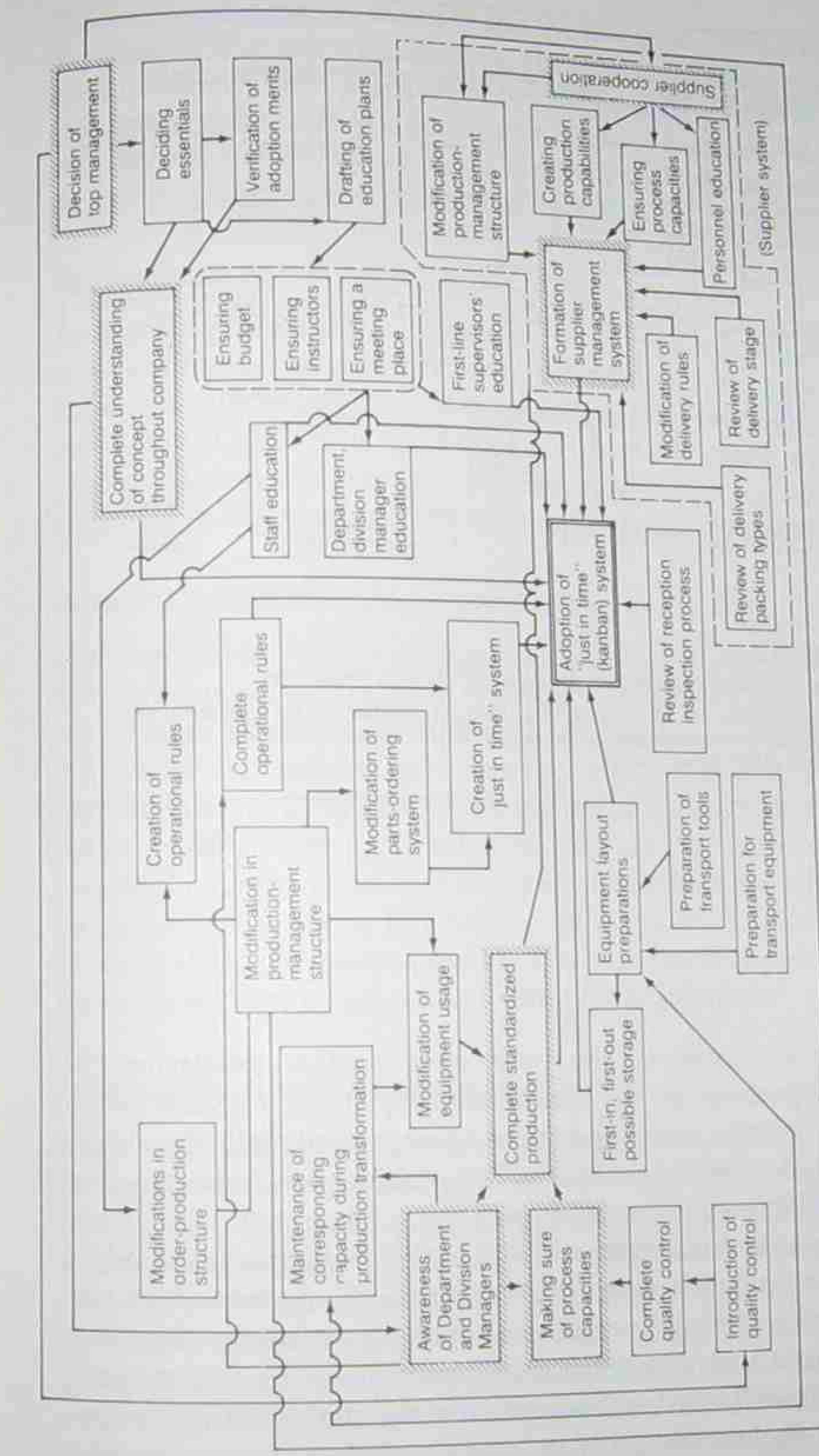


FIGURE 4.12
A relations diagram used in the introduction of new production methods

Source: From *Shin QC Nanatsu Dogu Katsuyo Jirei Happyokai Mondaishu (A Workbook on Case Studies on the Application of The Seven New QC Tools)* (Tokyo: JUSE Press, Ltd., 1978).

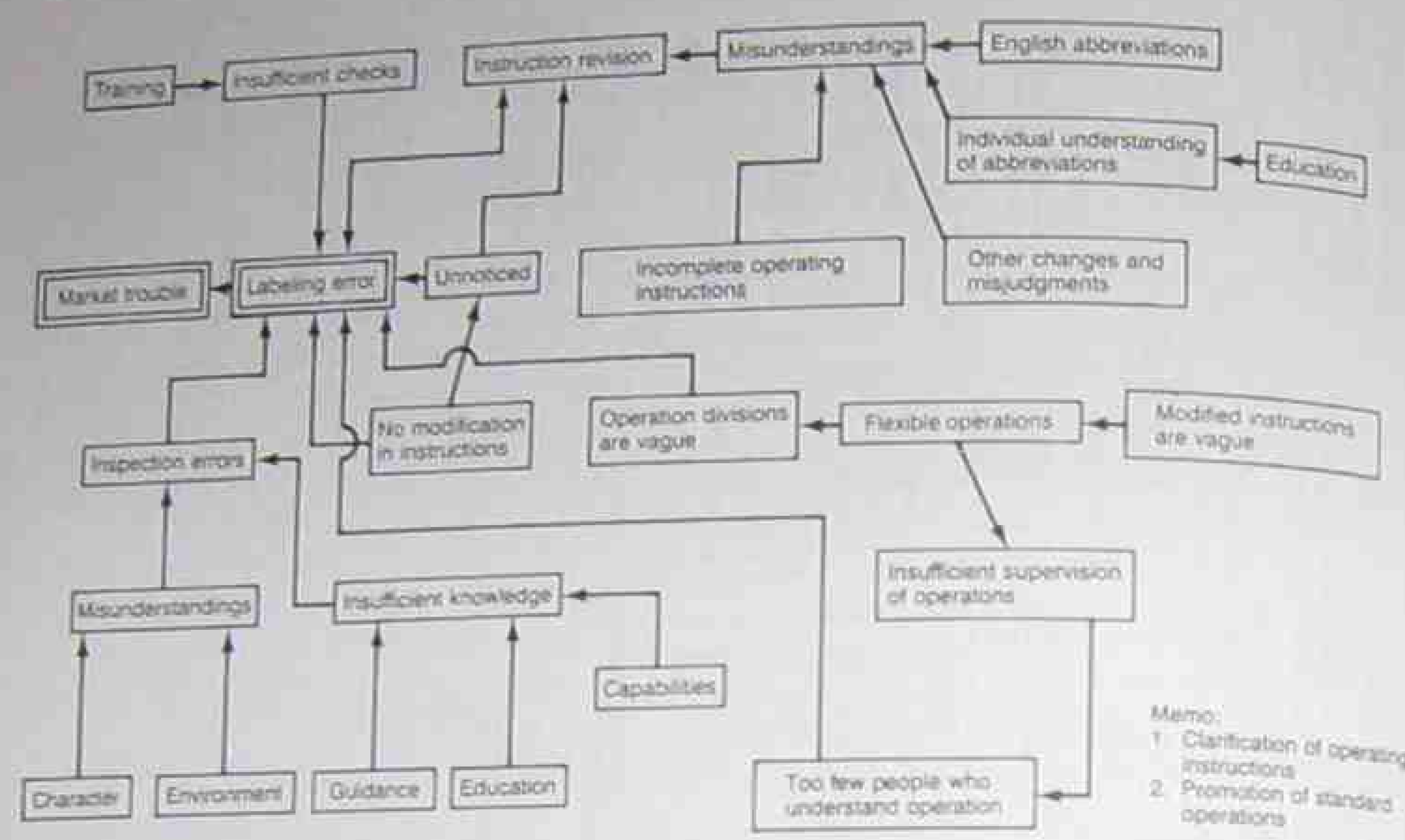


FIGURE 4.13
Use of a relations diagram in claim-prevention measures
Source: From *Shin QC Nanatsu Dogu Katsuyo Jirei Happyokai Mondaishu (A Workbook on Case Studies on the Application of the Seven New QC Tools)* (Tokyo: JUSE Press., 1978).

assembly line; however, the results were less than satisfactory. Behind this problem lay the interrelationships between such various factors as human issues, operations-management methods, and the methods of information sharing. The problem was reexamined using the relations diagram method.

Line-section managers and staff interviewed operators and collected their opinions about the reasons for the occurrence of nonconformities. During the process of creating a relations diagram, it was realized that nonconformities were caused by factors totally different from the earlier conclusion that the quality of the operators was low. Rather, the problem lay in such factors as incomplete

operations-management methods and the lack of an optimal equipment layout. With this understanding, a factor analysis was conducted. Appropriate corrective measures were taken, and the number of nonconformities per month was greatly decreased. Furthermore, the operators' morale was greatly enhanced. The third and final diagram used during this process is shown earlier in Fig. 2.3.

Reducing variations in product characteristics

An analysis process at the S Company was conducted to determine measures for reducing variations in the hardness of the product *A*. Various factors were arranged in a relations diagram similar to the one in Fig. 4.14 in order to devise a method of dealing with this problem. This type of relations diagram is best suited for situations where there are complex interrelationships among various factors.

Reducing receipt inspection nonconformities

The responsibility for nonconformities evident at the time of receipt is usually not placed on those receiving an order, but is more often the responsibility of the sender or the partial responsibility of both the sender and the receiver. However, the factors in this type of problem frequently have complex interrelationships and often cannot be readily resolved. A relations diagram may be an effective means of expressing the cause-and-effect relationships among the factors and of isolating key factors. Although the single-objective relations diagram could be used in this case, the multiple-objective format is perhaps better suited to this type of situation.

Improving the pace in the assembly process

A slow assembly process can be caused by material, operators, equipment, and other factors, all of which are intertwined in a complex relationship. This makes the production process extremely difficult to control, especially in the case of individual items. Figure 4.15 presents a relations diagram (the third) created to analyze the reasons for slow assembly processes at the Company M's electric motor assembly plant. During preparation of this diagram, many of the points were changed, and it was found that in order to make improvements in the assembly process, it was necessary to conduct

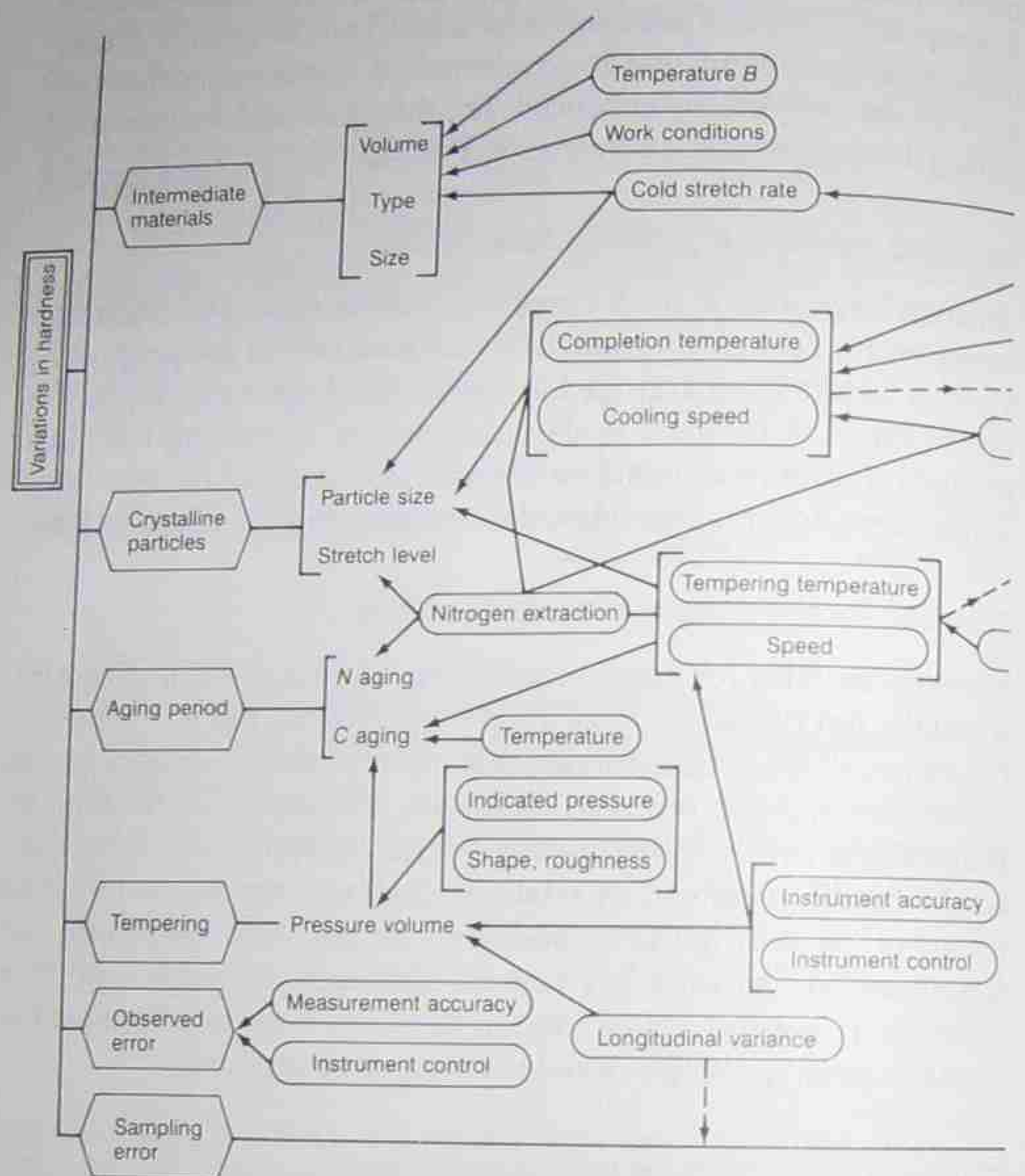


FIGURE 4.14
Relations diagram concerning variations in product characteristics

operations roughly according to a daily planning schedule. In this particular case, the relational analysis method also was used.[2]

Conclusion

This chapter presented the definition, application methods, and actual case studies of the relations diagram method. The relations diagram, in short, is the only effective QC technique for cases involving complex interrelationships. A problem cannot be solved simply by drawing up a relations diagram; nonetheless, through the revision process, which is usually carried out several times, the measures necessary to solve a problem are clarified, and in this sense, the relations diagram is truly effective.

Because the format of the relations diagram is unrestricted, it also can be used in QC circles. However, because it is a technique for use in problem solving involving large-scale issues having numerous factor interrelationships, it is best applied by managerial staff.

Here we would like to touch on relational analysis. This process is formally referred to as "management indicators relational analysis" and was developed by Professor Toshio Senju at Keio Gijuku University for application in industrial relations analysis.[3] In this analysis, problems having linked cause-and-effect relationships are numerically analyzed. Disadvantages of the method are that two sets of data are required for each factor; furthermore, the calculation process is time consuming. Nonetheless, this technique is sure to see future applications.

The relations diagram method is still a relatively new QC technique. Accordingly, it remains a technique with potential for expansion into new fields of application. Various problems that may have seemed too complex to tackle can now be challenged by teams of managers and staff using this method.

In conclusion, we would like to thank Professor Senju Toshio of Keio Gijuku University for his guidance and advice in the development of the relations diagram method. We also thank those who helped in the preparation of these diagrams.

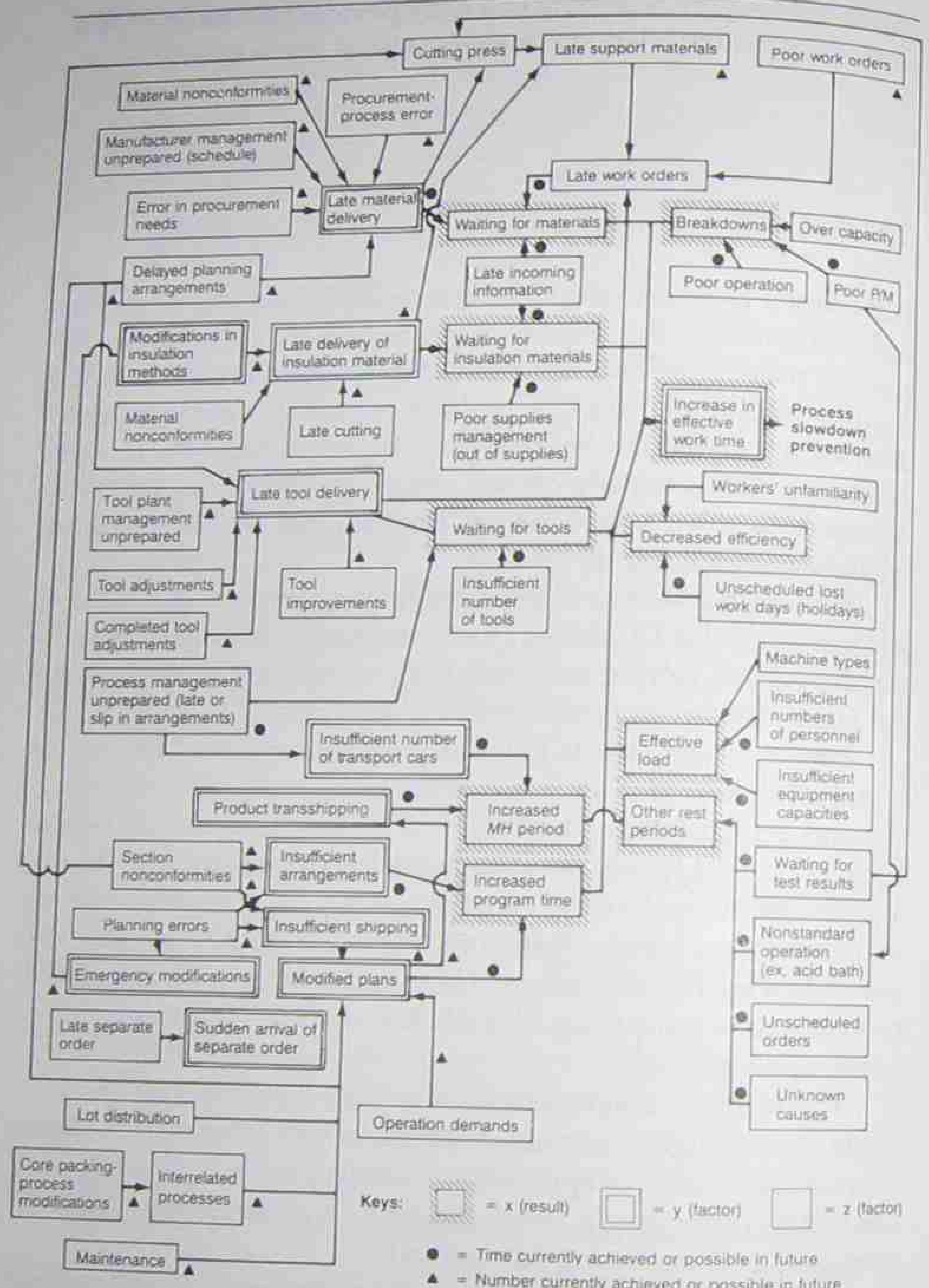


FIGURE 4.15
Relations diagram for a slow assembly process

Notes

1. Senju Toshio and Mizuno Yukiichi, *Hinshitsu Kanri no tame no Keizai Keisan* (Economic planning for quality control) (Tokyo: JUSE Press, Ltd., 1971); Senju Toshio and Fushimi Michio, *Keizaisei Kogaku* (Economic engineering) (Tokyo: Japan Management Association, 1967).
2. Yagishi Seichi and Kurabayashi Mikihiko, "Kanri Shihyokan no Renkan Bunseki Shuho" (Management indicator relational analysis), *Quality Control*, vol. 26 (May 1975), special issue, pp. 213–217.
3. Senju and Mizuno, *Economic Planning*, 1971; Senju and Fushimi, *Economic Engineering*, 1967.

Other references

Nagao Seigo, *Sangyo Renkan Bunseki Nyumon* (Introduction to industrial relational analysis) (Tokyo: JUSE Press, Ltd., 1970).

Nagao Seigo et al., "Relational Analysis," *Operations Research*, vol. 17 (September–December 1972).

Nagatani Shotaro and Shinohara Ko, "Zensha teki Hinshitsu no Donyu to Suishin" (Companywide quality control: introduction and advancement), *Quality Control*, vol. 28 (June 1977), special issue, pp. 5–9.

Fujita Shinji, *Kanri Gijutsu Koza* (Lectures on management methods) (Tokyo: Japan Standards Association, 1975).

Kumatani Tomoe, "Toyota Seisan Hoshiki no Tokushitsu" (Special characteristics of the Toyota production system), *Kojo Kanri* (Factory Management), vol. 24, no. 13 (1978), p. 155.

5

The KJ Method: Affinity Diagram*

The KJ method clarifies important but unresolved problems by collecting verbal data from disorganized and confused situations and analyzing that data by mutual affinity.

The KJ method and the seven new QC tools

The KJ method was developed and popularized by Kawakita Jiro. The principal tool in the KJ method is an affinity diagram. The seven new QC tools help solve problems not only through the repeated application of the affinity diagram, but also through the combined use of other diagrams. Table 2.1 shows the differences between the KJ method and the statistical method; Fig. 2.4 shows the similarities between the two methods.

*Permission to use material on the KJ method in this chapter granted by Kawakita Jiro, Kawakita Research Institute.

Description of the affinity diagram of the KJ method

The affinity diagram of the KJ method is designed to collect facts, opinions, and ideas about unknown and unexplored areas which are furthermore in a completely disorganized state. Like the first two stars in the sky at sunset, the data arrange themselves according to mutual affinity. Then the data areas are reduced to narrative rather than quantitative form.

Uses of the affinity diagram

Recognizing facts

Information about unknown and unexplored fields is difficult to obtain. It is important to collect facts about a field, one by one, to learn how it is organized. When we collect data in such a situation, it is important to do so with a completely open mind. Otherwise, we can make mistakes by interpreting these data in terms of previously formed opinions and hypotheses. This often occurs when we try to obtain information about an unknown. It also occurs when we try to understand consumer reactions or the reactions of manufacturers and retailers to highly unique products or to measure the effectiveness of a new factory or an existing factory operated by a new manager.

Forming ideas

When we have almost no information about a new field, we start from scratch, hoping to form our own ideas and concepts. In such a case, it is necessary for us to collect information about the target field, examine the opinions of other people, and write down our own opinions and ideas as they come to us. By organizing these data in an affinity diagram, we can build our own system of thought. We can use the same method when we are assigned to a new position and want to determine how to carry out the new assignment.

Breaking away from old approaches

We may face some difficulty in performing our duties if we are victims of our past experience. In a situation such as this, we must overcome the difficulty and formulate new ideas. Here, as in the previous example, the ultimate goal is also the formulation of new ideas, but unlike the previous situation, we must start by destroying the status quo. Old ideas and their systems are broken down into unstructured bits out of which we build a new way of thinking using the affinity diagram.

Adaptation

Any subject matter has a system of thought and a theoretical base. It is often useful to comprehend and then build upon that base by adding improvements and refinements. This process is analogous to destroying the status quo. First, we read books and papers and transfer pieces of information from these sources to cards; then we build a new concept by arranging the cards in an affinity diagram. In this way, we adapt the opinions and ideas of others and use them to build our own theoretical foundation.

Organizing a planning team

A diverse group of people brought together for the purpose of planning has no particular organization at first. Such people must come to an understanding and organize themselves into a team designed to take part in planning. Group members share experiences and exchange opinions in a brainstorming session, record these exchanges on cards, and individually arrange their thoughts into an affinity diagram. Members explain their thoughts using the diagram and hear the thoughts of other members. This is a way of promoting mutual understanding and teamwork. This method can be used to help form project teams for the promotion of total quality control or QA project teams, as well as to revitalize existing teams. It also can be used to create and revitalize teams in the workplace and in QC circles.

Thorough communication of management policies

Management cannot succeed in making employees understand its ideas and policies thoroughly through one-sided communication imposed from the top down. Management must be receptive to dialogue so that employees can understand management policy completely.

Management should participate in meetings with employees and hold brainstorming sessions with them on target subjects. Management personnel should allow employees to speak out freely about management policy, and the whole process should be recorded. Members of the managerial staff should not make direct rebuttals of opposing opinions expressed by employees; instead, management should also express its ideas and clarify its policies by accommodating opposing points of view as much as possible, in accordance with the rules of brainstorming.

As illustrated in Fig. 2.5 in chapter 2, an affinity diagram is constructed with the information gathered in these brainstorming sessions. The diagram can then be used to transmit management's ideas and policies to the employees. Thus management is able to study employees' opinions by hearing them, and management's ideas and policies are thoroughly understood by employees through brainstorming and the oral presentation.

In organizing teams to participate in planning, employees should be allowed to draw affinity diagrams. In other situations, however, only management should use affinity diagrams. This approach is effective in promoting a sense of participation in the planning process by employees.

Functions of the brain in relation to affinity diagrams

Impairment of all or part of the left half of the brain results in malfunctions in the right half of the body; impairment of all or part of the right half of the brain causes the left half of the body to suffer from motor disturbances. These phenomena are known from study of various mental and physical handicaps caused by cerebral hemorrhage due to high blood pressure, automobile accidents, or injuries to the brain during war.

The brain controls motor functions as well as cognitive and other functions. The left half of the brain controls the speech function, rational thinking, logical thinking, causal thinking, and analytical and discriminating thinking. The right half controls the ability to think instantaneously, intuitively, emotionally, and synthetically. The corpus callosum, located between the right and left halves, has the function of connecting the two halves.

Ordinarily, the thoughts and behavior of an adult are controlled by the left half of the brain. The left half is logical and critical, but not creative. Therefore, it is possible to develop the function of creative thinking by stopping the work of the left half of the brain and revitalizing the right half with more energy. In recent years, through research and training, techniques have been designed to increase creativity.[1] The KJ method is a creativity-mobilization technique that has been developed for practical purposes.

Of the seven new QC tools, six depend on the left half of the brain. The affinity diagram is the only one that depends on the right half of the brain (Fig. 5.1).

A relations diagram comes closest to an affinity diagram as a technique for arranging disorganized data. However, a relations diagram depends on the left half of the brain for logical thinking (the relationships of cause and effect), whereas an affinity diagram depends on the right half of the brain, which serves the function of feeling.

Drawing an affinity diagram

An affinity diagram is created through the following steps:

1. Selecting a theme
2. Collecting narrative data
3. Transferring narrative data onto cards
4. Sorting the cards
5. Labeling the cards
6. Drawing the diagram
7. Oral or written presentation

Many people, including Kawakita Jiro, have written in detail about the steps involved in preparing an affinity diagram for the KJ method.[2] This section will briefly discuss the most crucial points.

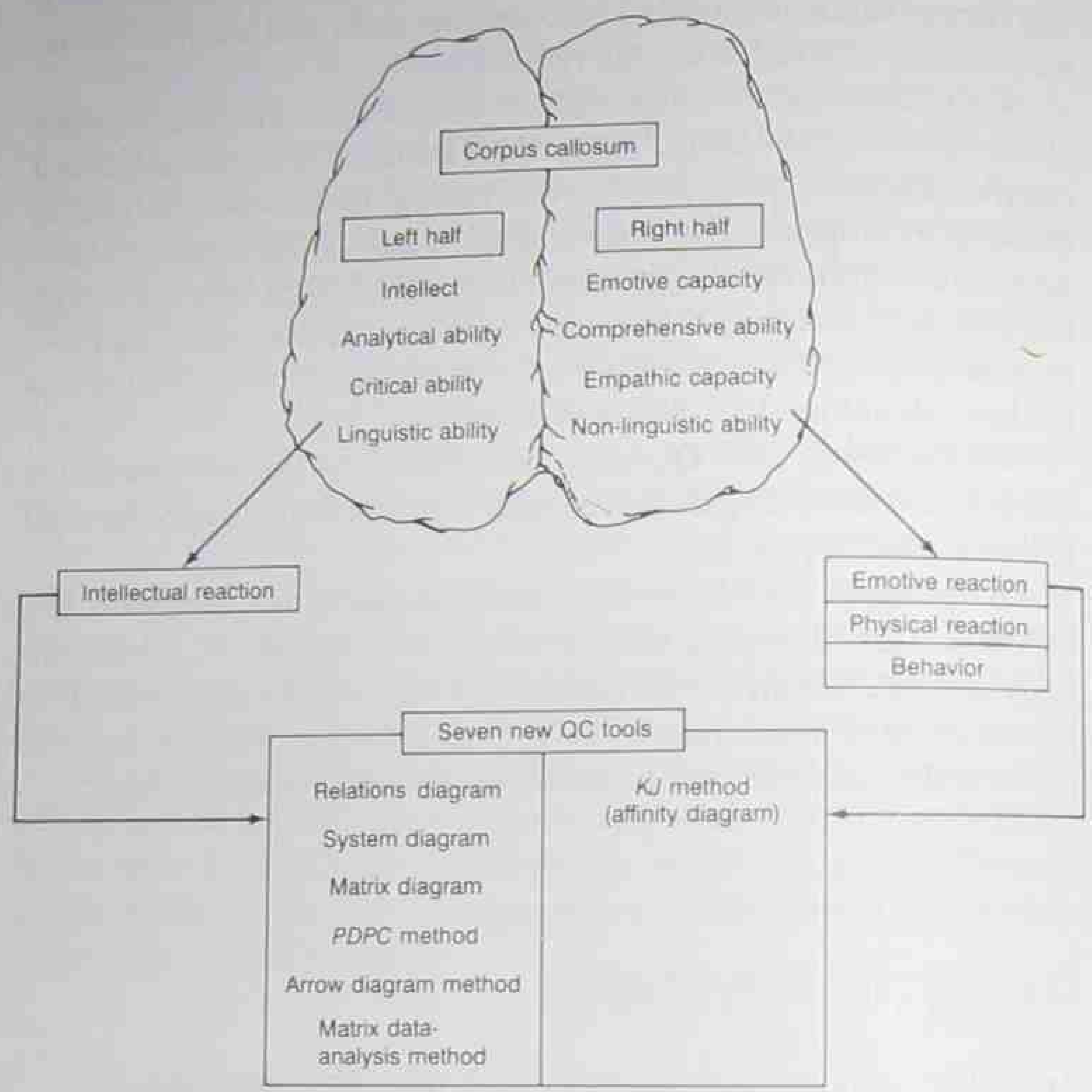


FIGURE 5.1
Function of the brain in terms of the seven new tools

Selection of methods

The seven new QC tools include six other techniques in addition to the KJ method. Accordingly, in selecting the appropriate method,

the main focus is on (1) who, (2) what, (3) when, (4) where, and (5) with what technique. From the point of view of the PDPC technique (see Chap. 9), the appropriate selection of an affinity diagram for problem solution is most important. The process of selecting the best technique for various stages of problem solving was discussed in the second section of Chap. 2 (see Fig. 2.14).

The affinity diagram, along with the relations diagram, is used at the plan 1 stage, where data are arranged and the problem is defined. It can be used in six different ways, as discussed in the third section of this chapter. Therefore, one must select the most appropriate technique for achieving the specific goal. The seven new QC tools constitute a system of diagraming and thinking. With the proper use of a diagram, a thinking pattern can be changed completely.

Management personnel and staff members, as individuals, have specific ways of thinking with both strong and weak points. Accordingly, the full use of the best technique for the job at hand is recommended. However, conscious attempts should be made to learn to use unfamiliar diagrams so that our abilities are increased.

The seven new QC tools also sum up patterns of thinking. As such, they are convenient for the development of *new* ways of thinking. When we give assignments to employees, we should be careful to bring together the right person with the appropriate technique.

Selecting a theme

Affinity diagrams are appropriate for use within the following limits:

- Facts are uncertain and hard to understand; they need to be grasped systematically.
- Thoughts are uncertain and disorganized; they need to be arranged.
- Pre-existing notions make it difficult to achieve a goal; current ideas must be eliminated and a new way of thinking must be adopted.
- The existing thought system and ideology need to be dismantled; a new system needs to be established.
- No unity exists in a group of heterogeneous people; teamwork must be promoted for mutual understanding.

- Management needs to listen to employees and clarify its ideas and policies.

As described in the preceding section, you are able to think with feeling with the help of affinity diagrams by reducing the dominance of the left half of the brain and revitalizing the function of the right half. Accordingly, use of an affinity diagram is appropriate (1) when there is a strong need for a solution of any form, (2) when an easy solution is not available, and (3) when much time is needed for solution of the fundamental problem. The affinity diagram should not be used for solving problems that require instant solutions (Fig. 5.2).

Collecting verbal data

Several methods exist for the collection of verbal data (Fig. 5.3). The following subsections discuss these methods.

Direct observation

We may do field work, making direct observations by seeing, hearing, and experiencing. Whereas quality control as "fact control" (management based on fact) emphasizes fact finding, the KJ method uses the term *external exploration* to emphasize the importance of fact finding in the field. Exploratory fact finding is important in quality control.

Interview and reference-search method

There are many methods of obtaining information, such as reviews of literature, interviews of people, and group sessions designed to elicit the opinions of many people. There are limits to the use of direct, on-the-spot, and personal fact-finding methods. Therefore, the indirect method is useful for the collection of data from various sources. Furthermore, these are the only methods of obtaining opinions and ideas of other people.

Individual thought

There are two types of thought: *recall*, in which past experiences are recalled and used as data, and *reflection*, in which one's inner

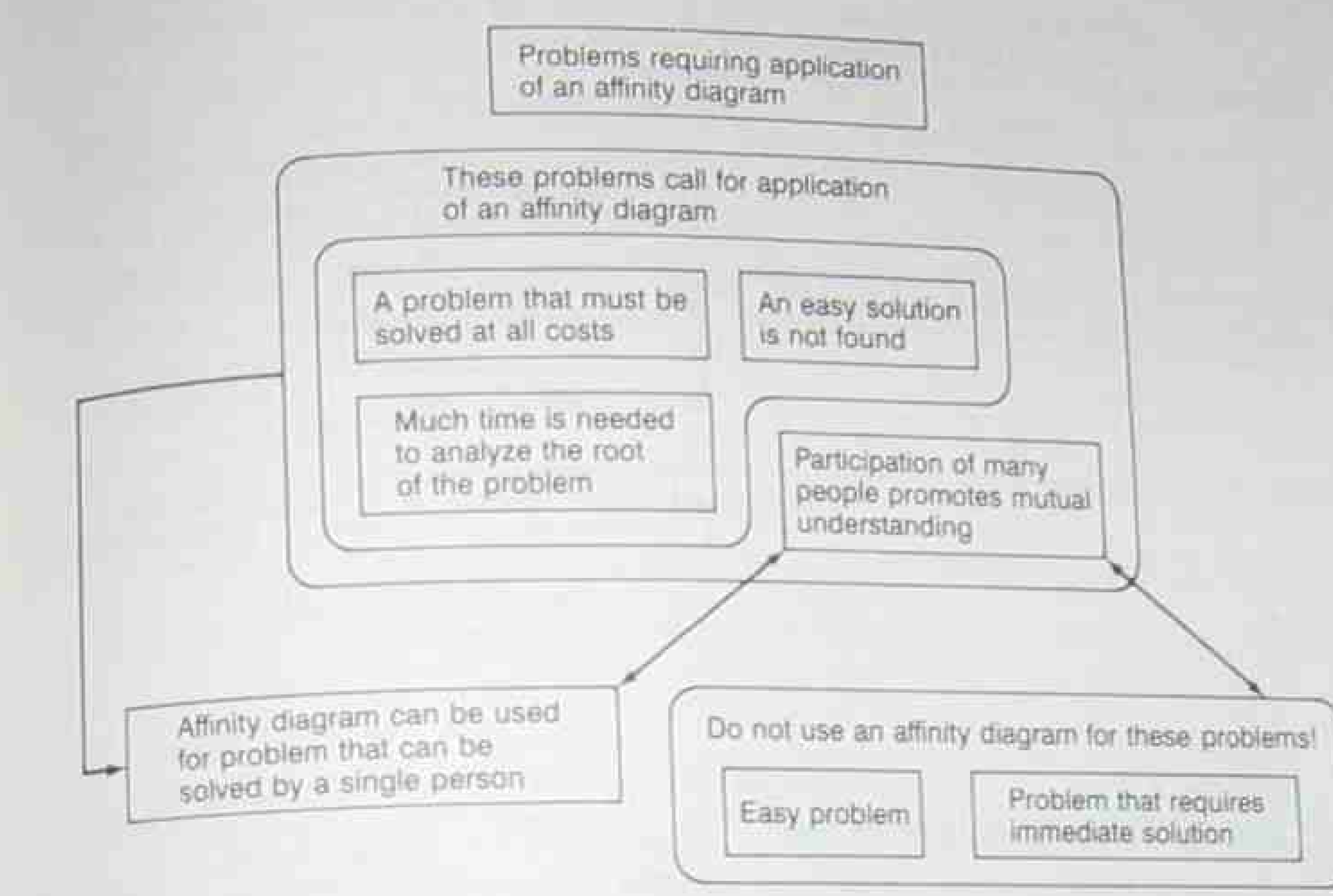


FIGURE 5.2
Problems that can use affinity diagrams

This diagram is drawn on the basis of material available in Japan Management Association (Eds.), *Introduction to the KJ Technique for Management: Creative Technique Born Out of Practical Experience.* (Tokyo: Japan Management Association, 1974). Appreciation is extended to the author.

thoughts about a theme are explored in depth. Brainstorming by a single person may make the recall method and the reflection method more effective. Kawakita calls this individual contemplation an "inner exploration."

The collection of spoken or narrative data varies with the application and purpose of the diagram (see Tables 5.1 and 5.2). The following subsections describe the various purposes of data collection.

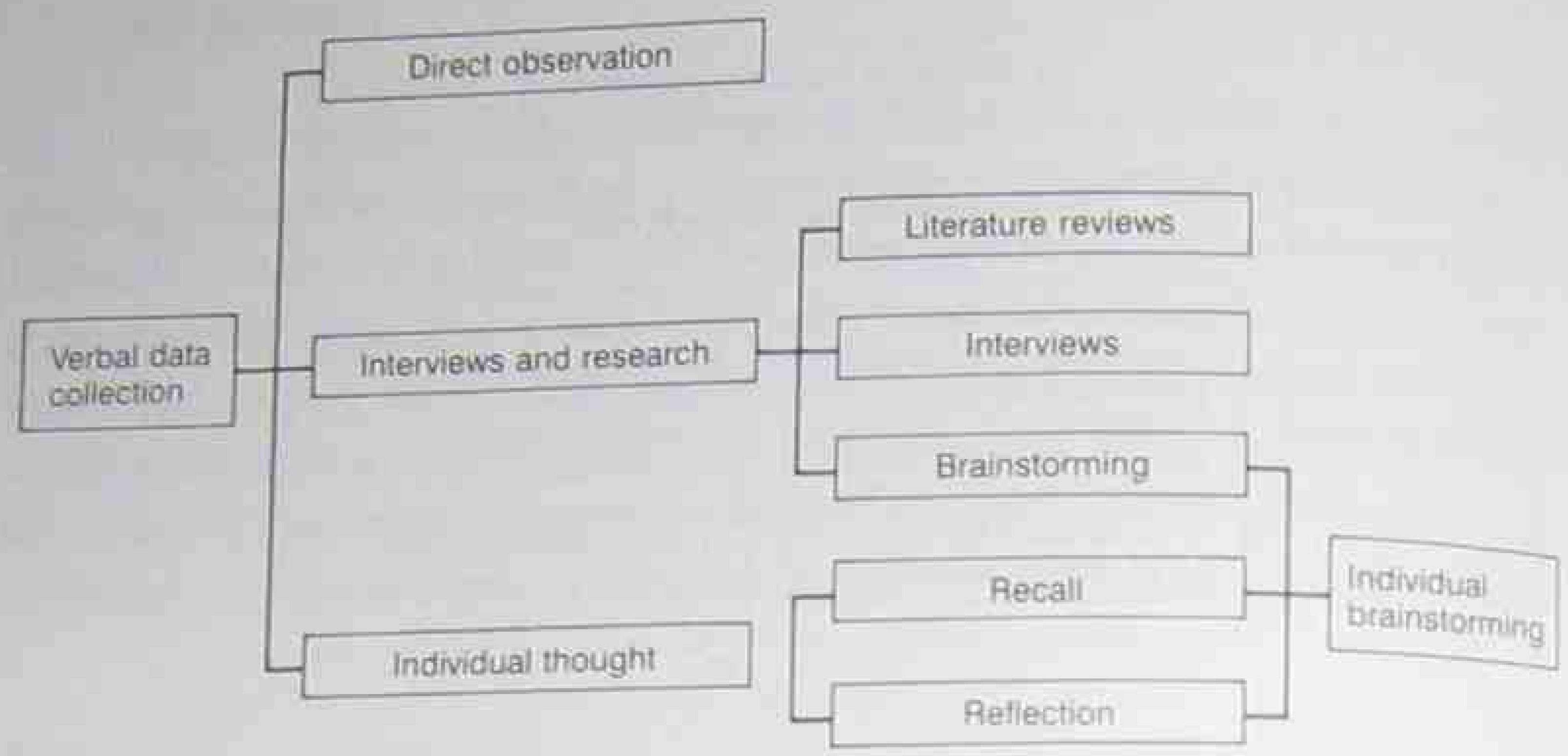


FIGURE 5.3
Methods of verbal data collection*

This chart is based on material appearing in Kawakita Jiro and Makajima Shiniichi, *Science of Problem-solving—The KJ Method* (Tokyo: Kodansha, 1970)

Recognition of facts

In this case, fact gathering is important. We attempt to prevent opinions, ideas, and preconceived notions from confusing the facts. We make direct observations in the field by using our own eyes and ears and simply by being there in person.

The indirect data-collection methods, such as literature reviews, interviews, or brainstorming, allow for the potential influence of subjective opinions and judgments on the part of informants. This method should be avoided if possible. If indirect methods are necessary, we should make every effort to collect only objective facts.

When observed facts are reduced to spoken or narrative data, they should not be stated in the abstract. Retain the original flavor if

TABLE 5.1
Selection of methods for collecting verbal data (shown by matrix)

Purpose \ Type of Data	Direct Observation	Reference Search	Interview	Brain-storming	Recall	Reflection
Recognition of facts	●	△	△	△	○	X
Thought formulation	○	○	●	○	○	●
Breakthrough	●	○	○	●	●	●
Adaptation	△	●	●	X	○	○
Organization of planning and participation	X	X	X	●	○	○
Policy implementation	X	X	X	●	○	○

● = often used ○ = used △ = seldom used X = not used

TABLE 5.2
Selection of collected data (shown by matrix)

Purpose \ Type of Data	Factual Data	Opinion Data	Thought Data
Recognition of facts	●	X	X
Thought formulation	○	●	●
Breakthrough	●	○	●
Adaptation	△	●	○
Organization of planning and participation	△	●	△
Policy implementation	△	●	○

possible. When data are collected in abstract form, they cease to be objective because of unavoidable observer bias.

Formulation of thought

If you must organize your thoughts from scratch, it is important to comprehend the facts while at the same time giving a great deal of thought to them. Useful data may be based on opinions and ideas as well as on quantifiable factual data. On some occasions, these two forms of data may be mixed. We should distinguish factual data from opinion data at the time of data collection.

Breakthrough thinking

If it is necessary to eliminate existing concepts, it is important to remember these points: (a) reexamine facts without prejudice; (b) make every effort to collect a variety of opinions; (c) the existing ideas need not be discarded, but they must be approached as data after they are analyzed; and (d) use the individual brainstorming method with all the preceding data in order to formulate new ideas (see the section on individual brainstorming below).

Organizing participative planning

In this case we collect data through brainstorming.

Transferring verbal data onto cards

Collected verbal data are broken down into individual thoughts units in the form of independent sentences with one clear, single meaning, and one sentence is recorded on one card. Abstract terms and complicated expressions should be avoided in these sentences; use ordinary everyday expressions. The use of abstract expressions makes the data almost useless at later stages.

Card grouping

The cards should be shuffled well to eliminate pre-existing order and then spread around so that every card can be easily seen. Each card should be read slowly two or more times. Cards containing similar items are grouped together and on the basis of their affinity. This sorting is done not on the basis of reason, but on the basis of feeling. The cards should not be grouped on the basis of the sorter's ideas and preconceptions or according to existing categories. The term "feeling" refers to a state that precedes logical consciousness. The cards should be grouped with the help of the *right* half of the brain.

What is desired is the "impression" that the cards group themselves. The cards are not to be gathered according to a certain classification scheme or on the basis of certain key words, because they are not supposed to be classified but simply to be grouped. (Use a

matrix diagram if they are to be classified; to connect them logically, use a systematic diagram and a relations diagram.)

After two or three cards are grouped, they are labeled. Sometimes five or six cards are collected, but this kind of collection often indicates that they did not group themselves but were classified into certain established categories. This is similar to the grouping of letters and postcards into prearranged boxes. When two cards are grouped, a procedure similar to making nameplates should begin.

Labeling the cards

The grouped cards should be read one more time and checked to see if they are properly grouped. Cards that are suspected of being "strange" or inappropriately filed are taken out of the group and returned to the presorting pile. The cards that appear to be properly grouped are given a label that represents succinctly the characteristic of the group; the label is written on a blank card. The essence of labeling lies in retaining the nuances of the cards with lively, ordinary expressions. The label is supposed to convey the meaning of the cards fully, but not to say more than they do. It is inappropriate to express the meaning in abstract terms. Once the label card is completed, it is placed on top of the cards it covers and a rubber band is put around them. The labeled groups of cards are treated as single cards, and the procedure of card sorting continues.

Drawing the diagram

Some cards are left out in the process of sorting and do not fall into a specific group. These are called "lone wolf" cards or "isolates." They should be left alone, and no attempt should be made to force them into a sorting scheme. The process of sorting continues with "lone wolf" cards counting as individual groups until the number of groups is about 10.

After the groups are arranged, a diagram of the groups is drawn. For this purpose, the groups of cards that are the last to be collected are positioned so as to show their mutual relationships. These cards are pasted on a sample sheet, and symbols are used to indicate their mutual relationships.

Oral presentation

The content of the diagram is then explained orally. The meaning of each card is explained along with ideas that come to mind at the time of the presentation.

Written presentation

A written report is composed along with the affinity diagram. In explaining the data, include new ideas that come to mind. Such expressions as "It is," "I think," and "It appears" are encouraged.

Brainstorming techniques

Twenty-five years have passed since the brainstorming technique was introduced in Japan in 1952. It is widely used and well known, but it is often not used properly. In order to make effective use of the brainstorming technique, one must follow some basic rules:

1. *Criticism forbidden:* No expression of criticism or opposition to statements by others is allowed.
2. *Complete freedom:* Everyone must have complete freedom to express thoughts.
3. *Accommodation of many ideas:* It is better to have many ideas.
4. *Combination and improvement:* Opinions of other people are adopted and improved; statements of other people are to be pursued for possible adoption.

First of all, these basic rules must be followed faithfully. Second, self-examination is needed when things do not go well. Among the problems that may be encountered are the following: There may be no sense of urgency or immediacy in solving the problem. Leaders must motivate members to participate. Participants may make little effort to formulate ideas about the collected data. It is frequently necessary to give members homework so that they are well prepared for active participation. Participants may be reluctant to express their opinions, hampering the free flow of ideas. To avoid this, each

member could be asked to make a presentation. There may be a sense of too little freedom; members may be too serious and depend too much on common sense. Statements should be kept as concrete as possible, using ordinary language. General and abstract statements do not comprise good data. Statements made without adequate awareness of the problem and with insufficient knowledge of facts and thought tend to become abstract.

Individual brainstorming method

The individual brainstorming method applies the principle of brainstorming to promote individual thinking. This may be assisted by taking notes.

It was pointed out in the section on method selection that a diagram is a method of thinking, and there are many types of thinking. One type of thinking is through language. Japanese-speaking people think in Japanese, and English-speaking people think in English. Language thinking is expressed in spoken language as well as in writing. "Note-taking thinking" is thinking with letters and writing. Albert Einstein, famous for his theory of relativity, said, "New ideas are born in my mind one after another while I am speaking."

Thinking is also done in images. According to Dr. Edward De Bono, famous for his concept of lateral thinking, one way of avoiding the rigidity of language is to think not with speech, but with images.

Note-taking thinking is taking notes about what goes on in one's mind while thinking. Reading these notes helps to develop ideas further, with hints from the notes. It helps to continue the thinking process while one "talks" with notes. This is why this method is called individual brainstorming.

Unless notes are taken, thinking may ramble, wander, and derail. Ultimately, the process becomes unproductive if this occurs. Note-taking helps the thinking process by preventing rambling and derailment. It may result in generation of a new thought.

The following is an example of an individual brainstorming session:

Book (thesis) is written with the KJ method.

It is called a KJ diagram.

It is called seven new QC tools along with other diagrams.

Is it different from the KJ method, even if it is called the KJ diagram?

No, it is not different; the essence of the KJ method lies in its diagram.

How is it done?

Take notes.

Take as many notes as possible.

Notes taken on any idea can be sorted out using the KJ method.

Don't pay attention to structure; simply take notes.

Arrangements of notes is the same as data collection.

Conduct external exploration completely; conduct internal exploration.

Try to change the surroundings from time to time.

Unless you change the surroundings, you may go around and around.

Many books have been written about the KJ method.

Is there any sense in writing a book about the subject?

It is the KJ method as seen by QC advocates.

Could the feature of the method be brought out by that method?

Think about it.

KJ method as seen by QC advocates.

KJ method for quality control.

The KJ method for quality control is a method that is supplementary to statistical techniques.

Statistical techniques and the KJ method are comparative theories.

Include new viewpoints and a way of thinking as opposed to the KJ method.

Let the KJ method serve useful purposes as a guide for those who want to use the method.

Introduction to the Study of Problem Solution – KJ Method Workbook is excellent as a guidebook.

Introduction to the KJ Method for Management is also a good book.

I wish to make this as valuable as the above-mentioned books.

This is a book about the KJ method; write a book by using the KJ method.

It is common practice in writing a book to begin with a table of contents (structure) and then write the text (content); actually a book will begin where the author feels like starting, and the entire book will be completed as a whole later. This is the way the KJ method is used.

Following the preceding guidelines, notes should be taken about anything that comes to mind. When a mental block occurs, relax and turn your attention to other things or work on something else. Frequently, going outdoors for a walk is helpful in producing a lead to follow.

KJ method case studies

How should research and development be carried out?

This case is based on a seminar conducted at the KJ Method Workshop (with Mr. Kawakita as a speaker) held at the Hakone Hotel in 1970 (Fig. 5.4 was drawn at that time). The workshop was conducted with a team composed of 10 engineers and engineering managers from different manufacturers, including Kataoka Hirō of the Japan Paint Company. Various affinity diagrams were drawn in the first and second rounds by faithfully following the rules of the KJ method. Figure 5.5 is an affinity diagram drawn in the first round that shows the presentation of a problem. Figure 5.6 was drawn in the second round and shows an understanding of reality.

These figures were drawn with the participation of various manufacturers, but it was a homogeneous group because it was

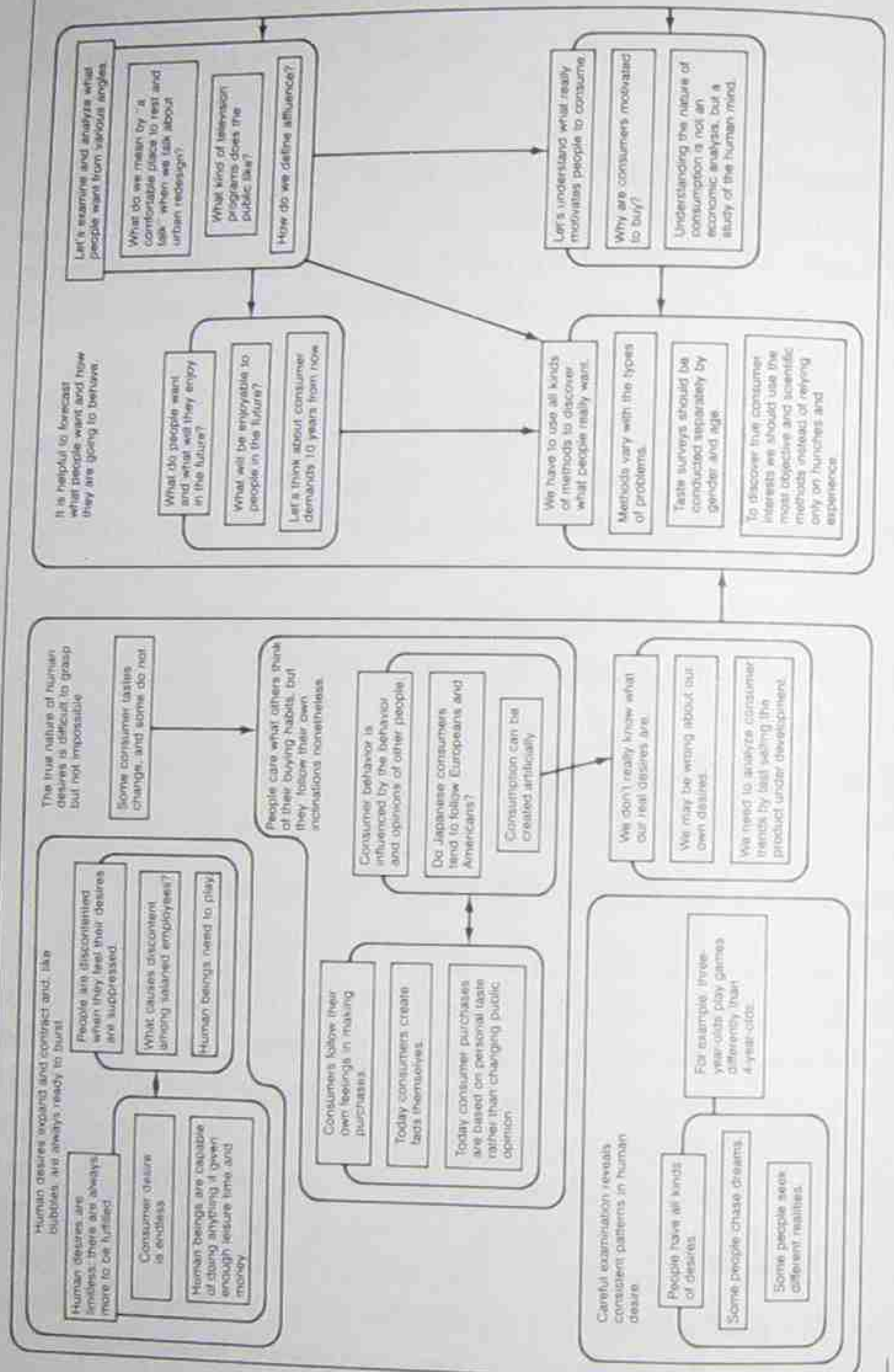


FIGURE 5.4
Example of an affinity diagram (one segment)

made up entirely of engineers. A group can be formed of people from the same industry as well as from different industries, such as banking, business, advertising, publishing, and film. The more people from different industries, the more lively will be the brainstorming and the more new findings will emerge in the affinity diagram.

How can QC circles be managed properly?

This case centers around the theme, "What is the best way to lead a QC group?" The case involves affinity diagrams drawn on the basis of data collected from brainstorming by a group of QC engineers representing different manufacturers.

Two diagrams (Figs. 5.7 and 5.8) were drawn on this theme because members were divided into two groups of 10 each and these two groups conducted separate brainstorming sessions. Two different teams were organized because the number of members in each group should be limited to a manageable brainstorming number and because the two teams were pitted against each other to increase productivity.

The same theme can produce entirely different data (verbal information on pieces of paper) if different members are involved. Furthermore, with different data, the same person could draw entirely different affinity diagrams. The drawing of similar diagrams would mean that the drawer probably predetermined the appearance of the diagram, and this kind of diagram cannot be expected to be truly creative.

What is the future of quality control?

With the same data (information on pieces of paper), two different drawers could produce entirely different affinity diagrams. First, brainstorming is done in groups, and each member shares data (e.g., fact cognition, opinions, ideas). Each member, using the same data, draws an affinity diagram independently, expressing his or her ideas. Then each member takes turns making oral presentations on

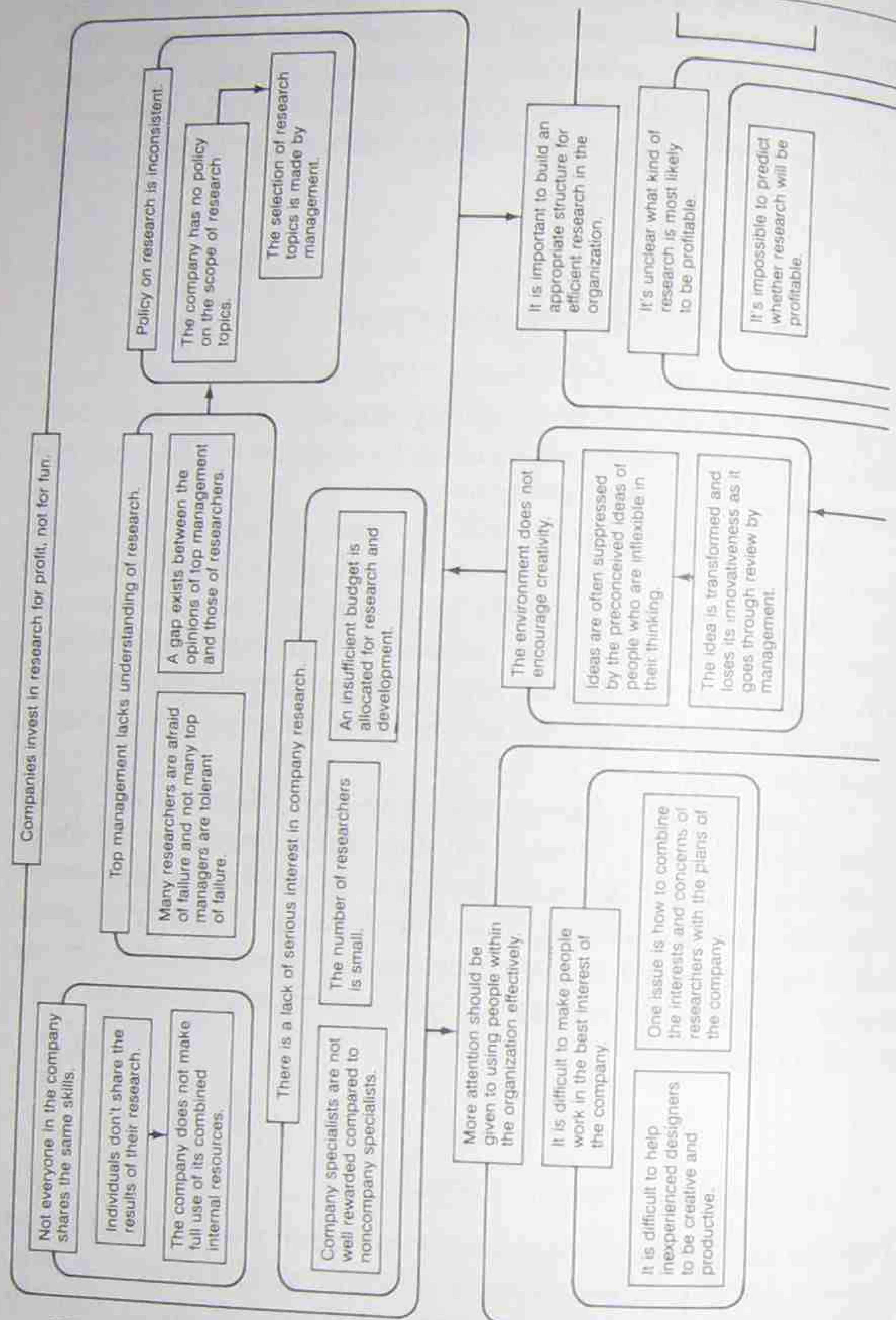
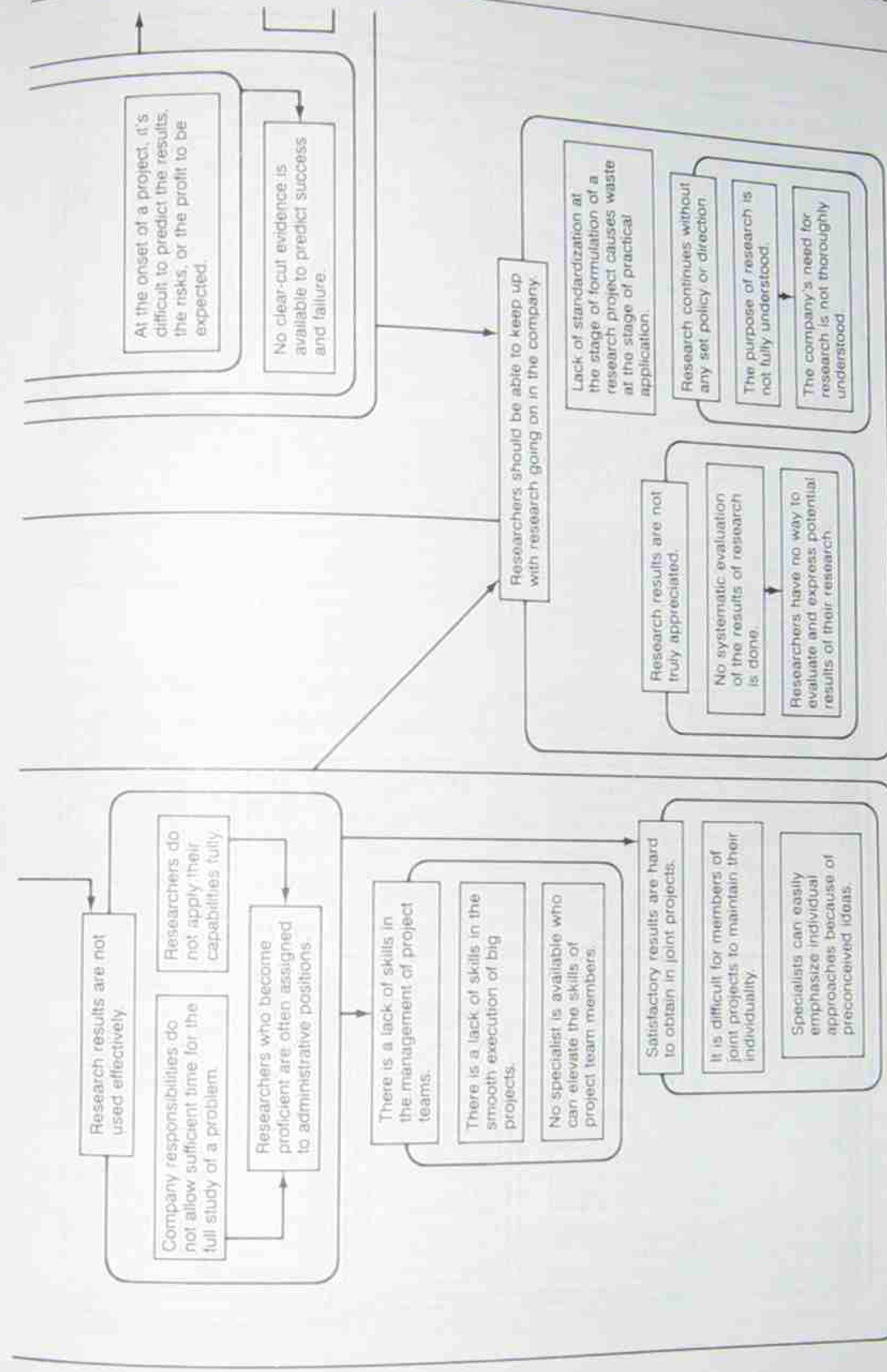


FIGURE 5.5
The process of developing research (the first round: asking questions)



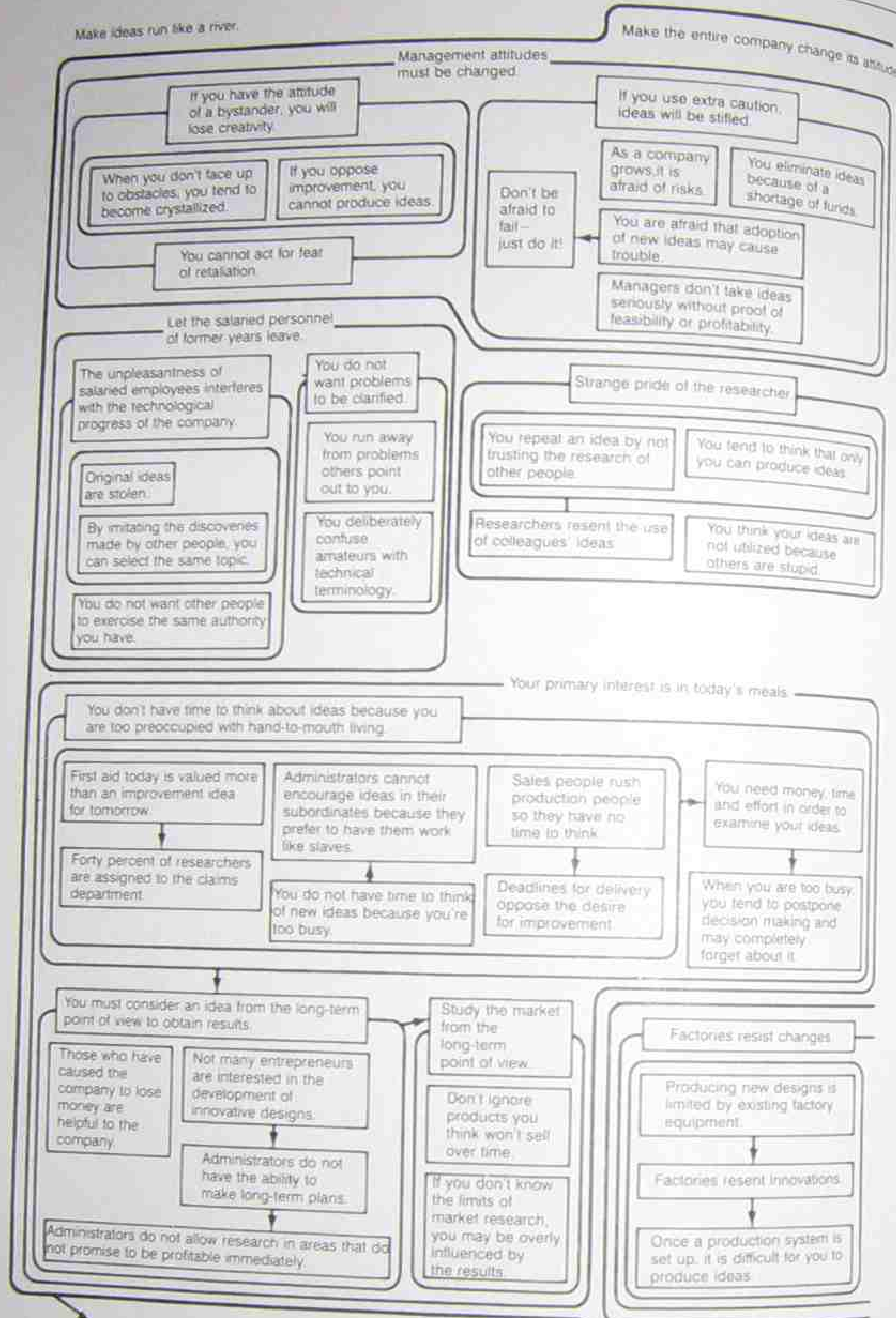
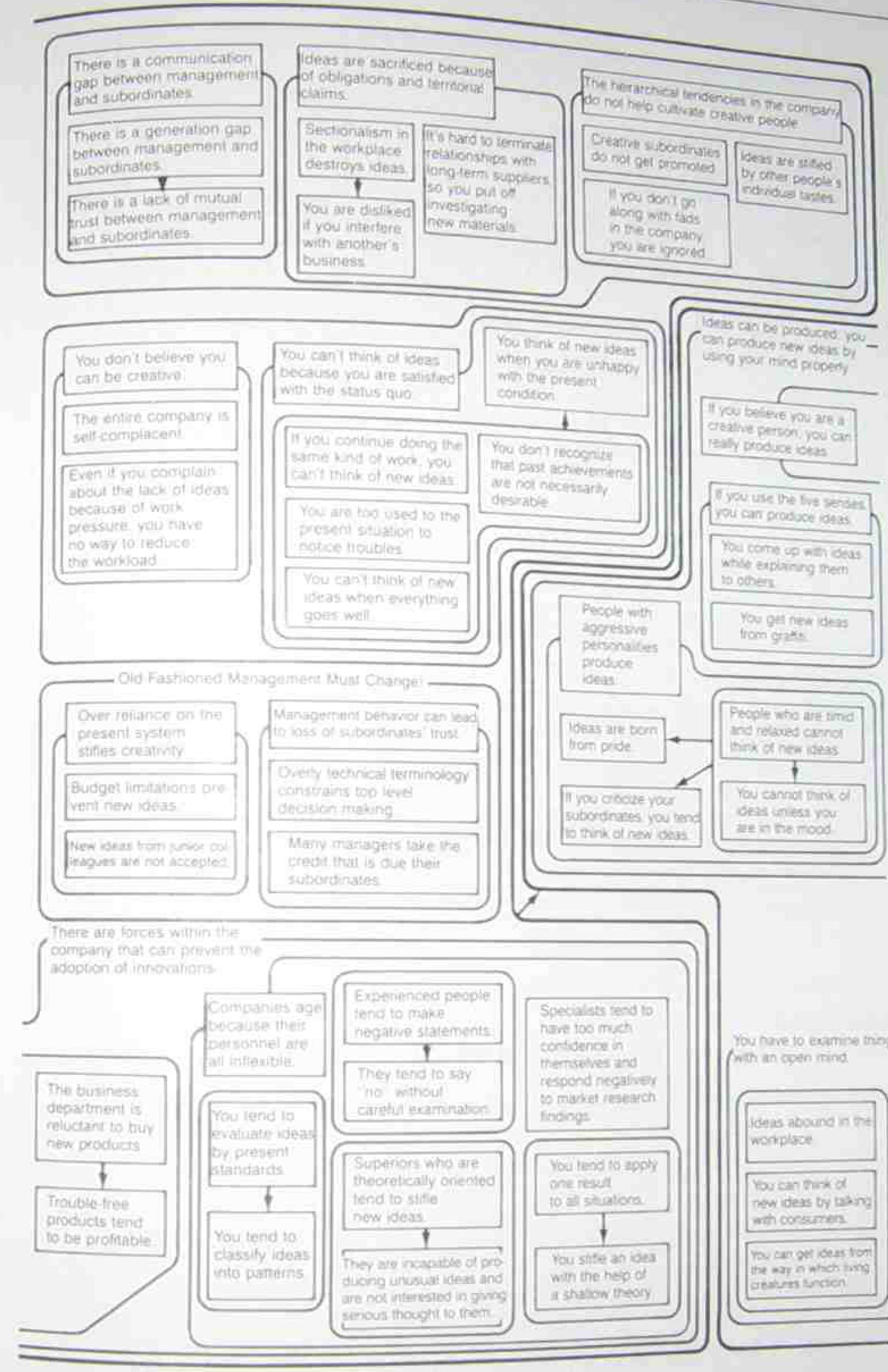


FIGURE 5.6
The process of developing research (the second round: understanding reality)



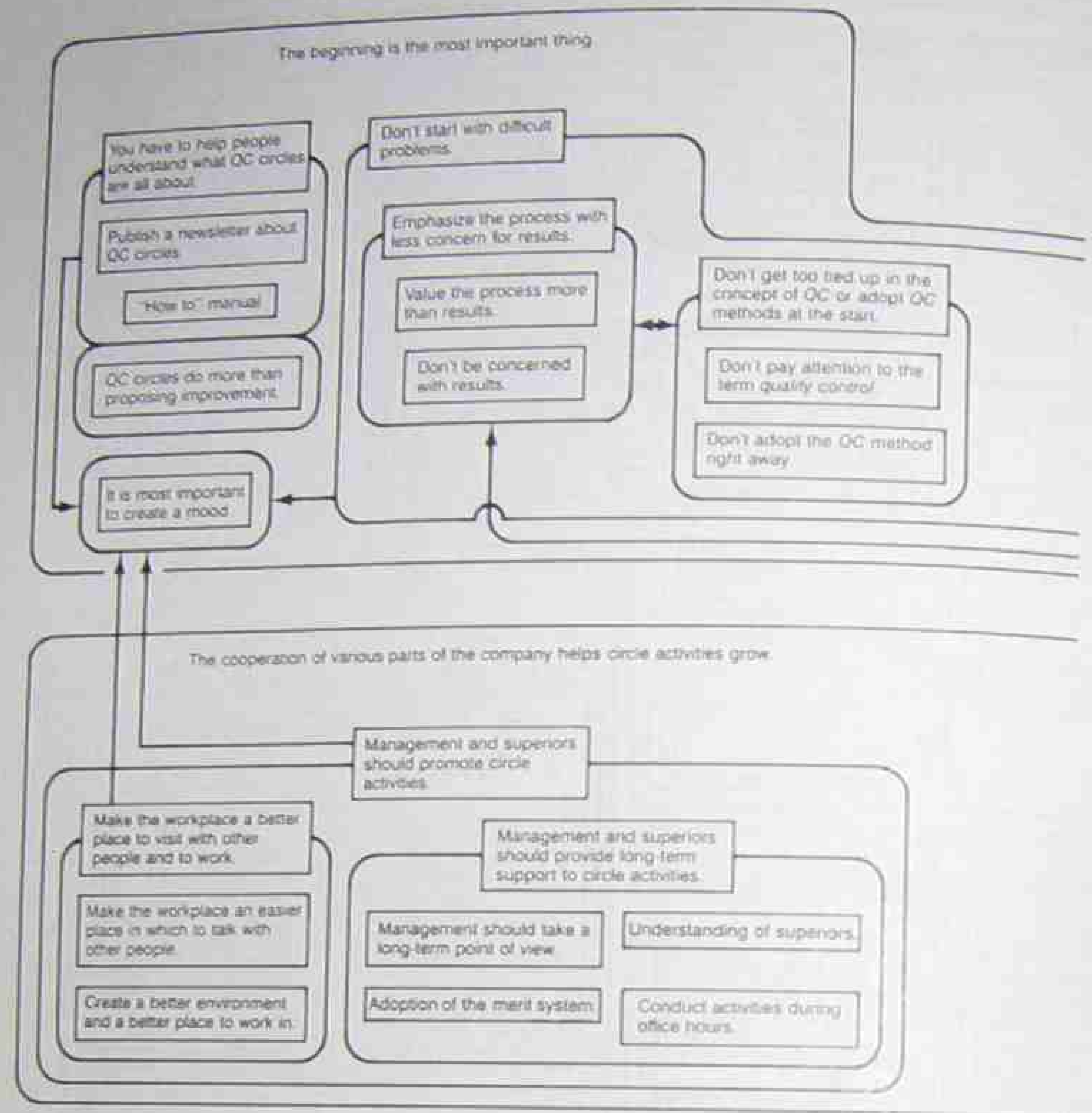
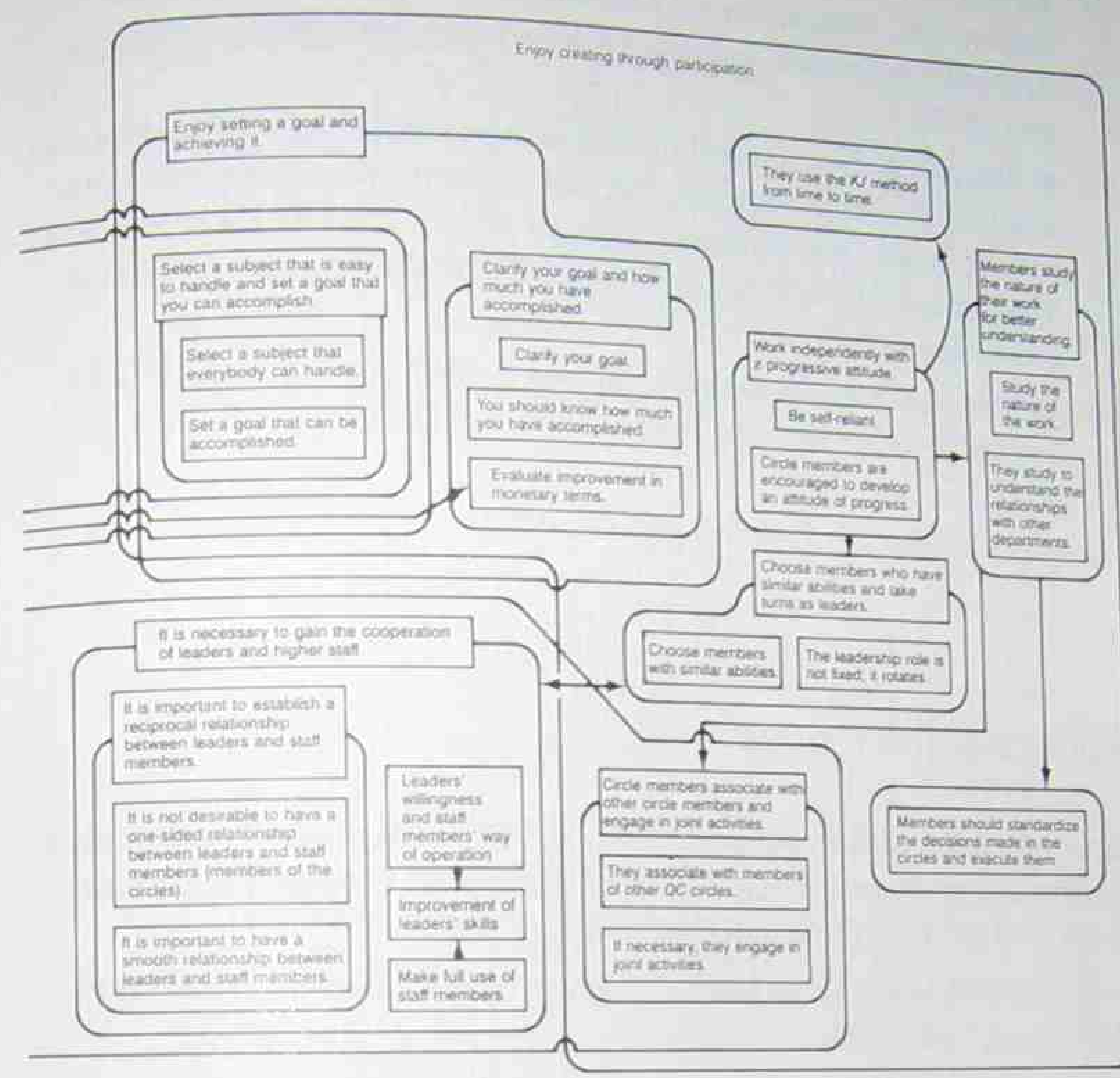


FIGURE 5.7
What is the best way to promote the smooth operation of QC circles?

the basis of his or her affinity diagram, sharing ideas with everybody else. At this stage, participants learn about the ideas of other members and may think of new ideas. Through these stages, thought productivity is enhanced to the maximum and the search for a solution to the problems is pushed forward.



Notes

1. Ko Takahashi, "Lecture Notes of the Kansai Self-Study Group – Luthe Research Society of Creativity and Development" (recorded by Kago), 1977.
2. *See* Kawakita Jiro, *Thought Method II* (Tokyo: Chuo Koronsha, 1970); Kawakita and Makajima Shiniichi, *Science of Problem-Solving – KJ*

Method Workbook (Tokyo: Kodansha, 1970); Japan Management Association, eds., *Introduction to KJ Method for Management—A Practical Approach to a Creative Technique* (Tokyo: Japan Management Association, 1974).

3. Material in this section and the following sections on card sorting and labelling, drawing diagrams, and making written and oral presentations is drawn from Kawakita, *Thought Method II*, 1970; Kawakita and Makajima, *Science of Problem-Solving*, 1970; and JMA, eds., *KJ Method for Management*, 1974.
4. Mita and Ishizaki, "What Will Happen to QA?" in *Summary of Tasks 3, 4, and 6—Seven New QC Tools Study Group* (Tokyo: JUSE Press, Ltd., 1978).
5. Mita and Ishizaki, *Summary of Tasks*, 1978.

Additional references

Japanese Union of Scientists and Engineers—QC Methods Development Division, *Proceedings of the QC Methods Development Division Conventions—Seven New QC Tools* (Tokyo: JUSE Press, Ltd., 1978).

——, *Introduction to Seven New QC Tools for Management and Staff* (Tokyo: JUSE Press, Ltd., 1978).

Futami Ryoji, *VE Manual* (Tokyo: Sogogiken Publishing Co., 1978).

Motoaki Hiroshi, "Science Illustrated #2: Psychology of Perception," in *Betsatsu Science* (Tokyo: Japan Economic Journal, 1976).

Tsukada Yuzo, "Science Illustrated #4: Brain," in *Betsatsu Science* (Tokyo: Japan Economic Journal, 1977).

Takahashi Hiroshi, *Training of Creativity* (Tokyo: Japan Industrial Training Association, 1970).

Kawakita Jiro, *Study of Party* (Tokyo: Shakai Shiso Sha, 1964).

——, *Teamwork* (Tokyo: Kobunsha, 1966).

——, *Exploration of Knowledge* (Tokyo: Kodansha, 1977).

Umezao Tadao, *Technology of Intellectual Production* (Tokyo: Iwanami Shoten, 1969).

Nakayama Masakazu, *All About NM Method* (Tokyo: Sanno Publishing Co., 1977).

Tokinori Norihiko, *Story of the Brain* (Tokyo: Iwanami Shoten, 1962).

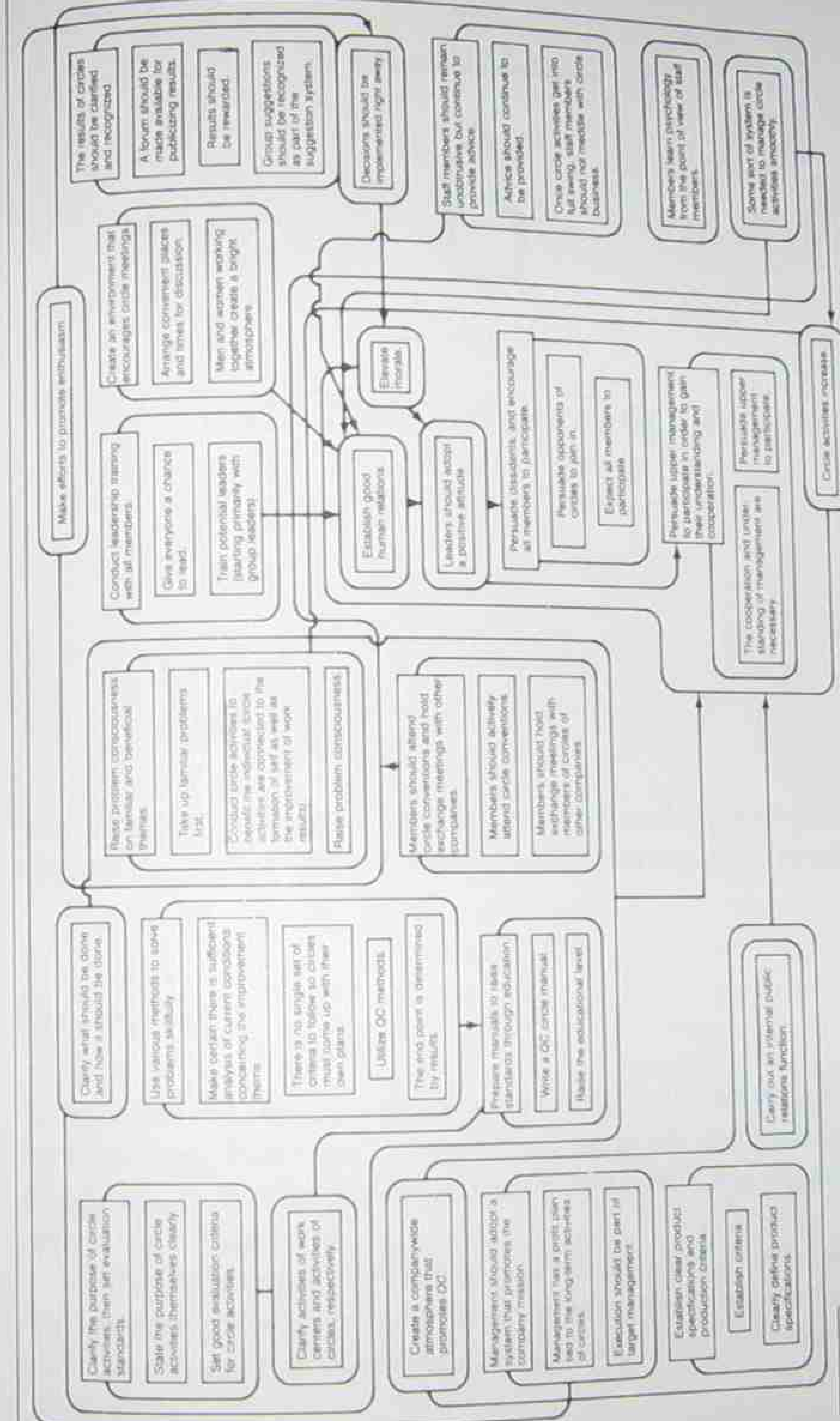


FIGURE 5.8
What should be done to operate QC circles smoothly (part 2)?

6

The Systematic Diagram

The systematic diagram is a technique developed to search for the most appropriate and effective means of accomplishing given objectives.

The method

The systematic diagram represents events in the form of a tree and its branches. This type of systematic diagram, sometimes called a dendrogram, has been used in family tree charts and organizational charts for a long time.

When means to achieve a goal are selected, secondary means are necessary to secure the primary means; thus the principal means become the goal of the secondary means. Figure 6.1 illustrates this relationship.

The systematic diagram displays the means necessary to achieve specific goals and objectives, clarifies the essence of the

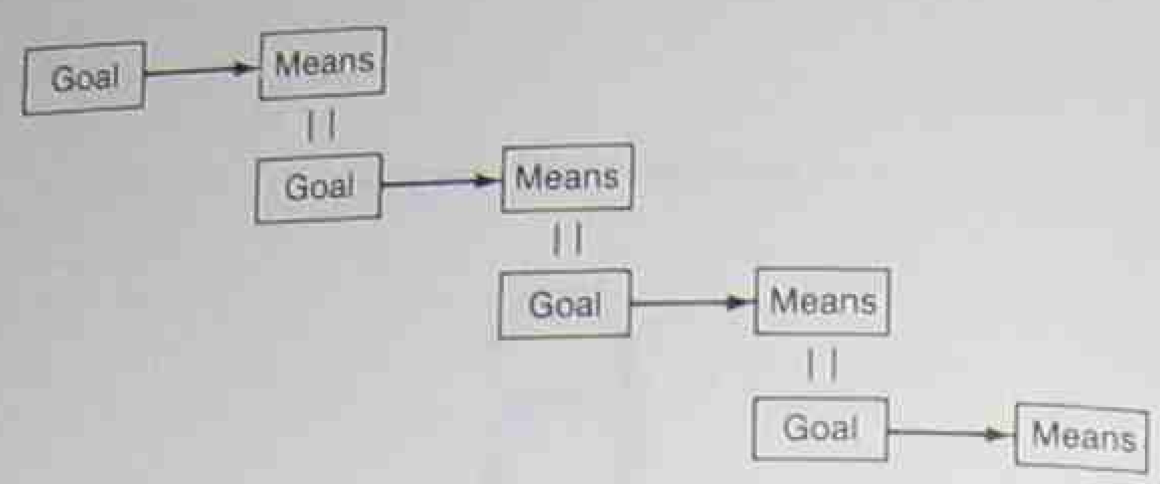


FIGURE 6.1
Conceptual systematic diagram

problem by making the subject matter visible, and searches for the most suitable means of realizing the objectives. This method is efficient not only in clarifying key control points in QC activities and developing effective improvement methods, but also in training business people to think in terms of means and objectives. Caught up in daily business activities, it is often difficult for business personnel to have clear notions of means and objectives. The systematic diagram alleviates this difficulty.

Uses of the systematic diagram

The systematic diagram can be applied in many phases of quality control:

- developing design quality for new products
- coordinating the development of quality assurance to the QC process diagram to ensure QA activities
- utilization of the cause-and-effect diagram

- developing solutions to various internal problems such as quality, cost, and production volume and delivery
- developing objectives, policies, and activity items
- clarifying departmental and control functions and promoting increased productivity

Constructing a systematic diagram

The diagrams used in the systematic diagram method can be classified into two major categories: the *component-development type,* which displays the component elements in terms of their purposes and means, and the *means-development type,* which systematically develops the ways and means for solving problems or achieving objectives.

Procedure for constructing a systematic diagram

To begin, obtain about 10 large (3′x4′) sheets of paper, 50 to 100 cards (3x5 in size), and several markers in two or three different colors.

Step 1: Establish objectives and goals

Clearly indicate on the cards the ultimate objectives or goals to be achieved. As a matter of principle, express goals and objectives in simple language, such as "Do X to obtain Y." If necessary, a short sentence or a phrase may be used. Most important here is that the expression be simple, direct and understandable to everyone. If there are conditions to be satisfied in achieving the goal, they must be clearly recorded. In establishing the goal, ask for what purpose the goal is to be attained. The answer to this question will help clarify a principal goal, which in turn will ensure that the proposed goal is an appropriate one.

Step 2: Deduce the means

The means thought to be necessary to achieve the established goal are deduced and recorded on paper, one by one. In deciding on the means, use the following guidelines: (a) start with the principal means and go to the secondary ones through association; (b) start with ideas of means considered to be least important and arrange these in groups, building up to more important ideas; and (c) simply pick up ideas of means without questioning their importance. Selection procedures will depend on the nature of the subject matter or the view about it. When collecting ideas for means, make sure they come from as many different points of view as possible. Make use of the divergent experience and knowledge of participating members through techniques like brainstorming. It is useful to express ideas for means in the form of a transitive verb and its object: "To (verb) to ______ (object)." The procedure here is basically the same as in step 1.

Step 3: Evaluate the means

The means deduced are evaluated for their adequacy before moving on to the next step. Constraints, if there are any, should be taken into consideration in sifting the means according to their adequacy. In evaluating means, the marks ○, △, and X are used. The mark ○ means "practicable"; the mark △ means "uncertain practicability without further investigation"; and the mark X means "impractical." Those marked △ need to be investigated promptly and should receive either an ○ or X marking.

In evaluating the means, it is important to take into consideration the following: (a) avoid superficial evaluations and hasty rejections of means; (b) an idea that at first seems impractical may be improved through incorporation with other ideas and further refinement; efforts should not be spared to help generate ideas and nurture them; (c) unusual ideas tend to be considered impractical, but when they are made practical, they sometimes bring great results; and (d) new ideas often occur during evaluations of other ideas; remain open to new ideas, because they may lead to good, practical ones.

Step 4: Prepare means cards

The means selected in step 3 are recorded on cards in large letters.

Step 5: Systematize the means

Place the goal cards made in step 1 in the left half of a sheet of paper. If there are constraints, record them below the goal cards.

Pose the following question with regard to each goal card: "To achieve this goal, what is the most necessary means?"

From the means cards made in step 4, sort out those containing an appropriate answer to this question and place them to the right of the appropriate goal cards. If more than one means card can be used to answer the question, place them side by side.

Now the relationship between the goal and the means necessary to achieve it has been made clear. This state of affairs, however, cannot by itself lead to concrete actions. Further secondary means must be found until the least and most practicable means are identified.

Next, pose the following question regarding each means card placed to the right of each goal card: "If this means is considered a 'goal,' what further means will be necessary to achieve this goal?" From among the means cards made during step 4, locate those which respond to this question and place them to the right of the means card in question.

Proceed by repeating the second question, locating the corresponding means cards among those prepared in step 4, and arrange them in order of their relevance to the goal and its means. When all the cards are arranged, connect them so as to display the goals and means. The result is a systematic diagram displaying the goal established in step 1 and all the means necessary to attain it.

In the process of constructing a diagram, new ideas other than those listed in step 4 sometimes occur. Needless to say, these new ideas must be added, and unnecessary ones should be weeded out.

This process of diagram construction is the most important part of the systematic diagram method. Hasty diagram construction, which merely arranges means in some systematic fashion, should be avoided. What is most important here is continual and careful scrutiny of the viewpoints from which a system of means is developed so that any omission or oversight is precluded.

Step 6: Confirm the objectives

Although the systematic diagram was constructed in step 5, we confirm whether the high-level means are adequate as the objectives of their corresponding secondary means. Starting with the least important means, the following question is asked: "Can the principal means ('goals') really be achieved by these means or by part of them?"

If the answer is "yes," raise the same question about the next levels, one after another. Confirm the attainability of the goal established in step 1 exclusively through the means developed in step 5. If the answer to the preceding question is "no," this reflects the insufficiency of the means developed in step 5 and the need for adding other means in order to realize the established goal.

The above-mentioned confirmation completes the development of the means to achieve the purpose and goal. This, in turn, completes the systematic diagram. Figure 6.2 illustrates these steps.

Application to quality designing

In the field of quality control, the literature defines the concept of quality designing as follows:[1]

Quality designing (in a wider sense) or quality planning
- Quality planning .... Quality of planning
- Quality designing (in a narrow sense) .... Quality of designing

Quality planning means deciding, as the primary policy, on the quality that will satisfy the users, i.e., the true quality characteristics. *Quality designing* (in a narrow sense) refers to replacing the quality desired by the users (the true quality characteristics) with a group of controllable substitute characteristics through inference, translation, and conversion.

Speaking in general terms, the concrete application of quality designing may be divided into two major categories: The first is transformation and development of the planned quality (the top policy), determined by quality planning, into the designed quality (controllable substitute characteristics) and buildup of an image of the product that satisfies the planned quality, i.e., *quality designing for*

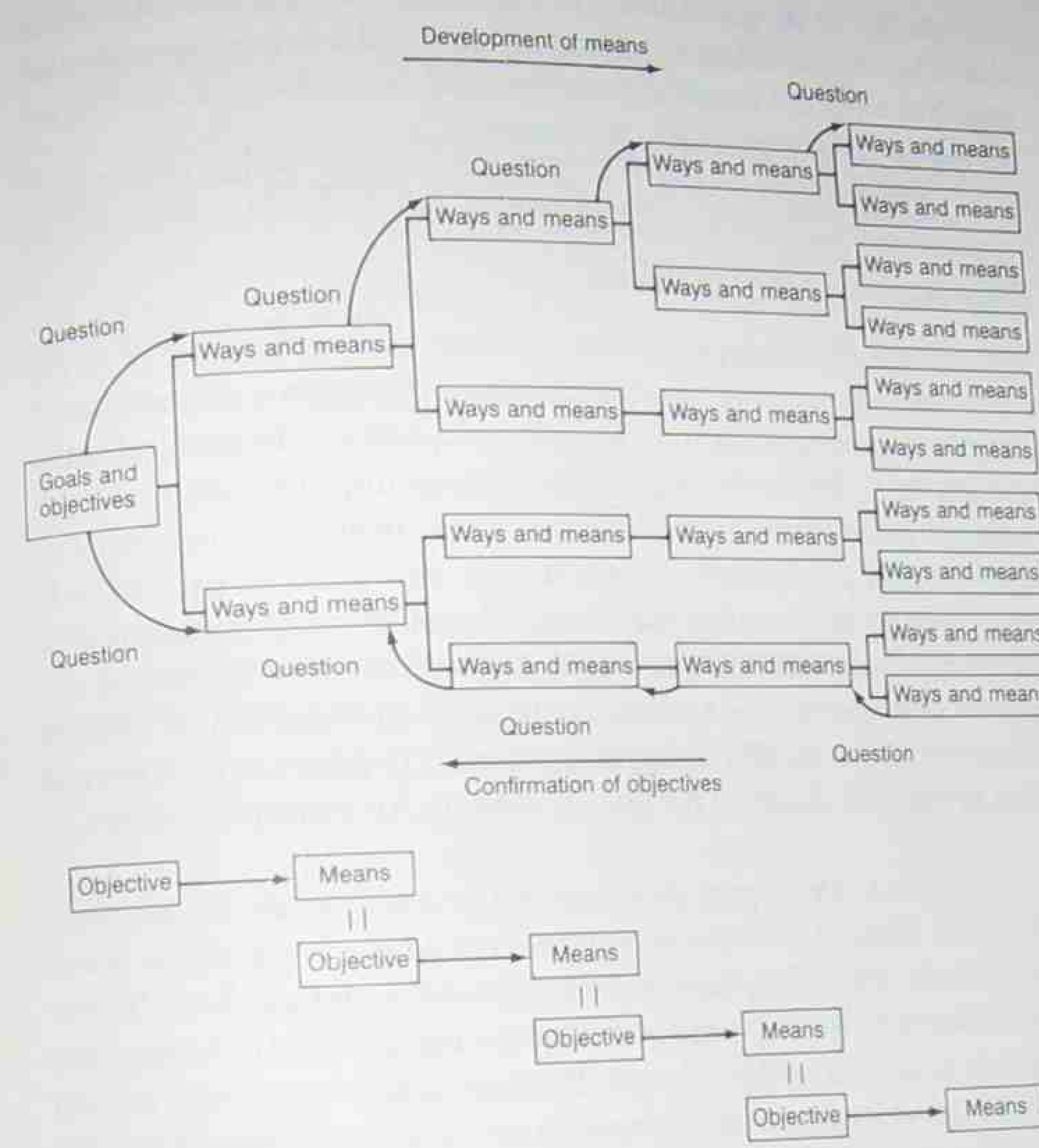


FIGURE 6.2
Constructing a systematic diagram

Note: In branching, *and/or* can be used when ideas are in the developmental stage. If a distinction between *and* and *or* is necessary, use the logical symbols employed in fault tree analysis (FTA).

the purpose of developing a new product. The second is transformation of the quality to be guaranteed to users into controllable substitute characteristics and determination of items and levels to be controlled together with the intracompany method of control, i.e., *quality designing for the purpose of quality assurance*.

Development of new products

The process of developing a new product can be thought of in the following manner. After the quality desired by users is defined based on research and speculation, the planned quality is designated as the primary goal. Then, through quality designing, planned quality is replaced (substitute characteristics). Next, from the viewpoint of function and production design, design and production methods are chosen so as to realize the designed quality most economically. Figure 6.3 illustrates this process as a system of corresponding objectives and means. The systematic diagram method can be utilized effectively in the quality designing of a new product, which we can call a systematic diagram for quality design. An example follows.

Example 1. Developing design quality for a new product, a television UHF tuner. When UHF broadcasting was promoted actively in Japan in March of 1969, Corporation M foresaw the demand for higher production and quality improvement for the all-channel television tuner. In response to this forecast, the company organized a project team to develop a new all-channel tuner. The goal assigned to the team was to develop, by October of the same year, a competitive, high-quality, high-productivity, low-cost product that would make its own contribution to the anticipated full-scale spread of all-channel TV receivers. The team members examined and evaluated existing products and those of competitor companies in terms of quality, capacity, productivity, and cost. Their discussions were guided by the criteria of practicability within the fixed period of time and by the certainty of securing full priority at the time of project completion. They constructed a systematic diagram for quality design (as shown in Fig. 2.7) and presented it to the development council. Having obtained the approval of top management, they started the actual development operations.

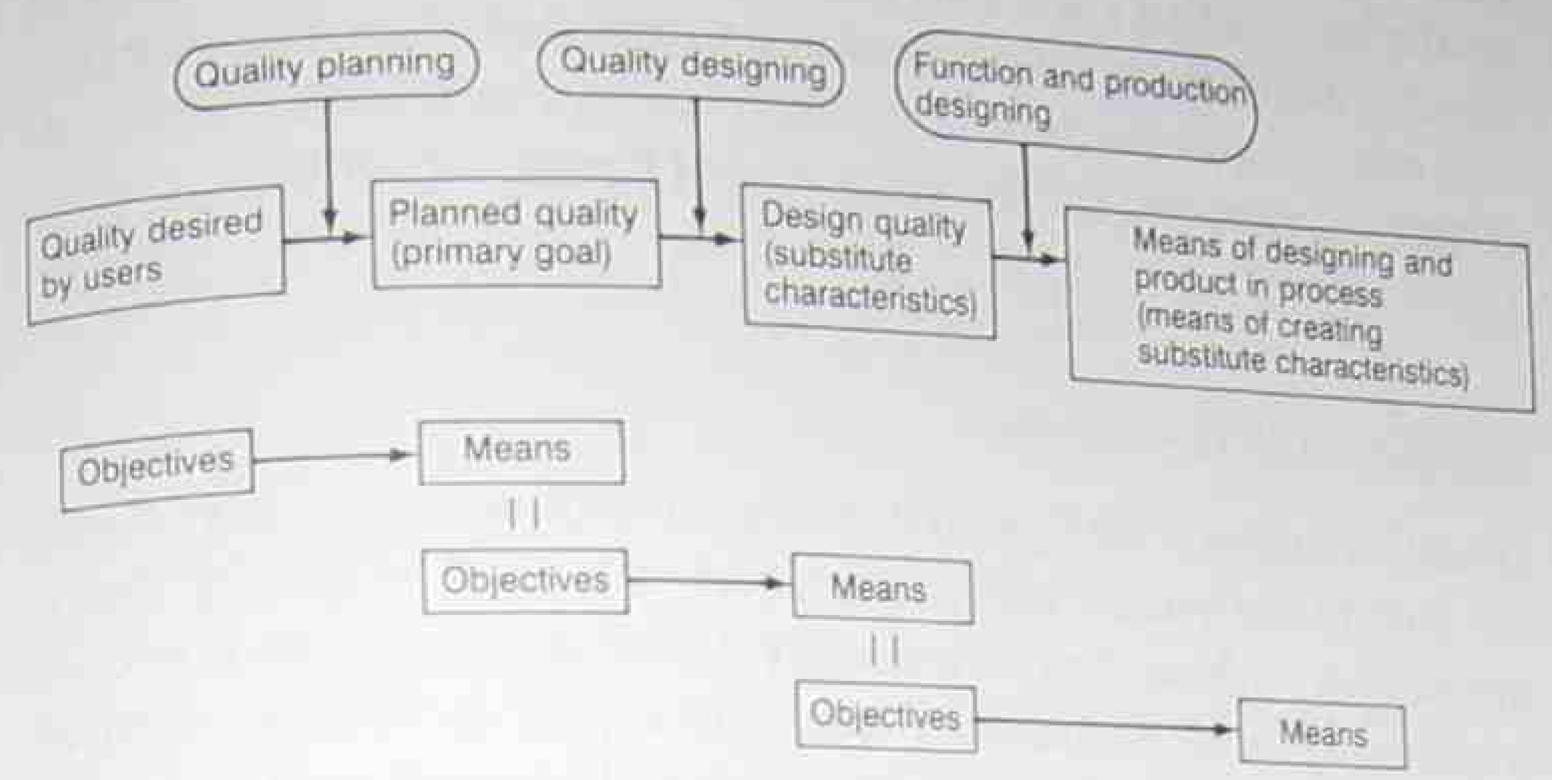


FIGURE 6.3
A systematic diagram for quality designing in new product development

In this manner it was possible to harmonize the plans of top management and the thinking of those charged with development by making a systematic diagram for quality design. Moreover, when changes were considered necessary in the development process, it was possible to incorporate them quickly along the established line of objectives-and-means relationships.

Figure 2.7 has a column designated "Undecided Items." When planning a new product, it is not always possible to develop all the means in detail to satisfy quality design. (As a matter of fact, it is generally impossible to do so.) In such cases, the items which cannot be detailed or which require further investigation or experimentation should be singled out and put in proper places in the development plan. Making a place for undecided items often brings about good results.

As soon as a development plan for a new product is decided on, a systematic diagram for quality design of the new product should be constructed; then the development process should be carried out on the basis of that plan. This procedure will clarify the direction of development efforts, help accommodate needed changes in the process rapidly, and preclude counterproductive opinions, thereby effectively promoting the development of the new product.

Application to QA activities

Quality-assurance (QA) activities are control activities that clarify the quality desired by users and secure the level of quality that is satisfactory to them. Since the quality demanded by users, however, is generally understood only qualitatively, it is hard to control. Therefore, it is necessary for QA activities to transform the quality desired by users into controllable quality characteristics.

In other words, the quality demanded by users is first transformed through quality design into substitute characteristics (designed quality). Then this quality is correlated with control quality characteristics and their level (in many cases, the standard specifications and characteristics of the product). Next, control specifications and methods are established to realize these control quality characteristics and levels. In this manner, it is possible to clarify the object and method of control to realize quality assurance.

This process also can be reduced to an objective-and-means system, as shown in Fig. 6.4 This systematic diagram used for quality assurance activities is called a quality assurance systematic diagram. An example follows.

Example 2. Developing assured quality for a television VHF tuner. Figure 6.5 is a systematic diagram for assured quality that shows how the quality of a VHF tuner can be developed as part of the QA activities for color television receivers. The diagram shows how the required quality of a VHF tuner was derived from the design quality developed by the maker. It also shows how the required quality of a tuner was correlated to its design quality, and then how this design quality was developed to the standard specifications and level of the product. Furthermore, it shows the design check items and control points in the process. Because such a systematic diagram for assured quality was constructed, it was possible, when such problems as market claims occurred, to proceed systematically to determine

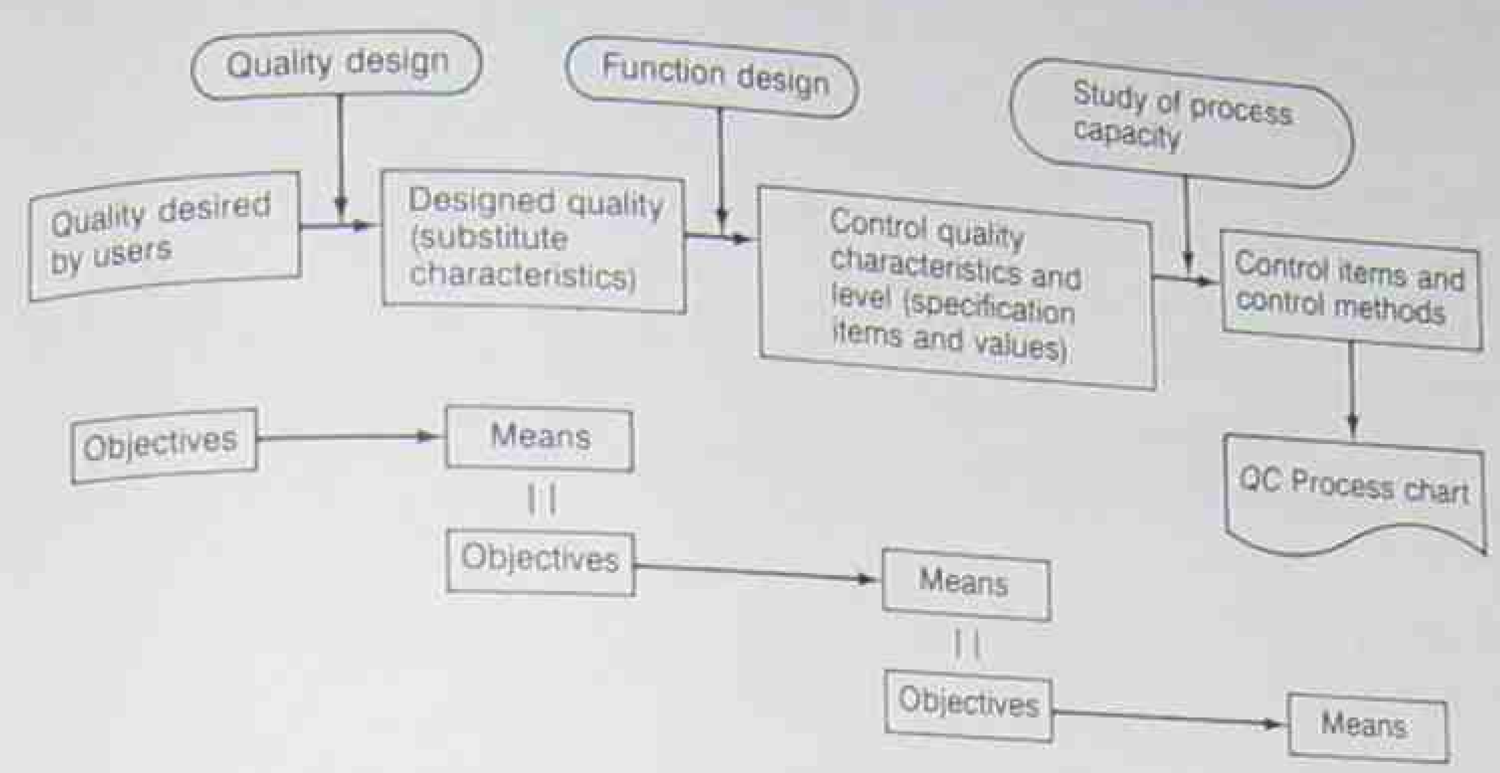


FIGURE 6.4
A systematic diagram for quality assurance

whether there had been slippage in control quality characteristics or whether the control quality level was too low. This determination could be conducted by examining the diagramed means to the product quality. Use of this method made it possible to develop appropriate countermeasures.

Example 3. Developing assured quality for automobile brakes. Corporation N, which specializes in automobile brakes, is extremely strict in securing the stipulated quality of its products. Moreover, ithas to ensure that its products can be assembled into the final product, i.e., automobiles, in such a way that the brakes function exactly as they are intended to.

Management established "a brake that works without failure" as the required quality. Based on this requirement, management derived secondary required qualities such as braking power, strength, durability, safety, operability, and feasibility of inspection

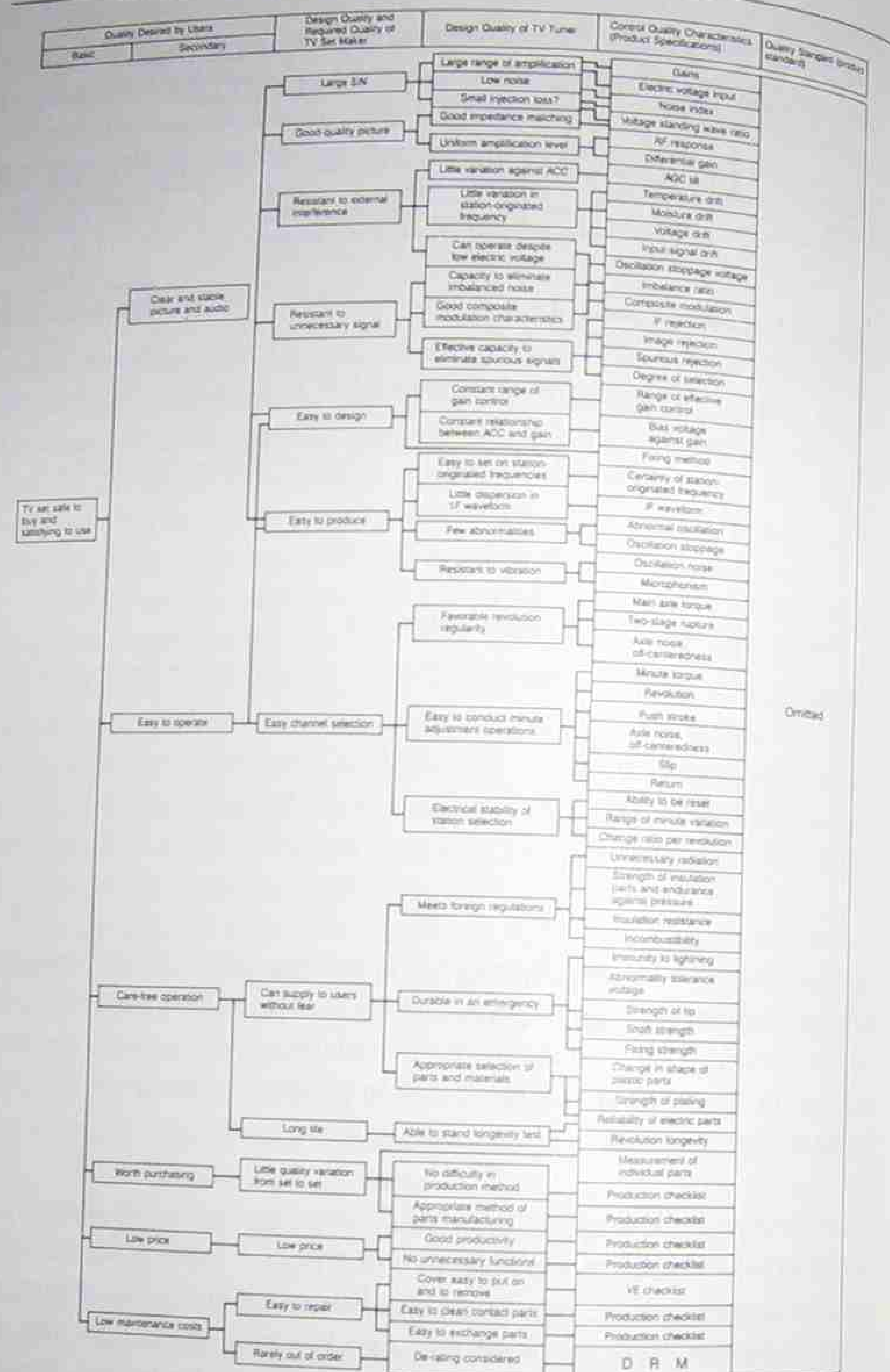


FIGURE 6.5
Systematic diagram for quality assurance of a television VHF tuner (a systematic diagram of the component-development type)[2]

and maintenance. Each of these items was then considered as a basic required quality, from which secondary and tertiary design-quality levels were derived and control specifications were established.

Because the corporation's production process was insufficient to guarantee control specifications, QA activities were carried out at every step in the production process, from the manufacturing of brake parts to their assembly in automobiles. Accordingly, the control specifications were examined in relation to the suppliers of parts, the production process, and inspection of the brake manufacturer and the automobile makers.

This process of examination made it possible to diagnose problems, provide guidance to the part suppliers for their quality control, incorporate appropriate details into the company's own QC process chart, and make requests to the automobile makers. Figure 6.6 shows a portion of the systematic diagram for automobile brake quality assurance, which was useful in establishing the QC system.[3]

Application to quality improvement

Used as a cause-and-effect diagram

Needless to say, a cause-and-effect diagram is simple and effective. However, we encounter minor inconveniences in the following cases:

1. When causes at the sample level need to be compared, examined, and evaluated.
2. When the influence of each cause is quantified and expressed in a diagram similar to that in number 1 above.
3. When the number of causes is very large.
4. When the lowest-level causes need to be examined in relation to necessary measures, procedural details, or a list of standards.

In such cases, one solution is to arrange causes and effects in the form of a systematic diagram. A systematic diagram that expresses

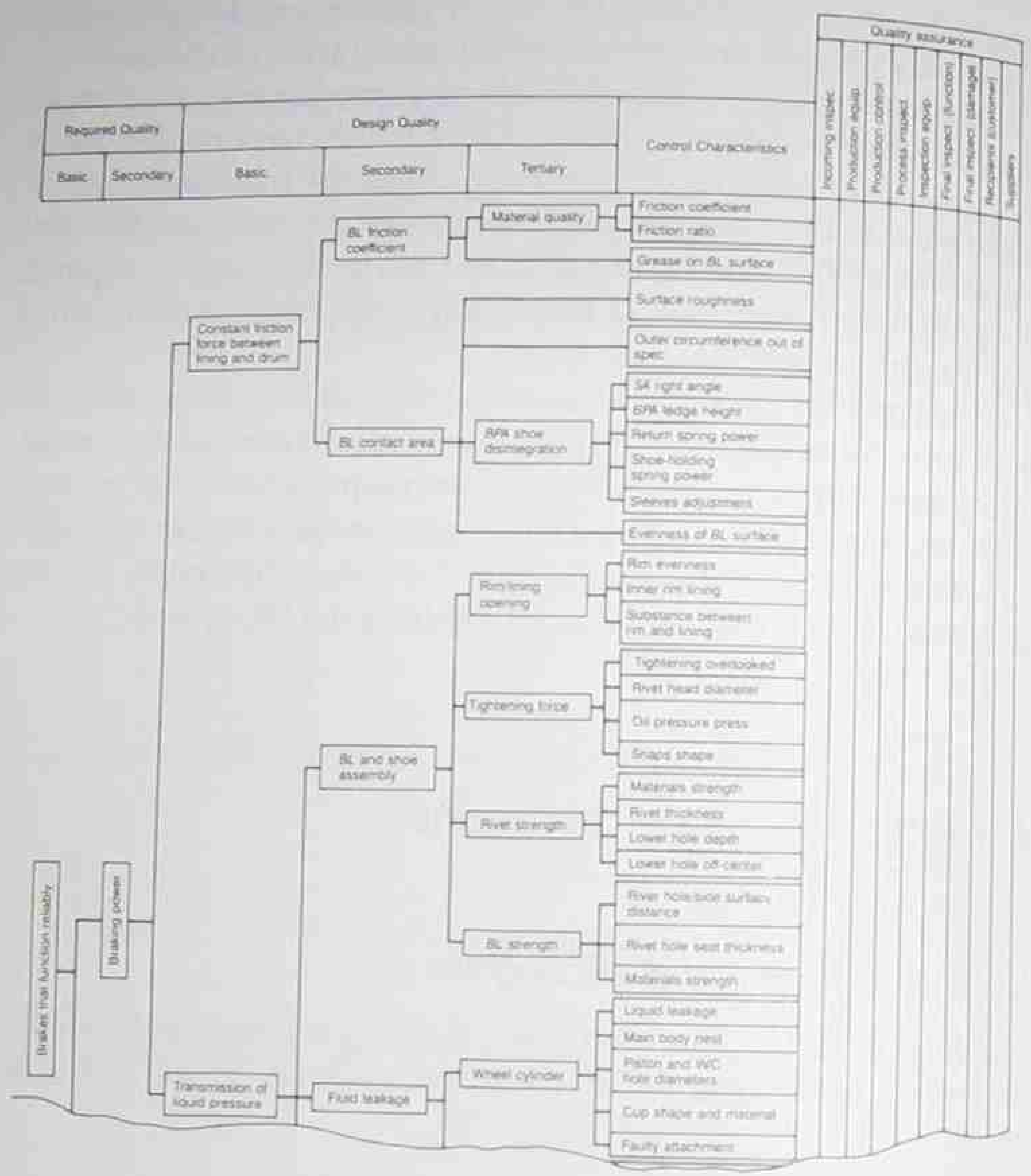


FIGURE 6.6
A systematic diagram for quality assurance for automotive brakes (component-development type)

causes and their effects is called a cause-and-effect systematic diagram.

As can be seen in Fig. 6.7, a cause-and-effect diagram and a cause-and-effect systematic diagram express an identical event or process in two different forms.

Application to the reduction of nonconforming items

Such problems as the presence of abnormalities and the inability to reduce them can be dealt with in one of two ways.

Case 1

Consider the situation where the necessary means of production have been determined and are in use, such as parts, materials, facilities, repair tools, operational procedures, and finishing requirements. The appearance of nonconforming items indicates that these means either have deteriorated or have deviated from the standards occasionally. Dispersion of the so-called four *M*'s (material, machine, method and man) includes use of substandard parts and materials, loss of precision of machinery, facilities, and repair tools, and nonobservance of operational standards. In such a situation, all the production means presently in use should be displayed in a systematic diagram and examined one by one. A priority list should be established for this process on the basis of the extent of each item's influence. In addition, those means with large dispersion must be located and corrected, since they are the source of abnormalities.

Case 2

Sometimes the production means currently in use are found to be inadequate for producing the quality characteristics in question. Such a situation may be due either to insufficient process planning or to the fact that while the required quality has become more and more exacting, the process has remained the same. When this is the case, it helps to reexamine and redefine the goals and collect ideas about effective means using a technique such as brainstorming with persons of diverse viewpoints. Such ideas may then be refined and developed into concrete measures for improvement.

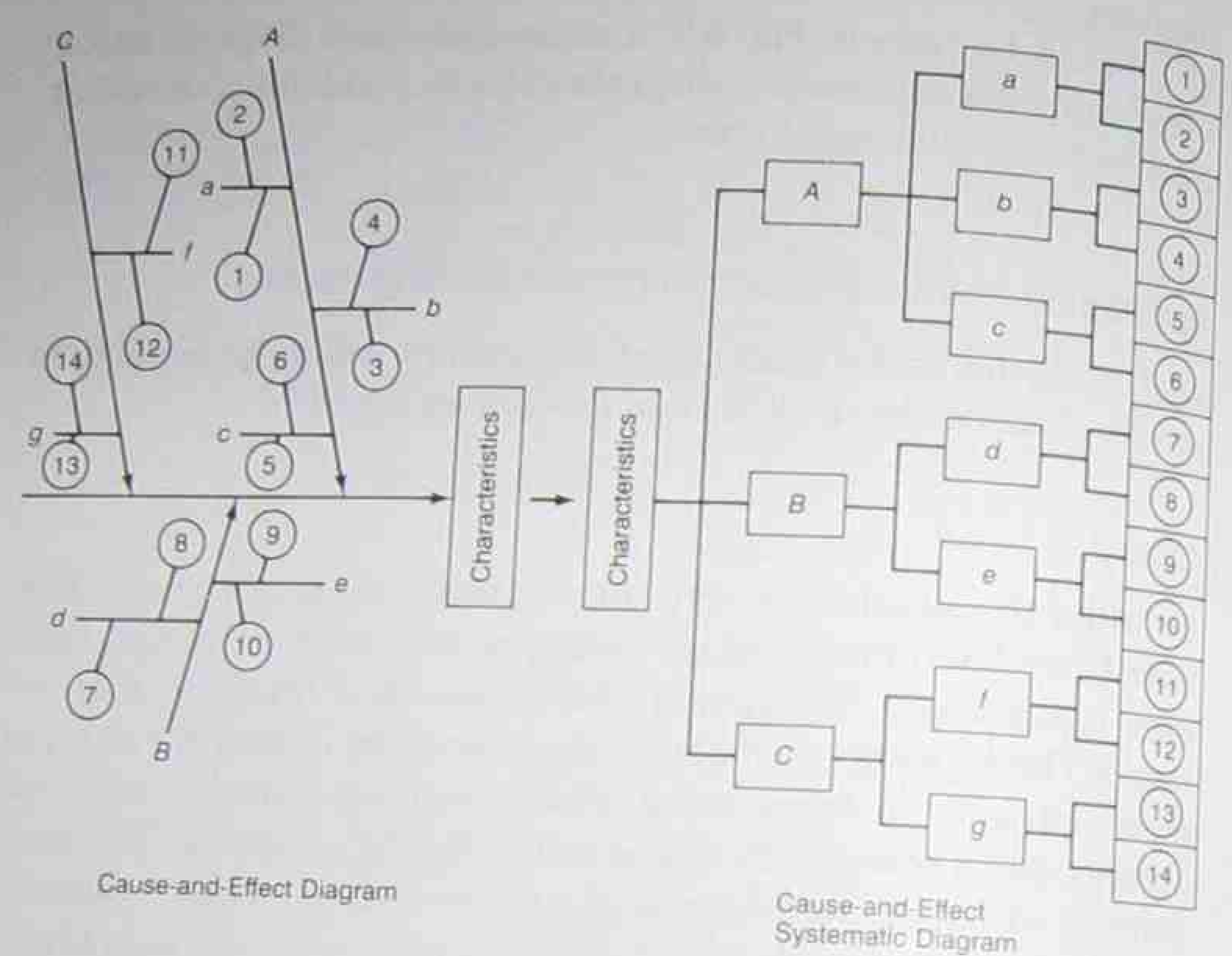


FIGURE 6.7
A cause-and-effect diagram and a cause-and-effect systematic diagram

In either case, the first thing to do is to clarify the quality characteristics (objectives) to be attained. Next, the means for reaching objectives must be developed and examined, and measures for reducing nonconformities must be specified. In this process of developing objectives and means, a systematic diagram will help you to gather and organize information, pursue the cause(s) of abnormalities, and devise countermeasures.

The existence of so-called chronic abnormalities means that despite various measures taken in the past, the situation has not improved. The systematic diagram method has often been effective in eliminating such chronic abnormalities.

Example 4. Reducing dispersion in the thickness of wall-covering materials. Corporation D, which produces wall-covering material, experienced persistently conspicuous fluctuations in the thickness of its product. Although various measures were tried, the capacity of the production process, C_p, dropped to 0.647. The company decided to go back to the starting point. The firm had all personnel concerned enumerate the factors that could have caused the problem and summarize the results in a systematic diagram. Figure 6.8 shows that diagram.

Discussions about the diagram were held by all personnel concerned, while examination of the degree of influence of each factor was divided and assigned to several people. As a result, it was found that factors that had been thought to be of little importance, such as a change of speed, had a surprisingly large effect. (A change of speed occurred when the operation was started or the product type was switched.) A countermeasure to this problem was devised and implemented. In addition, the method of adjusting the steam pressure was standardized and the roll balance was improved. As a result, the process capacity was improved to 1.20.

The cause-and-effect systematic diagram used in this example is similar to a cause-and-effect diagram of the dispersion-analysis type. Since the causes of the thickness irregularities were investigated with a systematic diagram, it was easy to compare the causes of dispersion with one another and to correlate them with the operation standards and control sources. Furthermore, it also became easier to determine the priority of the measures to be implemented and to follow up the results of that implementation. In this case, the merits of the systematic diagram are fully apparent.

Example 5. Reduction of loss caused by cross-sectional rupture of book-cover cloth. Corporation D, a producer of book-cover cloth, faced the problem of cross-sectional rupture, which occurred as often as 59 times a month on average during the production process. When such a rupture took place, the machinery had to be stopped to reconnect the ruptured parts, and each stoppage caused the loss of 80 to 90 meters of the product, resulting in an enormous overall loss.

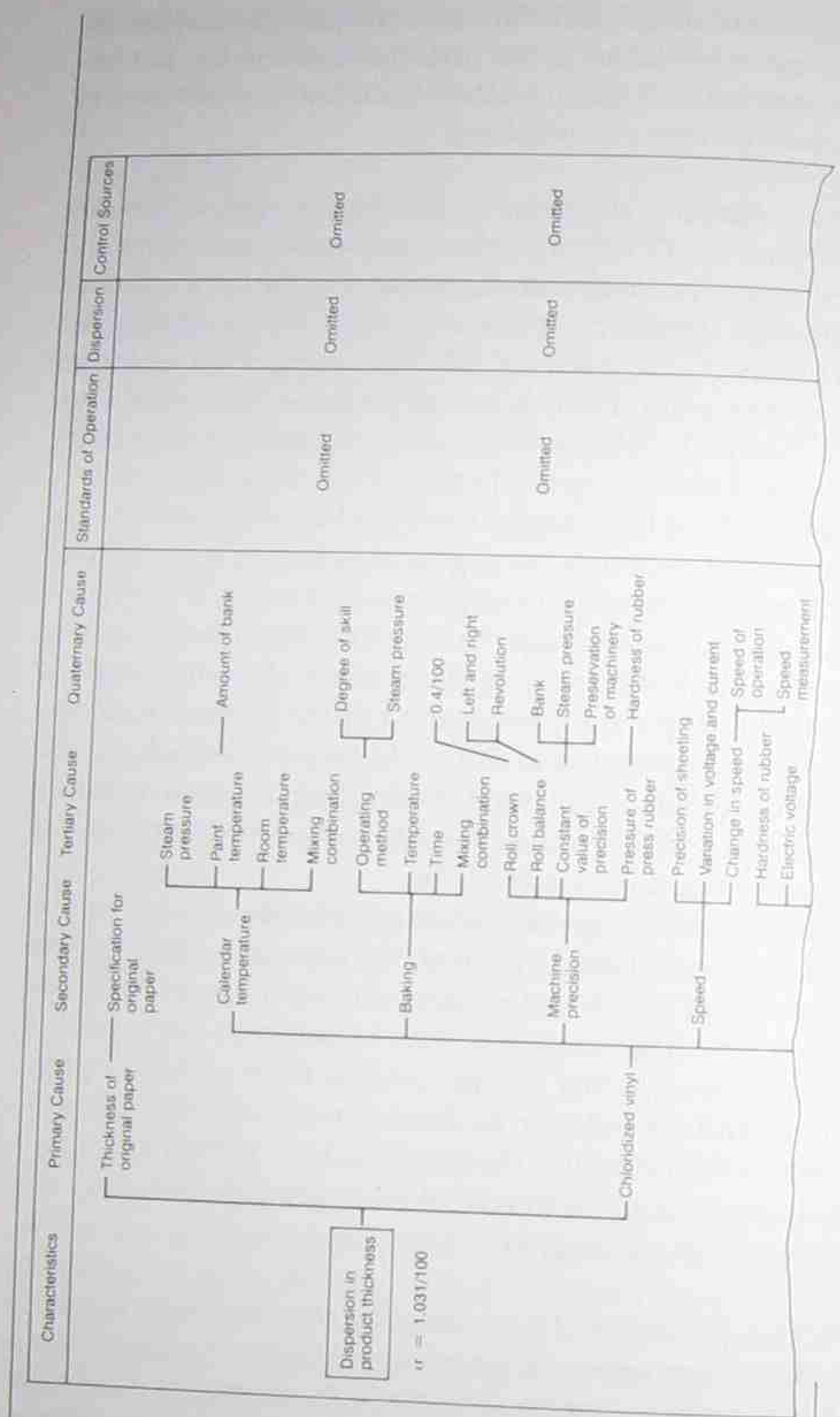


FIGURE 6.8
A cause-and-effect systematic diagram for the dispersion in the thickness of wall coverings[4] (component-development type)

Consequently, an objective of minimizing the loss was established. This objective was divided into two parts: one was prevention of ruptures, and the other was minimizing the loss caused by ruptures. Regarding each of these new objectives, suggestions were collected through brainstorming, and after sifting through the ideas, a systematic diagram was constructed, as shown in Fig. 6.9. The ideas at the bottom of the diagram were thoroughly evaluated, were given practicable forms with implementation directions, and were carried out one by one. After a period of 6 months there was a decrease in the number of ruptures to an average of 13.2 per month and the loss of the product was reduced to an average of 35 to 45 meters per rupture.

Other applications

The systematic diagram method helps us find the starting point in problem solving by arranging the branching phases of the problem into understandable relations between an objective and the means to achieve it. It is applicable not only to quality design and quality improvement, as was explained earlier, but also to other aspects of the manufacturing process. Several applications will be briefly mentioned below.

Developing policies and goals

If top-management policies or departmental objectives are to penetrate to lower-level personnel and be implemented by them, it is necessary to analyze problems in appropriate ways and develop concrete solutions. The development of goals and policies results in a system of goals and means, and it is here that the systematic diagram method becomes applicable.

Developing administrative functions

The systematic diagram method is effective in promoting efficient management because it clarifies both departmental and control functions within a company. In planning to increase departmental

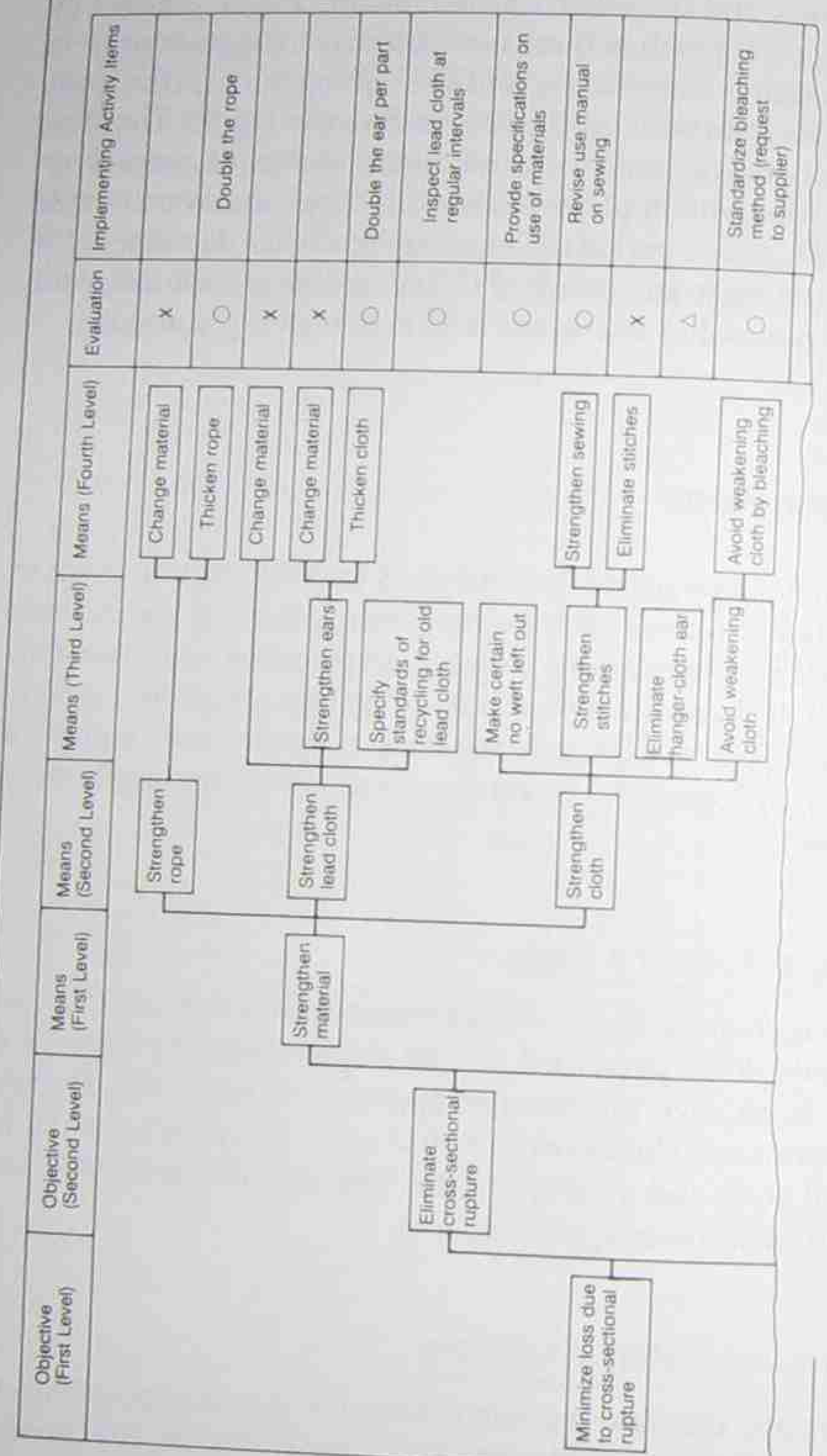


FIGURE 6.9
A systematic diagram for ideas to reduce the loss caused by cross-sectional ruptures[5] (policy-development type)

efficiency and reduce personnel, it is not desirable to analyze current assignments too deeply. One may encounter the unwelcome situation where close examination of inefficient or unattended tasks leads to a staff increase. A better strategy is to develop a systematic diagram of "target conditions" for work in each department and to clarify the purpose of that work. This will promote the discovery of new approaches to achieving departmental objectives as well as improvement in the administrative structure itself.

Other areas

In addition to the preceding examples, the systematic diagram method can be applied to various other problem areas, such as

- Developing means for promoting productivity
- Generating ideas for improving bookkeeping
- Developing implementation specifications for design standardization
- Developing ways to activate the QC circle
- Developing ways to reduce claims
- Developing practicable means for the effective promotion of research and development
- Developing means for effective project team management
- Developing ways to promote TQC activities
- Comparing production costs with rival makers through function comparison
- Functional analysis to improve the production process

Techniques employed with the systematic diagram method

Function analysis

Function analysis is used in value engineering (VE) to evaluate how well a product design fulfills user requirements in relation to cost. A brief explanation is given at the end of this chapter.

Correlation-tree method

This technique is used in planning assistance through technical evaluation of relevant numbers (PATTERN). It establishes an objective, arranges the means necessary to achieve the objective in several levels of the system, and helps visualize the relations between the goal and each means. This technique effectively promotes decision making in technological developments.[6]

Reverse PERT (program evaluation and review technique)

PERT is an effective technique used in developing a new product. It relates project plans and pertinent research to the development goal, clarifying where to start.

PERT networks are used in making day-to-day plans. When we plan a route to a final goal in this manner, however, we often become exhausted along the way. The reverse PERT proceeds backwards from the final goal, reversing the order of steps used to create a PERT network.[7]

A reverse PERT begins with the final goal of development and enumerates all the resources needed to attain that goal, such as technology, parts and materials, production methods, measuring methods, and so on, including resources that have not yet beendeveloped. It then arranges them in reverse order until the practical means that can be implemented are reached. In this manner, the route to the goal can be charted.

Decision tree

A decision tree is a technique used to facilitate decision making. It represents the subject matter in the form of a tree with numerous branches and twigs. A decision tree contains points that can be decided on by the decision makers (determined points of selection are indicated by □) as well as points that are decided on by outside elements (probable points of selection are indicated by ○). These points are connected one by one to form twigs and branches that represent various phases of the process, leading to a final result. Ifthe probability of phases or events and the gain derived from the anticipated results can be quantified, the most suitable "decision" can be calculated.[8]

Fault-tree analysis (FTA)

Fault-tree analysis as well as failure mode and effect analysis (FMEA) are techniques used to analyze and evaluate the reliability of systems. They were originally developed as techniques to analyze the causes of fatal accidents. Fault-tree analysis determines things that are unfavorable to an object system and their causes and connects them in parallel (*and*) or series (*or*) by logical gates according to their cause-and-effect relationship. It presents them all in the form of a tree, which helps visualize how the unfavorable phenomena occurred. Because fault-tree analysis arranges and analyzes occurrences and their causes by logical circuits, it is called logic diagram analysis (LDA).[9]

The YS technique

This technique was introduced by Yabiki Seiichiro as a method for solving problems that range from expansion of sales to the lifelong plans of newly employed staff members of a firm.[10]

The PC technique

This is a variation of the correlation-tree method that consists of four steps: consideration of purpose, prospective concepts, clearing of problems, and planning and communication.[11]

A few words on functional analysis in value engineering (VE)

The functional systematic diagram (FAST* diagram), used for functional analysis in value engineering, is probably the best systematization and application of the systematic diagram among control techniques. The concept of and procedure for functional analysis will be explained, in terms of the functional systematic diagram.

*FAST = Functional Analysis Systems Technique developed by Charles Bytheway in 1965.

Functional analysis and a functional systematic diagram

Functional analysis clarifies and evaluates the essential functions of a product and its component parts in order to discover ways to improve the product and its cost.

Functional analysis consists of a series of analyses: First, functions are *defined* in a way that clearly expresses the functions to be performed by the product and its component parts (sub-assemblies, units, parts, materials, etc.); next the functions of each component part are *arranged* systematically in the functional systematic (or FAST diagram) in terms of their goals-means relations to one another (see Fig. 6.10); finally the functions are *evaluated* to determine which parts of the product design can be improved to obtain the most significant results.

Steps of functional analysis

Step 1:

Determine the objective of functional analysis. Here the reason for conducting the analysis is established and the product structure and its components are clarified.

Step 2:

Define functions. The functions of a product and its component units are succinctly described in short phrases such as those shown in the box below. Keep in mind that there are functions as objectives and functions as means.

> Function = verb + noun
>
> To (*verb*) (*object*), e.g., "transfer torque," "connect components," "rotate assembly"

Step 3:

Prepare function cards. Each defined function is recorded on a small card (about half the size of a business card).

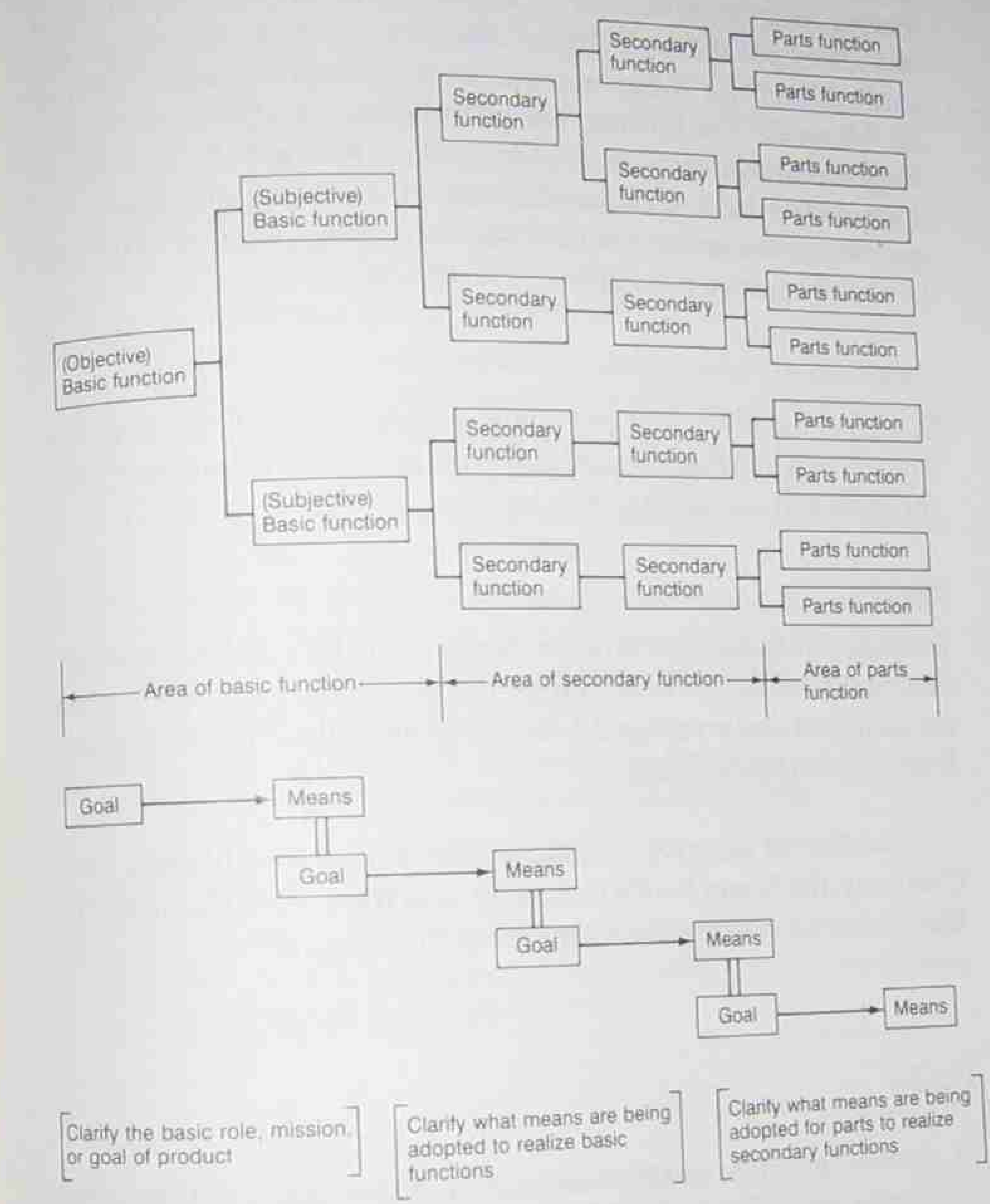


FIGURE 6.10
A functional systematic diagram

Step 4:

Determine primary- and secondary-level functions. Apply the following questions to each function card and determine its primary or secondary positions: (1) "What objective must the means on this card achieve?" (2) "What means is necessary to achieve the objective on this card?" The function answering question (1) becomes a primary-level function (objective) in relation to what the question is applied to on the card. In a similar manner, the function that answers question (2) becomes a secondary-level function (means) in relation to what the question is applied to on the card.

Step 5:

Construct the functional systematic diagram. This diagram is constructed by arranging the cards in terms of their primary or secondary levels and connecting them by lines.

Step 6:

Evaluate functions. Each function-series in the functional systematic diagram is evaluated: functions generating ideas for improvement are identified and investigated, and the potential effects of proposed improvements are evaluated. [12]

Finally, we are very grateful to the Matsushita Electric Parts Company, the Nissei Textile Company, and the Dainik Company for their permission to use the diagrams in this chapter.

Notes

1. Editorial Board for Quality Control (Ed.), *Understanding and Promoting Quality Design* (Tokyo: JUSE Press, Ltd., 1967).
2. Futami Yoshiharu, "Analysis of Quality Characteristics Using the Function Analysis Technique," *Hinshitsu Kanri* (Quality Control), vol. 26, Supplement (May, 1975), pp. 208–212; Tsutsuya Yutaka and Futami Yoshiharu, "Technique to Sustain Management During Low Growth Periods – The Technique and Implementation of Value Engineering," *Kojo Kanri* (Factory Management), vol. 21, no. 4 (1975), pp. 6–63.

3. For more on the development of quality and quality functions see Mizuno Shigeru and Akao Yoji (Eds.), *Hinshitsu kino Denkai* (Development of quality and function) (Tokyo: JUSE Press, Ltd., 1978).
4. Futami, *Quality Control*, pp. 208–212; Yutaka and Futami, *Factory Management*, pp. 6–63.
5. Futami, *Quality Control*, pp. 208–212.
6. *See* Mizuno Keizi, *Jishyu Gijutsu Kaihatsu* (Do-it-yourself technology development) (Tokyo: Sanno Publishing Co., 1971) and Makino Noboru, *Gijutsu Nyumon* (Technology forecasting for beginners) (Tokyo: Nikkan Kogyo Shimbunsha, 1969).
7. Makino, *Technology Forecasting*, 1969.
8. *See* Saito Yoshihiro, "An Approach Toward Decision (1): Rationale and Technique," *Operations Research*, vol. 12, (August 1967), pp. 16–21; Saito Yoshihiro, *Kettei no Hanashi* (Decisionmaking) (Tokyo: JUSE Press, Ltd., 1969); R. V. Brown, A. S. Carl, and C. Peterson, *Decision Analysis: From Introduction to Application* (translated by Fujita Tsuneo) (Tokyo: Sanno Publishing Co., 1974).
9. Shiomi Hiroshi, *Kosaiseki to Shindan* (Breakdown analysis and diagnosis) (Tokyo: JUSE Press, Ltd., 1977); Japanese Union of Scientists and Engineers, Division of Machine Reliability, eds., *FMEA and FTA Explained* (Tokyo: JUSE Press, Ltd., 1978).
10. Yabiki Seiichiro, *Handan to Kettei no Kagaku* (Science of judgment and decision) (Tokyo: Japan Management Association, 1975).
11. *See* Mizuno Keizi, *Genjo Daha no Shikoho* (Getting to breakthrough) (Tokyo: Dobunkan, 1973).
12. Tamai Masatoshi, *Kinobunseki* (Function analysis) (Sanno Publishing Co., 1967); Futami Yoshiharu, *VE Manual*, 1978; Tsutsuya and Futami, *Factory Management*, pp. 6–63.

7

The Matrix Diagram

The matrix diagram method clarifies problematic spots through multidimensional thinking.

About the matrix diagram method

The matrix diagram method is designed to seek out principal factors from a plethora of phenomena concerning a subject under study. As shown in Fig. 2.8, it indicates the relationship of L factors and R factors at the point of intersection. Factors $L_1, L_2, \ldots, L_i, \ldots, L_m$ are arranged horizontally, and factors $R_1, R_2, \ldots, R_j, \ldots, R_n$ are arranged vertically.

The matrix diagram helps to expedite the process of problem solving by indicating the presence and degree of strength of a relationship between two sets of factors. By using the intersecting points as starting points, the design allows us to (1) explore the problem under study from two points of view and (2) build a base for further two-dimensional problem-solving.

The systematic diagram method discussed in Chap. 6 is used to clarify a problem when its causes and the methods for solving it can be explained in one dimension. When there are two sets of factors and methods, however, the matrix diagram method, which can correlate these to each other, is more effective.

Uses of the matrix diagram method

Many kinds of quality control and management problems can be addressed through applications of the matrix diagram method. Some examples are as follows:

1. A new idea is needed for new product development or for product improvement, and this is accomplished correlating the hardware and software functions of the product system to each other.
2. The substitution of a feature affects many quality requirements and generates complexity, for this reason a systematic diagram cannot be used for quality development.
3. A system of quality assurance must be established by clarifying the relationship between required quality characteristics and the related control functions of the quality assurance department.
4. The system of quality evaluation must be strengthened by clarifying the relationships among quality characteristics, testing and measurement items, and testing and measurement equipment.
5. Many factors that contribute to the production of a nonconforming item should be eliminated by clarifying the relationships between the defect phenomena and their causes (when the defect phenomena have common causes).

6. Matrix diagram method is also the best method of organizing data in order to apply multivariate analysis.

Types of matrices

Successful matrix diagraming depends on proper expression of the subject matter in the matrix chart. Various patterns of matrix diagrams can be constructed; the choice of a particular pattern depends on its purpose.

L-type matrix

This is a basic matrix diagram with one set of data expressed in two dimensions employing rows and columns. Figure 7.1 shows that it is composed of *A* factors and *B* factors. The L-type matrix can be used for associating goals and the means to achieve them, as well as for drawing conclusions about the relationships between consequences and their causes.

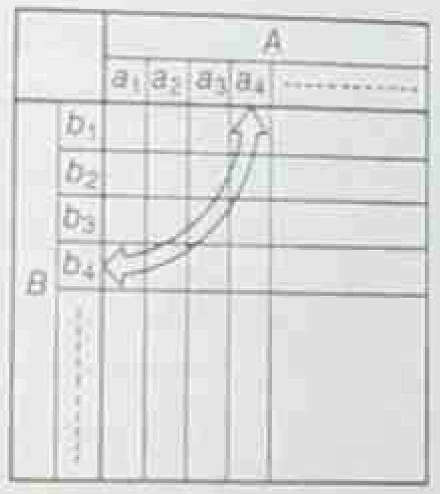


FIGURE 7.1
L-type matrix

T-type matrix

As Fig. 7.2 shows, the T-type matrix combines a matrix of *A* and *B* factors with a matrix of *A* and *C* factors. In other words, it is a matrix of *A* factors corresponding to *B* and *C* factors, respectively. The T-type matrix shown in Fig. 2.9, illustrating defect phenomena by cause and process, is a good analysis method for defect-reducing activities. Also, when exploring a new use of materials, this T-type matrix can be used to analyze ingredients and components of the material by characteristics and usage.

Y-type matrix

As Fig. 7.3 shows, the Y-type matrix is a combination of three L-type matrices (*A* factors and *B* factors, *B* factors and *C* factors, *C* factors and *A* factors). In other words, it shows how *A* and *B*, *B* and *C*, and *C* and *A* correspond to one another. For example, in its newspaper advertisements the Osaka Asbestos Slate Association used the Y-type

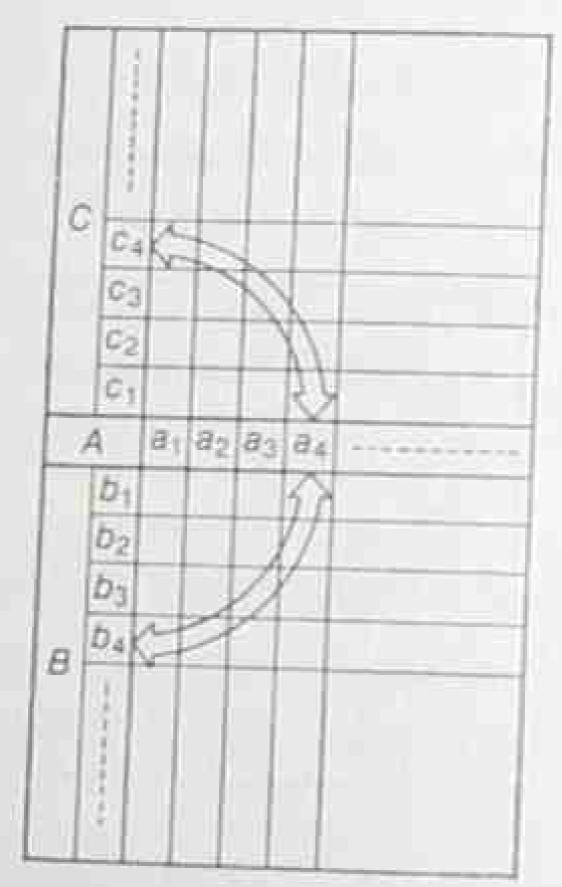


FIGURE 7.2
T-type matrix

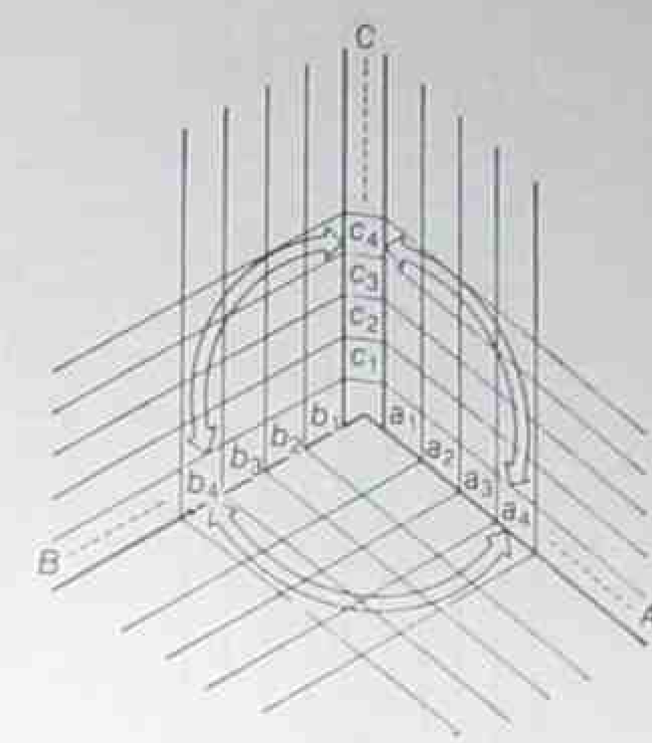


FIGURE 7.3
Y-type matrix

matrix to explain simply how material can be selected rationally by considering the place of its use, its features, and its shape.

X-type matrix

As Fig. 7.4 shows, this is a combination of four L-type matrices. This matrix shows the correspondence of four sets of factors, *A* and *B* and *AB* and *D*, *B* and *A* and *BA* and *C*, *C* and *B* and *CB* and *D*, and *D* and *A* and *DA* and *C*. Applications of this matrix are limited, however, as with the EDP (electronic data processing) system, it can be used to consider the correspondence of management functions, management items, output data, and input data.

C-type matrix

As Fig. 7.5 shows, the C-type matrix is expressed in a rectangular cube whose sides are represented by three elements, *A*, *B*, and *C*. The main feature of this cubic type of matrix is the "point of conception of the idea," which is determined by three elements of *A*, *B*, and *C* in three-dimensional space. This point of idea conception is difficult to

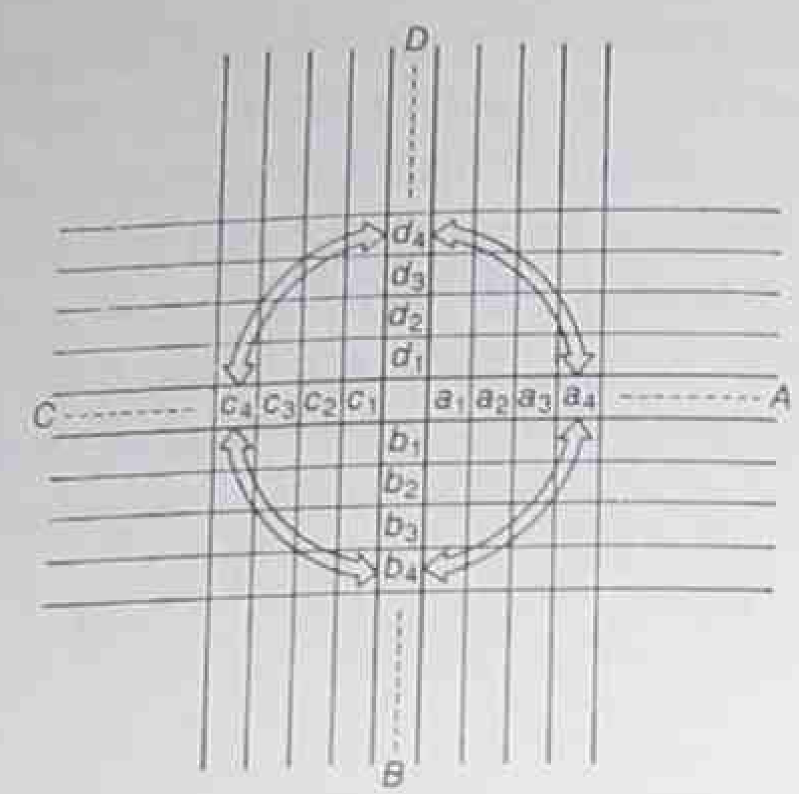


FIGURE 7.4
X-type matrix

show in Fig. 7.5, but it can be seen clearly in Fig. 7.6. This matrix may be better understood if the L-type matrices, containing the corresponding sets of *A* and *B*, *B* and *C*, and *C* and *A*, are combined with Fig. 7.5 or Fig. 7.6. Consider the relationship between Fig. 7.9 on the one hand and Figs. 7.8, 7.10, and 7.11 on the other.

Other uses of the matrix method can be designed by combining these five types of matrices. In short, the matrix to design is the one that is most appropriate for one's purposes.

The systematic diagram and the matrix diagram

The most important point in designing a matrix is deciding how to combine the sets of phenomena and the factors that correspond to them. There is no single way to combine sets of phenomena because the combinations depend on the nature of the problem under study. The examples presented later will illustrate some of the possibilities.

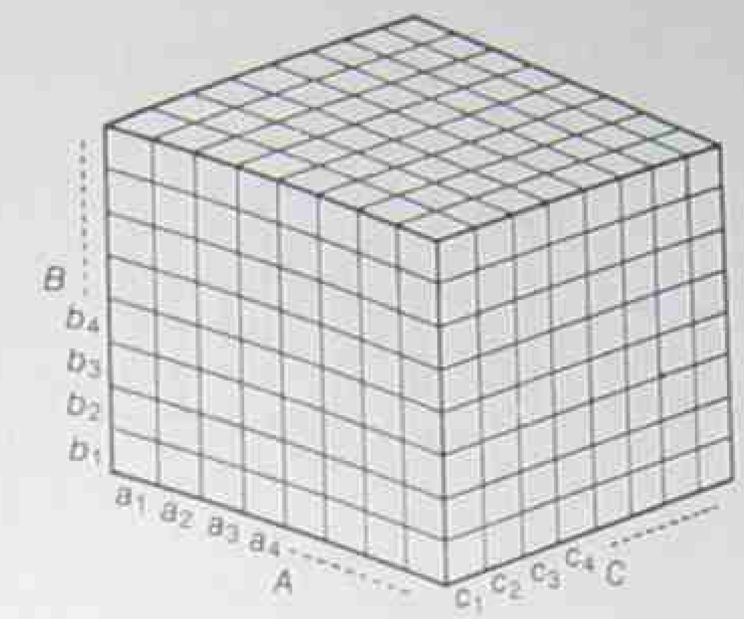


FIGURE 7.5
C-type matrix

Using a matrix diagram, the correspondence between phenomena is shown; then the relevant factors are developed to illustrate their level of significance. The systematic diagram is used to develop these factors. Another approach is to combine the systematic diagram with the matrix diagram, as shown in Fig. 7.7.

Applications of the matrix diagram

For a better understanding of matrix diagrams, a few examples of their application are introduced here.

Application of system products to functional design

Company Y planned to develop "a system of weight distribution of particles." The company decided to study the system to determine how to add new functions, how the company's own technology could be used to add new features, and how to increase added value.

The company analyzed the types of system requirements, drafted a new product plan, described the system in detail, and selected and defined one necessary function. The defined function

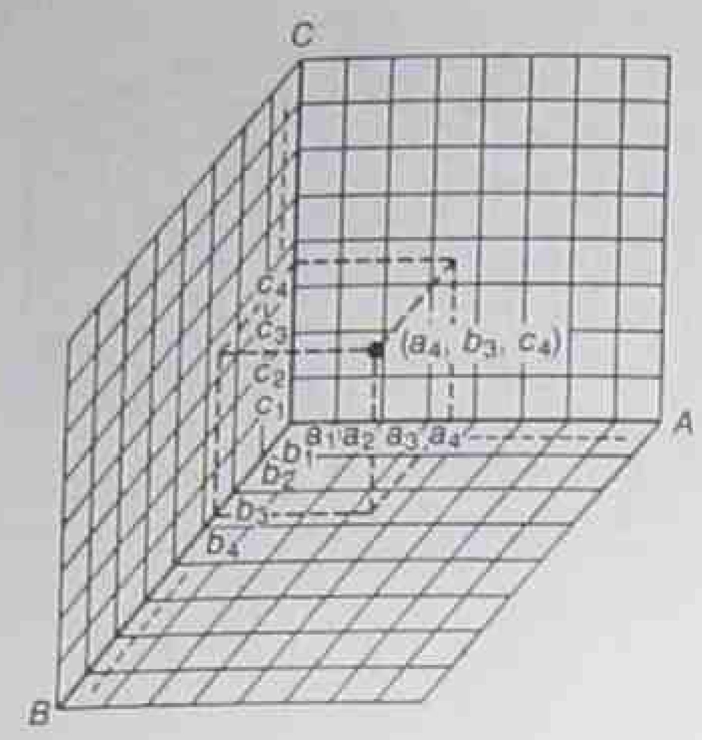


FIGURE 7.6
Developmental diagram of the C-type matrix

was divided into a hardware function and a software function; a functional system chart was drawn for each of the two functions. These functional system charts were combined to produce a matrix of the software-hardware functional system, as shown in Fig. 7.8. Furthermore, they decided to include function layout in the matrix, since the systems product layout function needed to be consistent with the current factory layout. They came up with the C-type matrix in Fig. 7.9. This type of matrix is called a function system matrix.

In the functional system matrix in Fig. 7.9, which is a C-type matrix, it is not easy to conceive of ideas, because the point of idea conception is a point in three-dimensional space. With the addition of Figs. 7.10 and 7.11, three L-type function system matrices were designed, including software-hardware, hardware-layout, and software-layout functions, and the point where they intersect was made the point of idea conception. The intersections marked ○ in the two charts are functions that specifically gave rise to many ideas. The intersections marked ● represent functions that correspond to the ● in the function system matrix of the entire system in Fig. 7.9.

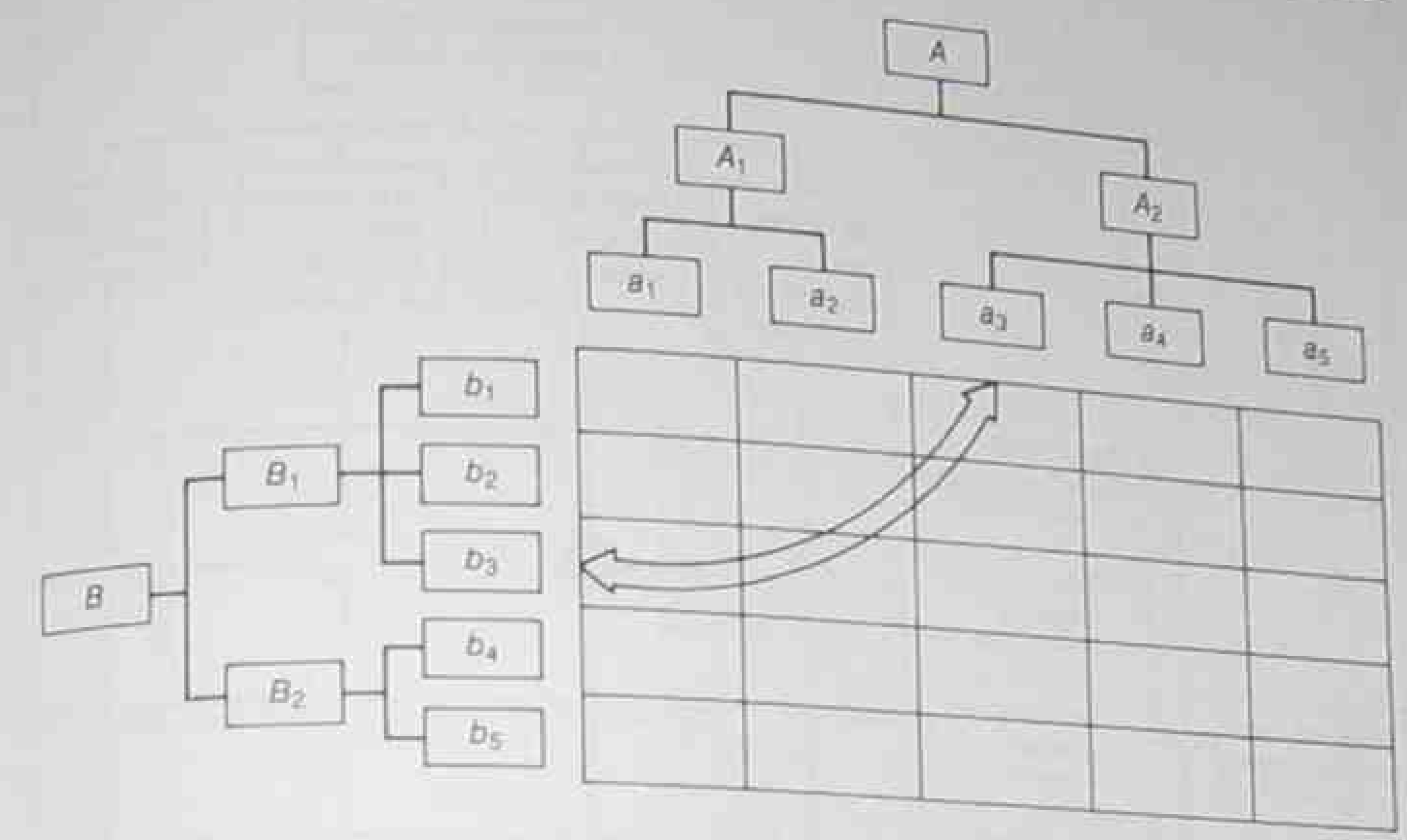


FIGURE 7.7
Combination of a systematic diagram and a matrix diagram

Although the content of the idea and its concrete manifestations cannot be discussed here, it is interesting to note that from this analysis, three new ideas led to patent applications.

When it is necessary to develop functions from various points of view, as in the case of a system product, designing a functional system matrix that uses the principle of matrix diagraming results in effective functional design and analysis.

Application to the quality development of component products

In the case of parts, a substitute feature may be used as a way of obtaining a required quality, and therefore, a systematic diagram can be used to explain its development. In this case, the process can be explained with a matrix, as in the following examples:

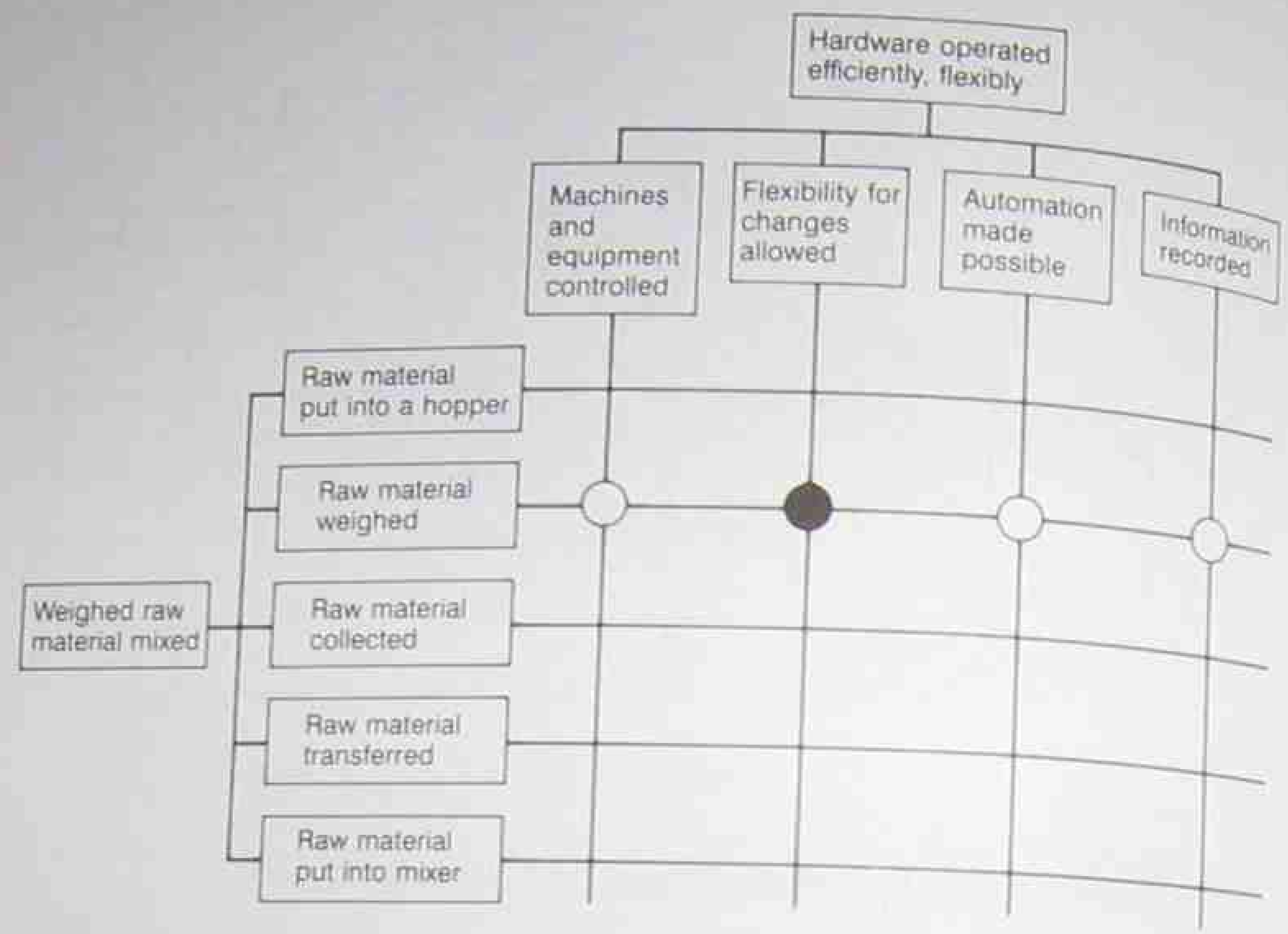


FIGURE 7.8
Soft-hard function system matrix[1] (L-type matrix)

In order to ensure a warranty for water-pipe fittings that had been put on the market recently, YK Company clarified the relationships among quality required by users, substitute product features, and manufacturing management items.

Connecting the required quality to substitute product features

First, two separate systematic diagrams were drawn for the quality of water-pipe fittings required by users and for the features of their substitute product, respectively, and these diagrams were arranged to correspond to each other in a matrix of required quality and

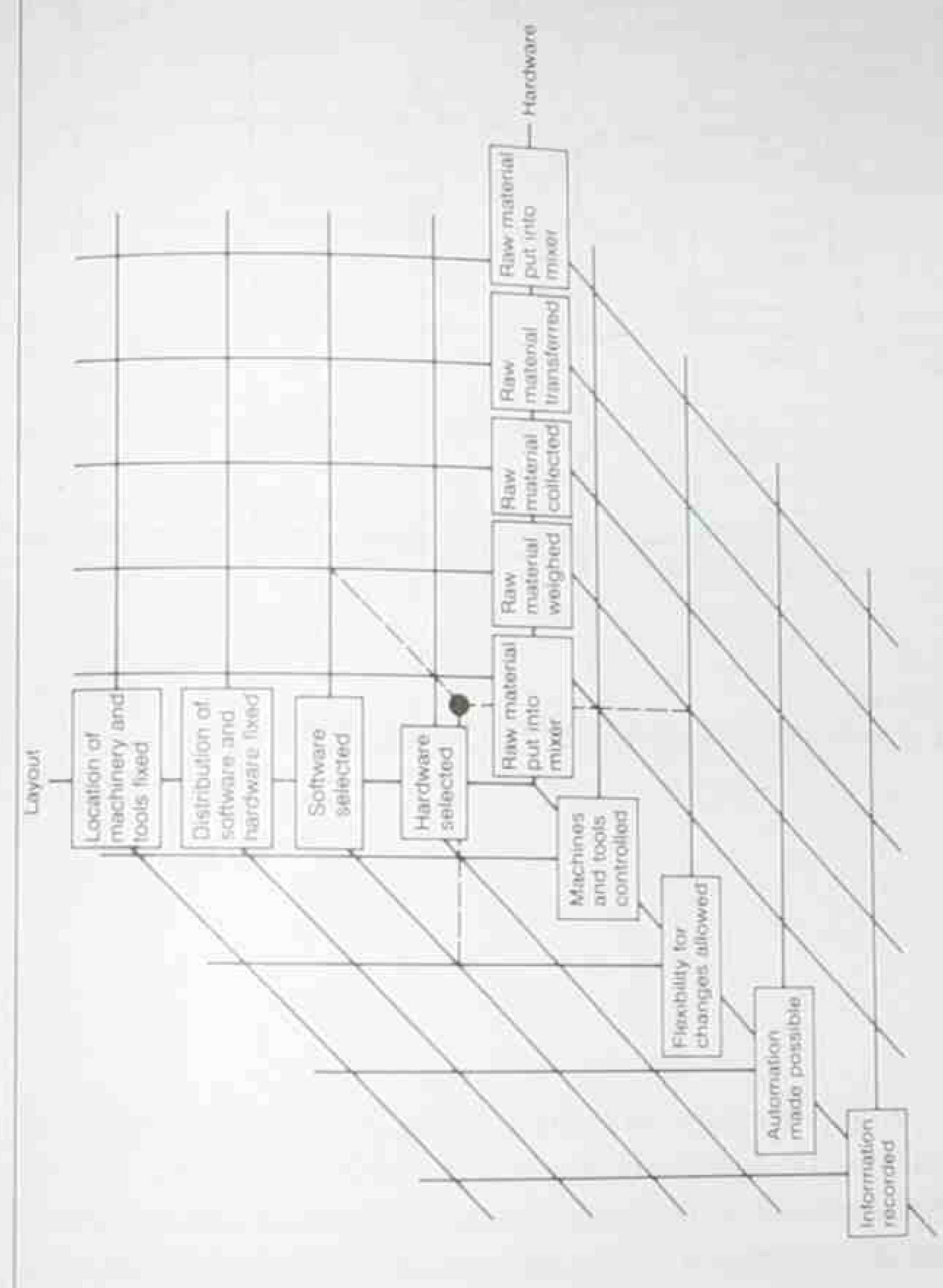


FIGURE 7.9
Function system matrix of particle weighing and combination system[2] (C-type matrix)

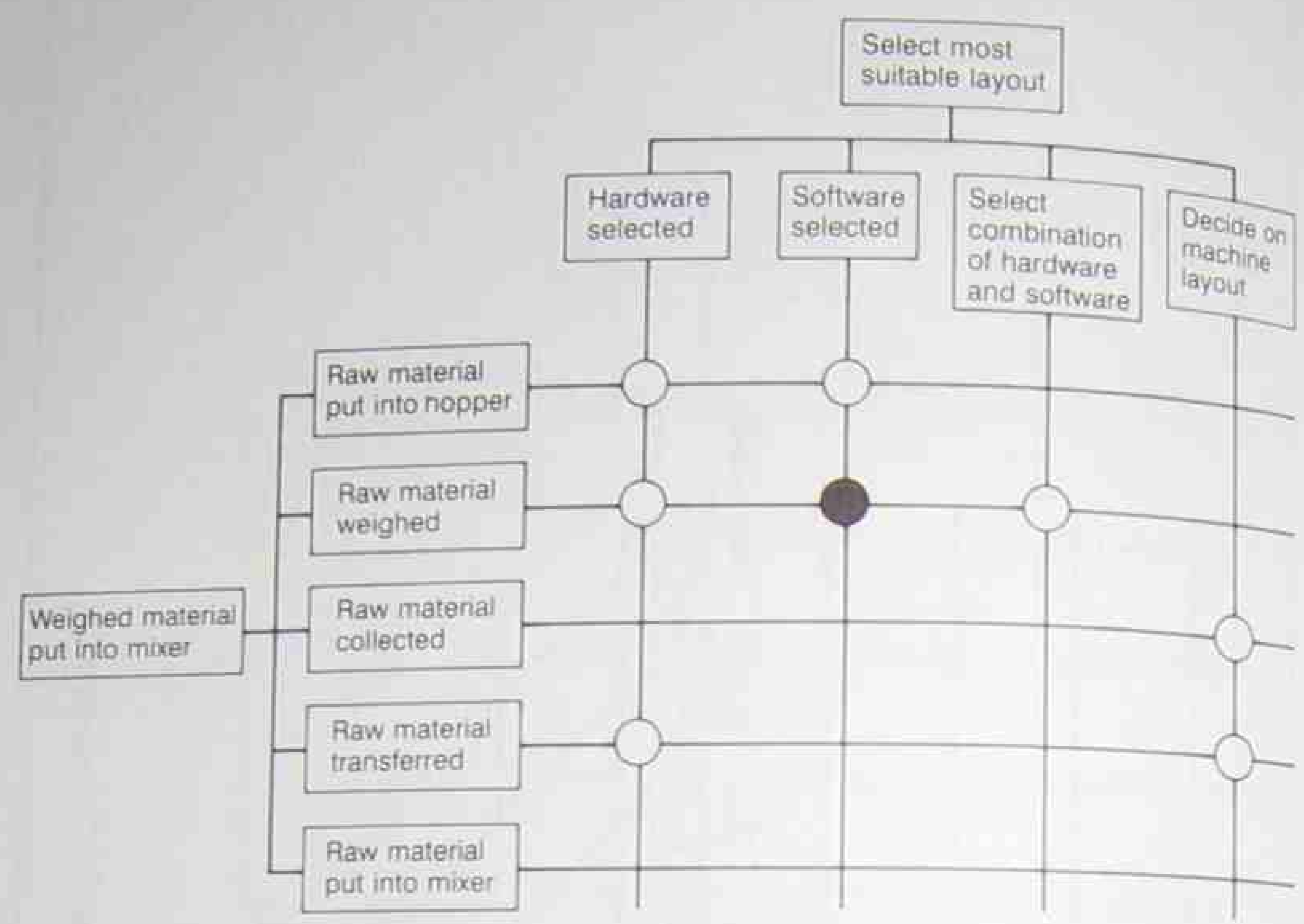


FIGURE 7.10
Matrix of hardware-layout function system[3] (L-type matrix)

substitute product features, as shown in Fig. 7.12. This matrix was used to evaluate the degree of correspondence between the quality required by users and the substitute product features. In addition, the degree of importance of the substitute product feature was sought. The acquired degree of importance was compared with the level of control, and measures were taken to eliminate the gap, if any existed between the two. Much improvement was made in terms of the large gap between the degree of importance and the current quality and management levels. The systematic diagram was used to develop a plan for this improvement.

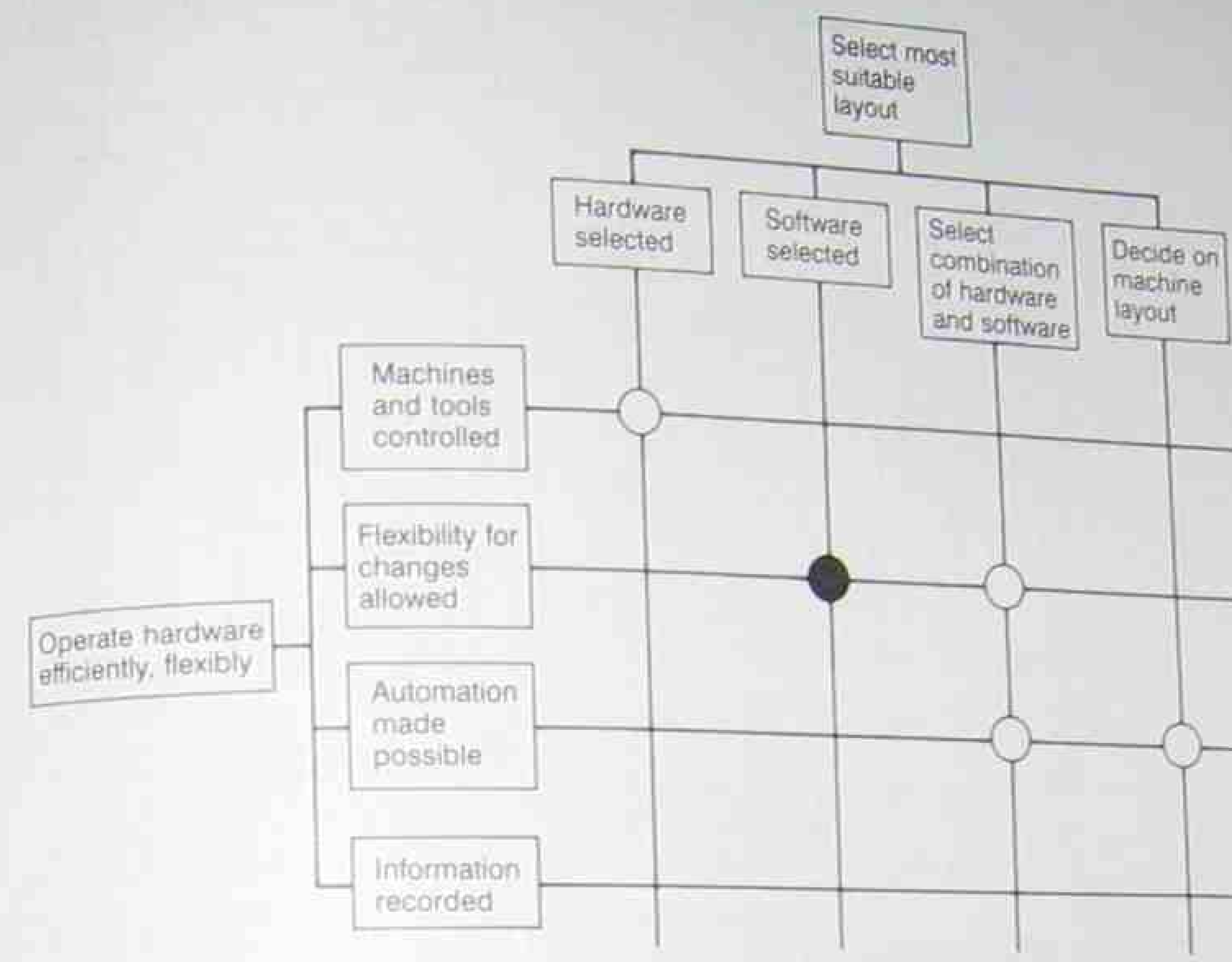


FIGURE 7.11
Matrix of software-layout function system[4] (L-type matrix)

Connecting substitute product features and manufacturing control items

Quality must be ensured during production. No matter how strict testing and inspection of a substitute product feature may be, they do not ensure quality. For this reason, a "matrix of substitute features and process control items" was constructed as shown in Fig. 7.13, which correlates substitute features with the process-control items of the QC process chart. In this matrix, the same degree of correspondence was evaluated, and the degree of importance of the manufacturing control item was obtained. As a result of reorganizing the

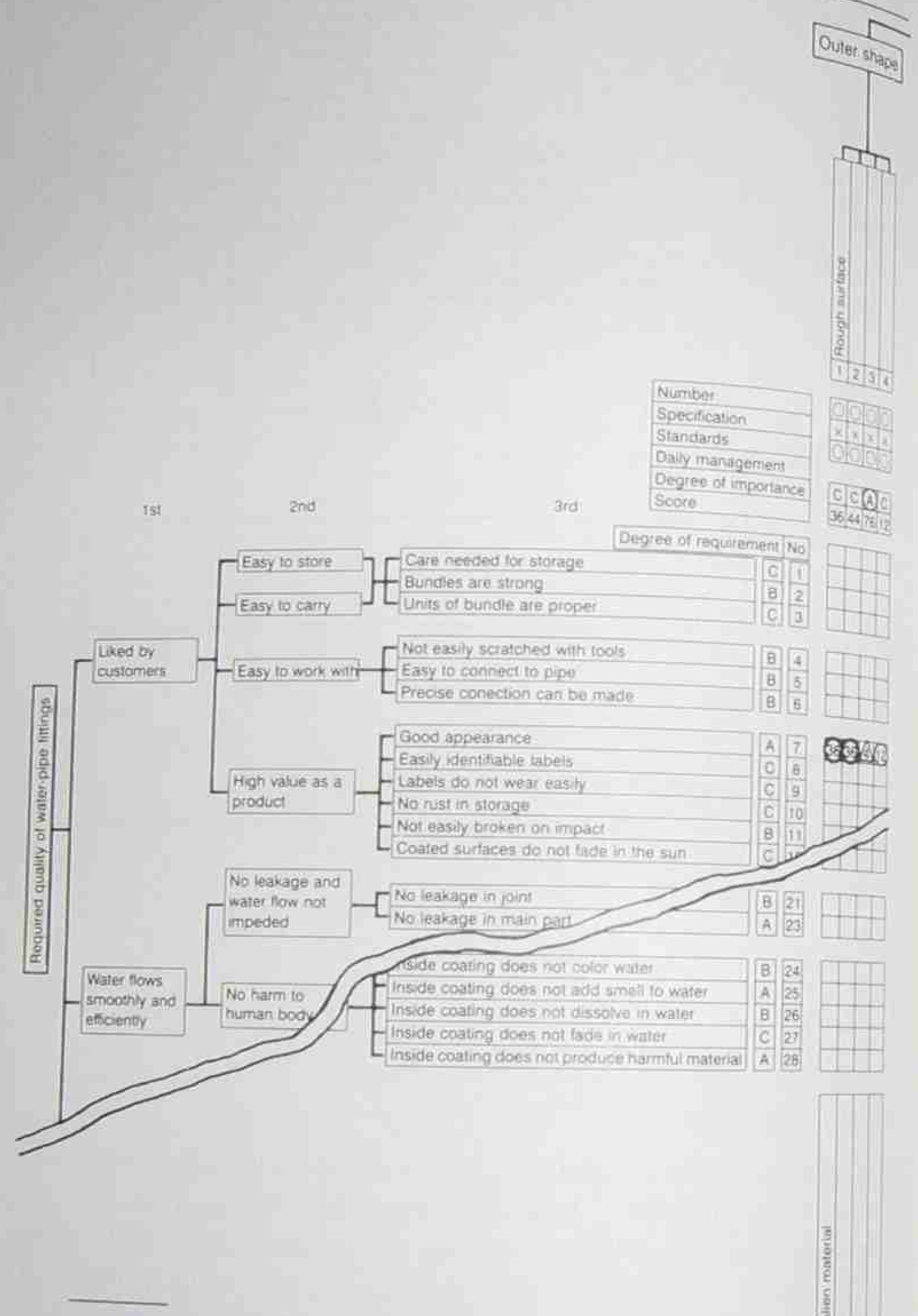


FIGURE 7.12
Matrix of required quality of water-pipe fittings and substitute features (systematic diagram and L-type matrix)

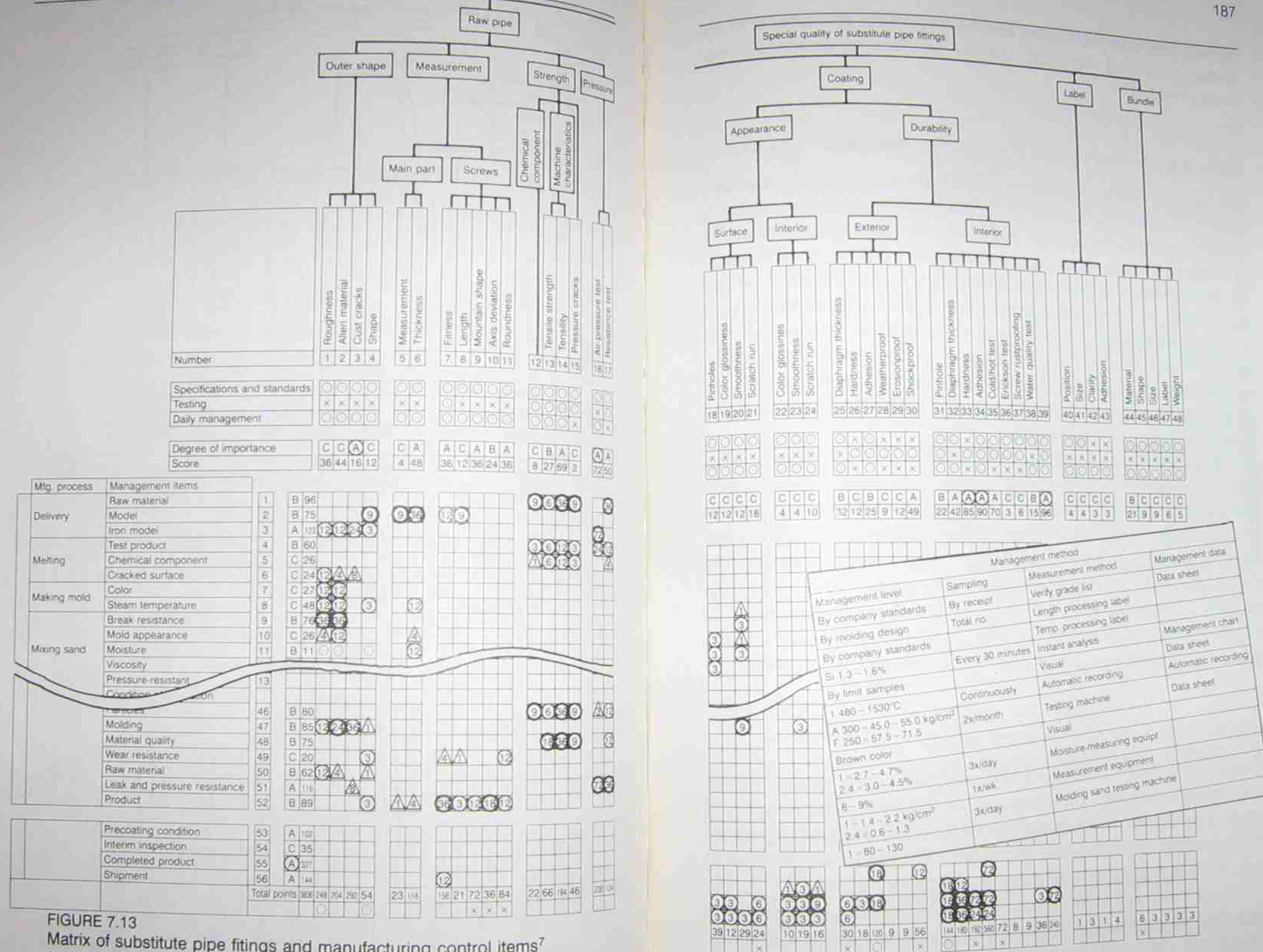


FIGURE 7.13
Matrix of substitute pipe fittings and manufacturing control items[7]
(systematic diagram and L-type matrix)

control system of production efficiency to reflect this degree of importance, control increased and the number of nonconforming items dropped.[5]

Application to improvement of the quality-evaluation structure

The testing department of Company N tests and inspects the quality features on many parts of an automobile brake system. The company undertook the current quality-evaluation program to strengthen the system and to make it more efficient because users' requirements for quality had become more strict. The development department had requested a greater number of tests on an increasing number of new products, and superior testing and measurement equipment was now available. In order to adequately plan for improvement of the system of quality evaluation, it was important to clarify the relationships among warranty features, test and measurement items, and testing and gauging equipment.

A warranty systematic diagram was used, and a T-type matrix was incorporated with the quality feature (control feature) corresponding to test and measurement items and test and measurement items corresponding to testing and gauging equipment, as shown in Fig. 7.14. Circumference rates were calculated by adding all the solid circles (●) in the rows and columns of the matrix; they were then arrayed in such a way as to provide a number of meanings and improvements, as shown in Fig. 7.15. From this, plans to make the quality system stronger and more efficient were implemented.

Application to a search for causes of poor quality

Company D manufactures book cloth and had a great deal of trouble with soiling. Recently, the company produced a large amount of light-colored cloth, which is easily soiled, and more than 10 percent of the product failed the final inspection. The company decided to find a way to reduce the soiling nonconformities by half.

Since there are many types of "soil," the company gathered inspection data and classified them into categories on the basis of the types of soil. These data were compiled into a pareto chart, as shown in Fig. 7.16. This pareto chart shows that one type of nonconformity

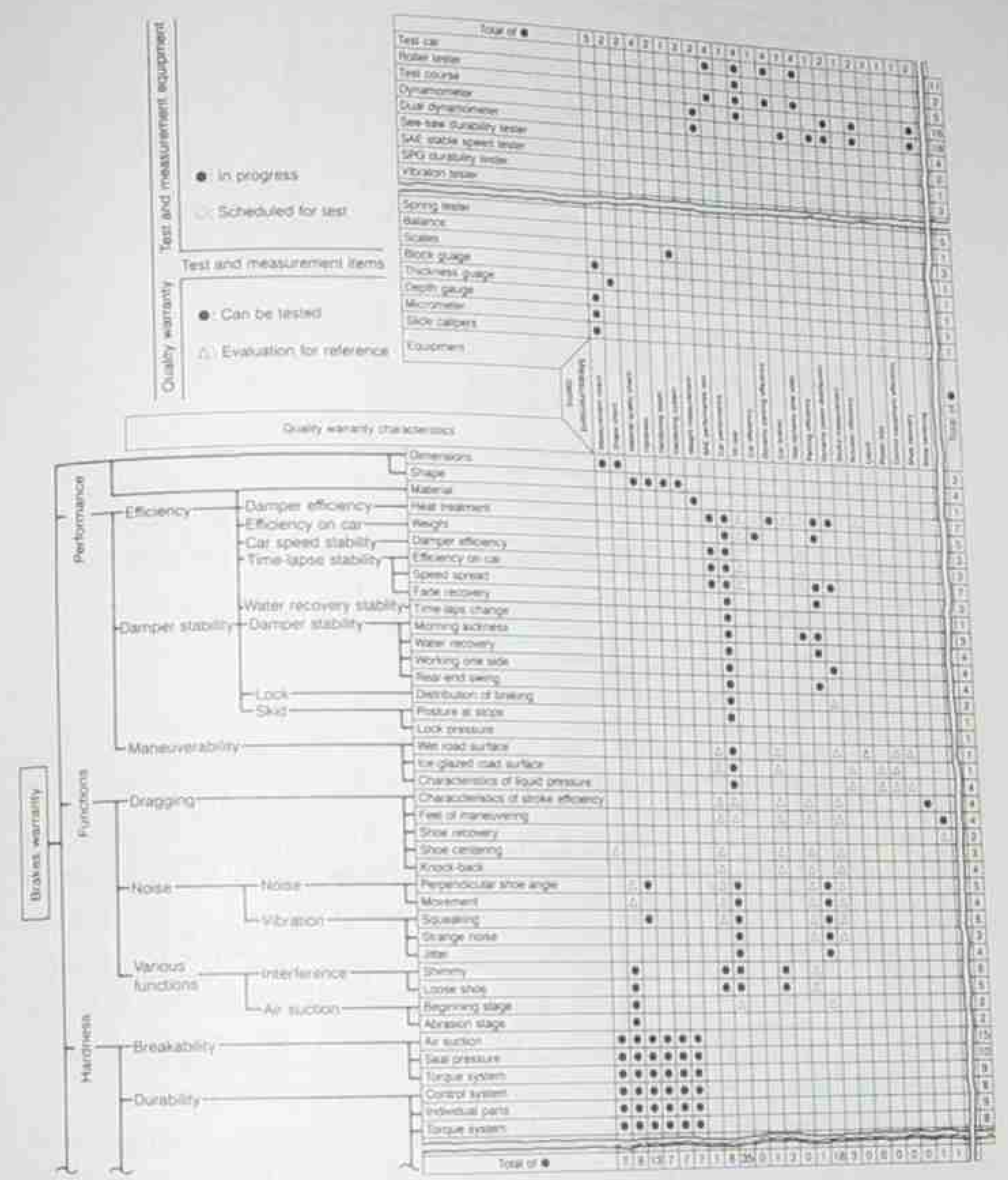


FIGURE 7.14
Matrix of car brake warrenty, test items, and test equipment (T-type matrix)[8]

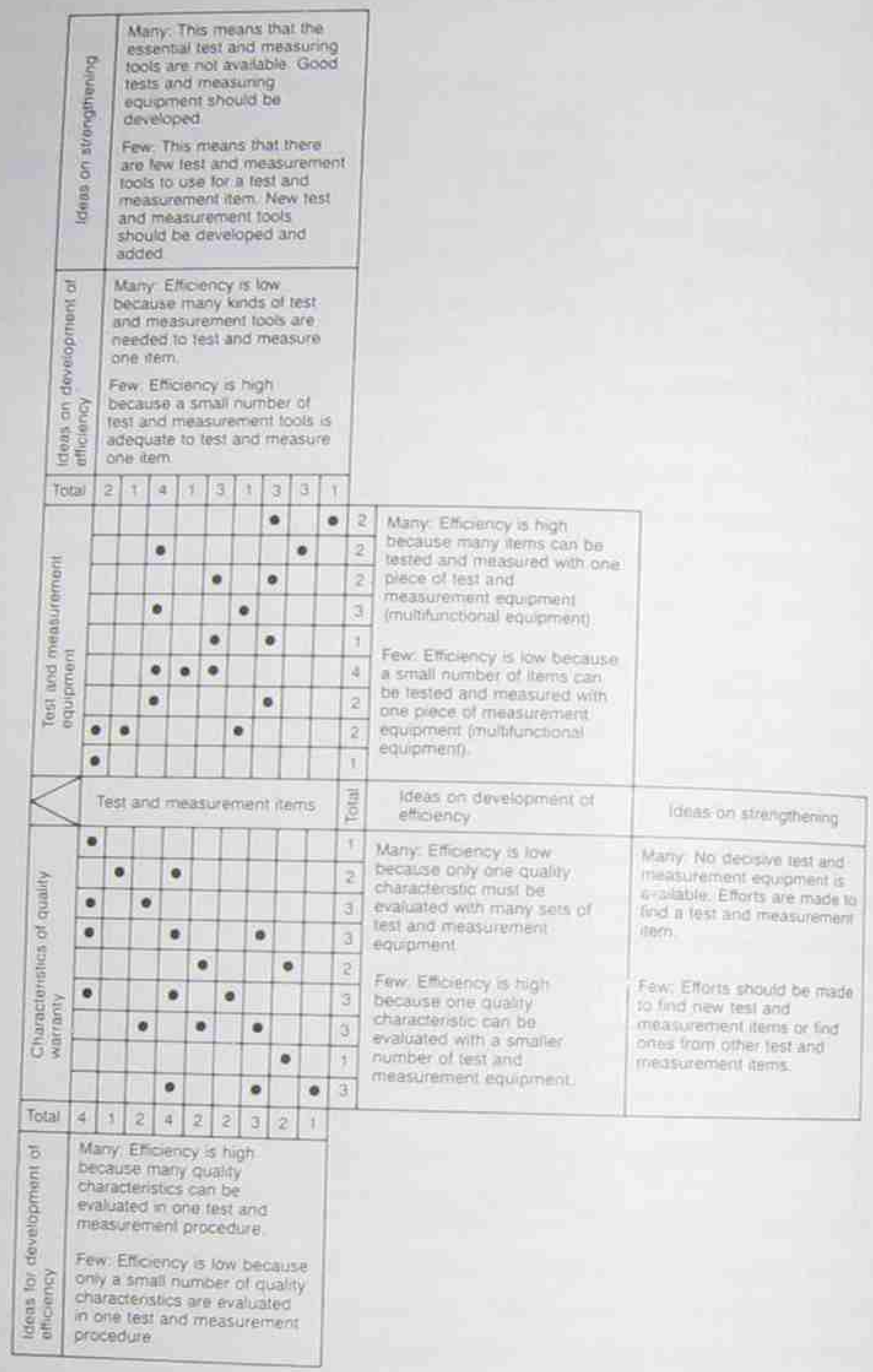


FIGURE 7.15
Ideas on improvement of quality-evaluation system

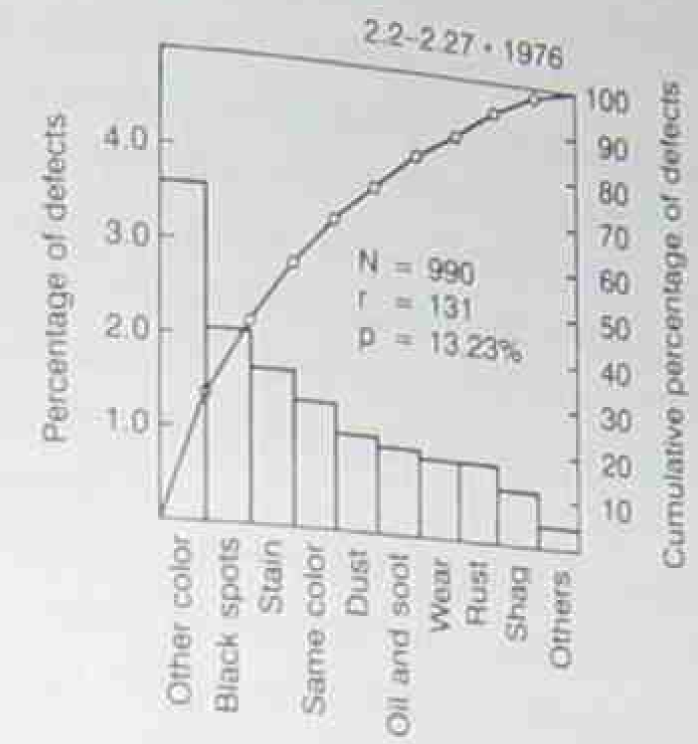


FIGURE 7.16
Pareto diagram of abnormalities due to soiling

represents the soiling event (e.g., unusual color, soiled spots, black spots), and the other type represents the causes of the soil (e.g., soiled with oil and soot or rust soil). The phenomena and causes of soil defects were separated, and an attempt was made to show the relationship between the two by correlating them on a matrix.

Furthermore, the relationship between the cause of soiling and the manufacturing process, which is the source of soiling, was shown in a matrix by drawing a correlation; a T-type matrix was designed to show the phenomenon, the cause, and the manufacturing process of the "soil defects," as shown in Fig. 2.9. After discussions about the problem possible solutions were considered and evaluated, and a list of countermeasures against soil defects was prepared, as is shown by Table 7.1. Notice the first countermeasure requires some time to examine and the third measure requires a great deal of discussion and does not include any concrete plans, but is simply a point of view.

TABLE 7.1
List of steps taken to prevent book cloth from soiling

Manufacturing process	Place	Causes	First countermeasure	Person responsible	Second countermeasure	Person responsible	Third Countermeasure	Person responsible
Paint mixing	Mixer	Oil	Clean blades					
		Rust	Replace cloth filter (2 × /mo.)		Replace drums with stainless steel			
		Soot	Cleaning (esp. blades)					
	Drum cans	Attendance (others)	Clean drum dents					
		Attendance (self)	Replace lids					
	Raw material	Dust						
Paints Machines Roll-out/roll-in		Oil	No oil lubrication with light-color paint		Elevate roller position			
		Rust						
		Dirt from hands						
		Cardboard	Recycle cardboard					
		Paper						
		Black paint						
		Static electricity	Place cloth on roll-out machine		Install device to vacuum cardboard debris			
	Pump hose	Rust						
		Soot						
		Dust						
	Roller coating	Rust	Moisten cloth					
		Crease			Scrape rust from gluing dryer Expand diameter of guide roller		Quick return device developed	
		Dew						
		Paint dregs						
		Black paint	Clean rubber roller with ammonia					
	Guide roller	Oil						
		Dew			Replace metal with bearings			
		Paint dregs						
	Cylinder	Soot	Clean air curtain (4 × /yr.)					
		Dust			Improve drain system		Cavity between joined parts narrowed	
		Paint dregs	Clean cylinder					
	Clip	Oil						
		Rust	Clean cover		Prevent particles from dropping on heater			
		Dew						

The company implemented the countermeasures identified on this list, and consequently, the abnormalities dropped to less than 5 percent and have continued to drop since then.

When the relationships between abnormalities and their causes are analyzed, many people think that the information must be expressed in terms of numerical data. With numerical data, quantitative and objective analyses are carried out with the statistical quality control (SQC) method. At least one to two months are needed to obtain numerical data, however, and a considerable amount of time and labor is required.

The matrix design shows the degree of relationship between nonconforming phenomena and their causes with such symbols as ○ (relationship) and △ (possible relationship) (see Fig. 2.9). Here the subjective opinions of the evaluators are likely to be involved, and the objective analysis possible with numerical data cannot be expected. On the other hand, if people with much experience are involved, they should be able to evaluate the data in a very short period of time because of their long experience. It is difficult to determine which method is better, but the kind of data obtained on the basis of years of experience and expressed in ○ and △ can be more effective than numerical data. The matrix design has the advantage of drawing highly reliable data from observations made over a long period of time, analyzing these data skillfully, and finding a solution for the problem under study relatively quickly.

Different techniques used with matrix diagraming

The benefits of matrix diagraming have been well integrated with many widely used techniques, some of which are introduced here.

Morphologic analysis

This is a technique for forming ideas. First, the existence of a problem is established, then variables in the problem are listed and analyzed at all possible levels, and these are combined to form an idea. In other words, an L-type matrix is designed with two variables and a C-type matrix is designed with three variables; the point where all theseva-riables intersect is a focal point for idea formation.

Matrix for determining objectives

This technique is one of many used for technological development. This technique sets an objective when technical forecasting is used as input and evaluation of a substitute plan is taken as output.[9]

The PESIC system

PESIC is an acronym for projects, elements, services, and information construction. It attempts to establish a system of management by arranging projects, products, elements, engineering, and services vertically, horizontally, and diagonally. This was developed by Tachiishi Electric Company, Inc., as a management system adaptable to all types of management.[10]

Finally, we would like to extend our deep appreciation to Yasukawa Electric Manufacturing Company, Inc., the Yoshitoshikaren Cast Iron Foundry, Inc., the Nisshin Textile Mill, Inc., and Dainik, Inc. for permitting us to use the charts reprinted in this chapter.

Notes

1. Futami Ryoji, "Uses of Matrix," *Kojo Kanri* (Factory Management), vol. 24, No. 9 (1978), pp. 6–15.
2. *Ibid.*
3. *Ibid.*
4. *Ibid.*
5. Yada Hiroshi, "A Quality Development System Applied at Sumikoto Joint," in *Proceedings of Regular Meeting of the Society for Study and Use of the Seven New QC Tools* (Tokyo: Japanese Union of Scientists and Engineers, 1978), pp. 3–22.
6. *Ibid.*, Futami, "Uses of Matrix," pp. 6–15.
7. Futami, "Uses of Matrix," pp. 6–19; *Proceedings*, 3–22.

8. Futami, "Uses of Matrix," pp. 6–19.
9. Makino Noboru, *Introduction to Technology Forecasting* (Tokyo: Nikkan Kogyo Shimbunsha, 1970).
10. Editorial Board for Handbook Study and Development, *Handbook on Research Development* (Tokyo: JUSE Press, Ltd., 1973).

Additional references

Mizuno Shigeru and Akao Yoji, *Development of Quality and Function* (Tokyo: JUSE Press, Ltd., 1978).

Futami Ryoji, "Using the Matrix to Come Up with Defect Causes and Countermeasures," *Kojo Kanri* (Factory Management), vol. 24, no. 3, (1978), pp. 15–19.

8

Matrix Data-Analysis

The matrix data-analysis method arranges the data presented in a matrix diagram so that the large array of numbers can be easily seen and comprehended.

The method

The matrix diagram arranges items in a column-row format, with the degree of correlation entered in the relevant column using symbols or numerical values. In Fig. 2.9, for example, the relationships among blemishes on printed cloth, the cause factors, and the place of occurrence or process are presented. Symbols are used in the relevant columns to indicate those factors believed to be important. In this case, concrete solutions can be developed through further examination of the causes of the blemishes, which, in this case, had been an area targeted for improvement.

Table 8.1 evaluates preferences among men and women, by age group, for 100 types of food items on a scale of 1 to 9, with 1 indicating the highest preference and 9 indicating the lowest. The numbers indicate the average preference values for each observed age group and food product. For example, the data show that while the preference for food product 1 varies by age group, the preference for food product 2 varies by sex. However, the data presented in Table 8.1 actually extend across a 10-row, 100-column matrix having 1000 items of data. Even if one were to examine these data thoroughly, the general extent of sex and age differences would still be unclear. Is there some way in which this can be organized in a clearer form?

One type of matrix data analysis is principal-component analysis. This technique is used in multivariate analysis and is detailed in the literature.[1] Unfortunately, because this technique is generally not well known, it is not used very often. Recently, however, an increasing number of Japanese companies have been introducing multivariate analysis methods, including principal-component

TABLE 8.1
Average Food Product Preference by Group

Group	Food Product 1	Food Product 2	⋯	Food Product 100
Men:				
15 years and younger	7.8	4.6	⋯	3.1
16–20 years old	5.4	3.8	⋯	2.8
21–30 years old	3.9	4.4	⋯	3.3
31–40 years old	3.5	4.0	⋯	3.0
41 years and older	3.0	3.5	⋯	2.5
Women:				
15 years and younger	8.1	6.2	⋯	3.9
16–20 years old	6.0	7.2	⋯	3.5
21–30 years old	5.4	7.5	⋯	3.0
31–40 years old	3.8	7.0	⋯	2.8
41 years and older	2.5	9.0	⋯	3.0

analysis, in both total company and organizational settings.[2] Furthermore, reports of the successful use of this method in the industrial sector continue to increase.

As stated in Chap. 1, even if it is impossible for all managers and staff to understand and use this method, it is still important because it is a method that is coming into greater use in the industrial sector.

Recently, along with the development of automatic calculation techniques, there has been an increase in the use of curves, rather than single values, to present data. Figure 8.1 shows a curve of the spectral distribution characteristics of a fluorescent lamp. Table 8.2 presents some of the spectral distribution characteristics of fluorescent lamps for each 10-nanometer interval. Similar readings taken for two other types of fluorescent lamps are also presented. Table 8.2 shows a 45-row, 3-column matrix data array similar to that in Table 8.1. When 94 types of spectral distribution characteristics of various fluorescent lamps, which vary in color quality, color, and luminescence, are entered, 45 rows and 94 columns of matrix data are obtained. For example, if one wishes to examine the spectral distribution shape characteristics for fluorescent lamps having good color characteristics using these data, the principal-component analysis method can be used.[3]

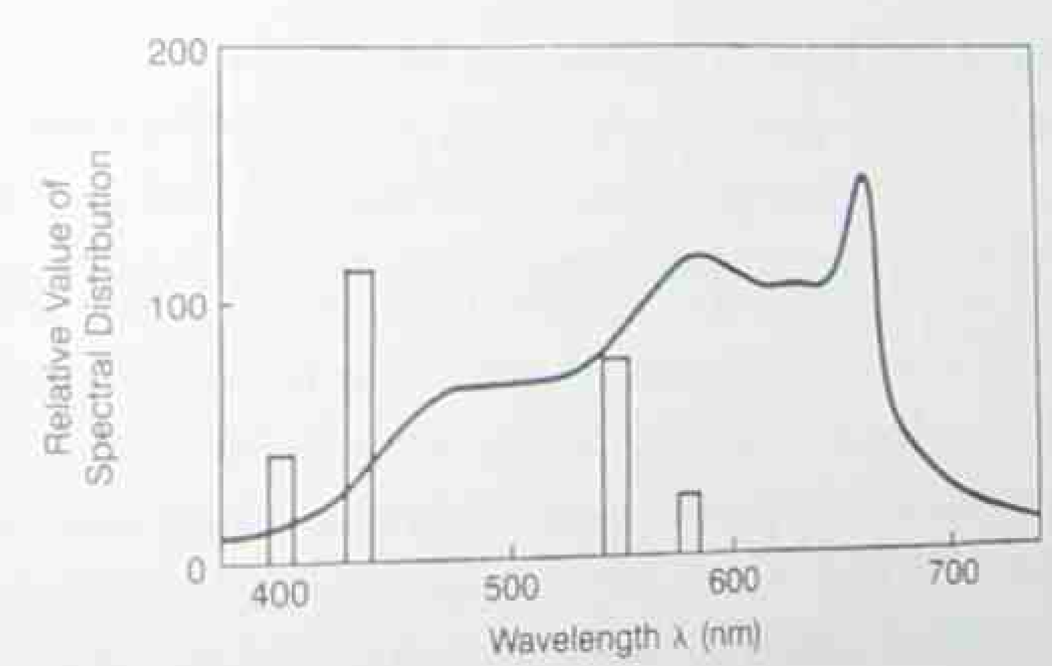


FIGURE 8.1
Example of spectral distribution of a fluorescent lamp (lamp F2 in table 8.2)

TABLE 8.2
Spectral distribution values for three types of fluorescent lamps

No.	λ (nm)	F1 S (λ)	F2 S (λ)	F3 S (λ)
1	380	5.4	10.7	23.0
2	90	5.6	12.0	27.5
3	400	5.8	13.9	33.4
4	10	6.1	16.8	43.6
5	20	7.4	20.8	55.0
⋮	⋮	⋮	⋮	⋮
18	550	82.3	88.4	99.8
19	60	100.0	100.0	100.0
20	70	113.2	110.4	101.1
21	80	125.7	116.0	102.7
22	90	112.9	115.3	102.7
⋮	⋮	⋮	⋮	⋮
38	750	1.0	6.6	11.0
39	60	0.2	5.2	9.0
40	70	0.0	4.0	7.3
41	80	0.0	3.1	6.0
42	404.7	27.2	42.3	77.7
43	435.8	84.0	112.1	182.4
44	546.1	77.7	77.7	100.8
45	577.8	23.7	23.0	29.1

Note: Values less than 404.7 nm correspond to the emission component.

Analysis of this type of matrix can be important in the fields of market research, new product planning and development, and process analysis. It is a technique that can be used to the greatest advantage in the PDCA plan and do stages when a large volume of data must be analyzed.

Calculation examples

In this section we will examine the results obtained through a calculation process, by computer, used in the principal-component analysis approach to a case study originally prepared by Toda.[4]

As shown in Table 8.1, an analysis was conducted based on average preference values presented in a matrix array for over 50

people categorized by sex and age. One-hundred food items, considered to be staple foods in the Japanese diet, were selected for this study. The breakdown of food categories selected is as follows: 19 types of staple foods, 4 types of soups, 10 types of meat dishes, 11 types of fish dishes, 3 types of side dishes, 14 types of drinks, 11 types of cakes and candies, and 4 types of fruit.

The purpose of this study was to determine how food preferences vary by sex and age. Should a difference in preference exist, group preferences for various types of foods should be made clear. In other words, the analysis did not determine whether each of the 10 age/sex groups preferred a particular food or not, but rather it illustrated the representative characteristics of these preferences instead.

Principal-component analysis was used to calculate representative characteristics. For example, using this method, the results of a science test, for subject i and student j, are expressed as x_{ij}. In this case, one representative characteristic of student j for the number of subjects taken, p, is expressed as:

$$w_j = \sum_{i=1}^{p} x_{ij} = x_{1j} + x_{2j} + \cdots + x_{pj} \qquad \text{(a)}$$

Usually, a weighted value (l_i), which varies by 1 for each subject, is added and the representative characteristic is calculated using the following formula:

$$w_j = \sum_{i=1}^{p} l_i x_{ij} = l_1 x_{1j} + l_2 x_{2j} + \cdots + l_p x_{pj} \qquad \text{(b)}$$

For example, just as a greater weight is placed on mathematics for a person who wishes to enter the technical section of a company, the representative characteristic allows for various selections to be made in response to a particular objective. Principal-component analysis is a selective measurement technique in which the representative characteristics can be mathematically calculated. These representative characteristics are independent. The weighted value l_i, which appears in the second equation, is referred to as a *characteristic vector*, and the representative characteristic corresponding to the base data for each representative characteristic is referred to as the *contributing ratio*.

An outline of the calculation process follows:

Step 1:

Each data item is assigned a term, x_{ij}, where $i = 1, 2, \ldots,$ and $j = 1, 2, \ldots, 100$. The term i corresponds to the group evaluated, and j corresponds to the food item.

Step 2:

The correlation coefficient matrix is then calculated for each observed group. These results are presented in Table 8.3. That is,

$$r_{ii}' = \frac{1}{n-1} \sum_{j=1}^{n} z_{ij} z_{i}' j \qquad \text{(c)}$$

where

$$z_{ij} = \frac{x_{ij} - \bar{x}_{i\cdot}}{\sqrt{V_i}}, \bar{x}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} x_{ij'}$$

$$V_i = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{i\cdot})^2, n = 100.$$

TABLE 8.3
Correlation matrix for each group

	Men					Women				
	15 years and younger 1	16–20 2	21–30 3	31–40 4	41 and older 5	15 years and younger 6	16–20 7	21–30 8	31–40 9	41 and older 10
2	0.871									
3	0.516	0.759								
4	0.370	0.604	0.852							
5	0.172	0.402	0.726	0.874						
6	0.938	0.821	0.517	0.358	0.208					
7	0.811	0.838	0.658	0.488	0.354	0.889				
8	0.615	0.709	0.698	0.620	0.523	0.746	0.894			
9	0.500	0.647	0.701	0.721	0.710	0.621	0.768	0.852		
10	0.330	0.457	0.558	0.632	0.748	0.493	0.642	0.773	0.911	

Note: The diagonal is 1; the upper-right portion was omitted for symmetry

Step 3:

The characteristic values and vectors are calculated using the correlation matrix. This part of the sample calculation was carried out using a computer. The results are presented in Table 8.4.

The analysis process is itemized as follows: First, as can be seen in Table 8.1, the question addresses the degree of preference for each food item by evaluated group. These data should first be collected into a smaller number of representative characteristic groups. The size of the characteristic value indicates that perhaps a roughly similar representative characteristic can be extracted.

Second, each characteristic value in Table 8.4 (i.e., 6.83, 1.76, and 0.75) expresses the degree of preference for a certain food item; the larger the number, the larger is the representative characteristic. The first item (the first principal component) has a characteristic value of 6.83, and because the total fluctuation in the case of a 10 × 10 correlation matrix is 10, the influence exerted (contributing ratio) by the first principal component (individual food items not yet entered) is 0.683 (68.3 percent). Similarly, the contributing ratios for the second and third principal components are 0.176 (17.6 percent) and 0.075 (7.5 percent), respectively. By combining three preference patterns, the total fluctuation in the base data is 0.934 (93.4 percent). This is

TABLE 8.4
Characteristic value and vector values

Evaluated Group	First Principal Component	Second Principal Component	Third Principal Component
1	0.286	0.446	0.194
2	0.331	0.240	0.336
3	0.323	−0.166	0.442
4	0.299	−0.359	0.375
5	0.261	−0.507	0.128
6	0.309	0.408	−0.084
7	0.344	0.253	−0.171
8	0.348	0.032	−0.290
9	0.346	−0.164	−0.322
10	0.303	−0.267	−0.522
Characteristic value:	6.830	1.760	0.750
Contributing ratio:	0.683	0.176	0.075
Cumulative contributing ratio:	0.683	0.859	0.934

shown in the cumulative contributing ratio row. The characteristic value in the case of 10 × 10 correlation matrix is generally 10; however, this is gradually reduced to first, second, and third characteristic values, and a large part of the total fluctuation can be explained in ordinary numbers.

Next, the significance of the three principal components is applied by characteristic vectors. In Table 8.4, the numbers in the left-hand column represent the observed groups given in Table 8.1. There are 10 values in the first principal component column. These are characteristic vectors. Each of these values shows the relationship between each observed group and the preference pattern (principal component). Regarding the first principal component, all the coefficients have the same sign and are roughly equal. In other words, this shows a common pattern of preferences for all the groups. In the second principal component column, the characteristic vector value decreases as it moves down the column from groups 1 to 5 and shows a similar pattern for groups 6 to 10. Here observed groups 1 to 5 are men, while observed groups 6 to 10 are women. In other words, the value of the characteristic vector changes from positive to negative in accordance with age for both men and women; thus the preference pattern appearing in the second principal component column indicates a difference in preference in accordance with age. Similarly, the third principal component column shows positive preference values in relation to men and negative values in relation to women. From this it is seen that for this component there is a difference in preference based on sex. The preceding results thus reduce the food item preferences to a general preference (contributing ratio 68.3 percent), a preference affected by age (contributing ratio 17.6 percent), and a preference affected by sex (contributing ratio 7.5 percent). Furthermore, the total fluctuation for the three principal components is 93.4 percent.

Finally, we shall attempt to compile a matrix array based on the preference patterns for each food item. This is calculated using the following formula which is available in a computer program:

$$w_{mj} = \sum_{i=1}^{10} l_{mi} z_{ij} \qquad \text{(d)}$$

With $m = 1, 2, 3$ indicating the three principal components, the principal-component score is calculated for each food item, j, which is assigned a value 1, 2, . . . , 100. The value of the vector corresponding to i is in Table 8.4. When the first and second principal component scores are plotted on the horizontal and vertical axes, respectively, a diagram such as that shown in Fig. 8.2 is obtained.

In this diagram, the food items generally preferred appear as one moves to the right along the horizontal axis, while those food items not generally preferred appear as one moves to the left. Furthermore, those items preferred by younger age groups appear as one moves up the vertical axis, while those foods preferred by older age groups appear as one moves down the vertical axis. By plotting the first and third principal component scores in a similar fashion, information on general preferences by sex can be obtained.

Principal-component analysis is a method in which matrix data, as presented in Table 8.1, can be expressed more clearly through numerical analysis, as shown in Fig. 8.2. In this sense, it is a diagramming method that includes analysis.

Conclusion

This section presented a working example of principal-component analysis. As is clear from the results presented, this is a method of organizing and categorizing raw data that are difficult to grasp. A topic for future study in this same area will most likely be how this information can be used in the development of food products.

0–1 data: investigating new material applications: product *A*

To develop and market a new material effectively, we must know, prior to entering the market, how the new material can be used. The opinion of a specialist is considered, and then a process begins in

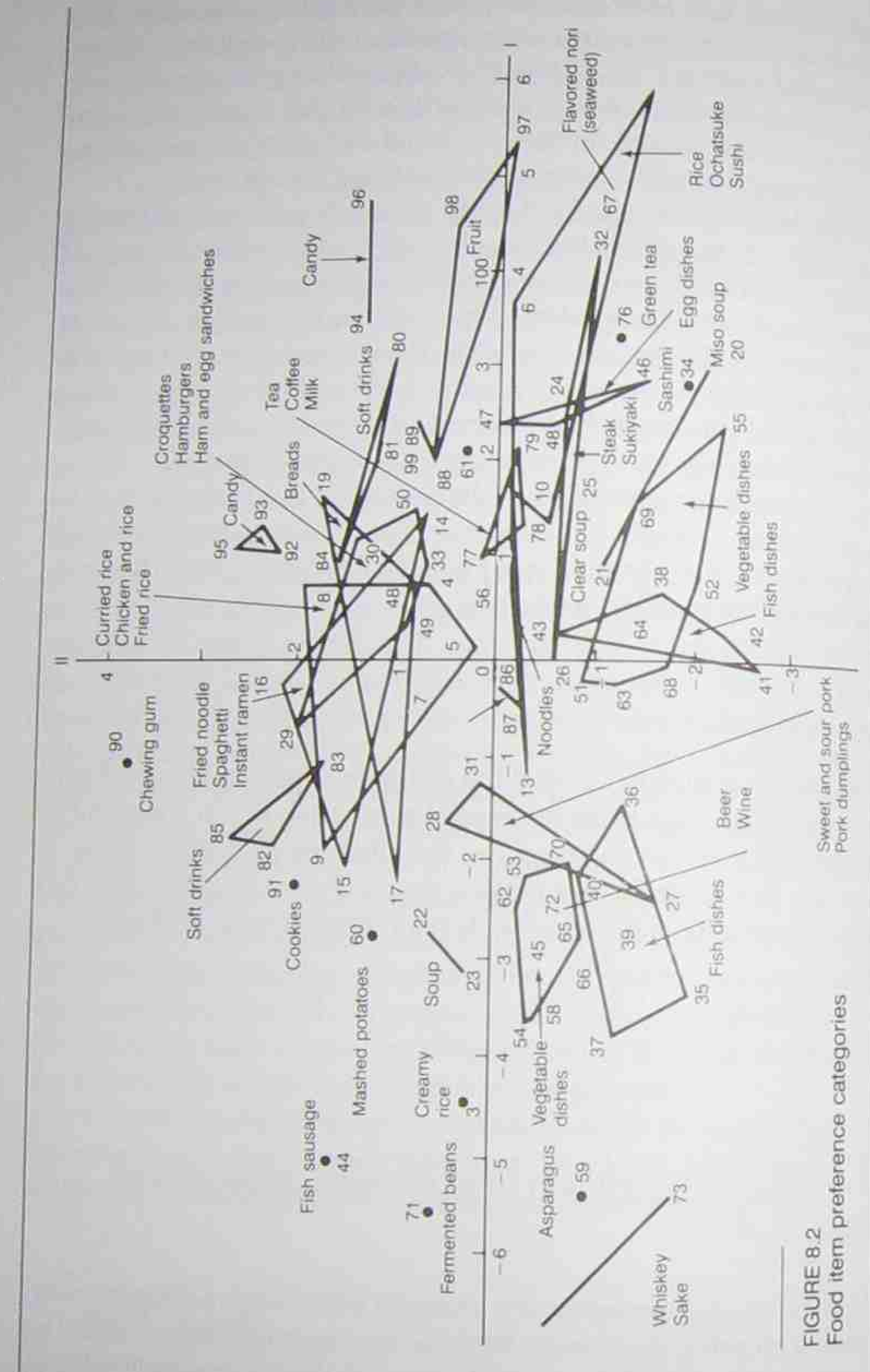


FIGURE 8.2
Food item preference categories

which trial products employing the new material are made and tested repeatedly. It is only after completion of this process that final decisions about applications of the new material are made.

Noguchi Hiroshi[5] has conducted various experiments involving a partially objective applications search process. Here we will provide an example of this method.

Prior to conducting an applications search for a new material, product *A*, a matrix array is compiled. This matrix shows the relationships among various applications, listed in the right-hand column, and desired characteristics of the material, across the top (see Table 8.5). In the table, a 1 is used to show whether the particular characteristic is desirable in relation to the application, while 0 indicates that the characteristic is not required. For example, in relation to men's summer wear, desirable characteristics include color-fastness in sunlight, stain resistance, ability to keep its shape, resistance to wrinkling, tears, and static, dryness, lightness, and breathability. The newly developed material *A*, which is listed in the last row of the table, satisfies these characteristics. Therefore, this could possibly be an area of application for the new material.

The principal-component method also can be applied to this type of 0–1 data matrix.[6] A diagram showing the locations of the first and second principal component scores has already been presented in Fig. 2.10. In this diagram, if the applications characteristics are similar, the scores will be closer, and conversely, if the difference in application characteristics is significant, the distance between the score locations will increase. Next, the principal component scores of the new material *A* are calculated. These are indicated in Fig. 2.10 with the symbol ○. As can be seen clearly in the diagram, applications for men's and women's summer wear, sports wear, pants, skirts, and raincoats are relatively close.

The desired characteristics shown in Table 8.5 also can be obtained using calculated data in addition to 0–1 data. Even if the matrix data appearing in Table 8.5 are complicated by having computed values and 0–1 values mixed together, principal-component analysis can be used.

TABLE 8.5
Various product applications and corresponding desirable characteristics

	Applications, Item *i* \ Desired Qualities, Item *j*	1 Color-Fast in Sunlight	2 Color-Fast in Wash	3 Color-Fast Perspiration	4 Color-Fast Dry Clean	5 Stain Resistant	6 Ability to keep shape	7 Wrinkle Resistant	8 Wash and Wear	9 Nonchafing	10 Nonpilling	11 Tear Resistant	12 Breathability	13 Warmth	14 Absorbency	15 Water Repellent	16 Lightness	17 Wind Resistant	18 Drape	19 Stretch Recovery	20 Elasticity	21 Slip Friction	22 Electric Charge	23 Nonflammability	24 Chemical Resistant	25 Gentle to Skin
A	1 Men's, summer	1	0	1	1	1	1	1	0	0	0	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0
A	2 Men's winter	1	0	0	1	1	1	1	0	1	1	1	0	1	0	0	1	1	0	0	1	0	0	0	0	0
A	3 Women's, summer	1	0	1	1	1	1	1	0	0	0	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0
A	4 Women's, winter	1	0	0	1	1	1	1	0	1	1	1	0	1	0	0	1	1	0	0	1	0	0	0	0	0
A	5 Skirts	1	0	0	1	1	1	1	0	1	1	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0
A	6 Trousers	1	0	0	1	1	1	1	0	1	1	1	0	0	0	0	0	1	0	1	1	0	0	0	0	0
A	7 Overcoats	1	0	0	1	1	1	1	0	1	1	1	0	1	0	0	1	1	1	0	1	0	0	0	0	0
A	8 Raincoats	1	0	0	1	1	1	1	0	1	1	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0
A	9 Office	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	10 Work	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0
A	11 Sports	0	1	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0
A	12 School	0	0	0	0	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
A	13 Home	0	1	0	0	0	1	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
A	14 Baby clothes	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	1	0	0	0	1	0	1
B	15 Dress shirts	1	1	1	0	0	1	1	1	0	0	1	1	0	1	0	0	1	0	0	0	1	0	0	0	1
B	16 Blouses	1	1	1	0	0	1	1	1	0	0	1	1	0	1	0	0	1	0	0	0	1	0	0	0	1
B	17 Sweaters	1	1	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0
B	18 Sports shirts	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1
C	19 Night wear	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1
C	20 Summer underwear	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	1	0	1	0	0	0	0	0	1
C	21 Winter underwear	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	1
C	22 Foundations	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	1
C	23 Baby wear	0	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1	0	1	0	0	0	0	0	1
D	24 Towels	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1
D	25 Handkerchiefs	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0
D	26 Hats	1	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
D	27 Mufflers	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	0
D	28 Ties	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0
D	29 Scarfs	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0
D	30 Gloves	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
D	31 Socks	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
D	32 Umbrellas	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
E	33 Carpets	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	1	1	0	0
E	34 Curtains	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
E	35 Tablecloths	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
E	36 Upholstery	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
E	37 Mats	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
E	38 Pillows	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
E	39 Cushions	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	Material *A*	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0

Applications in planning, development, and process analysis

An example of curve data analysis[7]

Fluorescent lamp spectral distribution characteristics are shown in Fig. 8.1 and Table 8.2. Data involving a total of 94 types of fluorescent lamps on the market that have various color and color-reproduction characteristics were collected. Then the data were arranged in a matrix comprised of 25 wavelengths and 94 lamps.

On the basis of these data, principal-component analysis results for the average and three characteristic vectors, $\bar{J}(\lambda_i)$, $S_1(\lambda_i)$, $S_2(\lambda_i)$, and $S_3(\lambda_i)$, as shown in Fig. 8.3, were obtained.

Using these results, it is clear that all possible fluorescent lamp spectral distribution characteristics can be expressed using the following formula:

$$J(\lambda_i) = \bar{J}(\lambda_i) + k_1S_1(\lambda_i) + k_2S_2(\lambda_i) + k_3S_3(\lambda_i) \qquad (e)$$

In this formula, the coefficients k_1, k_2, and k_3 were determined, using characteristic techniques, from the values of the calculated fluorescent color (x, y) and color-reproduction characteristics (R_a).

On the basis of equation (e), a projection of the spectral distributions of fluorescent lamps was prepared reflecting those developed to date and those yet to be developed. The color-reproduction characteristics and luminescent efficiency were tested. The results for fluorescent lamps having the same color and various color-reproduction characteristics are shown in Fig. 8.4. Part (b) represents a fluorescent lamp that provides color characteristics close to those of daylight. Because part (d) shows characteristics that exceed the fluorescent light range of coefficients k_1, k_2, and k_3, this lamp was found to have the special characteristic of color reproduction that is better than that of daylight, which is both an advantage and a disadvantage.

Forecasting fashion cycles

Forecasting fashions is a major concern in many industries. Kawasaki Kentaro and others[8] have reported some interesting results in this regard.

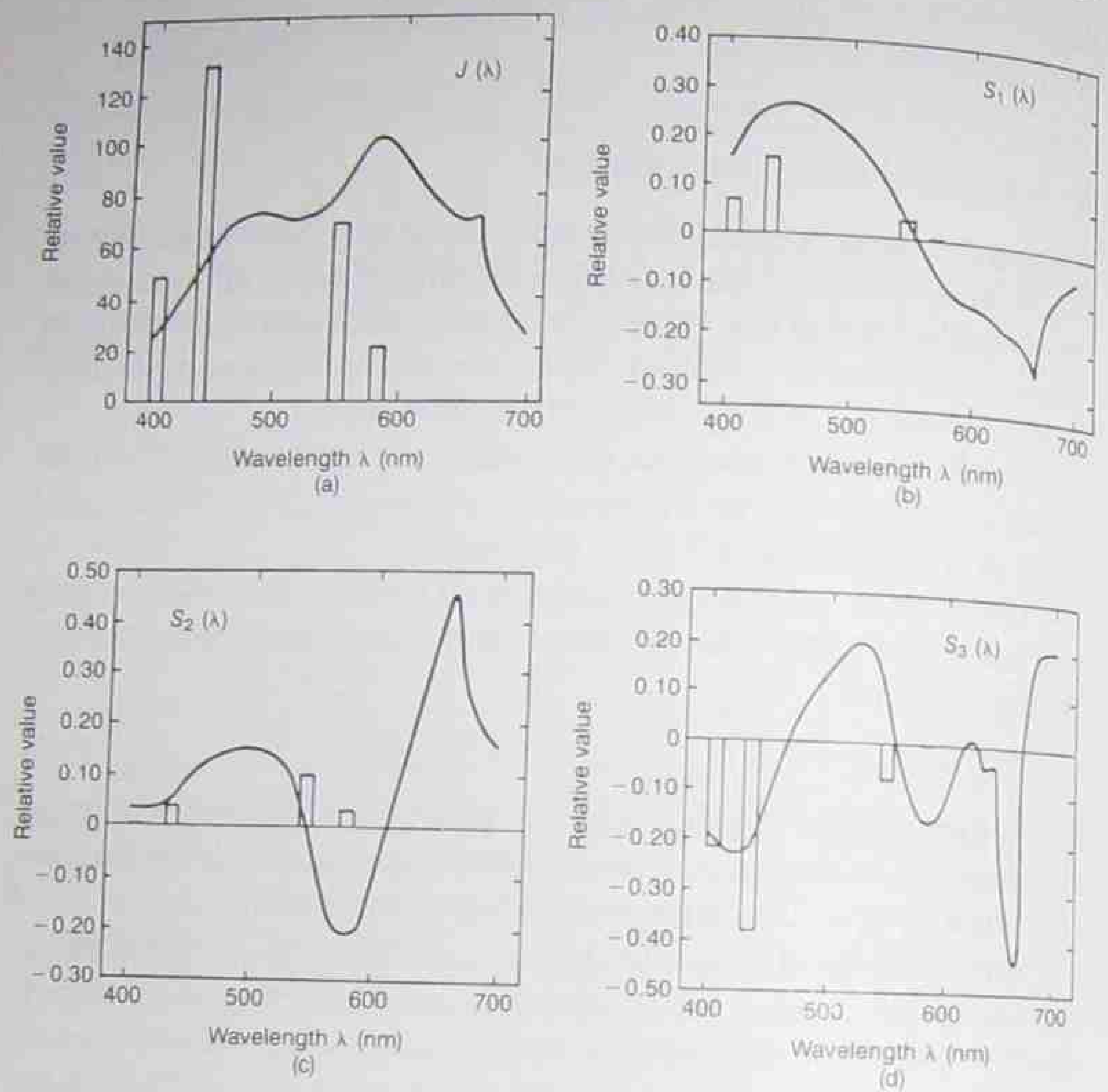


FIGURE 8.3
$J(\lambda_i)$, $S_1(\lambda_i)$, $S_2(\lambda_i)$, and $S_3(\lambda_i)$ distribution characteristics

Once, 53 representative fashions that appeared in *American Book* magazine were selected for each of the years from 1918 to 1974; these were then evaluated by 45 specialists employing 20 types of usage evaluation criteria. On the basis of these data, a matrix was constructed, the correlation array among the evaluation criteria was calculated, and principal-component analysis was conducted.

The result was that the following principal components were obtained: "contemporary elements," "feminine elements," and

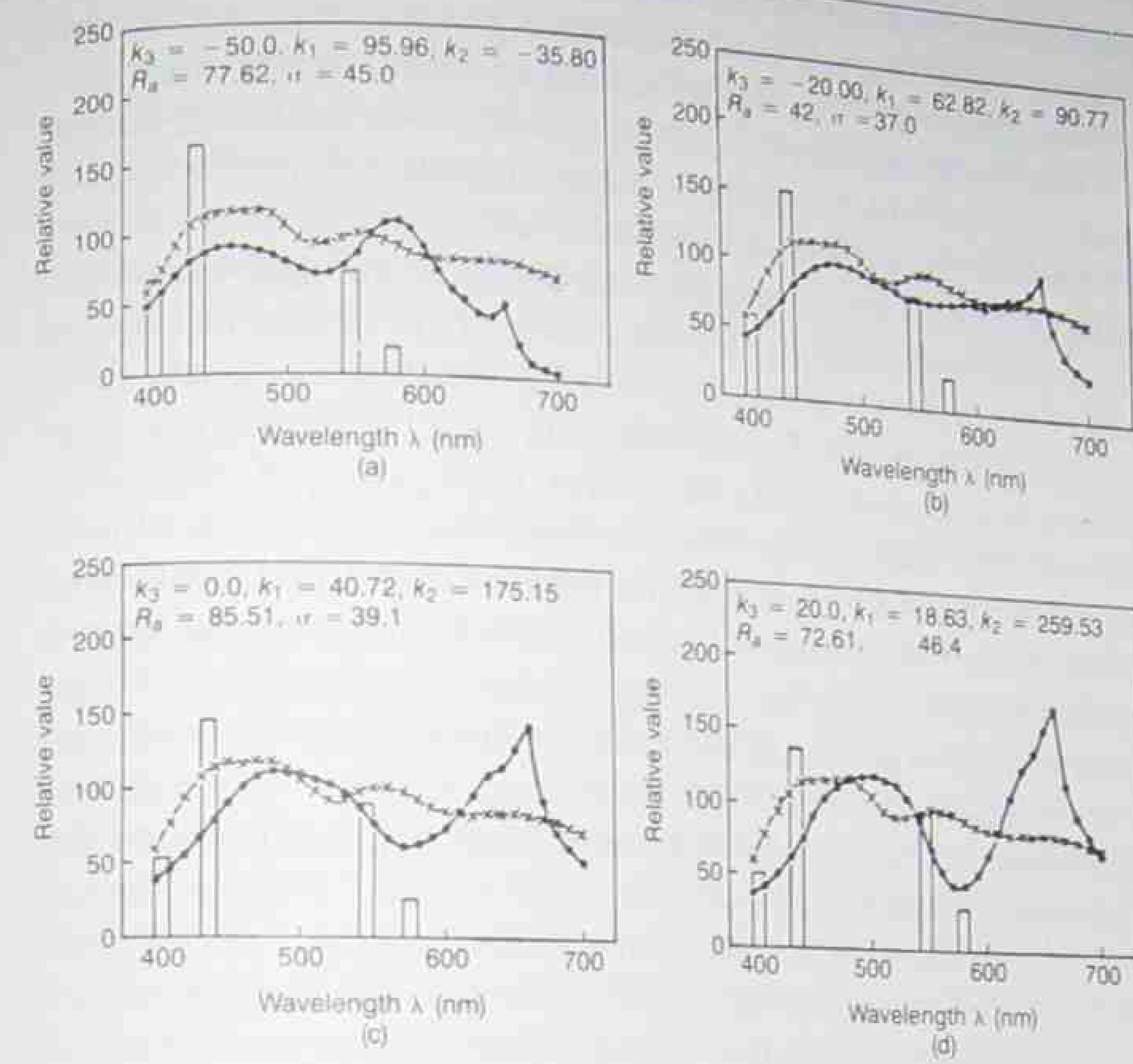


FIGURE 8.4
Spectral distribution for fluorescent lamps having daylight color and color-reproduction fidelity levels, R_a.

"unique elements." Forecasts based on simple insertion or recurrence of evaluation criteria and an evaluation of the next fiscal year were prepared. Using these, the following year's principal-component estimate score values were obtained.

Analysis of desired car styles

Tamanaka Hiroki[9] and others analyzed the responses of 1041 subjects on 12 items concerning 10 types of passenger cars that were being marketed, including car *A*, which, at the time, was receiving a great deal of public attention. It was discovered that the one essential element explaining the good evaluation given to this car was that structurally this car was seen as having unusual interior space.

In order to discover the relationship between the actual dimensions of the car's interior and the "psychological feeling of space," 26 different measurements of 29 types of cars, including car *A*, were taken. These data were then compiled in a 26-location, 29-car matrix.

A principal-component analysis was then carried out on the basis of a correlation matrix between the measurements and three principal components. The contributing ratio for these three components was approximately 54 percent. Each principal component was examined individually for items that reflected overall width, sitting room conditions, and superior front-seat and back-seat room. Finally, following a detailed analysis of the locations of the component scores, the reason for the high evaluation given to car *A* was that although in a geometric sense the interior of the car was average for domestically produced cars, the roominess of the back seat was emphasized at the expense of the front seat but was offset by the sedan shape of the car.

Press manufacturing wrinkle analysis

Sumimoto and Kamimura[10] prepared an analysis of a method to prevent wrinkles in automobile front bumpers when using right-side press. Although both the left- and right-side presses produced good results, when a slight difference in essential standards or operation occurred in the right-side press, a wrinkle would appear.

A series of 20ϕ circles were drawn at 39 prescribed locations on steel plates, and the degree of circle deformation after pressing was measured. Then 18 sample pressings of the left-side front bumper and 27 samples of the right-side bumper were produced for a total of 45 samples. These data were compiled into a 45-sample, 29-location matrix for examination.

On the basis of the correlation matrix between the locations, principal-component analysis showed that the source of wrinkles was not located where the circles had been drawn, but rather was affected by the degree of deformation that resulted at pressings at other locations.

Finally, an additional analysis was conducted. The machinery, materials, shapes, and other factors having an influence on the degree of deformation at specific locations were examined, and a process-management method was established as the factors contributing to nonconformities were removed.

This is a good example of the use of a combination of characteristic technology and this type of analysis.

Other examples

It has been reported that despite a variation in the number of factors, a stable factor structure can be detected by applying principal-component analysis to the metallic surface nonconformity cause groupings obtained using major recurrence analysis. In other words, the defects occurring on the steel surface were analyzed using the major recurrence method, as a special value related to the nine variables believed to affect this. In order to extract a few factors from the nine variables, eight, six, and five types of variables were selected and three types of major recurrence methods were calculated. However, the deviation recurrence value differed greatly for each formula, and in some cases, positive and negative values were reversed.

Thus principal-component analysis was carried out using all nine variables, four principal components were extracted, and the representative characteristics of the factors were calculated. The surface nonconformity and these four representative factor characteristics were calculated and analyzed using the major recurrence method. Consequently, the cause factors influencing surface defects were made clearer.

At a photo-processing plant it was discovered that there were large variations in a process using a two-dye mixture and that the product light-absorption curve varied with each batch processed. Because of the kinds of variations, it was assumed that impurities were being introduced into the mixture or that variations in dye composition occurred. To determine the cause, the problem was analyzed using principal-component analysis.[11] Spectral absorption curves were calculated for a sample group obtained using an experimental production process. These curves were then compiled into a matrix comprised of wavelength and sample rows and columns.

On the basis of these data, the horizontal sum and product sum matrices for the wavelengths were calculated, an examination was conducted on the basis of the characteristic values and vectors, and contributing ratios were obtained. The results showed that the spectral absorption curve fluctuations among the samples were a result of variations in the composition of the two-dye mixture. It was thus determined that the introduction of impurities was not the cause. Furthermore, spectral absorption curves were obtained from experimental results calculated in conjunction with the characteristics technique. These results showed that the composition of the dyes varied in substance from the characteristics established by technical standards.

Conclusion

This chapter showed that by using matrices as one method of analysis to arrange data, more information can be obtained than could be obtained through a study of the basic data.

This type of method has already been converted into a computer program[12] and certainly provides advantages in the understanding of input and output data. For this reason, principal-component analysis should not be thought of as complex. It can be actively applied to those problems for which it is suitable and understood through experimentation. Using this process, the individual's technique can be strengthened when combined with a mathematical foundation. We trust that this method, including the different multivariate analysis techniques, will see more use in the industrial sector.

Finally, we would like to thank Mr. Toda and Mr. Noguchi for their permission to use their tables and figures.

Notes

1. *See* for example, Okuno, Kume, Haga, Yoshizawa, *Tahenryo Kaisekisuho (Multivariate Analysis Method)* (Tokyo: JUSE Press, Ltd., 1971).

2. Takabatake Kunihiko, "Tokei Kaiseki no Suishin" (The advancement of statistical analysis), *Hinshitsu Kanri (Quality Control)*, vol. 29 (May 1978), special issue, pp. 194–196; Nakaso Nobutada, "Tahenryo Kaiseki no Taikenteki Donyu ni tsuite" (Concerning the experienced introduction to multivariate analysis), *Hinshitsu Kanri (Quality Control)*, vol. 29 (May 1978), special issue, pp. 197–202.
3. Noya, Kurioka, Shokugaki, "Enshokusei to Metamerizumu ni Kansuru Kenkyu, Sono 3" (Research on performance and *metamerizumu*, *Shomei Gakkai Zasshi* (Lighting Study Group), vol. 27 (1968), p. 190.
4. Toda Tanaka, "Shokuhin no Shiko Chosa" (Study of Food Preferences), *Dai 9 Kai Kanno Kensa Daikai Hobunshu* (Report on the 9th Sensory Inspection Conference), vol. 10 (1968), JUSE; Tanaka Toda and Jinbo, "Shiko kara mita Shokuhin Bunrui" (Categorization of food by preference), *Takeda Kenkyu Sho Nenpyo* (Yearly report of the Takeda Research Center), vol. 27 (1968), p. 190.
5. Noguchi Hiroshi, "Shinsozai Yoto Tansaku no Giho" (Method for investigating new material applications), in *Nippon Hinshitsu Kanri Gakkai Dai 10 Kai Kenkyu Happyo Kai, Kenkyu Happyo Yoshishu (Proceedings of the 10th Japan Quality Control Conference)* (Tokyo: JUSE Press, Ltd., 1977), p. 20; and Noguchi Hiroshi, *Kansai Tahenryo Kaiseki Semina—Tekisuto, Jishirei* (Kansai multivariate analysis seminar—text and case studies) (Tokyo: JUSE Press, Ltd., 1978), p. 37.
6. Mizuno, Nagae, Kamisugi, "Shitsugobyo no Sogoteki Hyoron Chi 0–1 score—Shuseibun Seki o Ooyoshite" (Aphasia evaluation 0–1 score—application of the principal-component method), *Gyodokeiryogaku* (Behavior measurement), vol. 3, No. 2 (1976), p. 1.
7. See note 3.
8. Kawasaki Kentaro et al., "Fashion Dynamics 3" *Seni Seihin Shohi Kagakkaishi*, (Textile product study group newsletter), vol. 17, no. 14 (1976).
9. Tamanaka Kiroki, *Kansai Tahenryo Kaiseki Semina—Tekisuto, Jishirei* (Kansai multivariate analysis seminar—text and case studies), (Tokyo: JUSE Press, Ltd., 1975).
10. Sumimoto and Kamimura, "Jidosha no Seizokotei ni okeru Tahenryo Kaisetsu Ooyorei" (Application of multivariate analysis in automobile manufacturing processes), *Hinshitsu* (Quality), vol. 6, No. 2 (1976), p. 59.
11. W. H. Lawton and E. A. Sylvester, Self-Modeling Curve Resolution, *Technometrics*, vol. 13 (1971) p. 617.
12. For example, Iguchi Haruhiro, *Tahenryo Kaisetsu to Komputa Purogramu* (Multivariate analysis and computer programming) (Tokyo: Nikkan Kogyo Shimbunsha, 1972), p. 61.

9

The Process Decision Program Chart (PDPC)

The process decision program chart (PDPC) method helps us select the best processes to use to obtain desired results by evaluating the progress of events and various conceivable outcomes.

The method

In the areas of policy control and total quality control, we make an effort to plan step by step in order to solve problems and reach our objectives. However, changing conditions often do not allow us to act as we anticipate, so we are forced to alter our plans. This occurs especially when we face the difficult problem of quality. Solution processes can be classified in two ways:

1. A solution can be anticipated based on prior knowledge.
2. A solution cannot be anticipated at this stage due to insufficient knowledge and unexpected changes in conditions or events.

Much of the research and development of new products belongs in the second category. In the investigative process, we often obtain new knowledge and encounter totally unexpected phenomena. For example, in the case of an extensive oil leak at an oil refinery or a train derailment, we may face totally new matters that are not anticipated under normal operating conditions. These cases also belong to the second category. To resolve problems involving new circumstances and unprecedented occurrences, we must be guided in the right direction (i.e., toward completion of research, prevention of accidents). Therefore, whenever new information is obtained, we review our plans and consider possible alternatives.

The process decision program chart is one of the methods used to solve problems in operations research and is introduced here to deal with problems that occur during the process of total quality control. The PDPC directs progress toward a desired goal in the planning stage, or during the design stage in research and development, by anticipating undesirable conditions and results. Moreover, the PDPC is an effective means of moving promptly toward one's goal at the very stage of the process where we encounter an unexpected problem. Therefore, the PDPC is used to define the solution process when we are dealing with problems that have more than one possible outcome. The PDPC is effective in preventing serious accidents; therefore, it is known as a chart for the prediction of serious accidents.

PDPC in operations research

The PDPC method is introduced here as a qualitative model in operations research. The definition of the qualitative model is a "change in a phenomenon and the reciprocal correlation between the substance and the nature of elements related to that phenomenon." Qualitative models are expressed in many cases with geometric figures and tree charts. If a certain condition occurs as a result of a certain action, and if the condition is not desirable, a proper solution should be sought. Suppose, however, that the solution causes various new conditions. In such a case, a chart is used to express a chronological and systematic picture of a phenomenon through the

entire process.[1] Fig. 9.1 shows a PDPC for a Gemini space flight. The chart graphically illustrates the whole process from takeoff to recovery. For example, various alternatives are planned in the event of death of the astronauts right after ignition during the first stage. Of course, matters may not progress as anticipated. In that case, the PDPC should be redesigned. For example, if an emergency not included in the original plan in Fig. 9.1 occurs, the method of rescuing the astronauts without any serious harm, despite the disastrous situation, is promptly determined through communication between the astronauts and their base.[2]

The PDPC method and quality control

Now that the era of stability and prosperity has begun, competition among enterprises has intensified. To deal with this situation, our day-to-day operations have become more diversified and complex than ever before. The same can be said about problems of quality control. Conventional QC methods, where consequences are analyzed to determine actions, often provide the solution too late. Even if the process is initiated as soon as problems occur, small environmental changes might force the entire process to be redesigned. Without the ability to deal promptly with such situations, it is frequently too late to do anything, especially in dealing with problems of product liability.

In developing new products, it is important to introduce into the market in a timely manner products with high added value that respond to customer demand. For this reason, features desired by customers should be well integrated during the planning stage of new products. In the planning stage, it is also important to consider users and the environment in which they use the product in order to minimize adverse impacts. In addition, the process of planning production should be as effective as possible because of limited production periods and cost. In reality, however, unpredictable matters occur frequently and circumstances are likely to change over time. Conditions relevant to alternative solutions to any problem, both social and economic, change gradually. Under such circumstances, previous ways of dealing with problems, in which only

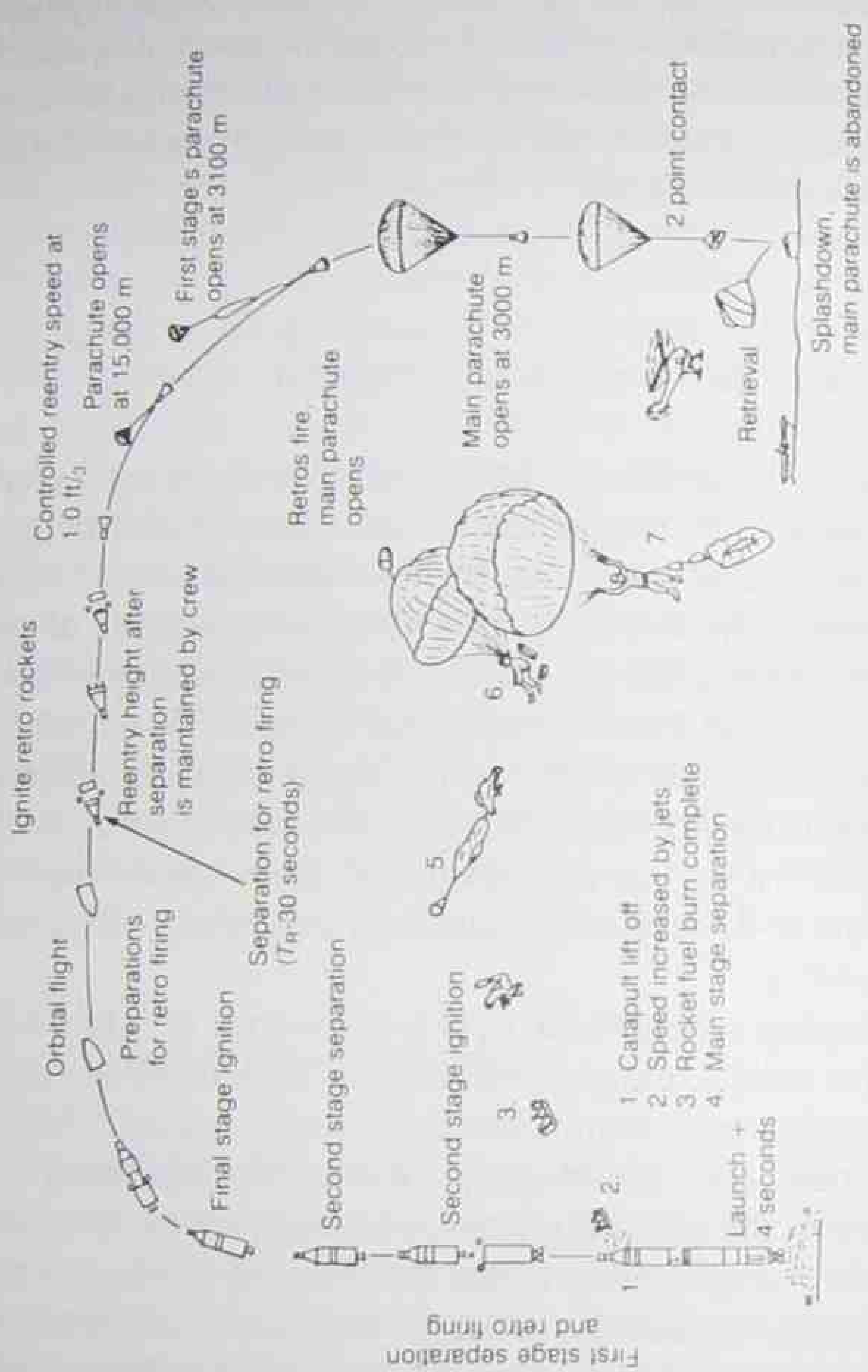


FIGURE 9.1
Graphic PDPC for a Gemini space flight
Source: From Jiro Kondo, *Operations Research* (Tokyo: JUSE Publishing Co., 1973), p. 131.

one solution is determined during planning, are not effective. We might not see a problem in the beginning, but later, in the face of changing conditions or the occurrence of unexpected results, we are forced to abandon our objectives because of slow response or lack of a proper solution.

As illustrated earlier, problems of development or product liability frequently put us in unanticipated or uncertain situations. To deal with these kinds of situations, a method that enables us not only to anticipate solutions, but also to respond to changing conditions is especially desirable. In light of this, the PDPC method is considered well suited. The PDPC method has no definite structural rules: Start with the present, suggest a possible solution under a conceivable future, anticipate undesirable outcomes, present a means of reaching a better result, and then decide on a course of action. If we are forced to face a new problem, information, or dimension that occurs during the process without previous warning or indication, we must review the process in a flexible manner, considering changes that accommodate the new conditions. Then, based on this review, we can proceed to redesign or correct the plan. Therefore, in the new era of quality, the PDPC method is quite useful for solving TQC problems.

Characteristics of the method

PDPC is a dynamic method

To solve a problem with defined parameters (e.g., building a house within a certain period of time), difficulties that occur during the process can be readily anticipated because of the building plan, which begins with planning and ends with the actual steps to reach the goal. However, problems often have to be solved under uncertain conditions. For example, if the elimination of defective products is set as a goal, it is impossible to determine exactly the process necessary to reach such a goal. Were it possible to do so, the problem would already have been solved earlier in the process. Therefore, the proper method for reaching a solution to the problem of product quality improvement remains to be discovered through new information and analysis at each step of the process. When the PDPC method is used for quality control, especially with problems related to important qualities, there are two steps:

Step 1

To apply the PDPC method during the planning stage, it is necessary to clearly describe the problem verbally, through previously established analysis, prior experience, and the proper techniques. It is important to describe all problems in the planning stage and to anticipate their solutions. For example, Fig. 9.2 indicates the process of moving from the present point A_0 (high percentage of defectives) to Z (tolerable percentage of defectives). In the first step of planning, a line from $A_1, A_2, A_3, \ldots, A_p$ is considered as a possible means to reach from A_0 to Z. However, when dealing with the problem of defectives, the matter is not so simple. As a result of discussion, A_3 is determined to be too technically difficult to implement. Thus, if not A_3, there is another line from A_2 to $B_1, B_2, \ldots, B_q$ to reach Z as a possible solution. If the two lines described above are not possible, there are others, such as $C_1, C_2, C_3, \ldots, C_r$ or $C_1, C_2, C_3, D_1, D_2, D_3, \ldots, D_s$, even though they are more costly than previous methods of reaching the goal.

So in the first step, to increase the probability of reaching the goal, not only one, but several lines of approach toward the desired condition Z should be considered. In actual practice, it is possible that each of the alternatives may be put into successive use, or if time is limited, several may be tried simultaneously.

Step 2

When dealing with the perpetual problem of quality, the first step described above is not always adequate because there is always the possibility that unforeseen technical problems will appear during the process of implementing a solution. On the other hand, there also exists the possibility that not only problems but also new perceptions will arise. For example, in Fig. 9.2, suppose that stages A_3, B_1, and C_3 were all reached, respectively, by processing through each line simultaneously and that a couple of months already have passed since this was started. If so, the possibility of success in progressing from A_p, B_q, C_r, and D_s to Z and associated problems becomes

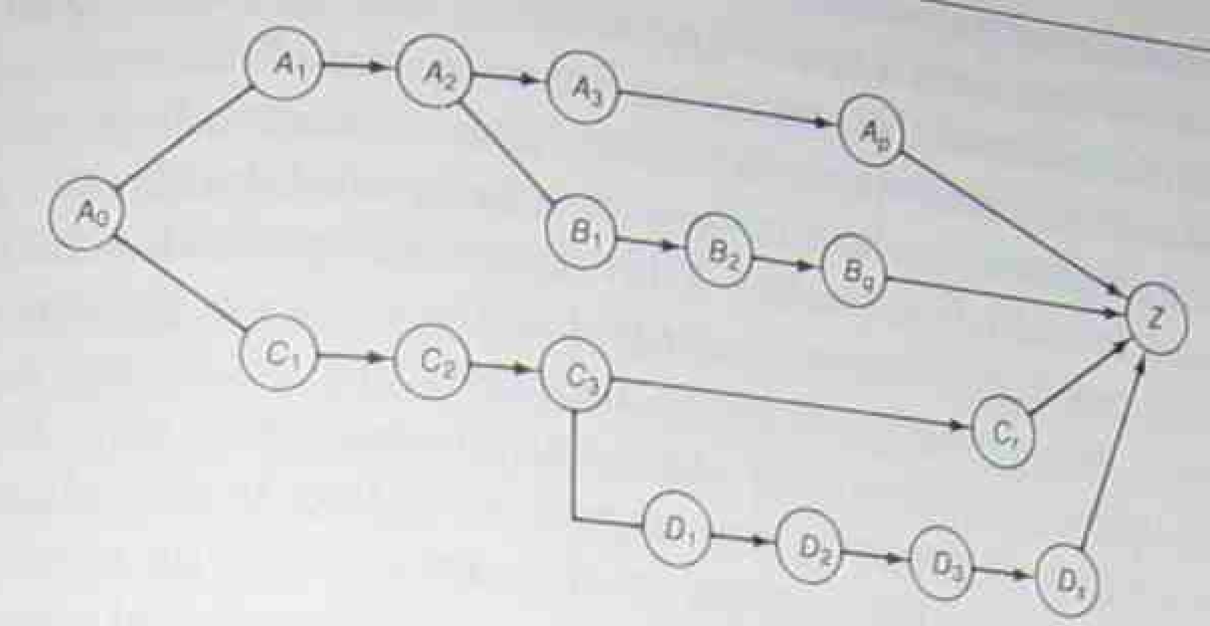


FIGURE 9.2
How to think in the PDPC method

clearer. Should it be proven that the lines described above are not adequate, it would be necessary to set new lines E, F, etc. based on the information learned. In this way, the second step is to be done every couple of months to increase the possibility of success in reaching goal Z through information learned. In quality control, it is said that the object of evaluation is not the result, but the process. Using a chart similar to Fig. 9.2 at each step, information about the adequacy of the original plan, the means of obtaining information, and the timeliness of response during the process is given as data. These data contribute largely to successive decision-making processes.

In the whole process of problem solving, the PDPC method promotes every possible means of solution. It also efficiently deals with conditions that work against a solution. For this reason, the PDPC method is characteristically quite dynamic in nature. Moreover, the PDPC graphically expresses the entire process for reaching a goal. Therefore, the PDPC method encourages ready summarization for executives and many new and inventive ideas from participants.

To understand the character of the PDPC method, let us compare it with other techniques that employ charts, their subjects, and their methods of use. The PDPC is similar in form to the tree chart. As lines progress from concept to actuality, the PDPC is certainly like a goal-method tree chart. But the tree chart is static because events progress within a fixed line of goal-method. On the other hand, in the PDPC, events in the process are linked by flow lines in chronological order; therefore, this method is dynamic. Arrow diagrams, as we shall discuss later, also arrange events in chronological order and are used for schedule control. With the PDPC method, schedule control is possible if dates are inscribed on the events in the process. However, the subjects of arrow diagrams are, in many cases, events under comparatively static conditions. Each step of an arrow diagram is precisely determined according to the planned date of completion. For example, in the case of building construction, based on past experience, each stage and its projected completion date can be determined quite precisely. When dealing with quality control, however, we often cannot visualize a solution from beginning to end. In these cases, the PDPC is more effective than an arrow diagram.

In the field of reliability engineering, two methods, failure mode and effect analysis (FMEA)[3] and fault tree analysis (FTA) are representative methods used to resolve system failures. FMEA starts from each function of subsystems or parts that make up a system or product and from the potential failure mode that might occur and evaluates its influence on the larger system and its significance. In other words, FMEA is oriented toward a bottom-up succession of analyses. FTA selects one "undesirable situation" and seeks the cause in a top-to-bottom fashion, in the manner of the branches of a tree, until it reaches the principle failure element to obtain a solution. In other words, FMEA deals with function and FTA deals with existing phenomena.[4]

As these examples illustrate, both FMEA and FTA are often used under static conditions. Unlike FMEA and FTA, the PDPC method can reveal not only logical phenomena, but also phenomena expressed in the light of new ideas. As a result of a detailed comparison of the PDPC method and FTA, PDPC can be characterized as follows:

1. Because the PDPC comprehends the actions of a system as a whole, we are able to make a comprehensive summarization (effective for identifying significant errors and problems, but not individual details).
2. We can view the progression of processes in the system in chronological order.

3. In the dynamics of the system, the relationship between initiating influence and terminating influence is shown. Therefore, causes of undesirable conditions can be followed through the course of their actions; undesirable conditions can be discovered by following the movement of actions starting from a certain initiating influence.
4. Because events are the basis of the PDPC, anyone who has a basic understanding of the system can use it easily. Thus, within the many possible directions of change and development of conditions, we are sometimes able to point out unanticipated fundamental problems. It is particularly effective in dealing with human factors engineering and with complex interactions between systems.

For instance, were the FTA chart in Figure 9.3 switched to PDPC, the following problems could be represented: 1) A fire prevention monitor inadvertently turns off the power to the alarm system; consequently, there is no alarm in the case of fire. To deal with this, switches such as those in the alarm system are to be set separately at night. 2) Poorly located lockers deflect the fire and delay the alarm system reaction until it is too late to do anything. To prevent this, notices are posted that nothing is to be placed around the sensor.

The PDPC enables us to readily discuss those problems influenced by human action or changes in the environment. In this way, it is possible to anticipate the consequences of major accidents and to prepare to deal with them in the planning stage. In this way, PDPC can increase the reliability and safety of system and products.

Decisions concerning reliability and safety are developed effectively through the combined methods of FMEA, FMECA, and FTA. Adding the PDPC enables the company to increase safety markedly.[5]

Patterns of PDPC

The PDPC method has no definite rules. One may distinguish the following two patterns.

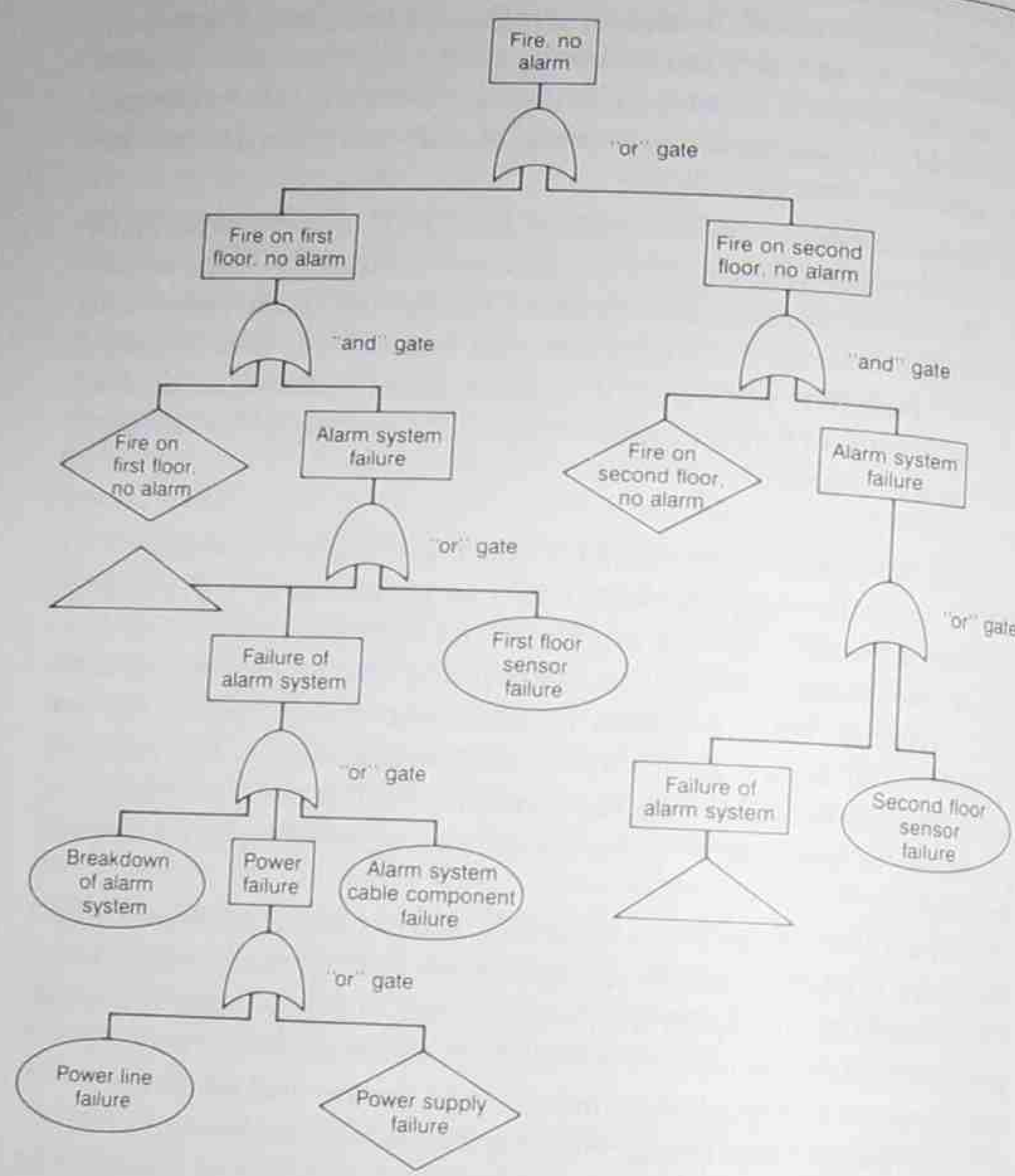


FIGURE 9.3
FTA fire warning system failure tree chart

Source: From Mizuno Shigeru (Ed.), Prevention Plan for Product Responsibility (JUSE Publishing Co., 1975).

Pattern I

In Fig. 9.2, the pattern I process starts with condition A_0 and proceeds to the desired condition Z in an organized manner. The process goes

through $A_1, A_2, A_3, \ldots, A_p$, starting at A_0 and ending at Z. If as a result of the process A_3 is not possible, $B_1, B_2, \ldots$, etc. are considered as replacements. Although the preference of the planner is $A_1, A_2, A_3, \ldots, A_p$ or $A_1, A_2, B_1, B_2, \ldots, B_q$, lines C and D are set up as alternatives in case both these preferred lines fail. In this way, we plan to proceed from starting point A_1 and end with desired condition Z, because A_p, B_q, C_r, and D_s are closer to goal Z, they are more uncertain and can be described in less and less detail. We describe each event $A_0, A_1, A_2, \ldots, D_3$ in a brief sentence enclosed in a circle. Twenty to thirty of these circles are sufficient. Because of increased uncertainty associated with events far from event A_0, it is acceptable in the early stages of planning to provide less than complete details.

In Fig. 9.2, Z is set as a goal to be accomplished, such as establishment of a new technology, completion of research into a new product, prevention of claims, or reduction of the defective rate. But if Z becomes an undesirable condition, such as a major accident, it is necessary to cut the line between the present points A_0 and Z. This pattern is in the same category as pattern I.

For example, in Fig. 9.4, A_0 is set as the normal operating condition of the system. Suppose that through various influences on the system, normal condition A_0 changes to A_1, and then through other influences it finally reaches B_i ($i = 1, 2, \ldots, n$; where B_i is an undesirable condition). In this case, we develop policies to deal with each B_i, considering the risk of occurrence and its effect, if it occurs. For example, B_p represents cases of serious accidents and undesirable conditions. In this situation, we have to prevent the occurrence of $A_0, A_1, A_2, \ldots, A_p, B_p$. To accomplish this, it is necessary to stop that sequence of events and guide the process to A_3 or A_5. This is the way to use a PDPC for the prevention of major accidents. Figures 9.2 and 9.4 are described as pattern I.

Pattern II

In the second pattern, first the goal Z is set as a desirable or undesirable condition. Then the process from Z to the beginning point A_0 is developed with the inclusion of various alternatives from many points of view. In this process of development, it is necessary to note the nonlinear elements, especially the actions of human beings. If the goal Z (or undesirable condition Z) is linked to A_0, then we will

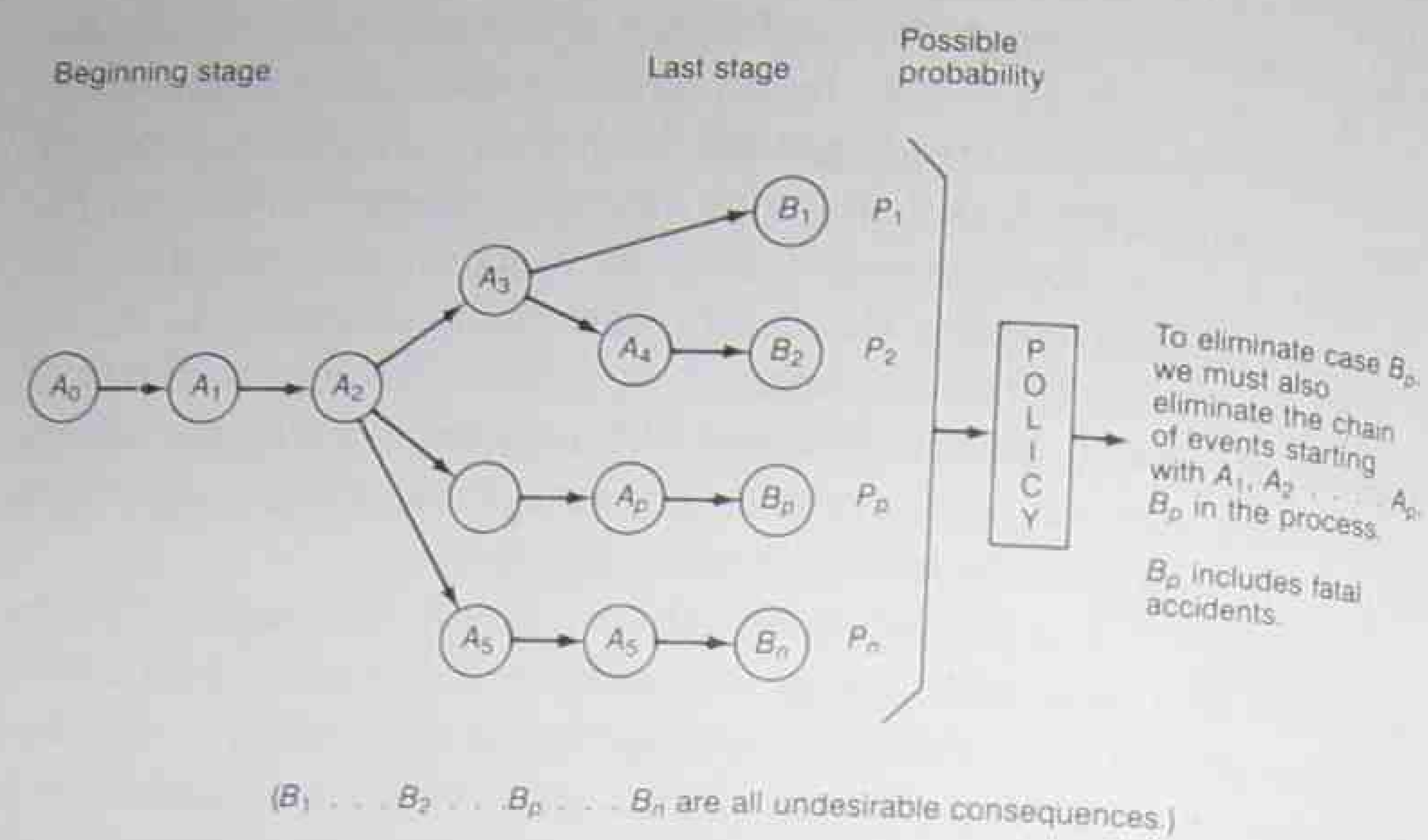


FIGURE 9.4
PDPC pattern I (anticipation of serious accidents)

determine a solution through detailed examination of the process, as with Pattern I. This is called pattern II and is shown in Fig. 9.5.

Where Z is desirable, we start with Z and attempt to link it to the beginning condition A_0. Should Z be an undesirable condition, we try to determine every possible chain of events that results in that condition and cut them off.

How to prepare a process decision program chart

Step 1

Participants in the project (preferably from various fields) meet to discuss the issues. To stimulate discussion, the leader can present one basic line as a possible solution.

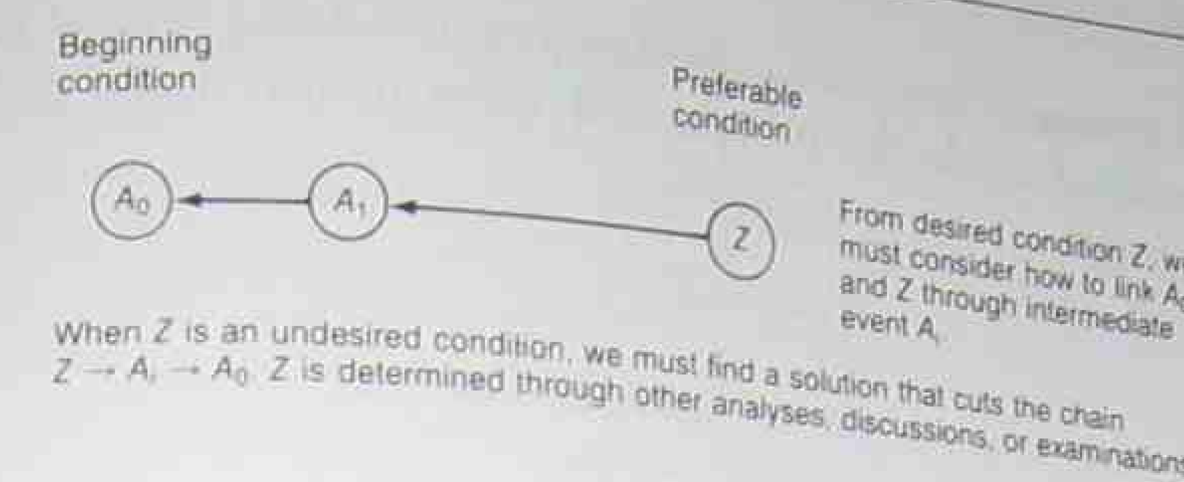


FIGURE 9.5
PDPC pattern II

Step 2

Discuss which issues must be examined.

Step 3

Once the issues are identified, consider and note down all the anticipated results. At this stage you have identified issues and some solutions; weigh the feasibility of each solution proposed and investigate alternative solutions.

Step 4

Classify each issue according to its urgency, number of operations required, likelihood of occurrence, and difficulty. Consider the anticipated results and alternative solutions related to issues that must be addressed immediately, and link the items with arrows to the desired goal.

Step 5

Prioritize the different issues and consider them all together. Information related to one set of possibilities can influence another set; related items should be linked with a broken line.

Step 6

If the department that will handle a process involving several lines is determined, circle the process and write in the name of the department.

Step 7

Set a date for completing the examination period.

When the PDPC method is put into practice, it is possible that new information and problems will emerge at each stage. Therefore, it is a good idea for participants to have regular meetings to check progress in terms of the original PDPC. At such points, participants can correct or add to the plan in light of newly emerging problems or rewrite the chart starting at that point.

To understand how to prepare a process decision program chart here are two examples.

Example 1 (Pattern I). Suppose that a major defect has emerged in the mass-production test stage for a new product and that it remained totally undetected during the short-run test process. The PDPC to solve this problem is shown in Fig. 9.6. After various observations, the major defect is considered to be caused by a heat reaction of substance *MB*, so the product research center, the central research center, and the production department decide to cooperate to solve the problem. At the central research center, the defective part is analyzed and the substance causing the defect is sought. At the product research center, the ability of substance *K* to withstand heat must be confirmed. Therefore, if substance *K* can sustain heat in a satisfactory manner, the defect must be caused by the combination of substance *K* with other substances. In that case, the composition of the base material will be reviewed, and if that is the cause, eliminating it will solve the problem. However, if substance *K* itself is the cause, a substitute must be found. At this point, substances *L* or *M* can be considered even though they are more costly than substance *K*. In this case, if *L* or *M* are tested without defects, the problem is resolved.

On the other hand, the production department must confirm results at each job site, and analyzed data at the central research

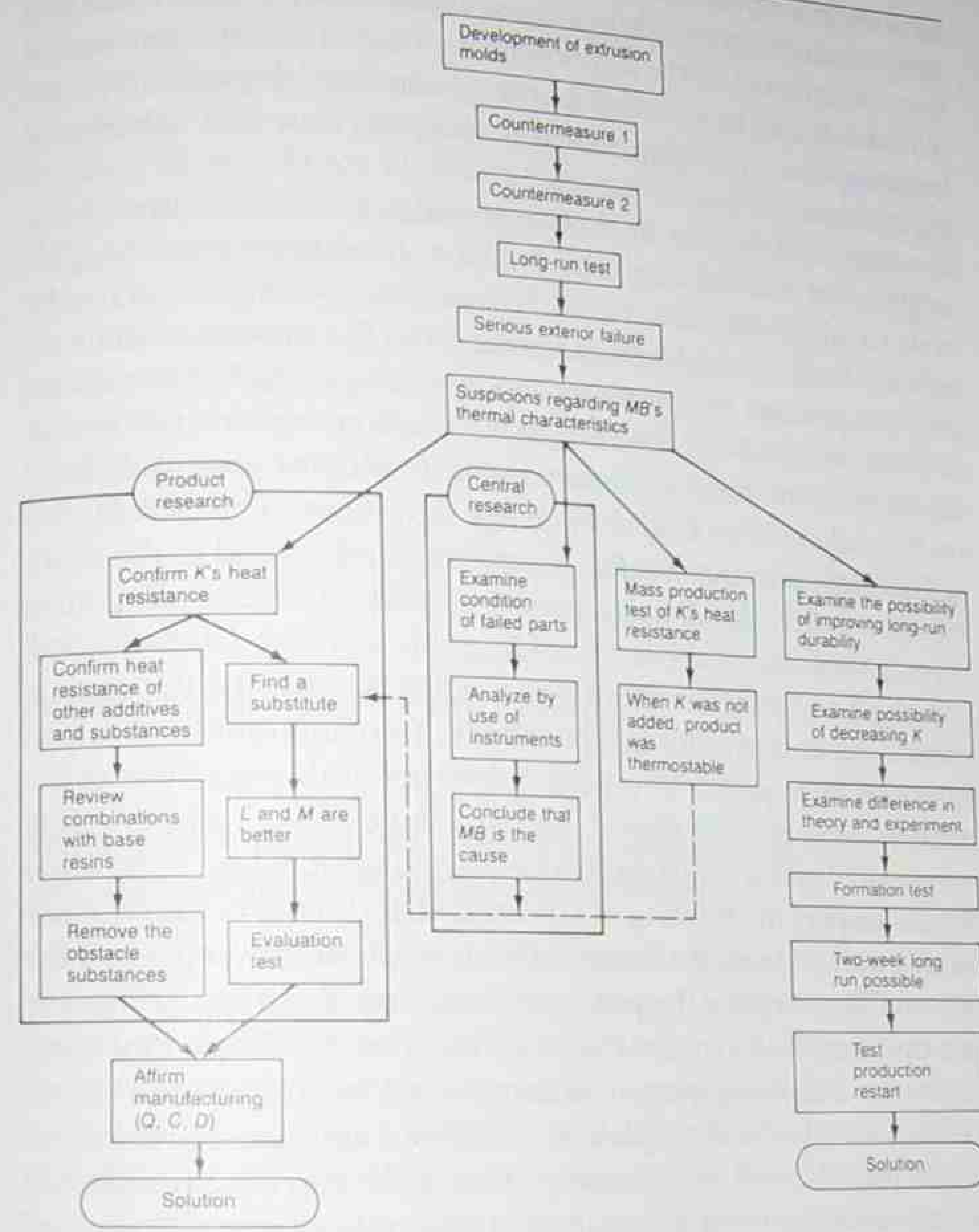


FIGURE 9.6
PDPC chart in research development (pattern I)

center must be passed to the production department. (This is indicated by a dotted line in the diagram.) After these three divisions have examined the problem simultaneously, it is determined that the problem lies in the additive amount of the material *K*, which had been determined at the development phase. Thus the amount of material *K* can be reduced in the production department, contributing to a successful two-week-long test run and solution of the problem.

Example 2 (Pattern II). Let us consider a PDPC analysis for a fragile crate that is being exported to a developing country. This crate should not be turned over. To make things simple and for the sake of clarity, we will limit examination to the process of delivery after the package has been unloaded (see Fig. 2.11). Let us assume unfavorable condition *Z*, in which the package is delivered upside down. First, condition *Z* is unavoidable if there is no warning label on the package. If an English language sign states "This Side Up" in red ink, provided that English is generally understood in the destination country, the delivery person should be able to determine correct handling methods by reading the warning label. Second, suppose a person who cannot read English is to handle the crate. A picture can be drawn for that possibility. Two different kinds ofpictures may be required if the person does not understand one of the pictures. One of them should display a wine glass, indicating that wine would spill if the package were turned upside down; the other should display the hoisting of a chain so that down and bottom can be determined from the picture. Third, what would happen if the person did not notice the pictures? Condition *Z* can be assumed in this case also. If so, handles can be added. Fourth, although it is too heavy for a delivery person to carry, what might happen if that person were to decide to deliver it without assistance? He or she would probably roll the crate over. This would lead to condition *Z*. If so, what can be done to ensure that the crate is not rolled over? Consider using a pyramid-shaped package. When developing a PDPC chart for this idea, we can see the possibilities of connecting A_0 to *Z* and the countermeasures that must be taken in order to avoid unfavorable situations.

Different cases may be considered in order to connect A_0 and *Z*. Express these variable conditions in a brief sentence, which then is enclosed in a frame or is illustrated along with an explanation. Then link each condition with an arrow. When considering these countermeasures, surprisingly good results may be obtained if, developing related events, we also consider economic or psychological factors rather than being dependent only on specific technological knowledge.

In this example of crate transport, the four methods considered are shown below to illustrate "This Side Up": (1) instructions in English, (2) two kinds of illustrations, (3) adding handles, and (4) improvements in packaging design, such as making a pyramid shape or adding support to prevent the package from being rolled over. The chart for this example is illustrated in Fig. 2.11.

The method for drawing a chart for Pattern I (forecast of critical accidents) is basically the same as for Pattern II, but it may be enhanced by enclosing data within a double frame if the events in condition *Z* occur only under circumstances that could endanger human lives.

Applications of the PDPC method

Interdependent methods used to accomplish goals (an application for productivity improvement)

Suppose we want to improve current productivity from A_0 to *Z* by 20 percent (Fig. 9.7). At one point in process *A*, a 10 percent improvement is expected due to the increased use of automation. At this stage, new failures occur due to the increase in speed, and these require implementation of policy A_4 to cope with the problem. Then plan B_1 must be enforced, since continued machine failures can be expected. In addition, improvement in operating rates is planned in process *C*, from which the remaining 10 percent improvement is expected.

To improve operating rates, it is decided to work through lunch breaks. To adopt this, labor-management negotiations (C_3) are needed; the change (C_4) must be implemented and staff secured (D_1 and D_2). If negotiations (C_3) fall through, then newly developed strategies (D_2) must be employed. Furthermore, as a more permanent plan, a plan for partial improvement of machinery must be added to process *A*.

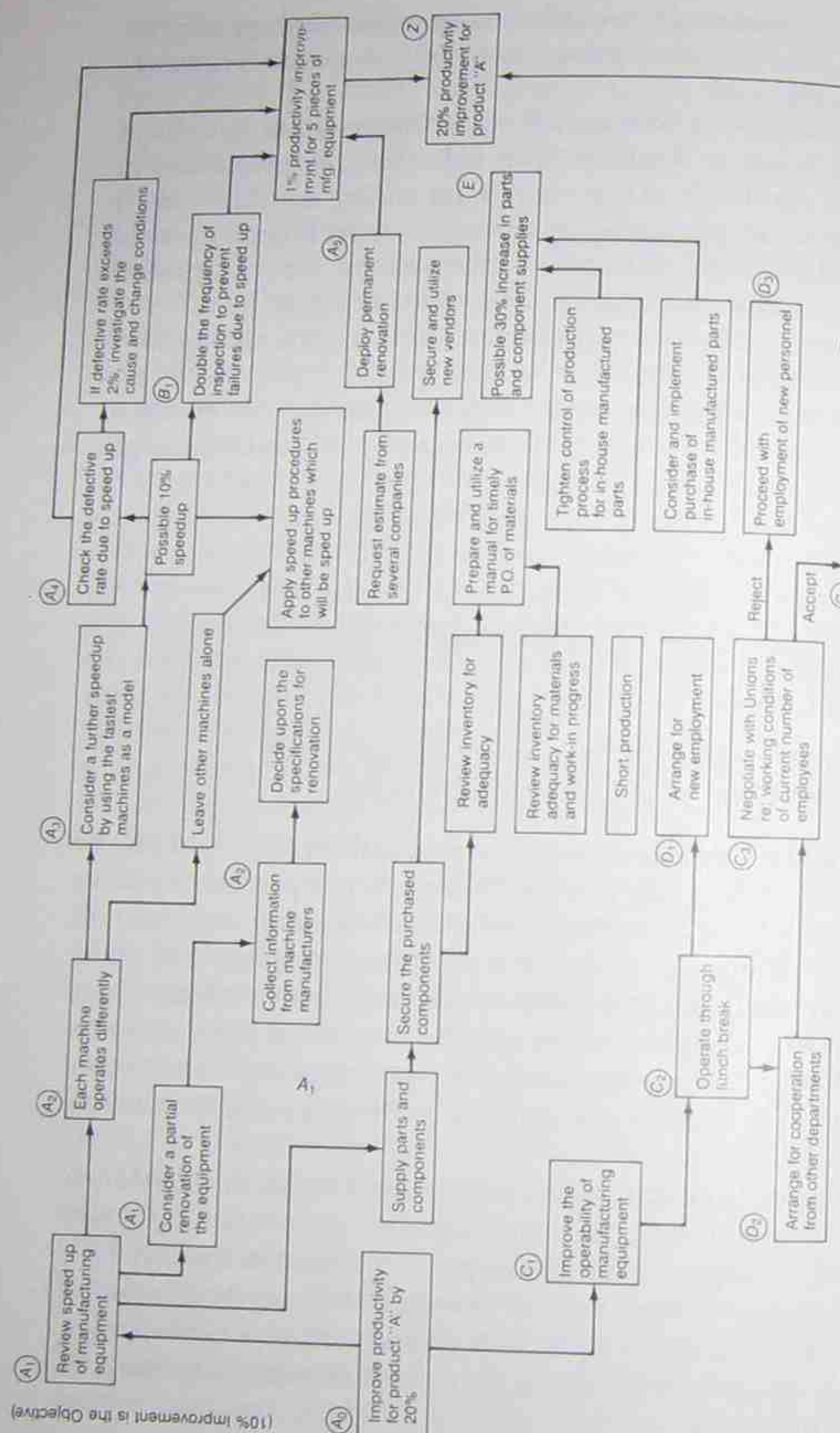


FIGURE 9.7
PDPC chart for productivity improvement (pattern I)

Therefore, corrective measures must be developed while maintaining organic links between processes A, C, and E. Along with increases in other areas of productivity, various problems related to the amount of raw material used and work-in-process inventory will result in implementation of more countermeasures.

When independent means exist and the order of implementation can be determined among them (an application for the prevention of environmental pollution)

Consideration was given, and policies studied, for reducing NO_x from the current 150 index of 100 or lower within a designated period. It was known from the beginning that the goal could be achieved if 20 million yen were invested in equipment. Avoidance of the investment was preferred. However, consideration was given to possible future investment in denitrating equipment. A total of five methods were identified and considered as capable of contributing to the reduction of NO_x; these included three boiler combustion methods and two equipment-improvement methods. After examining them from the aspect of cost-effectiveness and completion date, the goal was achieved, in cooperation with the boiler manufacturer, by implementing two methods for improving the boiler combustion system. In addition, the estimated investment cost was greatly reduced.[6]

When Fig. 9.8 is compared with Fig. 9.2, we can see that each process $A_3, A_4, \ldots, B_1, B_2, \ldots$, and $C_1, C_2, \ldots$ is independent and that the priority of implementation is determined from a standard. We can understand that the reduction is accomplished if the first-priority processes $A_3, A_4, \ldots$ lead to achievement of the goal. If they do not achieve the goal, then processes $B_1, B_2, \ldots$ will be implemented. Depending on the results, it may be necessary to adopt processes $C_1, C_2, \ldots$.

Implementing several almost independent methods for achieving a goal simultaneously (an application to technical development)

At a factory, goods developed by a research center were found to have a serious flaw that was not discovered at the research level. A

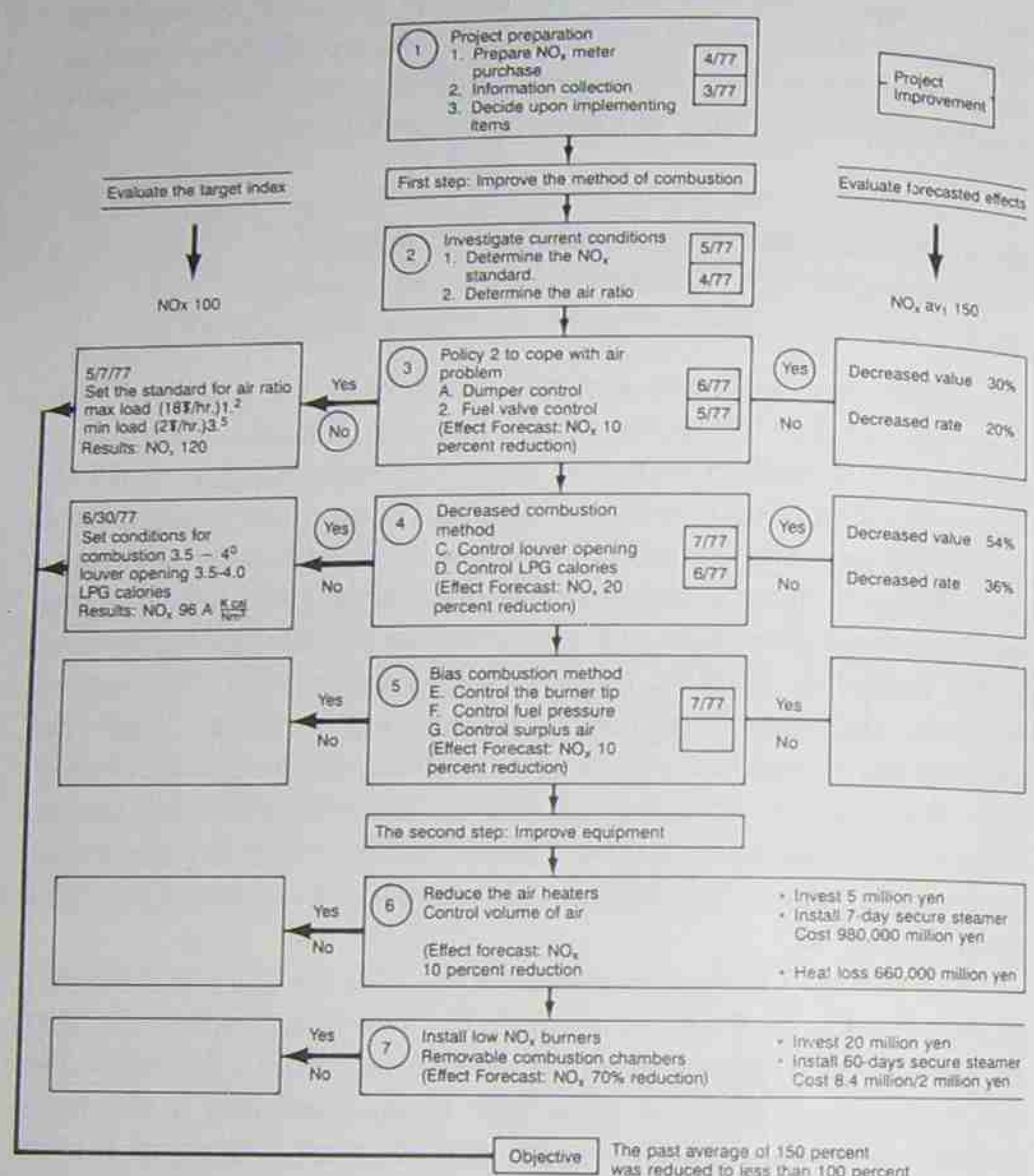


FIGURE 9.8
NO_x Reduction Plan Using PDPC (Pattern I)

Source: Eguchi, Kagoyama, and Kishimoto, "Nitrate Oxide (NO_x) Reduction Activities in the LPG Combustion Chamber," quoted from *Examples from Application Study Meetings for the Seven New QC Tools*, (Tokyo: JUSE Press, Ltd., 1978).

source reports that competitors are going to introduce a similar product in the market. We would like to solve the problem within

two months and commence line production. Therefore, four departments, two research centers, the engineering department, and the production department, meet and prepare a process decision program chart.

Consequently, it will likely be possible to correct the critical flaw with independent plans. On the basis of a PDPC each department is to examine and execute the process within its jurisdiction. Each department then corrects its PDPC after periodically discussing its progress with other departments. As a result of these interchanges, the plans developed by the production department are found to be superior to the others in cost-effectiveness and ability to meet the due date. So from that time on, all other departments join forces with the production department and solve the problem (see Fig. 9.9).

In this case, the two research centers and the engineering department were in charge of processes *4 and *5, design for a metal mold, which is illustrated in parts (b) and (c) of the figure, as well as of the decision-making process noted in part (a). The production department was in charge of other implementation items. In comparison with the model (Fig. 9.2), each of the processes $A_3, A_4, \ldots, B_1, B_2, \ldots$, and $C_1, C_2, \ldots$ corresponds to the two research centers, the manufacturing department, and the engineering department activities. The three parts of the figure show how the PDPC is successively modified in accordance with the project's progress, adding new steps where new types of problems are revealed. The strengths of all four departments were concentrated on one point and the goal was achieved despite the short timeframe.

Establishing a new plan after consolidating complicated processes (an application for accident prevention)

A serious accident occurred at a company three years ago, but emergency measures were applied and production was continued. There was, however, a minor failure in terms of quality. Various studies in different areas during the last three years have attempted to promote stable production by introducing a permanent policy. Finally, all other plans were narrowed down to one permanent plan. It was discovered that the plan required too many intricate processes for implementation. The engineering department determined that in order to implement the plan, it was important for all persons within

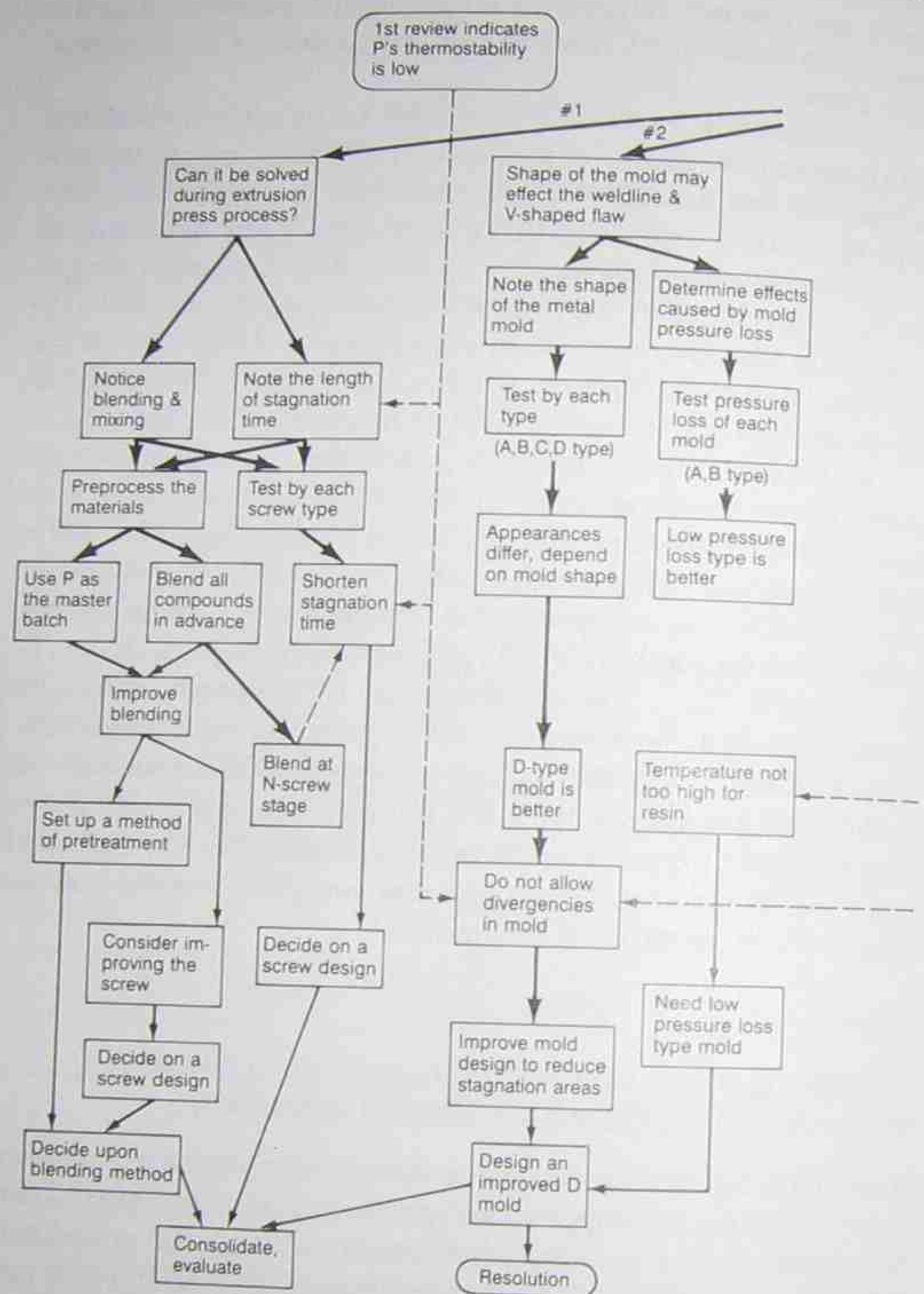
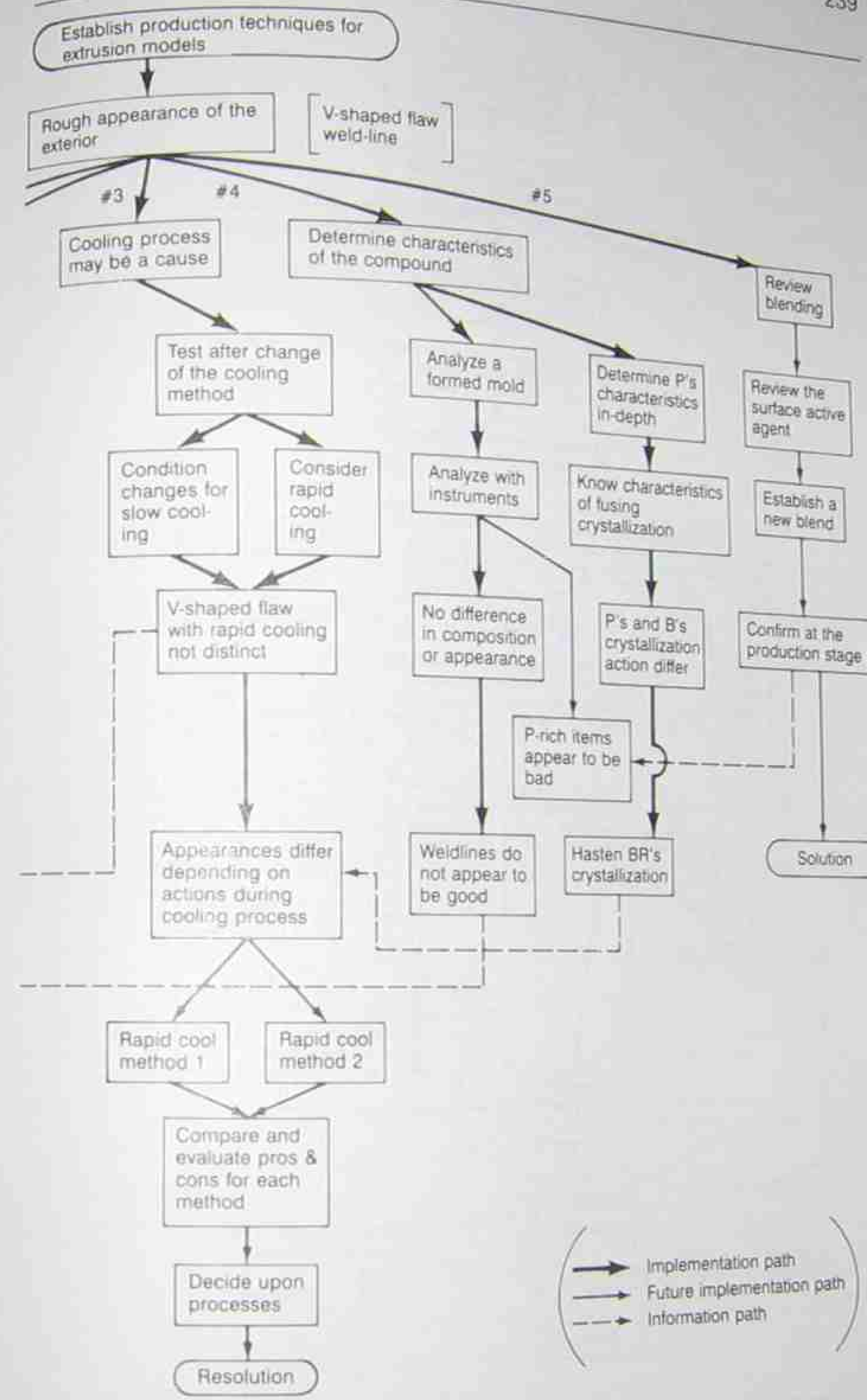


FIGURE 9.9 (a)
PDPC charts for technique review (pattern I)



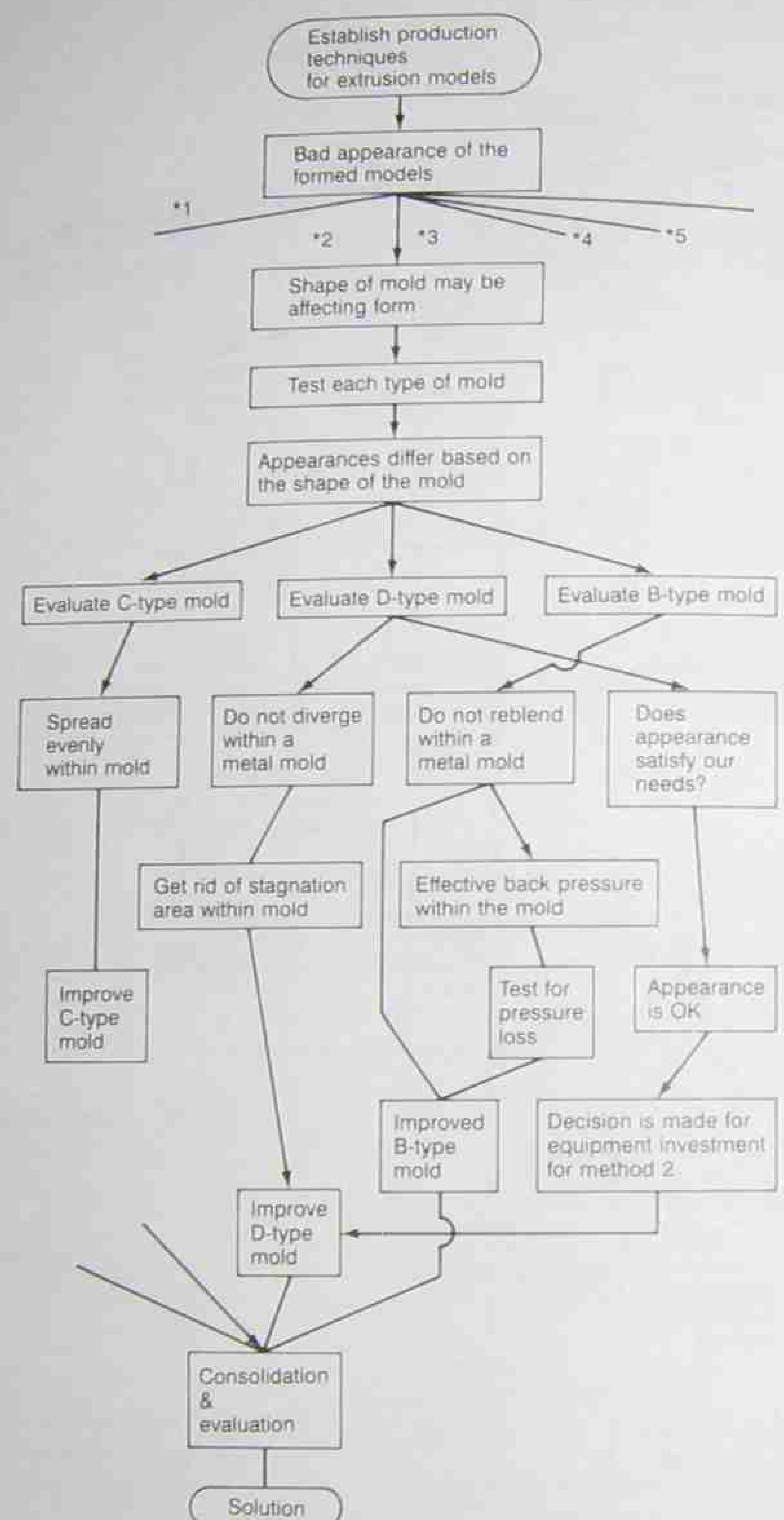


FIGURE 9.9 (b)
PDPC charts for technique review (pattern II)

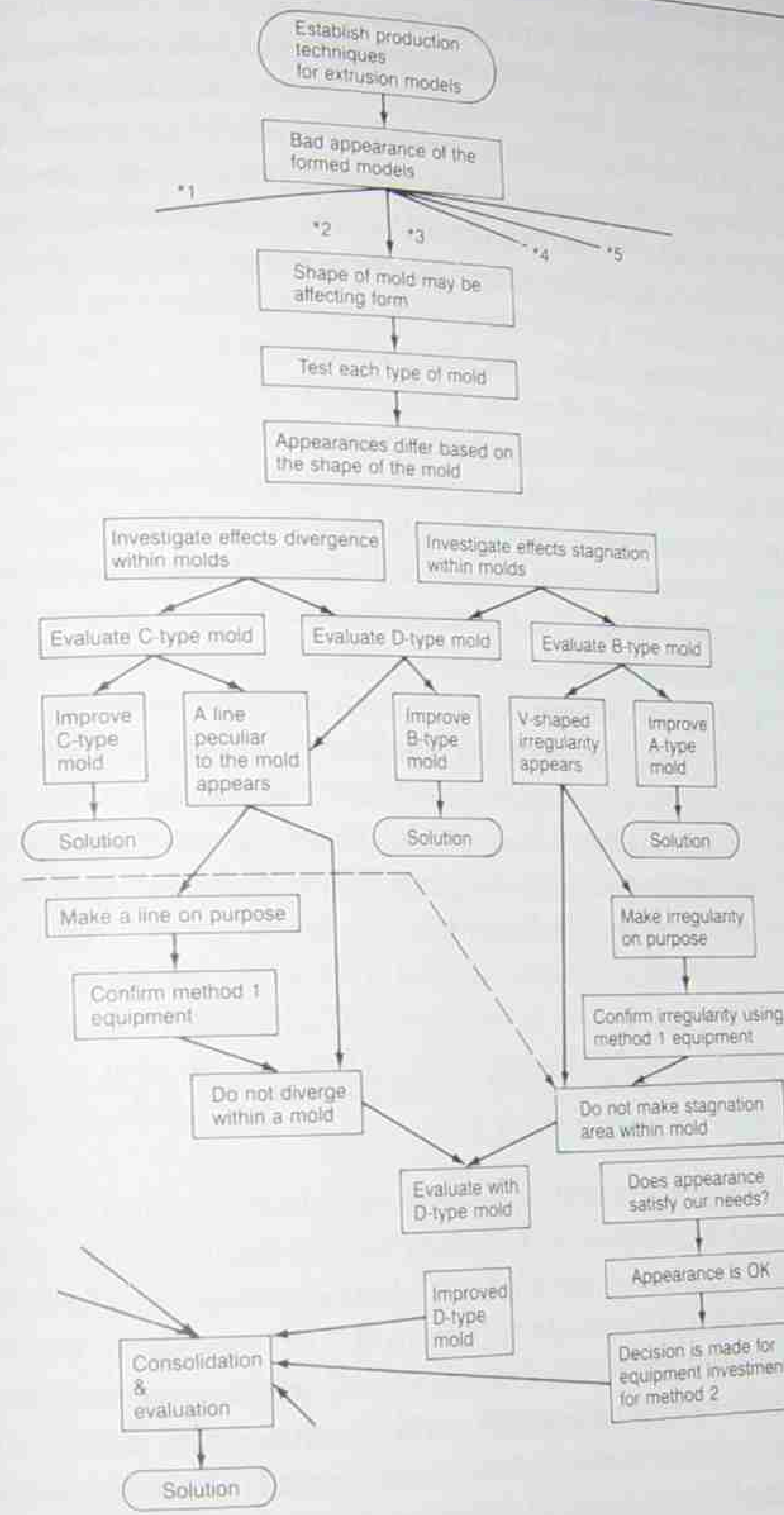


FIGURE 9.9 (c)
PDPC charts for technique review (pattern III)

the departments, including superiors, to have a good understanding of the group's past study processes.

The staff could quite easily understand past problems when the study processes were arranged in a process decision program chart. As a result of discussions, a plan that was initially rejected during the early stages now appeared to be more applicable because of technology and data developments and more economic than the current permanent plan. After three months of reexamination, the new plan became the new permanent plan (see Fig. 9.10).

When the model in Fig. 9.10 is compared with the model in Fig. 9.2, the processes follow $A_3, A_4, \ldots, B_2, \ldots, C_1, C_2, C_3, D_1, D_2, \ldots$, but during the investigation, all processes except A were terminated or abandoned. This left plan A_p as the only possible choice. When examination processes were consolidated in a PDPC chart, it became apparent that C_4, which had been discarded as impossible to implement, would, in fact, be the most effective in light of new data, cost, and operability. The last item for completion was a reevaluation. All these processes correspond with those noted in the model. As you can see, a plan that has been abandoned at an early stage can be easily forgotten as time passes. When the problems are complicated or the study requires a long period of time, a complete view of the processes, arranged according in a PDPC increases understanding and helps to prevent oversights.

Predicting critical accidents (an application for a safety study of a railroad car) (reference 9.3).

As an example of a system's safety examination, the process of designing countermeasures will show how to determine the possibility of danger to passengers if a train's brake system fails. The PDPC in this example corresponds to pattern II. Let us examine phenomena A_i $(i = 1, 2, \ldots, n)$, which links A_0 with Z, presupposing that A_0 is an incident involving falling rocks and Z is the critical accident of aderailment and overturning of the train. Several different causes can be considered for the case of falling rocks leading to derailment and overturn. Here we prepared a process decision program chart by assuming that the derailment and overturn had occurred due to an air brake failure in the first and second carriage (see Fig. 9.11), which

was caused by equipment failure. Physical phenomena, such as an operational failure or an engineer's error, are all considered in A_i. The process to reach Z and the process to stop the train are both considered in this PDPC.

When we study how an engineer would cope with an emergency situation in which the air brakes did not work because falling rocks damaged the brake system near the wheel unit between the first and second cars, we can consider two opposite cases: (1) derailment with overturn, and (2) a safe stop.

The objective is to learn from the PDPC how to increase the probability of obtaining a favorable state (stopping the train) and how to decrease the possibility of obtaining an unfavorable state. As shown in Fig. 9.11, we could develop countermeasures (1) and (2) for the first results and countermeasure (3) for the latter result.

Countermeasure (1) took into account the simplicity of inspecting the carriage and changing the location of part of the braking system. Countermeasure (2) considered the psychological difficulties accompanying shifting the air brake handle from the release position to the brake position as the engineer struggles to stop the train in an emergency situation. Thus it helped to develop the idea of an electrically controlled circuit that would perform the job just by pushing a button with the same effectiveness as moving the air brake handle by hand from the release to the brake position. Countermeasure (3) is a security brake to be used if countermeasures (1) and (2) do not operate. With this, the engineer is able to perform the last action of the process (located on the right side of the figure), which would normally follow.

Conclusion

These are just a few of the practical applications of PDPC. We hope that the PDPC method (as well as FEMA and FTA) will be more widely used in the future in the area of product liability planning, where the safety of goods or systems must be closely monitored.

Furthermore, since it is necessary to understand problem and design solutions at each step, it is important to utilize not only the conventional QC methods, but also value engineering, operations

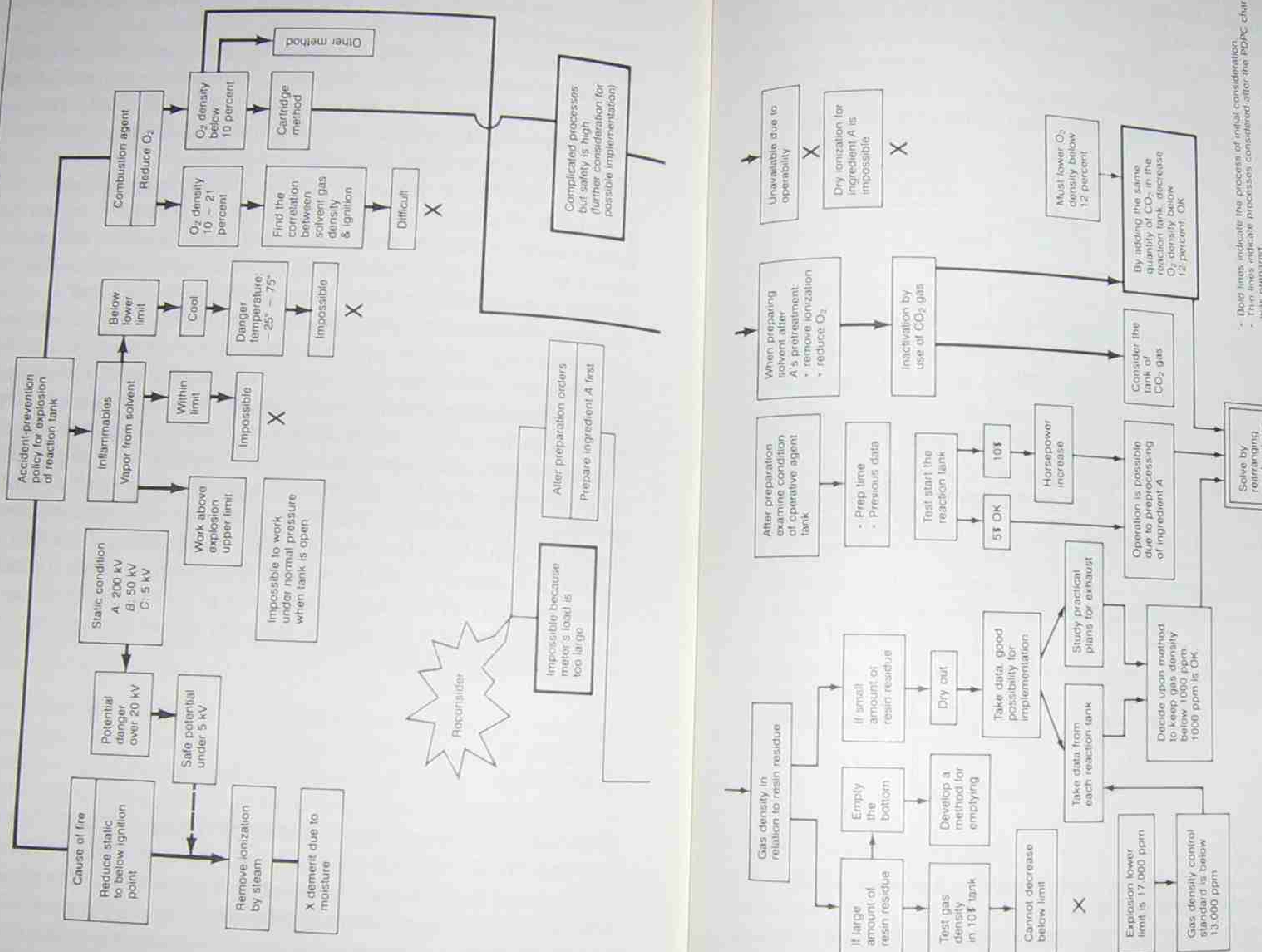


FIGURE 9.10
Example in which a solution was discovered by writing a PDPC chart for the process of a study

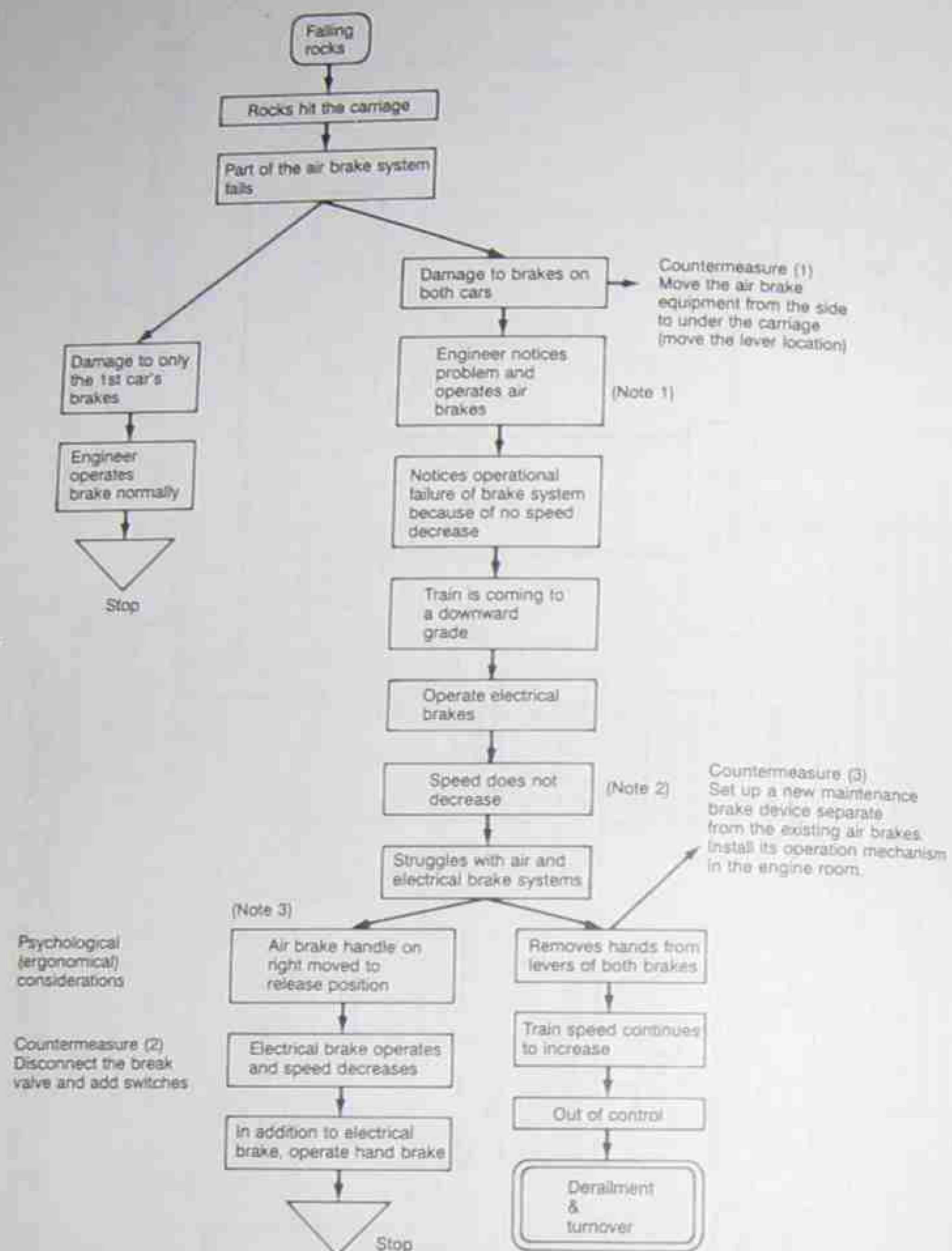


FIGURE 9.11
Application of PDPC to system's safety (derailment & overturn due to falling rocks and loss of control)

Note 1: There are two types of systems to stop trains: the air brake and the electrical brake. There is a separate handle to operate each.
Note 2: There is an order of priority set up for these systems because of peculiar technical reasons. When both handles are operated, the electrical brake does not work because the air brake has priority. In this type of emergency situation, the train cannot be stopped unless a system is developed where the electrical brake automatically works first.
Note 3: As stated in note 2, the air brake has priority when both systems are in operation; however, there is no braking action because of a failure in the air brake system. If the air brake handle were in the release (nonoperating) position, then the electrical brake would be effective for stopping the train.

research, and the seven new QC tools. The range of application for the PDPC method has been widened to include critical accident forecast charts, and it is now ready to be used in other areas of quality control. We hope that this method, as well as others, will be applied in many new areas.

Our thanks to Dr. Kondo and other experts advocating and promoting the application of the PDPC method for allowing us to refer to their works.

Notes

1. Kondo Jiro, *Introduction to Mathematics for the Social Sciences* (Tokyo: Toyo Keizai Shimposha, 1973).
2. Kondo Jiro, *Operations Research* (Tokyo: JUSE Press, Ltd., 1973).
3. Division of Machine Reliability (Eds.), *FMEA and FTA Explained* (Tokyo: JUSE Press, Ltd., 1978).
4. Mizuno Shigeru, (Ed.), *Prevention Plan for Product Responsibility* (Tokyo: JUSE Press, Ltd., 1975).
5. Yagi Juichi and Naya Yoshinobu, "Application of PDPC in Quality Control Systems," Japan Quality Control Academy, 10th Research Presentation Conference, 1976; – – – –, "PDPC Method," in *The Seven New QC Tools Summary for Managers and Staff* (Tokyo: JUSE Press, Ltd., 1978).
6. Eguchi, Kagoyama, and Kishimoto, "Nitrate Oxide Reduction Operations for LPG Boiler," from *Seven New QC Tools: Case Study Presentation Meeting* (Tokyo: JUSE Press, Ltd., 1978).

Additional references

Sano Motohiko, "PDPC Method and Technology Development," from *Hinjitsu Kanri* (Quality Control), vol. 29 (November 1978), special issue, pp. 167–171.

10

The Arrow Diagram Method

The arrow diagram method establishes the most suitable daily plan for a project and monitors its progress efficiently.

What is an arrow diagram?

In addition to quality, timing is a vital management consideration in QC activities. Planning schedules and controlling their progress are very important to such management activities as planning and production of quantities of items within a designated schedule. Such concerns as new products, manufacturing startup dates, delivery dates, and project promotion plans and their progress require scheduling management.

Gantt charts have often been used for planning schedules and project management. The Gantt chart is an excellent method for rough plans and simple work instruction, but it cannot indicate subordinate relationships. In addition, with Gantt charts it is difficult (1) to make a thorough plan, (2) to review a plan at the planning stage, (3) to cope with changes in plans or situations after the initial plan has been implemented, (4) to obtain accurate information promptly concerning influences of a delay of one part of the process on the whole project, (5) when the project becomes larger, to gain an overview of the entire project, and (6) to judge the priorities of the control processes.

Methods such as the program evaluation and review technique (PERT) and the critical path method (CPM) are used to supplement the Gantt chart and control processes effectively by creating an optimum plan. The arrow diagram illustrates schedule planning when PERT or CPM are used. An arrow diagram displays every job necessary for promoting a project and its subordinate relationships through the use of an arrow network.

The Gantt chart is located at the top of Fig. 10.1, and the arrow diagram is at the bottom. When one job is delayed, it is difficult to evaluate with the Gantt chart whether or not the whole construction schedule will be affected or which jobs have time margin requirements and which ones do not. By contrast, we can see that the arrow diagram in Fig. 10.1 provides more information about these areas.

A wealth of literature is available on PERT and CPM generally.[1] In this section, schedule planning and management methods that use the arrow diagram for PERT and CPM will be referred to as the arrow diagram method. We encourage its application to QC activities and consider it one of the seven new QC tools.

How to prepare an arrow diagram

Rules on preparation

Symbols and terms

Elements used in the construction of an arrow diagram are shown in Fig. 10.2.

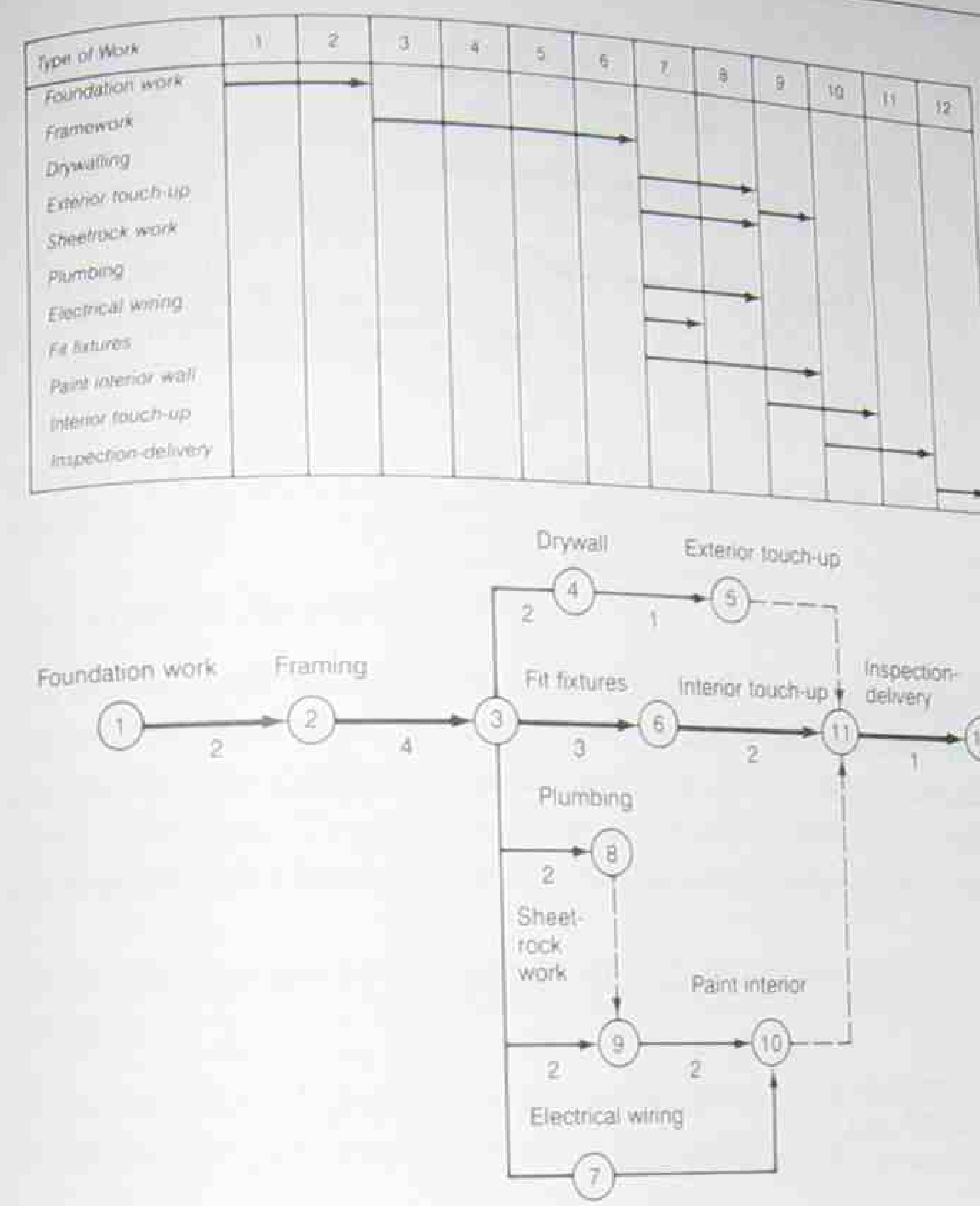


FIGURE 10.1
Gantt chart and arrow diagram

Preceding jobs and succeeding jobs

When there is a correlation between jobs *A* and *B* such that job *A* must be completed before commencing job *B* or job *B* can be started when job *A* is finished, *A* is *B*'s preceding job and *B* is *A*'s succeeding job. This relationship is illustrated in Fig. 10.3.

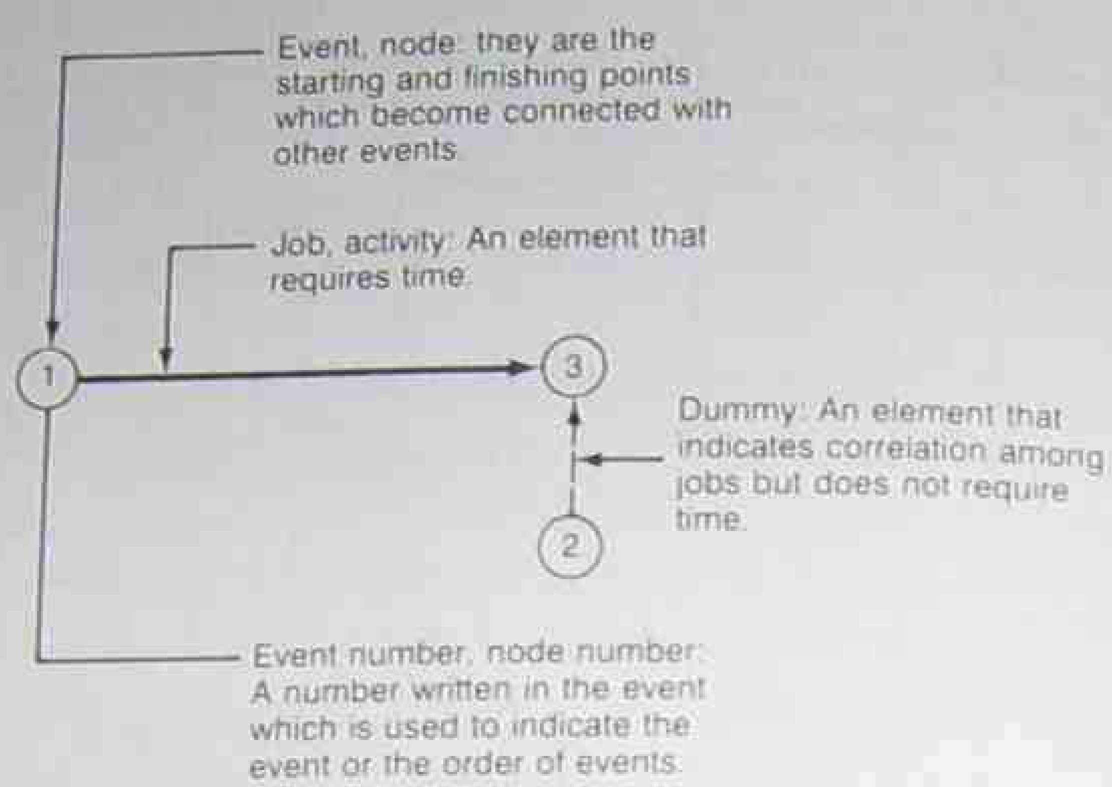


FIGURE 10.2
Symbols and their names

FIGURE 10.3
Preceding and succeeding jobs

Parallel jobs

When jobs *A* and *B* can be performed simultaneously, or when it is arranged so that jobs *A* and *B* are parallel, *A* and *B* are parallel jobs. This is illustrated in Fig. 10.4.

How to use a dummy (part 1)

No two events should be connected by more than one activity. For example, job *A* is described as job (1, 2). It is difficult to understand whether job (1, 2) is job *A*, *B*, or *C*, however, if an expression such as

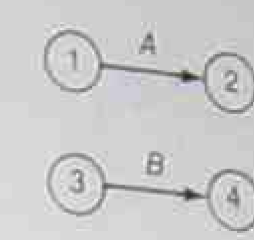


FIGURE 10.4
Parallel jobs

the one in Fig. 10.5 is used. We must use one of the expressions from Fig. 10.6.

How to use a dummy (part 2)

The use of dummy activities will be helpful when it is difficult to express the correlation between jobs by using the job elements only. For example, use a dummy if the four jobs, *A*, *B*, *C*, and *D* are correlated as follows: *C*'s preceding jobs are *A* and *B*, and *D*'s preceding job is *B*. This can be expressed with a dummy as shown in Fig. 10.7.

The same job cannot be used in more than one place in the diagram

The same job, a job performed at the same place and at the same time, should not appear more than once in an arrow diagram.

Do not use a loop

The loop shown in Fig. 10.8 should not be used for jobs *B*, *C*, and *D*.

Do not use unnecessary dummies

The use of a dummy in Fig. 10.9(a) is not a mistake; it only adds to the complexity of the arrow diagram. The diagram should be developed as shown in Fig. 10.9(b).

How to number a node

A number should be inserted at each node in the arrow diagram for identification purposes. The node number must be a positive whole

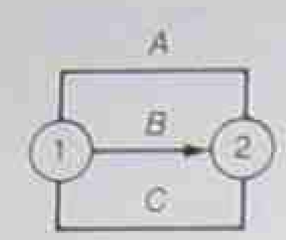


FIGURE 10.5
Example of graph that should not be used

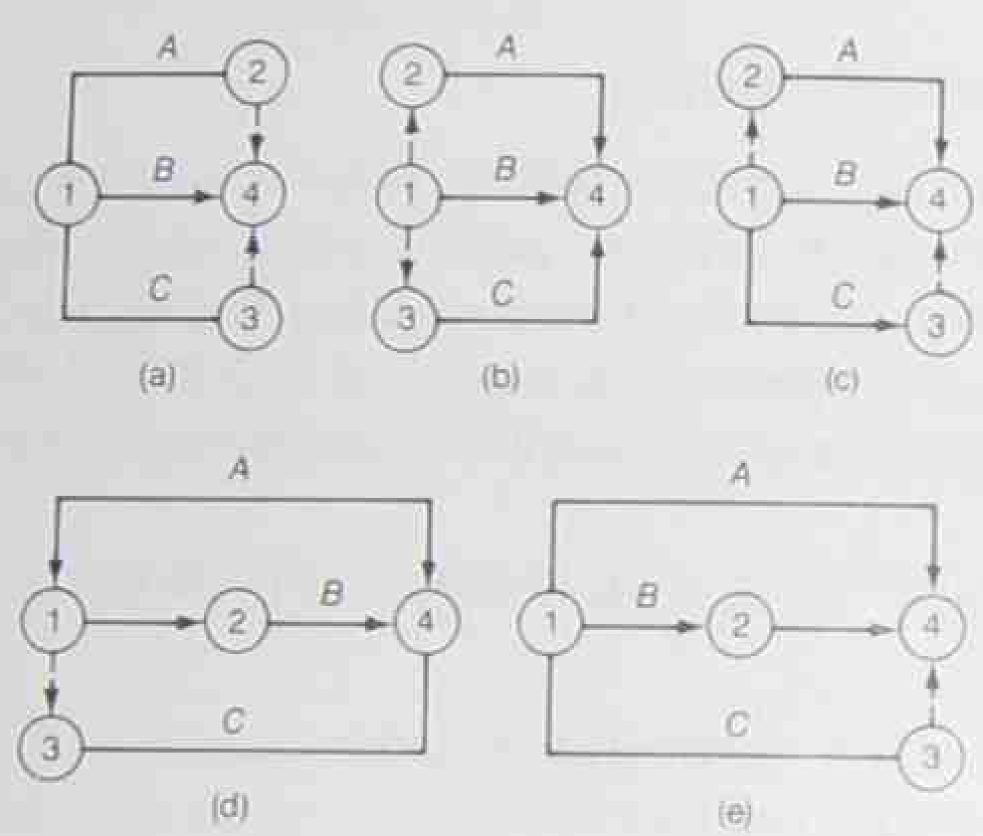


FIGURE 10.6.
How to use dummies (part 1)

number. A smaller number must be used at the job starting point than at the completion point of a job. Thus $i < j$ in Fig. 10.10.

Preparing an arrow diagram using cards

When several people gather in groups or project teams to prepare arrow diagrams, arrow diagram preparation cards and the procedures listed below should be used.

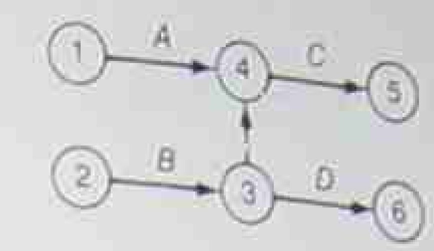


FIGURE 10.7
How to use dummies (part 2)

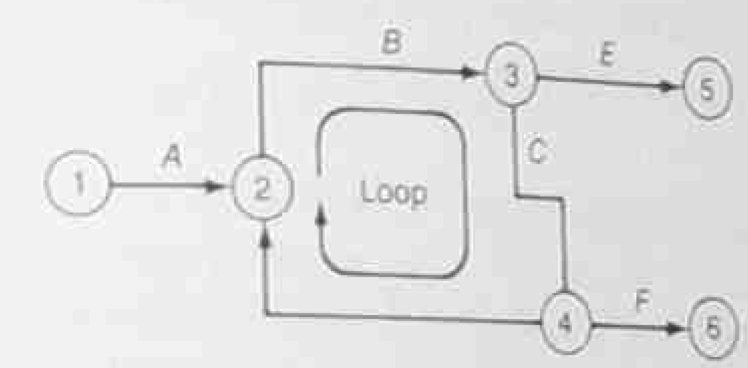


FIGURE 10.8
Arrow diagram with loop

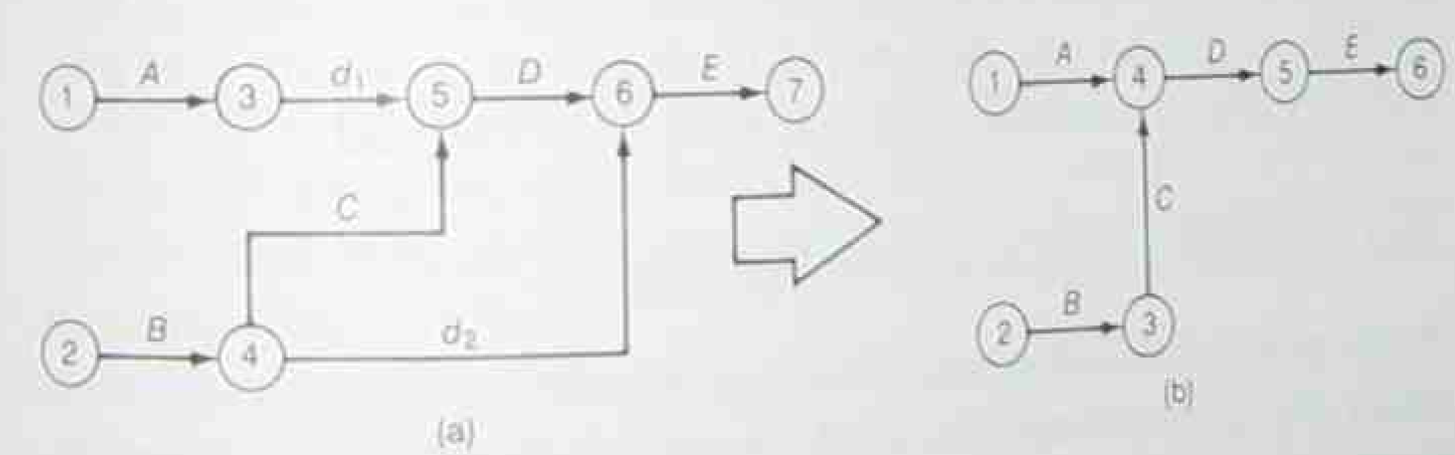


FIGURE 10.9
Arrow diagram with unnecessary dummy

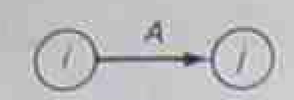


FIGURE 10.10
Node number

To prepare, obtain five or so large pieces of construction paper, 50 to 100 cards (business-card size), and several marking pens.

Step 1: Abstract jobs.

By discussion, list the necessary jobs for project completion and then write them on the construction paper.

Step 2: Preparation of job cards.

After all the necessary jobs are listed, draw a straight line across the center of a card, as shown in Fig. 10.11, and write the type of work above the line. Do not write below the line, since this space will be used for noting the number of days required for the job. The cards or writing can be color coded to identify the person or department in charge of each job.

Step 3: Correlation of job cards.

When all the job cards are completed, arrange them on a large piece of construction paper according to whether they are preceding, succeeding, or parallel jobs. Remove all cards for unnecessary jobs and add the cards of necessary jobs that have been omitted.

Step 4: Determine the location of the cards.

Position the cards using the following criteria: (a) Find the process where the greatest number of job cards can be placed in a series. Position the job cards that have a preceding-succeeding relationship along this process with an interval large enough for a node to be placed between them (about 30 mm). (b) Job cards having parallel

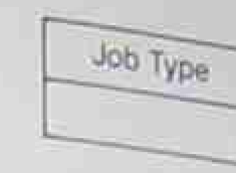


FIGURE 10.11
Job card

relationships must be positioned appropriately relative to the cards in step 4(a). Lightly pencil in the nodes and arrows. Decide on the final position for all cards and affix them to the paper.

Step 5: Preparation of the arrow diagram.

The arrow diagram is finished by connecting the final arrangement of nodes and job cards with arrows. An arrow should not branch off or join with other arrows. Branching and joining must occur only at a node.

Step 6: Survey to establish time requirements.

Make a survey of the amount of time required for each job in the arrow diagram, and place the information below the center line on the job cards.

Step 7: Calculation of node placement.

Calculate the earliest and latest node time for each of the nodes in the arrow diagram, and place them near the associated node. Calculating node times and expressing them will be explained in the next section.

The schedule of the project is completed by the use of an arrow diagram. (An example is given in Fig. 2.13.) If stricter time control is necessary, the work schedule and the number of marginal days must be calculated with the method that will be explained in the nextsection. It is, however, quite effective to perform steps 1 through 7.

Calculation of schedule

Node time and its calculation

Early and late node times are the two types of times for nodes. *Earliest node time* is the earliest possible time that the operation commencing at node i may be started and is expressed by t_i^E. *Latest node time* is the latest possible time that the operation ending at node i may be completed and is expressed as t_i^L. The expressions t_i^E and t_i^L must be indicated within the diagram frame near the node, as shown in Fig. 10.12.

The earliest node time is calculated as follows: In the arrow diagram, the earliest node time is at the starting point (node 1) and is equal to 0.

$$t_1^E = 0$$

When there is only one job followed by node j, the succeeding node, its earliest node time t_j^E can be found by using the following formula:

$$t_j^E = t_i^E + D_{ij} \qquad \text{(a)}$$

In this formula, t_i^E represents the earliest node time for event i, which precedes event j. D_{ij} represents the number of days required for the job (i, j). When there are more than two jobs that have event j as a

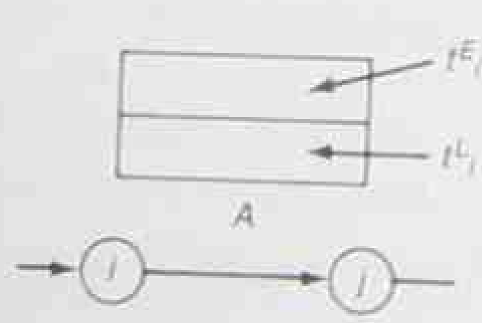


FIGURE 10.12
How to indicate node time

succeeding node, the earliest node time t_j^E can be found by using the following formula:

$$t_j^E = \max\,(t_i^E = D_{ij}) \qquad \text{(b)}$$

The latest node time is calculated as follows: In the arrow diagram, the latest node time for the last event (node n) must be the same value as the earliest node time for that event. That is,

$$t_n^L = t_n^E \qquad \text{(c)}$$

When there is only one job that has event i as its preceding node, its latest node time t_i^L is found using the following formula:

$$t_i^L = t_j^L - D_{ij} \qquad \text{(d)}$$

In this formula, t_j^L represents the latest node time for node j, which succeeds node i, and D_{ij} represents the number of days required for job (i, j). When there are more than two jobs that follow preceding event i, the latest node time t_i^L is found by using the following formula:

$$t_i^L = \min\,(t_j^L \cdot D_{ij}) \qquad \text{(e)}$$

The following correlation exists between the earliest node time t_i^E and the latest node time t_i^L of the same event:

$$t_i^E \leq t_i^L \qquad \text{(f)}$$

The *critical path* is the longest process on the arrow diagram from the start to the final point and is a series of jobs that are vital for scheduling controls. In the critical path,

$$t_i^E = t_i^L \qquad \text{(g)}$$

A bold, thick line is drawn in the arrow diagram for the process after calculating t_i^E and t_i^L.

Slack means the marginal time at node i and is expressed as SL_i by using the difference between the latest node time and the earliest node time:

$$SL_i = t_i^L - t_i^E \qquad \text{(h)}$$

This slack is considered as a rough standard for marginal time when calculating the earliest and latest node times. When managing a more strict schedule, total float and free float, which are explained on page 263, also must be calculated.

An example of the calculations for earliest and latest node times is shown in Fig. 10.13. The bold line is the critical path.

Job schedule and float time

It is possible to control schedules with the use of node times; however, if more indepth schedule control is required, then it is essential to calculate the earliest and latest start times, the earliest and latest finish times, and the float time. Definitions for these terms and their calculation methods are given below.

Job schedule and calculation

The *earliest start time* indicates the earliest possible time when job (*i*, *j*) can be completed, and ES_{ij} is used to represent this. ES_{ij} is found

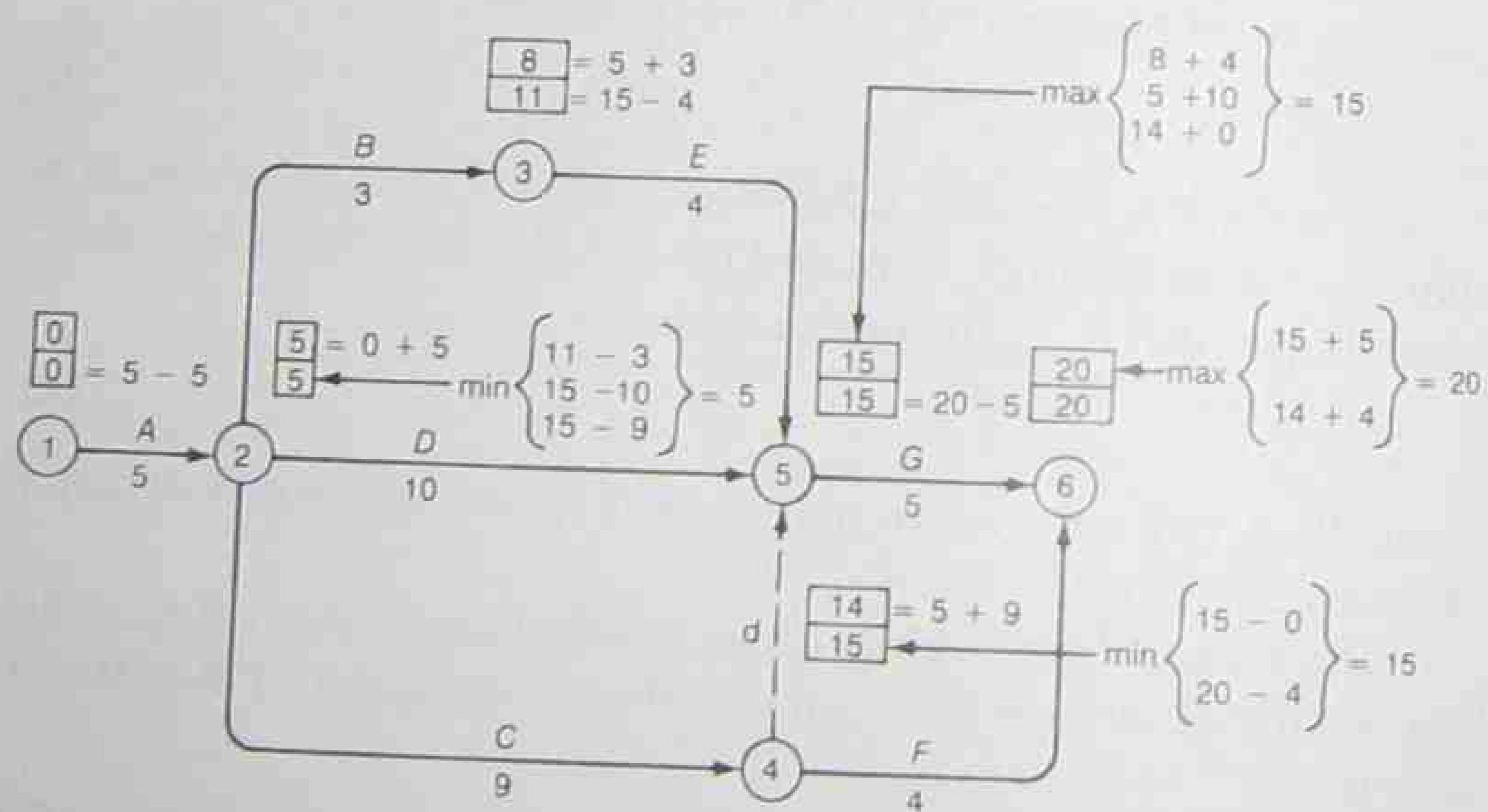


FIGURE 10.13
Calculation of Node Time

by using the following formula, since it is the same as the earliest node time t_i^E at even *i*:

$$ES_{ij} = t_i^E \quad \text{(i)}$$

The *earliest finish time* indicates the earliest possible time when job (*i*, *j*) can be completed, and EF_{ij} is used to represent this. EF_{ij} can be found by using the following formula, since job (*i*, *j*) is started at the earliest start time ES_{ij} and finished after D_{ij}:

$$EF_{ij} = ES_{ij} + D_{ij} \quad \text{(j)}$$

The *latest finish time* indicates the latest time by which job (*i*, *j*) must be completed, and LF_{ij} is used to represent this. LF_{ij} can be found by using the following formula, since it is the same as the latest node time, t_j^L:

$$LF_{ij} = t_j^L \quad \text{(k)}$$

The *latest start time* indicates the latest limit of time when job (*i*, *j*) must be commenced, and LS_{ij} is used to represent this. LS_{ij} can be found by using the following formula, since it is the latest possible time to commence job (*i*, *j*) so that it will be completed by the latest finish time LF_{ij}:

$$LS_{ij} = LF_{ij} - D_{ij} \quad \text{(l)}$$

Float time and critical path

Total Float is the total marginal time for job (*i*, *j*), represented by TF_{ij}. TF_{ij} is the difference between LF_{ij} and EF_{ij} (that is, $LF_{ij} - EF_{ij}$) when job (*i*, *j*) is started at ES_{ij} and finished at EF_{ij}. It is also the difference between LS_{ij} and ES_{ij} (that is, $LS_{ij} - ES_{ij}$) when job (*i*, *j*), scheduled to be finished by LF_{ij}, can be started at LS_{ij}. Thus TF_{ij} can be found by using the following formula:

$$TF_{ij} = LS_{ij} \cdot ES_{ij} = LF_{ij} \cdot EF_{ij} \quad \text{(m)}$$

Free float indicates the marginal time of job (*i*, *j*) independently; it is represented by FF_{ij}. FF_{ij} is the marginal time generated when the succeeding job (*i*, *j*) progresses to job (*i*, *k*). This starts at the earliest starting point ES_{jk} (= t_j^E) after job (*i*, *j*) at EF_{ij} is finished provided that ES_{jk} is later than EF_{ij}. Free float can be found by using the following formula:

$$FF_{ij} = ES_{jk} \cdot EF_{ij} \quad \text{(n)}$$

A job without any total float is called *critical work*. The process created by the critical work is called the *critical path* and is represented by *CP*; that is, *CP* is the sequence of work that satisfies the condition

$$TF_{ij} = 0 \quad \text{(o)}$$

The critical path is the longest process on the arrow diagram and is indicated by a thick, bold line. The jobs along the critical path influence the finish time of the whole project if the critical path jobs exceed the projected schedule period. Careful management of these jobs is essential.

If all the float time is used for one job that has total float, other succeeding jobs will not have any float time at all. Total float is the margin that must be carefully applied by the project manager throughout the entire schedule. If project members spend all the float time without any coordination, there could be a delay in completion of the entire project. On the other hand, free float is the marginal time that does not affect other succeeding jobs, even if all the float time is spent, as long as it is within a certain limit. It is considered to be the number of marginal days that can be used "freely" under the project members' control.

Correlation between float time and the work schedule

The correlation for the following float times is shown in Fig. 10.14: the earliest and latest start times, the earliest and latest finish times, total float, and free float.

The work schedule and float time for the arrow diagram in Fig. 10.13 are calculated and shown in Table 10.1.

Estimation of required number of days

The number of days (or time period) required for each job must be estimated to complete the schedule after the arrow diagram is prepared. The items noted below must be considered when estimating the required number of days.

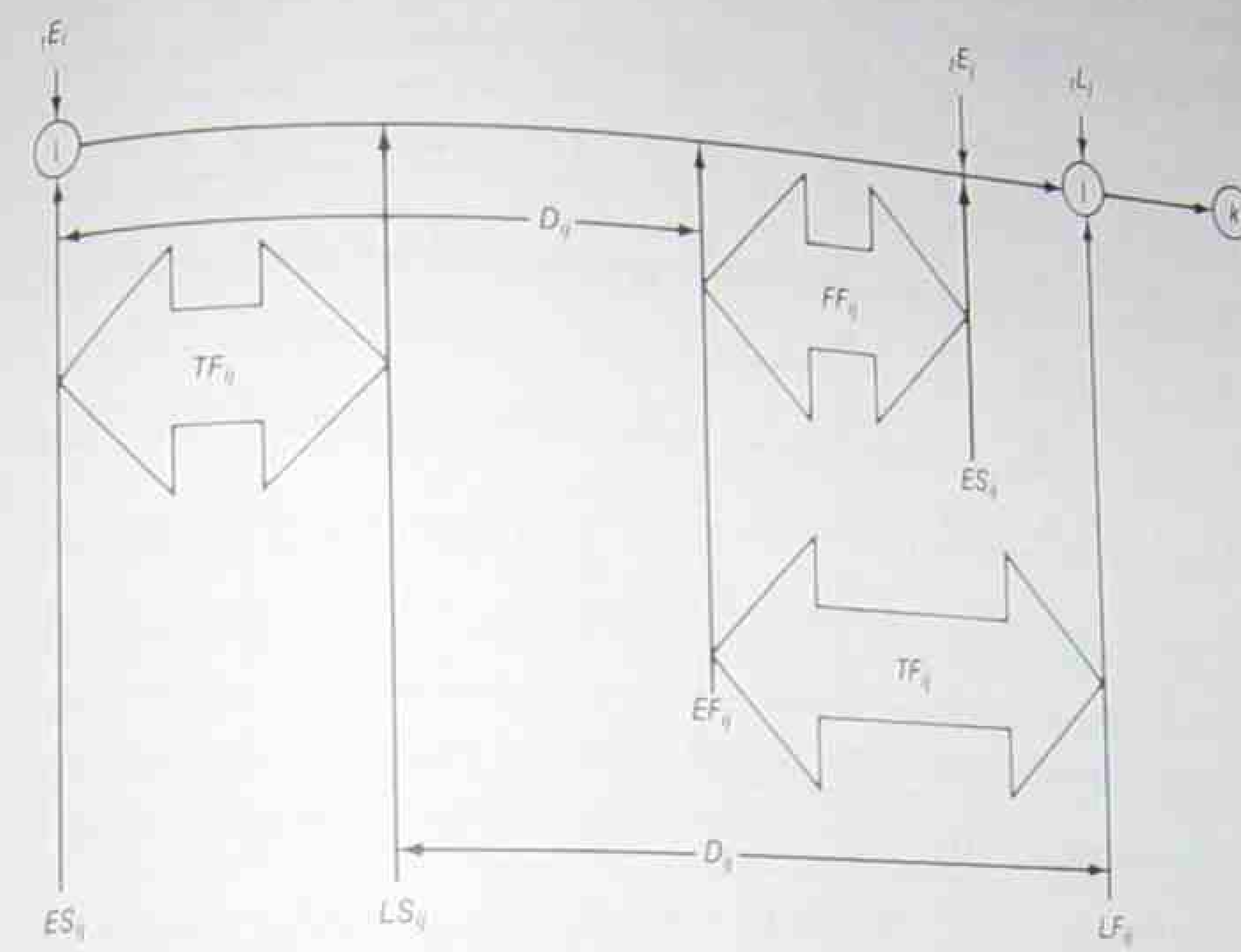


FIGURE 10.14
Correlation of each job schedule and float time

Tendency to overestimate number of days required

There is a tendency, in the interests of safety, to increase the number of days as the project becomes more long range. When estimating the number of days required, it is important to remember that "it can be done within this number of days" is the member's responsibility to the whole project.

Required number of days independently estimated for each job

When estimating the required number of days for each job, we must not relate the job to either preceding or succeeding jobs. Estimation

TABLE Table 10.1

Job i, j	Job type	Required Time D_{ij}	Earliest Start ES_{ij}	Earliest Finish EF_{ij}	Latest Start LS_{ij}	Latest Finish LF_{ij}	Total Float TF_{ij}	Free Float FF_{ij}	Critical Path CP
(1.2)	*A*	5	0	5	0	5	0	0	
(2.3)	*B*	3		8	8	11	3	0	
(2.4)	*C*	9	5	14	6	15	1	0	
(2.5)	*D*	10		15	5	15	0	0	
(3.5)	*E*	4	8	12	11	15	3	3	
(4.5)	*d*	0		14	15	15	1	1	
(4.6)	*F*	4	14	18	16	20	2	2	
(5.6)	*G*	5	15	20	15	20	0	0	
Calculation Sequence			①	③	④	②	⑤	⑥	⑦
Calculation Formula			t_i^E	$ES_{ij} - D_{ij}$	$LF_{ij} - D_{ij}$	t_j^L	$LS_{ij} - ES_{ij}$	$ES_{jk} - EF_{ij}$	$TF_{ij} = 0$

of the number of days it will take to execute that particular job must be done strictly independently.

Consideration of weekends and holidays

There are some types of work, such as aging and drying of paint, that can progress even on nonbusiness days. The duration of jobs can sometimes be drastically reduced by scheduling processes that can advance during the weekend throughout the entire project.

Remember to consider the weather

Every day may not be a sunny day. We must remember to take the weather into consideration if there is a type of work that can or cannot be done during inclement weather.

Application of average number of work days

The average number of work days is taken from past records of the actual number of work days required to complete a job. This average can be utilized in making estimates.

Applications of the arrow diagram method

How to shorten planned time

Process of shortening time

The possibility of shortening a project's time is one feature that is easy to examine when a schedule plan is prepared using an arrow diagram. A flowchart is used in Fig. 10.15 to summarize the methods for shortening planned schedules.

Examples of shortening time

Example 1. The entire schedule was shortened by decreasing the number of days required for jobs on the critical path. The early-plan's schedule is shown in Fig. 10.16. Right after the plan was prepared, the entire project had to be shortened by about a month. It was further shortened, beginning with the jobs on the critical path that required more days.

A total of 29 days was cut by reducing job (M_1, M_3) by 27 days, from 131 to 104 days, and by reducing job (M_5, M_8) by 2 days, from 10 to 8. The critical path, however, shifted to

$$1 \rightarrow E_1 \rightarrow E_2 \rightarrow 3 \rightarrow 4 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow 9 \rightarrow 10 \rightarrow 11$$

so that the project was only shortened by 11 days after all.

Job (4, 6) on the shifted critical path was cut five days, from 45 to 40 days, and job (6, 7) was cut three days, from 10 to 7 days. A total of seven days was taken from the schedule, but the critical path shifted to

$$1 \rightarrow 2 \rightarrow M_1 \rightarrow M_4 \rightarrow M_6 \rightarrow M_7 \rightarrow 7 \rightarrow 8 \rightarrow 9 \rightarrow 10 \rightarrow 11$$

so that only 20 days were cut.

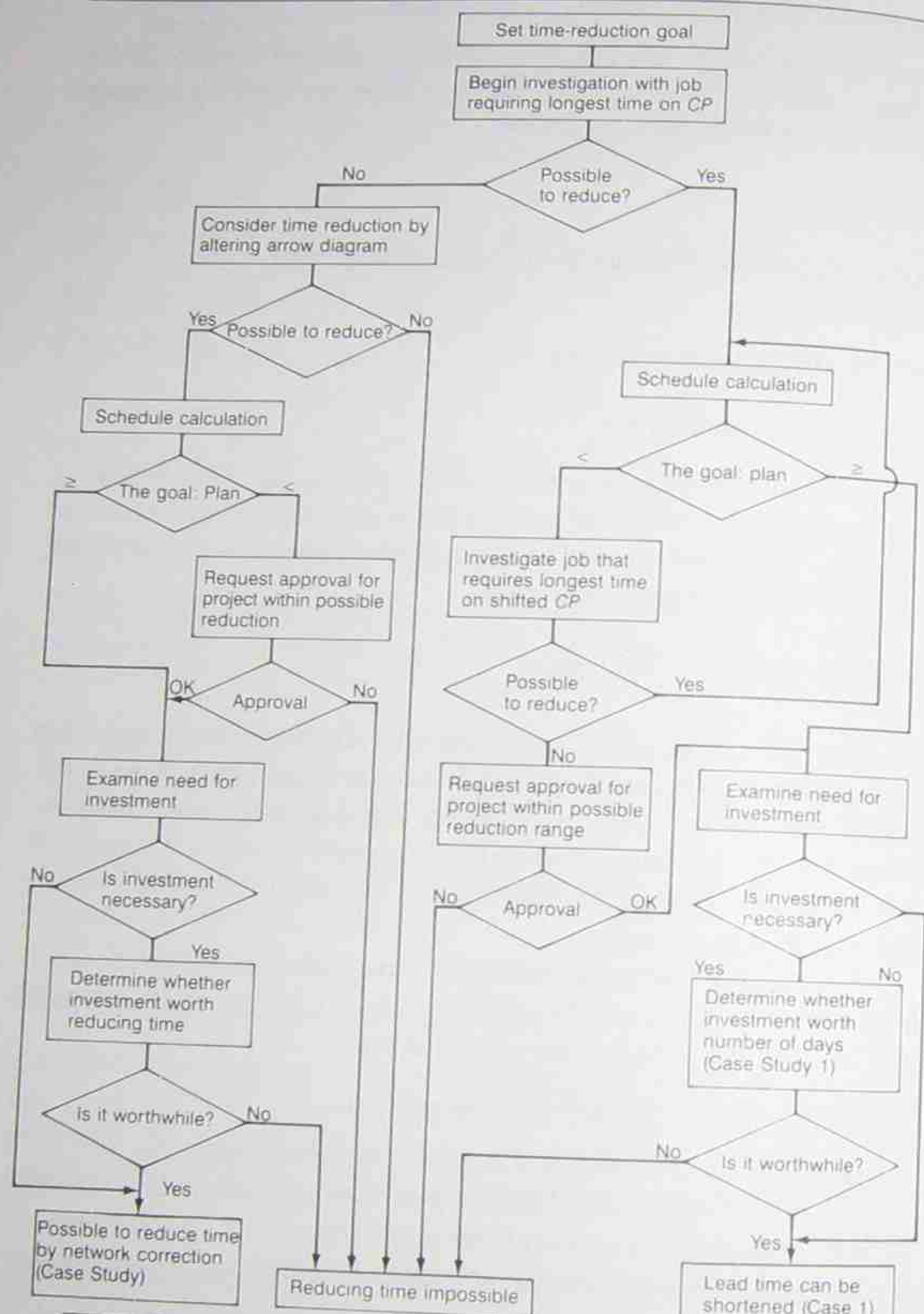


FIGURE 10.15
Examination procedures to shorten a project schedule

Finally, job (M_1, M_4) was cut nine days, from 123 to 114 days. The result, shown in Fig. 10.17, was that the completion schedule was reduced by about one month (29 days) by forming four critical paths.

As you can see in the example, the objective cannot be met when the critical path shifts, although the required number of days for the first plan was shortened to meet the reduction goal. The reduction goal can only be met by successively shortening each job of the newly formed critical path. Planning by arrow diagrams benefits the examination process in this manner.

Example 2. The entire schedule time was shortened by altering the arrow diagram. The initial plan is shown in Fig. 10.18. The requirement to shorten the entire schedule was identified shortly before the advance arrangement conference, event M_3. After studies were conducted, shortening the time required for a job on the critical path, as developed in Example 1, was found to be impossible. After many discussions and the use of an arrow diagram, a suggestion was made to use a handcrafted chassis for an engineering trial test in job (7, 8) and to use a metal mold from the mass production trial in job (10, 11). The arrow diagram was redesigned to accommodate this idea. The result was that the critical path shifted as shown in Fig. 10.19, for a reduction of 11 days. The new job (M_3, M_{17}) involved a trial of 70 handcrafted chassis units, at an additional expenditure of $350 ($5 × 70 units), thereby increasing overall expenses. This added expense averaged out to around $32 per reduction day. After the added expense was approved, the project progressed as shown in Fig. 10.19.

As you can tell from this example, when it is impossible to shorten the time required for a job on the critical path, it is easier to shorten the time by determining an alternative path using the arrow diagram.

Convenience of arrow diagrams

An arrow diagram can be used to express objectives in light of such considerations as size, type of project, management objectives, etc. The time-scale arrow diagram, which efficiently employs features of the Gantt chart is especially useful for small-scale projects or partial projects whose whole schedule period is relatively short.

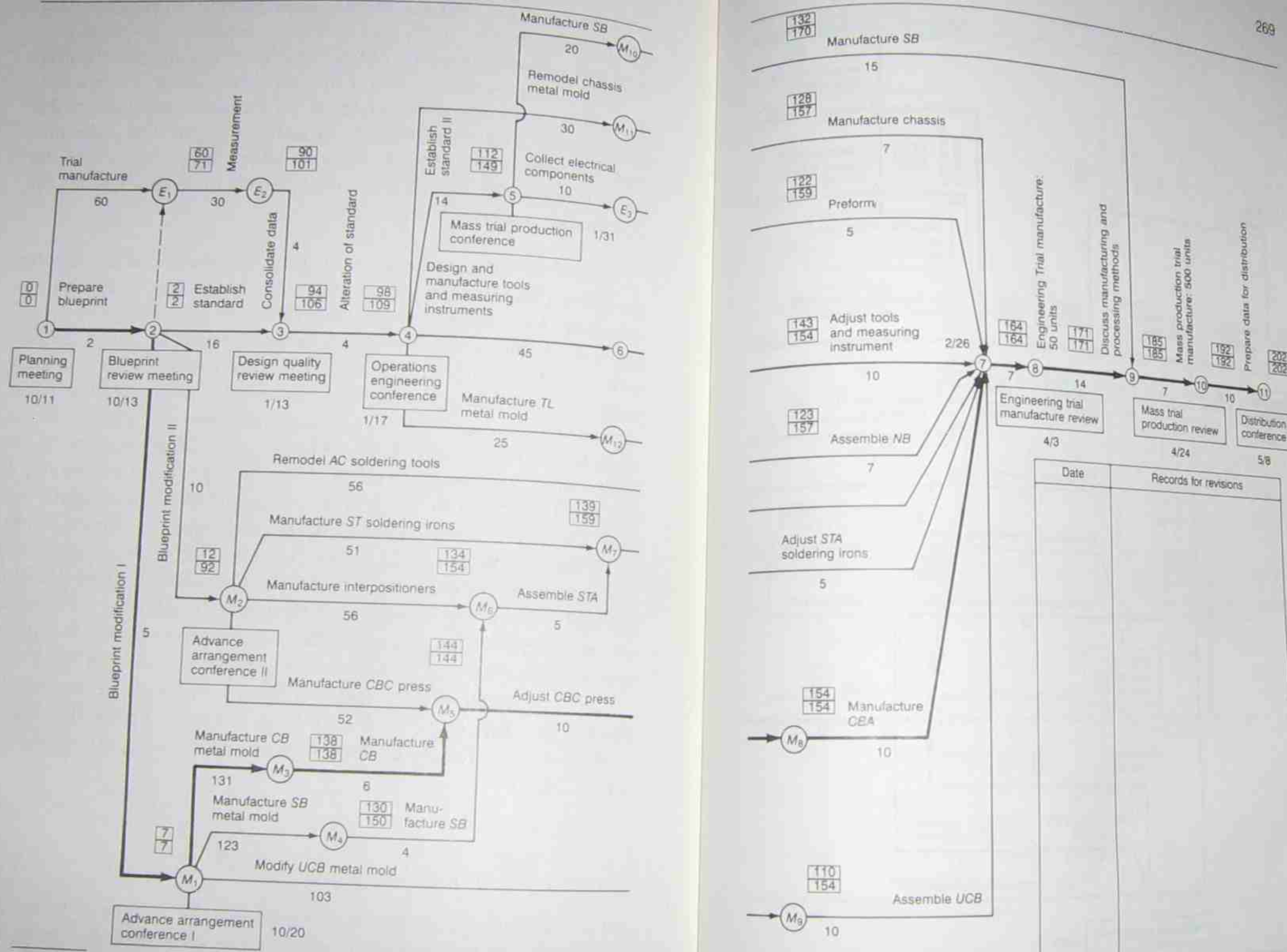


FIGURE 10.16
Electronic apparatus model 52 development plan (initial plan)

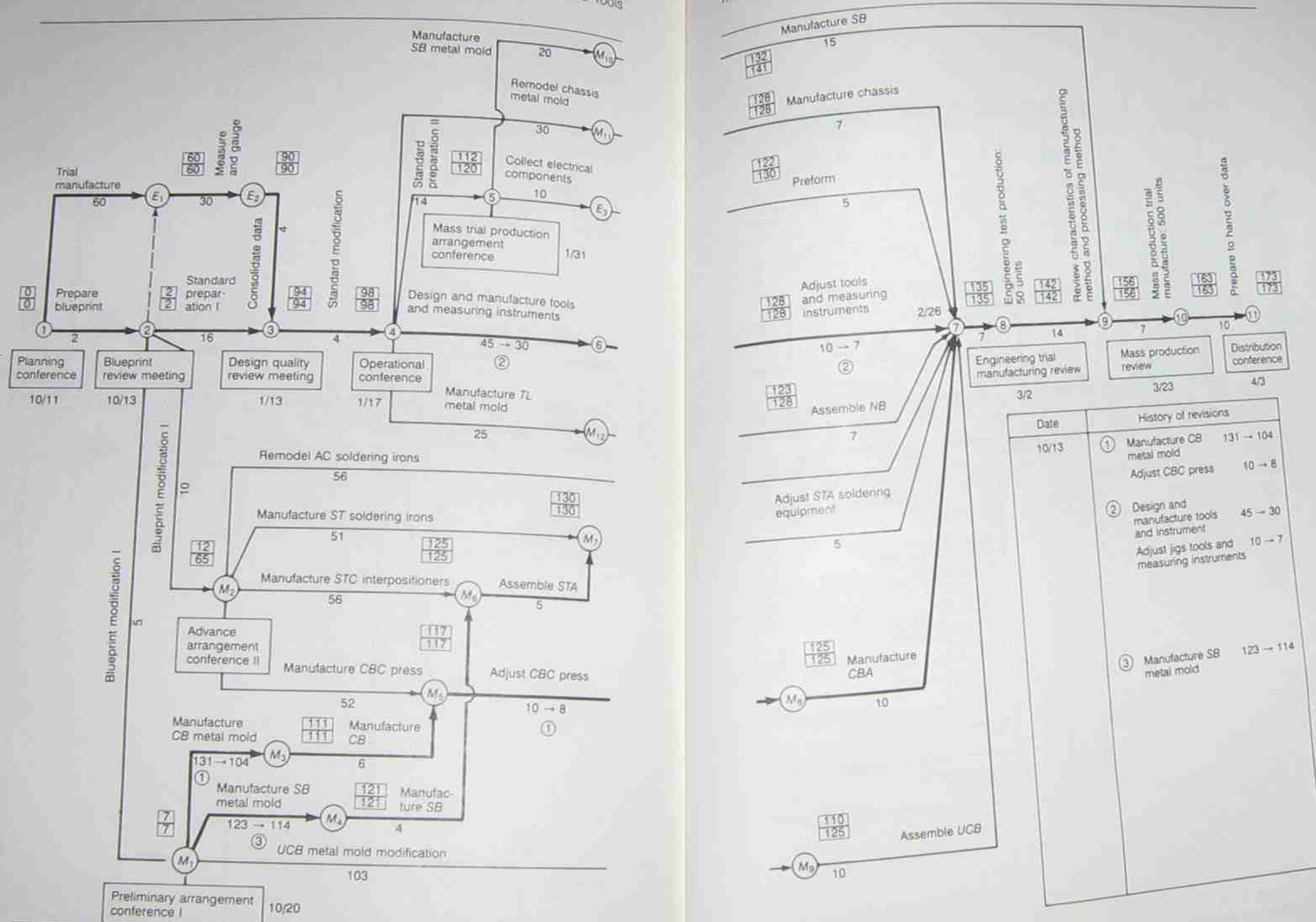


FIGURE 10.17
Electronic apparatus model 52 development plan (time-reduction plan)

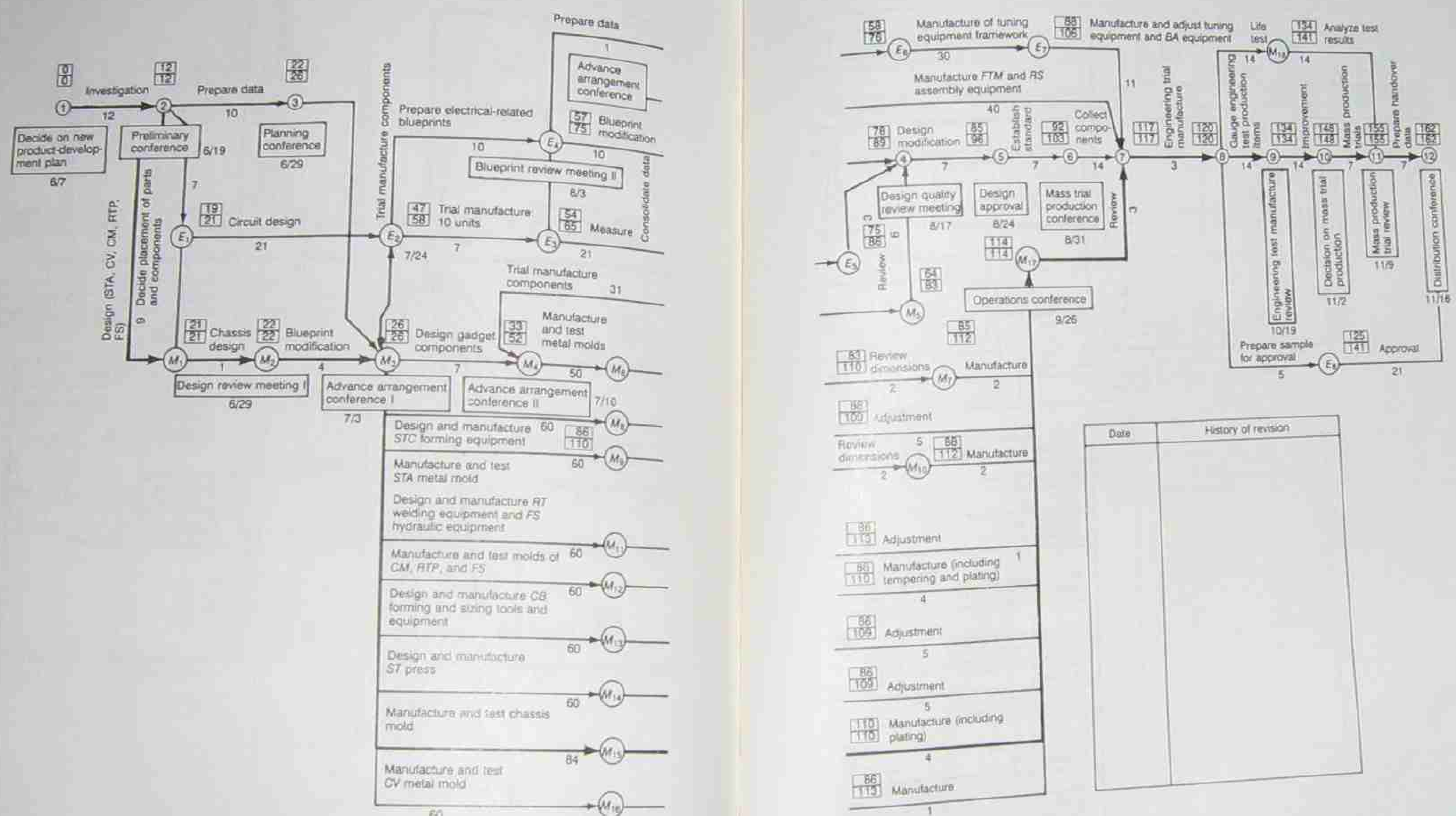


FIGURE 10.18
Electronic apparatus model 60 development plan (initial plan)

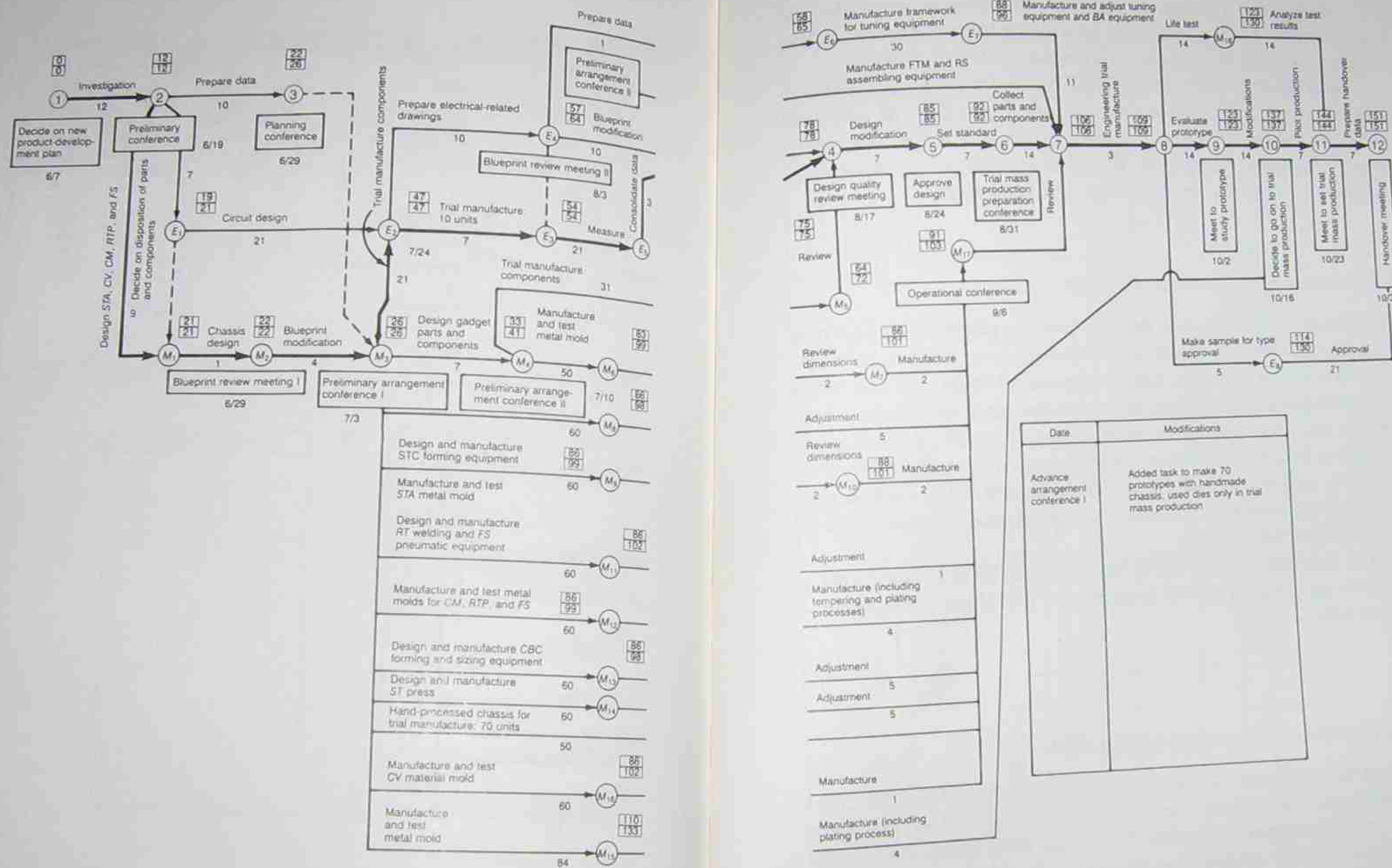


FIGURE 10.19
Electronic apparatus model 60 development plan (time-reduction plan)

In the time-scale arrow diagram correlations of jobs are described by nodes and arrows in reference to precedence, succession, and parallels, as in the arrow diagram. In addition, arrow length corresponds to the time required for a job, as in the Gantt chart. It is simple to use and there is no need for calculation of time.

The time-scale arrow diagram can be used to prepare a plan that does not need a full-scale arrow diagram to calculate its schedule and when the Gantt chart cannot be used because it does not satisfy requirements. An example of a plan that was prepared with a time-scale arrow diagram is given in Figs. 10.20 and 10.21.

How to express a plan with uncertain elements

It is difficult to develop jobs uniformly at the planning stage when there are uncertain elements involved, e.g., new product development. In this case, it is impossible to express all items with the nodesand arrows of an arrow diagram as was described earlier. An arrow diagram cannot be used if we do not have all the details. As the job progresses, however, the plan can be prepared by using alternative judgment nodes—decision boxes (represented by *a*)—when one process can be chosen from among several alternative processes that are considered in advance.

The decision box shown in Fig. 10.22 can be used to express a choice to go to jobs *R* or *S* or to go to jobs *T* or *U* after job *Q*'s completion point.

Synchronization of time and quality control

Methods to manage schedule control, in conjunction with non-time-related factors, were developed by the U.S. Navy in 1960 and have been put to practical use in PERT/cost and PERT/reliability. A relationship to quality control was advocated in 1963 in an article entitled "PERT and QC," by Professors Mizuno and Furukawa.[2] This theory has been successfully applied at Toyota Motor Company, Mitsubishi Heavy Industries, Inc., and Suzuki Motors Corporation under the guidance of Professors Mizuno and Furukawa.[3]

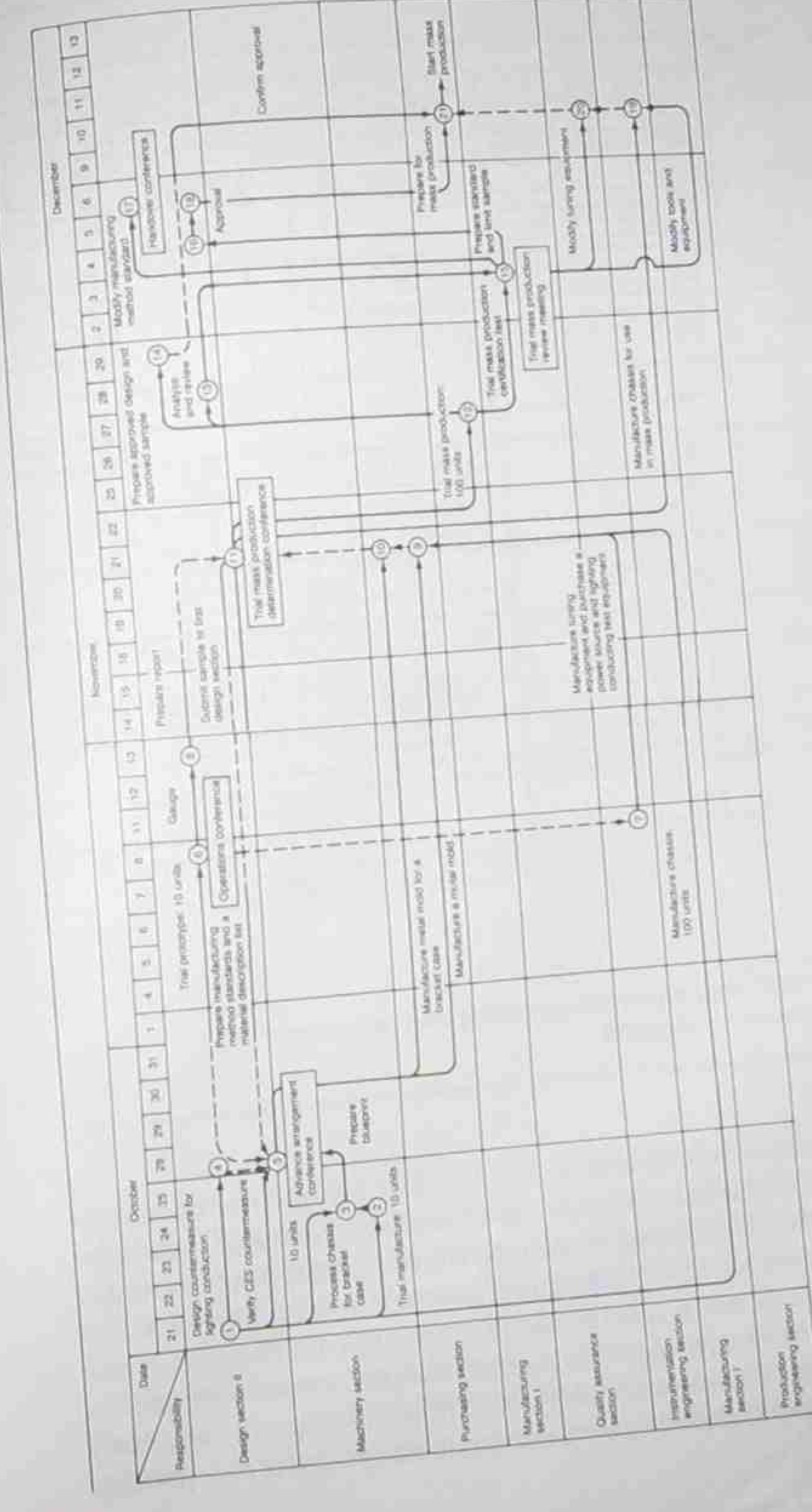


FIGURE 10.20 Model T-827 improvement plan (an example of a time-scale arrow diagram)

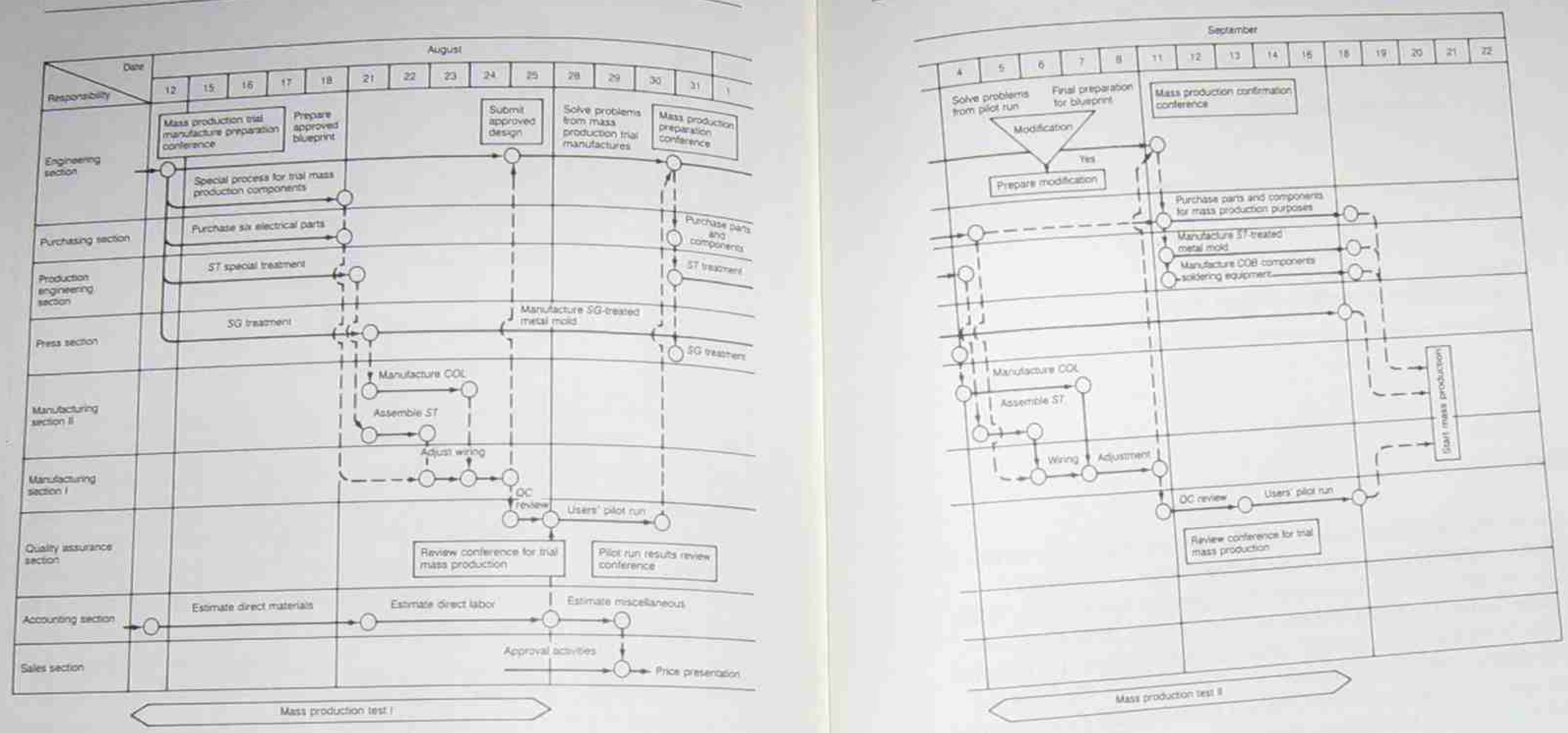


FIGURE 10.21
Model 62Q mass production promoting plan (an example of a time-scale arrow diagram)

Because the schedule and delivery date are also considered as points of quality, in a broad sense, they are essential for QC activities related to schedule control. Moreover, the relationship between synchronization of schedules and quality control is very important.

The critical path is used to identify schedule bottlenecks. It has also been used to surface problems in the implementation of quality improvements. Furthermore, the use of PERT/QC (quality and cost) to monitor the progress of new product development is also gaining

recognition. The PERT/QC not only monitors progress against schedules, but also helps in timing the building of quality into products, in the preparation of quality standards charts, and in controlling developmental and manufacturing costs.

In conclusion, we acknowledge our gratitude to Matsushita Electronic Components, Inc. and Yasukawa Electrical Manufacturing Company, Inc. for allowing us to publish their figures.

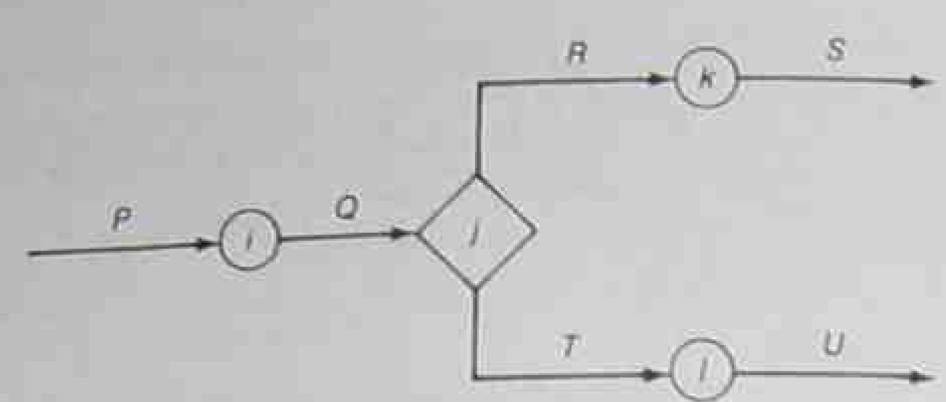


FIGURE 10.22
An arrow diagram used with a decision node

Notes

1. *See*, for example, Mori Tatsuo, *PERT – The New Way to Consolidate Jobs* (Tokyo: Japan Management Association, 1964); ———, *PERT II – Effective Application* (Tokyo: Japan Management Association, 1965); Sekine Tomoaki, *PERT/CPM Introduction* (Tokyo: JUSE Press, Ltd., 1965); Tone Kaoru (Ed.), *PERT Lecture I: Basics* (Tokyo: Toyo Keizai Shimposha, 1966); Kato Shokichi, *New Techniques for Planning and Management – Theory of PERT/CPM and Its Application* (Tokyo: Industrial Engineering Association, 1964).
2. Mizuno Shigeru and Furukawa Yasushi, "PERT and QC," *Hinshitsu Kanri* (Quality Control) (special issue), vol. 14 (November 1963).
3. Tone Kaoru (Ed.), *PERT Lecture IV: Applications* (Tokyo: Toyo Keizai Shimposha, 1967); Mizuno and Furukawa (Tokyo Engineering University) and Suzuki, Numano, Takimura, and Sato (Toyota Motor Company), "PERT and QC – Practical Examples in the Automobile Industry (2nd Report)," *Hinshitsu Kanri* (Quality Control), vol. 15 (fall, 1964), special issue, pp. 198–201; Mizuno and Furukawa (Tokyo Engineering University) and Mitsubishi Heavy Industries, Inc. and Mihara Manufacturing Corporation, "PERT and QC – Practical Examples in-the Field of Industrial Machinery (2nd Report)," *Hinshitsu Kanri* (Quality Control), vol. 15 (fall 1964), pp. 201–206; Mizuno, Furukawa,

and Oba (Tokyo Engineering University) and Watase (Suzuki Motor Corporation), "PERT and QC – Quality Evaluation in the Development of New Automobile Models," *Hinshitsu Kanri* (Quality Control), vol. 15 (fall 1964), special issue, pp. 207–211.

11

Education to Introduce the Seven New QC Tools

The seven new QC tools like other QC methods, have proven effective from the first as we have applied them to solve problems. Therefore, we hope that you will freely use each of the techniques illustrated in this book. The new tools are more effective however, if users are first given adequate education about and practice in their use. In this chapter we discuss three examples of education about and introduction of the seven new QC tools. Frst, we introduce the approach to education and training in classes held by the Japanese Union of Scientists and Engineers. Second, we discuss the popularization and implementation of the seven new QC tools by QC personnel from each factory in a class hosted by the central QC department of Company M. Finally, we discuss the progress toward universal use of the seven new QC tools in the TQC process at Company S. You will be headed in the right direction if you begin your education and training with one of the methods described in the first two sections and afterwards if your whole company or whole branch cooperates in their actual use, as described in the last section.

Education and training classes

Education methods

In this section, we introduce a method for educating participants to use the seven new QC tools more effectively. For corporations, day-to-day operations may require that the plan be flexible.

A comprehensive schedule is shown in Fig. 11.1. Classes are held once a month, over a period of nine months. Overall, 10 days are scheduled:

Practice: Participants divide into groups to solve a given problem using the method taught in the morning. Each group works together and presents its solution.

Homework: Participants apply the method taught that day to actual company problems and present their work during the next class.

Study project with company examples: Participants continue to work on departmental problems during the period of study. They prepare both interim and final reports.

Symposium: This is held on the final day of class and is open to nonparticipants. Examples of solutions found are introduced.

Each meeting's timetable is to be set up according to the daily schedule (see Fig. 11.2). Information discussed will differ according to the theme.

Recommended methods

Training site: Typically the number of participants range between 30 and 50. The meeting place can be either on or off company premises. Some care should be taken in considering location since the KJ method is presented on an overnight. Although group size can vary,

	A.M.	P.M.	
	Lecture and explanation of method	Study and practice	Notes
1st Month Date:	What are the new seven QC tools?	Tree charts, orientation: Projects, homework, etc.	Homework 1. Tree charts and graphs
2nd month Date:	Relations diagram	Practice with relations diagrams	Homework 2. Practice with relations diagrams; submit project topics
3rd month Date:	PDPC	Practice with PDPC, direction on projects	Homework 3. PDPC
4th month Date:	Matrix diagram	Presentation of homework and discussion; first project interim report	Homework 4. Matrix diagram
5th month Date:	Interpretation of matrix diagram data, implementation of policies	Direction on projects, interim report 2	Homework 5. Policy implementation
6th month Date:	KJ method	Practice with KJ method	One night for study
7th month Date:	Arrow diagram method	Practice with arrow diagram method, direction on projects	Homework 6. Arrow diagram method
8th month Date:	Presentation of projects; questions and answers General discussion		
9th month Date:	Symposium: 1. Explanation of the seven new QC tools 2. Presentation of examples; debate 3. Special lecture		Symposium is open to nonclass members.

FIGURE 11.1
Program (example)

if the training room is large enough, there is no need for small rooms for each group. Generally, it is sufficient that the place chosen be quiet and not subject to outside noise. Make certain that adequate supplies for the training are available, for example, poster paper, cards, and markers.

Time	Subject	Number of Data	Instructor
9:30–11:00 a.m.	Matrix diagram		
11:10–12:15 p.m.	Homework examples • Introduction and discussion		
	—Lunch break—		
1:15–4:35 p.m.	Presentation of homework • Project interim report— 25 minutes per person (presentation: 10 minutes; discussion: 15 minutes)		
4:45–5:30 p.m.	Class projects • General discussion		

FIGURE 11.2
Example of timetable

Note concerning the training period: Outside distractions should be avoided as much as possible during this period. In addition, to create a friendly atmosphere and to promote better communication among students belonging to the same company, participants should have time to relax at scheduled breaks. This increases the effectiveness of their work.

Lecturer: At JUSE-sponsored trainings either an experienced member of the company or an outside consultant serves as the lecturer. Each case has its own advantages and disadvantages. In some cases, however, it is more effective if outside people present the material. Depending on the size of the group several instructors are needed during practice sessions, to make the rounds and work with each group.

Extensive use of the seven new QC tools in company M

Company M has already begun to promote the extensive use of the seven new QC tools inside the company. The content of its initiative is not particularly outstanding; it serves as a simple example.

Staff participation in a seminar held outside the company

The QC department of the main office informed the QC department of each factory that seminars regarding the seven new QC tools would be held outside their company, and that at larger meetings the results of their use would be presented.

Starting the class inside the company

A class of volunteers was gathered, hosted by the QC department of the main office. The purpose of this regular meeting was to give lectures and practical training about the seven new QC tools to QC personnel at each factory so they could put the training into actual use at each factory. The agendas of these meetings follow:

First meeting

Overview of the seven new QC tools: If a lecturer is invited from outside, then the top executives will attend the meeting.

Lecture and practice: Practice with relations diagram and PDPC charts.

Homework: Study of a practical problem in which each participant is involved until the next meeting.

Second meeting

Lecture and practice: Tree diagram and matrix diagram.

Presentation of examples: Introduction of good examples inside or outside the company.

Homework: Study of a practical problem in which each participant is involved until the next meeting.

Third meeting

Presentation and critique of homework: Reports on homework; questions and answers, discussions, critique of participants' applications of the tools.

Report of representatives of each factory concerning their applications of the tools.

Fourth and succeeding meetings

Plan and adjust lectures, practices, reports, etc. at each meeting.

On the basis of Company M's experience the following observations can be made:

- One full day is needed for the initial lecture and practice of a single method.
- One half day is needed for further practice, including the presentation of results.
- A lecturer from outside is more effective with the participants than one from inside the company.
- The topics for practice should be chosen from among common and relatively simple problems known to the participants.
- Practice exercises should be carried out by groups of four to five members.
- The practice should not be dominated by particular members (participation should be relatively equal).
- The interval between meetings should be two to three months.

Other activities

The lectures about the seven new QC tools were promoted as activities during "Quality Month." The results of the establishment of the new tools were announced to the whole company through regular communication systems within the company. The seven QC tools were also promoted within the QC circles. The staff from the main office publicized the positive results achieved through use of the system and encouraged its use in each factory. The company backed the presentation outside the company of outstanding results achieved through use of the seven new QC tools. In doing so, it increased the participants' consciousness of quality control.

Introduction of the system to Factory A of Company S

General details

The activities of total quality control at Company S started when Factory T was awarded the Deming Prize in 1975. At the same time, Factory A started reviewing quality control. At that time, Factory A had several problems in the process of implementation. Among these were how to select critical points within the production processes of higher-level activities and how to coordinate production processes generally. These problems were solved by using relations diagrams and systematic diagrams from the seven new QC tools.

In addition to these steps, the factory used an arrow diagram for schedule control and a matrix diagram to grasp the relationships among processes during actual production. In research and development involving many uncertain factors and comparatively complex work, the PDPC method was used and produced several remarkable results.

The relations diagram, systematic diagram, and the arrow diagram were included in the QC manual, as shown in Fig. 11.3, and were put into actual use. In the area of technological research, program decision process charts were developed on all subjects and reviewed by engineering committees assigned to solve the problems.

Important points for implementation

The following points where found to be important during implementation. First, top management must have a thorough understanding of the method. At Factory A, the factory manager himself proposed the policy, provided guidance on how to use the method, and was eager to lead other members. Second, these methods must be used consistently during the development and establishment of process control. These methods should not be used in isolation. Third, through regular study meetings, all participants have become very conscious of the methods.

There was no definite procedure for the introduction. The steps taken at Factory A were as follows:

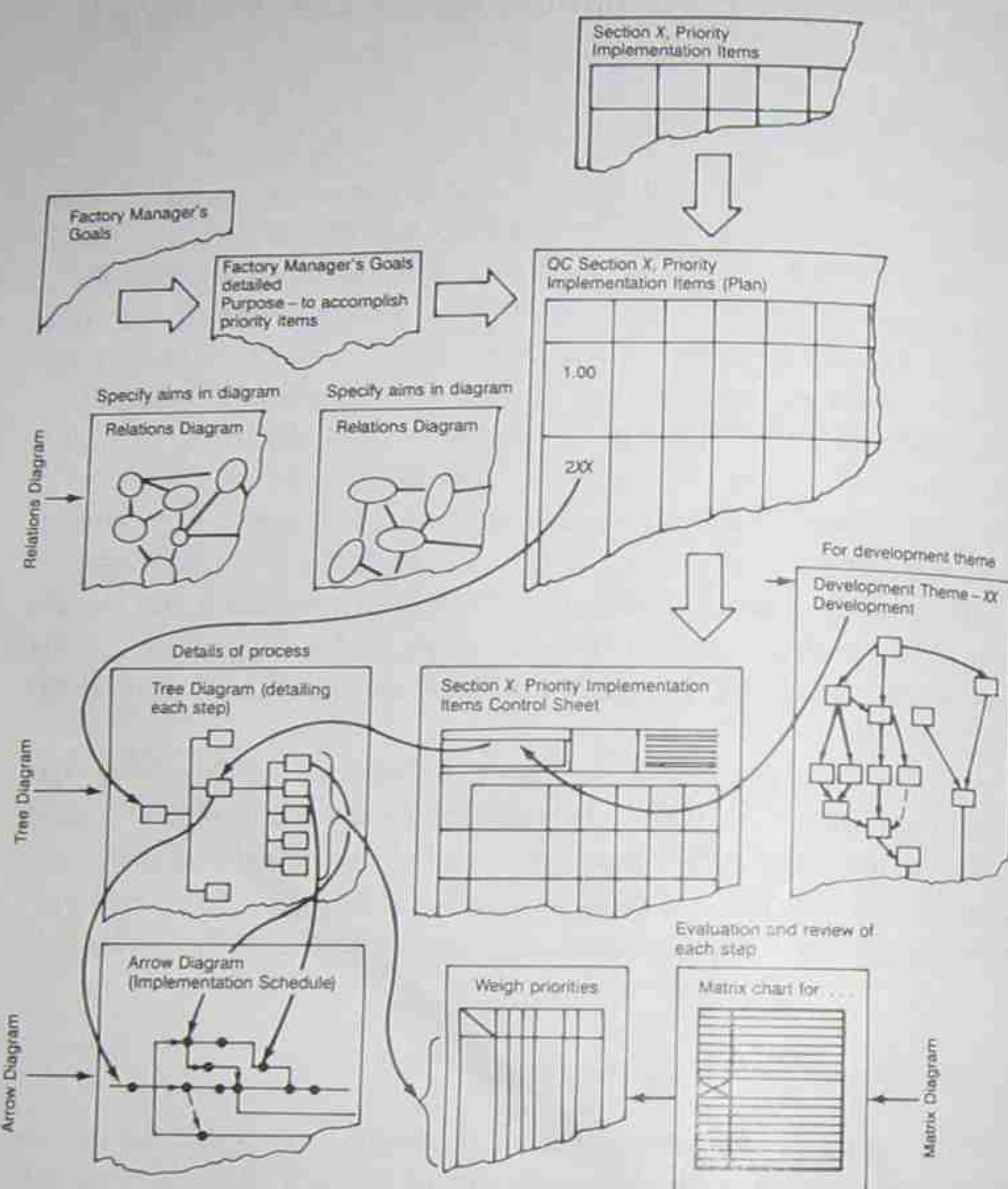


FIGURE 11.3
Process control and the seven tools for quality control

1. Introductory procedures. The introductory study and practice in the company was provided through on-the-job training. The presentation of examples was made at monthly meetings. The director of each department or division reported the problems of day-to-day

operation and the process of their solution. Then discussions followed. After this meeting, a review meeting was scheduled. At the review meeting, participants pursued the improvement of matters pointed out at the monthly meeting and reviewed any new topics. The membership consisted of directors of divisions and higher staff. Then group study meetings were set up. Small groups of four to five people discussed how to apply the matrix diagram method and PDPC method to actual operations. This was mainly for staff members.

2. Seminars outside the company. Factory A staff participated in the following seminars (participants included section managers, chiefs, and staff members): Seven New QC Tools Symposium, sponsored by JUSE; Multiple Variable Analysis Lecture, by JUSE.; and the Seven New QC Tools for QC Study Meeting.

Conclusion

Future of the seven new QC tools

We have tried to introduce simply the ideas, methods, and examples of actual use of the seven new QC tools. Since the first meeting of the QC Methods Research Division at the Osaka office of the Japanese Union of Scientists and Engineers, study and use of the seven new QC tools has been considerable. We hope to see this trend continue. Here are some of our hopes for the future of the seven new QC tools:

1. *To supply more case studies of the seven new QC tools.* In order for more people to understand and use the seven new QC tools we need to provide many examples of their application in different fields. Over the years JUSE has promoted this goal by sponsoring symposia and conferences to present case studies prepared by members as well as nonmembers in many different fields. Since many examples involve proprietary information, it is unfortunate that more case studies cannot be made available. For the future, however, we are

content to hope that we will see more examples that effectively combine both the old and new QC methods.

2. *To incorporate the seven new QC tools into the QC system.* In the second section of Chap. 3, we reported the results of the application of the seven new tools in process control. We hope to see similar efforts made in each system of quality assurance, new product research, product liability, and so on. These efforts, combined with a union of new and traditional QC methods should facilitate more efficient establishment of corporate quality control systems.

3. *To use the seven new QC tools more comprehensively.* We have observed over the years that isolated use of individual QC tools does not produce significant results. Like the seven traditional tools, the seven new tools will become essential parts of quality control as the time and effort is taken to apply them companywide and to apply them consistently and in effective combinations with other tools.

4. *To promote the seven new QC tools to top management.* Top management understanding of the new tools and leadership is very important. To realize the first three goals, the chief executive officer of each corporation or division and the person in charge of corporate quality control must firmly direct the application of the seven new tools along with the traditional tools in the development of their processes and in every new product developed.

5. *To see more use of the relations diagram method and the PDPC method.* Some Japanese observers have commented that drawing up these charts is particularly difficult, because there is no formalized procedure for their preparation. Because of the very freedom with which they can be drawn, however, we anticipate wide and successful application. We are promoting this goal through our continuing course offerings and research.

6. *To discover and add new, effective methods to the seven new QC tools.* Our current concept of seven new tools does not preclude the development of additional tools. This means that we will continue to seek new methods applicable to the various control techniques, such as operations research, value engineering, creative techniques, management, and economics. In addition, we will initiate their application to quality control. The Osaka office of the Japanese Union of Scientists and Engineers has established the N Seminar, which grew from the QC Methods Research Seminar in April of 1978. We are

studying in the techniques (as described above) and their systematic use (as shown in item 2).

7. *To introduce the QC circle to the seven new QC tools.* The relations diagram, systematic diagram, matrix diagram, and program decision process chart (PDPC) do not have complicated rules; hence, they are not difficult to use. We promote the introduction of these methods specifically to QC circles as industrial problem-solving tools.

Index

A

R

S

T

CEDAC
A Tool for Continuous Systematic Improvement
by Ryuji Fukuda

CEDAC, or Cause and Effect Diagram with the Addition of Cards, is a modification of the "fishbone diagram," one of the standard QC tools. One of the most powerful, yet simple problem-solving methods to come out of Japan (Fukuda won a Deming Prize for developing it), CEDAC actually encompasses a whole cluster of tools for continuous systematic improvement. They include window analysis (for identifying problems), the CEDAC diagram (for analyzing problems and developing standards), and window development (for ensuring adherence to standards). Here is Fukuda's manual for the in-house support of improvement activities using CEDAC. It provides step by step directions for setting up and using CEDAC. With a text that's concise, clear, and to the point, nearly 50 illustrations and sample forms suitable for transparencies, and a removable CEDAC wall chart, the manual is an ideal training aid.
ISBN 0-915299-26-7 / 142 pages / $49.95 / Order code CEDAC-BK

The Visual Factory
Building Participation Through Shared Information
by Michel Greif

If you're aware of the tremendous improvements achieved in productivity and quality as a result of employee involvement, then you'll appreciate the great value of creating a visual factory. This book shows how visual management can be used to make the factory a place where workers and supervisors freely communicate and take improvement action. It details how to develop meeting and communication areas, communicate work standards and instructions, use visual production controls such as kanban, and make goals and progress visible. Over 200 diagrams and photos illustrate the numerous visual techniques discussed.
ISBN 0-915299-67-4 / 256 pages / $49.95 / Order Code VFAC-BK

JIT Factory Revolution
A Pictorial Guide to Factory Design of the Future
by Hiroyuki Hirano / JIT Management Library

Here is the first-ever encyclopedic picture book of JIT. With 240 pages of photos, cartoons, and diagrams, this unprecedented behind-the-scenes look at actual production and assembly plants shows you exactly how JIT looks and functions. It shows you how to set up each area of a JIT plant and provides hundreds of useful ideas you can implement. If you've made the crucial decision to run production using JIT and want to show your employees what it's all about, this book is a must. The photographs, from Japanese production and assembly plants, provide vivid depictions of what work is like in a JIT environment. And the text, simple and easy to read, makes all the essentials crystal clear.
ISBN 0-915299-44-5 / 227 pages / $49.95 / Order code JITFAC-BK

Canon Production System
Creative Involvement of the Total Workforce
compiled by the Japan Management Association

A fantastic success story! Canon set a goal to increase productivity by three percent per month — and achieved it! The first book-length case study to show how to combine the most effective Japanese management principles and quality improvement techniques into one overall strategy that improves every area of the company on a continual basis. Shows how the major QC tools are applied in a matrix management model.
ISBN 0-915299-06-2 / 251 pages / $36.95 / Order code CAN-BK

TQC for Accounting
A New Role in Company-wide Improvement
by Takashi Kanatsu

TQC for accounting means more than streamlining office procedures or upgrading financial analyses. It requires, instead, a linking of the basics of marketing with the fundamentals of accounting through the medium of TCQ. This book is a guide for top and middle managers who wish to turn their companies around by redesigning the roles played by the accounting, sales, and marketing departments. The book's format offers detailed examinations of accounting TCQ in relation to a company's business plan, accounting department, and specific statistical methods. Its use will help to create the "awareness revolution" that is imperative in turning around a factory or any type of company. (Winter, 1991)
ISBN 0-915299-73-9 / 176 pages / $45.00 / Order code TQCA-BK

TQC Wisdom of Japan
Managing for Total Quality Control
by Hajime Karatsu, translated by David J. Lu

As productivity goes up, the cost of quality comes down. And as quality improves, the cost to produce comes down. Karatsu, winner of a Deming Prize who has been involved with the quality movement in Japan since its inception, discusses the purpose and techniques of Total Quality Control (TQC), how it differs from QC, and why it is so effective. There is no better introduction to TQC than this book; essential reading for all American managers.
ISBN 0-915299-18-6 / 136 pages / $34.95 / Order code WISD-BK

20 Keys to Workplace Improvement
by Iwao Kobayashi

This easy-to-read introduction to the "20 keys" system presents an integrated approach to assessing and improving your company's competitive level. The book focuses on systematic improvement through five levels of achievement in such primary areas as industrial housekeeping, small group activities, quick changeover techniques, equipment maintenance, and computerization. A scoring guide is included, along with information to help plan a strategy for your company's world class improvement effort
ISBN 0-915299-61-5 / 264 pages / $34.95 / Order code 20KEYS-BK

Total Manufacturing Management
Production Organization for the 1990s

by Giorgio Merli

One of Italy's leading consultants discusses the implementation of Just-In-Time and related methods (including QFD and TPM) in Western corporations. The author does not approach JIT from a mechanistic orientation aimed simply at production efficiency. Rather, he discusses JIT from the perspective of industrial strategy and as an overall organizational model. Here's a sophisticated program for organizational reform that shows how JIT can be applied even in types of production that have often been neglected in the West, including custom work.
ISBN 0-915299-58-5 / 224 pages / $39.95 / Order code TMM-BK

Poka-Yoke
Improving Product Quality by Preventing Defects

compiled by Nikkan Kogyo Shimbun, Ltd./Factory Magazine (ed.)
preface by Shigeo Shingo

If your goal is 100% zero defects, here is the book for you—a completely illustrated guide to poka-yoke (mistake-proofing) for supervisors and shop-floor workers. Many poka-yoke devices come from line workers and are implemented with the help of engineering staff. The result is better product quality—and greater participation by workers in efforts to improve your processes, your products, and your company as a whole.
ISBN 0-915299-31-3 / 288 pages / $59.95 / Order code IPOKA-BK

Achieving Total Quality Management
A Program for Action

by Michel Perigord

This is an outstanding book on total quality management (TQM)—a compact guide to the concepts, methods, and techniques involved in achieving total quality. It shows you how to make TQM a company-wide strategy, not just in technical areas, but in marketing and administration as well. Written in an accessible, instructive style by top European quality expert, it is methodical, logical, and thorough. A historical outline and discussion of the quality-price relationship, is followed by an investigation of the five quality imperatives (conformity, prevention, excellence, measurement, and responsibility). Major methods and tools for total quality are spelled out and implementation strategies are reviewed.
ISBN 0-915299-60-7 / 384 pages / $39.95 / Order Code ACHTQM-BK

Zero Quality Control
Source Inspection and the Poka-yoke System

by Shigeo Shingo, translated by Andrew P. Dillon

A remarkable combination of source inspection (to detect errors before they become defects) and mistake-proofing devices (to weed out defects before they can be passed down the production line) eliminates the need for statistical quality control. Shingo shows how this proven system for reducing defects to zero turns out the highest quality products in the shortest period of time. With over 100 specific examples illustrated. (Audio-visual training program also available.)
ISBN 0-915299-07-0 / 328 pages / $70.00 / Order code ZQC-BK

Productivity Newsletter

Productivity Newsletter has been helping America's most effective companies improve quality, lower costs, and increase their competitive power since 1979. Each monthly issue contains detailed case studies, articles on important innovations and world trends, book reviews, and much more. Subscribers save money on Productivity conferences and seminars. To subscribe, or for more information, call 1-800-888-6485. Please state order code "BA" when ordering.

Shingo, Shigeo. **A Revolution In Manufacturing: The SMED System**, Translated by Andrew P. Dillon
ISBN 0-915299-03-8 / 1985 / 383 pages / $70.00 / order code SMED

Shingo, Shigeo. **The Sayings of Shigeo Shingo: Key Strategies for Plant Improvement**, Translated by Andrew P. Dillon
ISBN 0-915299-15-1 / 1987 / 208 pages / $39.95 / order code SAY

Shingo, Shigeo. **A Study of the Toyota Production System from an Industrial Engineering Viewpoint** (rev.)
ISBN 0-915299-17-8 / 1989 / 293 pages / $39.95 / order code STREV

Shingo, Shigeo. **Zero Quality Control: Source Inspection and the Poka-yoke System**, Translated by Andrew P. Dillon
ISBN 0-915299-07-0 / 1986 / 328 pages / $70.00 / order code ZQC

Shinohara, Isao (ed.). **New Production System: JIT Crossing Industry Boundaries**
ISBN 0-915299-21-6 / 1988 / 224 pages / $34.95 / order code NPS

Sugiyama, Tomo. **The Improvement Book: Creating the Problem-Free Workplace**
ISBN 0-915299-47-X / 1989 / 236 pages / $49.95 / order code IB

Suzue, Toshio and Akira Kohdate. **Variety Reduction Program (VRP): A Production Strategy for Product Diversification**
ISBN 0-915299-32-1 / 1990 / 164 pages / $59.95 / order code VRP

Tateisi, Kazuma. **The Eternal Venture Spirit: An Executive's Practical Philosophy**
ISBN 0-915299-55-0 / 1989 / 208 pages/ $19.95 / order code EVS

AUDIO-VISUAL PROGRAMS

Japan Management Association. **Total Productive Maintenance: Maximizing Productivity and Quality**
ISBN 0-915299-46-1 / 167 slides / 1989 / $749.00 / order code STPM
ISBN 0-915299-49-6 / 2 videos / 1989 / $749.00 / order code VTPM

Shingo, Shigeo. **The SMED System**, Translated by Andrew P. Dillon
ISBN 0-915299-11-9 / 181 slides / 1986 / $749.00 / order code S5
ISBN 0-915299-27-5 / 2 videos / 1987 / $749.00 / order code V5

Shingo, Shigeo. **The Poka-yoke System**, Translated by Andrew P. Dillon
ISBN 0-915299-13-5 / 235 slides / 1987 / $749.00 / order code S6
ISBN 0-915299-28-3 / 2 videos / 1987 / $749.00 / order code V6

TO ORDER: Write, phone, or fax Productivity Press, Dept. BK, P.O. Box 3007, Cambridge, MA 02140, phone 1-800-274-9911, fax 617-868-3524. Send check or charge to your credit card (American Express, Visa, MasterCard accepted).

U.S. ORDERS: Add $4 shipping for first book, $2 each additional for UPS surface delivery. CT residents add 8% and MA residents 5% sales tax.

INTERNATIONAL ORDERS: Write, phone, or fax for quote and indicate shipping method desired. Pre-payment in U.S. dollars must accompany your order (checks must be drawn on U.S. banks). When quote is returned with payment, your order will be shipped promptly by the method requested.

NOTE: Prices subject to change without notice.

THE 7 NEW QC TOOLS

Now you can manage any size quality project on any level of complexity! While the original seven QC tools analyze and solve quality problems, the seven *new* tools promote a higher level of QC activity — total coordination of the work-force and creative, forward-looking planning for QC. This management system, incorporating existing tools from other fields and specially designed tools, is now being applied in quality improvement with great success.

- [illegible] — clarifies causal relationships
- [illegible] — organizes problem-solving data
- [illegible] — discovers appropriate means to accomplish objectives
- [illegible] — pursues multi-dimensional thinking
- [illegible] — arranges numerical data in a [illegible]
- [illegible] program
- [illegible] daily plans

The Seven [illegible] Society for QC Technique Development [illegible] Japanese Scientists and Engineers [illegible]

ISBN 0-915299-29-1

90000

Second Edition

Best Practice in Inventory Management

Tony Wild

THE INSTITUTE OF OPERATIONS MANAGEMENT
BH

Best Practice in Inventory Management

Second edition

Tony Wild

THE INSTITUTE OF OPERATIONS MANAGEMENT

BUTTERWORTH HEINEMANN

AMSTERDAM BOSTON HEIDELBERG LONDON NEW YORK OXFORD
PARIS SAN DIEGO SAN FRANCISCO SINGAPORE SYDNEY TOKYO

Elsevier Butterworth-Heinemann
Linacre House, Jordan Hill, Oxford OX2 8DP
30 Corporate Drive, Burlington, MA 01803

First published by John Wiley & Sons 1997
Second edition published by Elsevier Ltd 2002
Reprinted 2003, 2005, 2006

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication Data
A catalogue record for this book is available from the Library of Congress

ISBN 0 7506 5458 9

For information on all Elsevier Butterworth-Heinemann publications visit our website at www.bh.com

Composition by Genesis Typesetting, Rochester, Kent
Printed and bound in Great Britain by Biddles Ltd, Kings Lynn, Norfolk

Contents

Preface

In the distant past when I ascended from being a scientist to the heights of inventory analyst, I was surprised at the lack of helpful written work on the subject. Although there are now many books, manuals and papers, there does not seem to be much written on the very important subject of 'how to do inventory control faced with the real inventory, customers, and an assortment of inventory records and styles of suppliers'.

Yet better control of stock (inventory) can give major benefits to the profitability of all companies. This book is aimed at showing how good inventory control can be used in practice. It is a result of working continuously on inventory control with a large number of companies over many years. It contains the distilled techniques which have been tried out and proven to work.

This book contains two main ingredients, namely:

- the basics of inventory management as covered in the Institute of Operations Management (IOM) Diploma syllabus
- the application of these techniques to real inventory management.

The applications are what really count, since the knowledge is not at all interesting without application. The experience of the author has been to use the basics to provide powerful changes in inventory and profitability. In several cases millions of pounds of inventory value have been saved and in others customer service has been greatly improved. The topics here have been tried, tested and approved. They are all a matter of common sense, but which piece of common sense to use in a particular situation is a matter that needs deeper understanding.

The book is a driver's manual for Inventory Controllers. It covers the working of the engine (how inventory control techniques work) how to use the controls (what the techniques do and how to manage them), and how to get the best out of the vehicle (how to optimize inventory). Understand the text and it will show the way to guaranteed improved inventory control, reduced stock levels and higher availability.

In preparation for this book there have contributions from the many thousands of people to whom I have presented courses on these ideas, and who have helped to hone the concepts. My thanks also to those companies who have let me loose on their inventory, carrying out consultancy projects which have extended my knowledge and proved the effectiveness and scope for different techniques.

Central to the development of inventory techniques has been the Institute of Operations Management (formerly BPICS) which has been a channel for professional debate, friendship, and technical information. This book has been created with the support of Richard Turner as Secretary of the Institute.

The evolution of this book has been a two-handed operation. The concepts have been blended into a logical text through the work of myself and Elaine Duckworth, who has shared the task of developing a useful and readable handbook.

Dr Tony Wild

Introduction

- Why we need professional inventory management
- Structure of the book
- Focus on developing theory that works in practice

Everyone is an inventory controller, at home and at work. We all keep food, clothes, domestic items, paper, pens and many other goods. We also have shortages and emergency purchases. Some people regularly have to throw out the contents of the refrigerator because they have been there a while and changed in character. So inventory control is a natural occupation that everyone undertakes, some more successfully than others.

There are many other activities that people do, such as sport, carpentry, music and medicine, and for these it is generally accepted that a level of technical expertise is needed. In carpentry and music there are people who become highly skilled through their natural talents. Success appears to come from a mixture of talent, determination, practice and knowledge. In medicine, particularly in surgery, we are more likely to rely on those who have studied the subject and have a deep knowledge of both theory and practice. Lawyers cannot function without knowledge; often a common-sense approach does not work for them.

As with any profession, inventory control (or the subject of Professional Inventory Management) has a body of techniques and knowledge which differentiates the professional from the do-it-yourself (DIY) enthusiast. This expertise is the result of 100 years of development and refining. Now there are professional inventory managers with recognized qualifications, including Diploma in Production and Inventory Management (DPIM) and Certificate in Operations Management (COM). There are still colleagues who believe that inventory can be managed using a clear brain and common sense. It is unlikely that they would be happy with the same competency in their doctor or financial adviser.

If you want to do better you need to train.

This book is for the inventory control practitioner. With the techniques described many people have been able to manage their stock of inventory so that their customers are happier and so are the accountants. The reduction of inventory value, the avoidance of unnecessary work and the improvement of customer service can be accomplished at the same time through simple application of the techniques discussed. Improved inventory management has been shown to halve stocks and improve service at the same time.

The techniques described are the basic concepts which inventory controllers should have at their fingertips. The more complex theoretical approaches can be left to those developing sophisticated systems, or those with excellent inventory control (and over the years I have not met any of these!). Inventory practitioners should be able to use this book to understand the best approaches and then to apply them to their own circumstances. Simple application of the methods is most successful, while modifications usually result in less effective outcomes.

The text covers the syllabus for the technical qualifications including DPIM and COM. It discusses good practice, inventory theory and the practical application of the techniques. The structure of the book is to start on inventory structure with target setting in Chapters 1 and 2, and how to structure inventory in Chapter 3. This chapter shows how a big impact on inventory levels can be made easily. The approach continues to give major benefits in tightening control. Continuing the theme of restructuring stock leads to the ultimate in stock management – the zero inventory approach – just in time (JIT), discussed in Chapter 4. Gradual absorption of the philosophy throughout inventory management is proving a great benefit, and it is worth examining how to utilize JIT in each stock environment. The rate of progress is often governed by the attitude and/organization in a company, and the development of lean inventory management practice is reviewed in Chapter 5.

The discussion then gets down to the detail of item-level inventory management, the essence of real inventory management. Chapters 6 to 8 are the elixir of item-level stock control, the secret recipe for success. Chapter 6 concentrates on how to keep out of trouble – a rational approach to the risks of stockouts, Chapter 7 shows how to use this information to set stock levels and Chapter 8 discusses the options for ensuring suppliers are agreeable and flexible.

The third aspect of inventory management, one which is usually carried out poorly, is forecasting. Forecasting methods are relatively well advanced,

but there is less practical application of these in most companies. This is an opportunity, discussed in Chapter 9, while the basic techniques are discussed in Chapter 10. Chapter 10 also includes the techniques used to ensure that forecasts produce sensible results. In practice there are items where it is worth managing the stock in more detail, so in Chapter 11 there are some more techniques which are commonly used and provide good forecasts.

The one item at a time approach to inventory management is really not appropriate in many situations. Users often require a range of items to do a job. For example, servicing the car requires filters, plugs, oil, etc., all available at one time. Inventory controllers should therefore balance the stocks to meet the structure demand. Unfortunately, the usual techniques do not help, and so the approach described in Chapter 12 is applied. The technique, developed over many years in manufacturing, is an important tool for general inventory management, although the application does not need the complexity used in manufacturing.

The inventory manager who has a good knowledge of these topics can then look round for new challenges. Industry has not been slow to provide them. In addition to the increasing pressure on inventory value and availability levels, there are new dimensions being added. The new challenges are in logistics and supply chain management. The balancing of inventory in one store is not sufficient; now the inventory, distribution costs and manufacturing efficiency have all to be optimized at the same time. Chapter 13 reviews the opportunities in this area.

Chapters 3, 6 and 7 are the bedrock of inventory practice. From experience, inventory management as described in this book works well, and the purpose of these pages is to enable more individuals to use the methods with confidence.

Readers outside the UK may find that a little translation is required. For instance in America the term 'inventory management' is applied to the whole subject, whereas in the UK it is normally known as 'stock control'. The terms 'inventory' and 'stock' are both used in the text to mean the same (in the UK inventory has a financial flavour and stock an operational context). It is hoped that this and other nuances of translation do not cloud the understanding of this technical and fascinating subject, whatever it is called.

1

The basis of inventory control

- The conflicts which beset inventory controllers.
- What inventory is there to achieve?
- How good inventory management will improve profitability.
- Analysing current stockholding.

The role of inventory management

The success of a venture depends on its ability to provide services to customers or users and remain financially viable. For an organization which is supplying goods to its customers, the major activity is to have suitable products available at an acceptable price within a reasonable timescale. Many parts of a business are involved in setting up this situation. Initially these are the marketing and design departments. Then purchasing, and in some cases, manufacture is involved. For an item which is already in the marketplace, the main activity is providing a continuity of supply for the customers.

Inventory control is the activity which organizes the availability of items to the customers. It co-ordinates the purchasing, manufacturing and distribution functions to meet the marketing needs. This role includes the supply of current sales items, new products, consumables, spare parts, obsolescent items and all other supplies.

Inventory enables a company to support the customer service, logistic or manufacturing activities in situations where purchase or manufacture of the items is not able to satisfy the demand. Lack of satisfaction could arise either because of the speed of purchasing or manufacturing is too protracted, or because quantities cannot be provided without stocks.

Stock control exists at a crossroads in the activities of a company (Figure 1.1). Many of the activities depend on the correct level of stock being held, but the definition of the term 'correct level' varies dependent upon which activity is defining the stock. Stock control is definitely a balancing act between the conflicting requirements of the company, and the prime reason

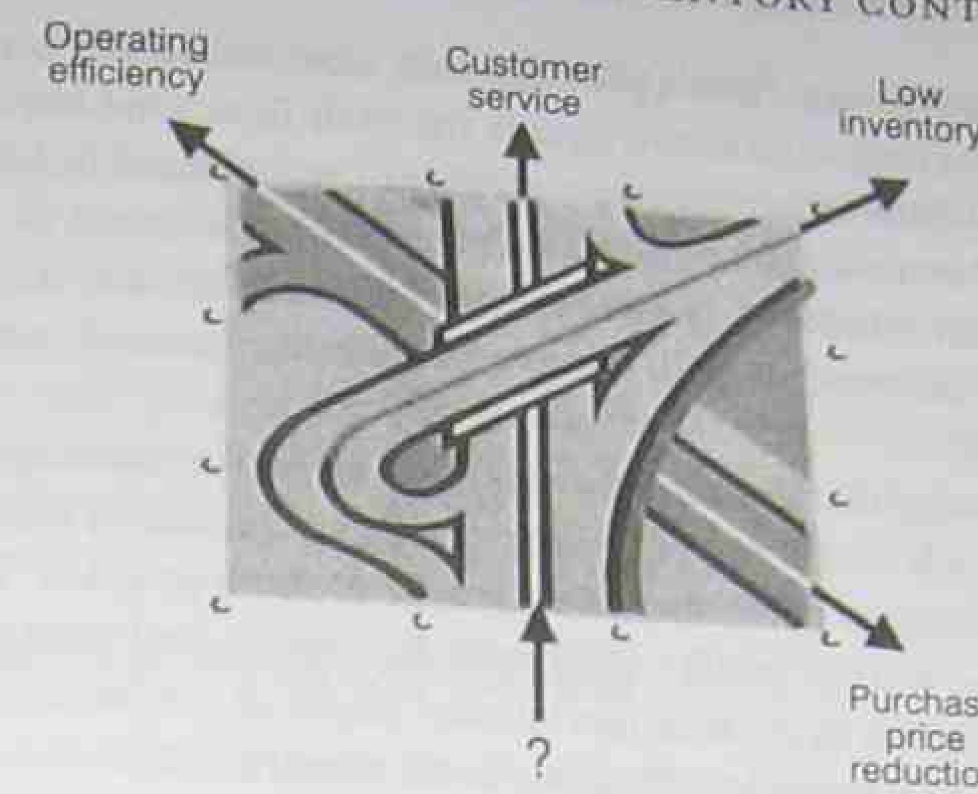


Figure 1:1 Which direction?

for the development of inventory management is to resolve this conflict in the best interest of the business. A conventional supply organization will have many departments including sales, purchasing, finance, quality assurance, contracts and general administration. In some cases there will also be manufacturing, distribution or support services or a variety of industry-specific activities. Each of these has a particular view of the role of stock control. Note the terms 'inventory' and 'stock' are used to mean the same thing in this book.

Sales consider that good stock control enables the company to have available any item which will meet immediate sales for as large a quantity as demanded: this requires large stock. For service companies where parts service is involved, the control of stock at the customer interface is traditionally left with the person carrying out the service, and this has led to overstocking and poor control. Similarly in distribution, the effect of bulking up shipments will lead to high stock levels, and a compromise has to be reached.

Inventory control keeps balancing conflicting requirements.

Purchasing consider that stock control provides the opportunity for goods to be purchased so that optimum prices can be obtained. Buying items in bulk often reduces the purchase price and it improves the efficiency within the purchasing department. The store is a means of keeping the bulk purchase items after buying advantageously.

Finance departments have a problem with stock because it consumes vast amounts of working capital and upsets the cash flow. One benefit of stock from a financial standpoint is that provisions can be made in case the stock turns out to be unsaleable, and this value can be adjusted to modify the profit figure in times of good or bad financial results. However, the existence of these provisions in the first place is detrimental to the finances of the company.

Quality management normally has the effect of slowing down the progress of stock while the necessary checks are made. This means that quality and inventory personnel effectively work in opposition. The gradual introduction of formal quality standards for supply and manufacture has reduced this conflict in most organizations. Supplier conformance has been improved enabling intrusive checking to be minimized.

General management see stock control, rightly, as a source of information. Some managers consider that they should be able to use stock control to give an immediate supply of information, statistics and forecasts. This can result in large amounts of unstructured work collecting, analysing and providing information.

The traditional view of manufacturing companies has been that large batches reduce the direct production costs. Manufacturing management tends to aim more for plant and labour efficiency and allow high stocks in order to avoid the disruptions caused by shortages, breakdowns and changing customer demand.

Good stock controllers (or materials management) will keep the stock down as long as they are responsible for stocks of everything, including raw materials, finished and semi-finished stocks, consumables, tools, work on the shop floor and, in fact, all inventory items.

Do not confuse high stock with good availability.

These views are now subject to question. Those who agree with these ideas should be considering a rethink and change to a consensus on stock management. The development of stock control practice has been given impetus by the move toward total quality management (TQM) and JIT concepts. These have focused on the possibility that good management, particularly, is both desirable and profitable, and has brought into question all the attitudes which are presented above as alternative views of stock control.

The responsibility for maintaining the correct balance is normally left to inventory management. (The actual name for the activity varies but the task remains the same: the controller could be materials or inventory manager, stock planner or controller, or even considered as a buyer or logistics controller.)

Demand for the product results from changes in market and financial forces and the amount and type of stock control varies in tune with this. Stock control is a dynamic activity, and the successful inventory manager has to ensure that the balance is kept right. This requires both communications skills and professional inventory techniques. Operating methods should be continuously revised to reflect the changes and systems should be altered to suit new situations and operating policies.

Objectives for inventory control

The first consideration is the overall objective of the work of stock control. Like all other activities in the company, inventory management has to contribute to the welfare of the whole organization. The logistic operation must aim to 'contribute to profit by servicing the marketing and financial needs of the company'. The aim is not to make all items available at all times, as this may well be detrimental to the finances of the company. The normal role for stock control is to 'meet the required demand at a minimum cost'.

Our aim towards long-term profitability has then to be translated into operational and financial targets which can be applied to our daily operations. The purpose of the inventory control function in supporting the business activities is to optimize the three targets:

- customer service
- inventory costs
- operating costs.

The most profitable policy is not to optimize one of these at the expense of the others. The inventory controller has to make value judgements. If profit is lacking, the company goes out of business in the short term. If customer service is poor, then the customers disappear and the company goes out of business in the longer term. Balancing the financial and marketing aspects is the answer: the stock controller has a fine judgement to make.

The first target, customer service, can be considered in several ways, depending on the type of demand (discussed in Chapter 2). In a general stores environment the service will normally be taken as 'availability ex stock', whereas in a supply to customer specification, the service expected would be delivery on time against customer requested date.

The second target, inventory costs, requires a minimum of cash tied up in stock. This target has to be considered carefully, since there is often the feeling that having any stock in stores for a few months is bad practice. In reality, minimizing the stock usually means attending to the major costs; very low-value items are not considered a significant problem. Low inventory can also be considered in terms of space, or other critical resource. Where the item is voluminous, or the stores space restricted, the size of the items will also be a major consideration.

The third target, avoiding operating costs, has become more of an issue as focus has been placed on inventory management. The prime operating costs are the stores operations, inventory control, purchasing and the associated services. The development of logistics, linking distribution costs with inventory, has added this new set of transportation costs to the analysis.

Optimizing the balance of these three objectives is the focus of stock control. The better the balance the greater the profits provided for the company (as will be shown in the next section, profit through inventory management). The improvement in stock control has been slow and gradual, created by new technology, financial need and competitive pressure. Those companies who can tighten their control faster than the average will flourish, but those which do not keep up with the average, even if they are improving, will gradually dwindle. The way to achieve much is relatively simple, and is outlined in the remainder of this book. The trick of the good inventory controller is to meet the objectives simultaneously, not one at a time, and of course 'the better the control the smaller the cost, the lower the stock levels, and the better the customer service'.

In reality: more inventory = worse delivery on time.

One of the dichotomies of inventory control is that at item level, the more stock the better the availability. (This is discussed quantitatively in Chapter 6, in the section on evaluation of safety stocks.) However for the whole inventory, experience has shown that the businesses with the highest stock are often those which have the worst availability. These observations are not in conflict if the causes of this are considered. Stockouts result from holding too little stock for the offending lines, because the forecasts, monitoring or controls are inadequate. High stock levels arise because too much stock has been purchased, through bad forecasting, monitoring or controls. High stock and poor availability are caused simultaneously as a result of poor control. The problem rests with the inventory controller, and the solution is in improved techniques. By using the techniques discussed in the remainder of this book the reader can reduce stockholding by 25 per cent even where the control is already reasonable.

Profit through inventory management

There is obviously a link between financing stock and profits. This can be seen most clearly using an example.

An example of the effect of stock on profitability

There are two companies in the same industry, M. Tight Ltd. and The Slack Company. Each has annual sales of £5 million and employs eighty-five people. They both have fixed assets of £2.5 million in plant and buildings, and operate in an industry where the value of customer debtors is balanced by the credit owed to suppliers.

In their organization the two companies differ. Tight is a small, independent company and Slack is a small part of a multinational group. Although they manage to achieve similar customer service, M. Tight Ltd. has concentrated on inventory control and has reduced stocks to £500 000, while the Slack Company have poorly defined responsibilities in this area and hold £2.5 million worth of stock.

The Slack Company have to finance an extra £2 million inventory, and they can borrow this from their head office. However, the company expects a 15 per cent return on investment and so charges Slack £300 000 per annum in interest. They also have the costs of controlling, holding and organizing the extra stock which is stored in an adjacent warehouse. These warehousing costs are shown in Table 1.1.

Looking at company finance, the size of expected profit will depend on the size of the company. Therefore, to compare how well different

Table 1.1 Warehousing costs

Cost element	Cost (£ p.a.)
Stores: extra area rent, heating etc.	500
Stores, equipment; trucks and racking depreciation	1 000
Maintenance	1 000
Stock obsolescence	5 000
Storekeeper's wages and employment costs	14 000
Annual stocktaking	500
Control systems and computer time	1 000
Total	23 000
Plus financing charge £2 million @ 15%	300 000
Total extra cost	323 000

businesses are doing, the profit has to be scaled to company size. The two most common ways of doing this are:

$$\text{Return on sales} = \frac{\text{Profit}}{\text{Annual sales}}$$

$$\text{Return on assets} = \frac{\text{Profit}}{\text{Assets employed}}$$

The assets include both:

- fixed assets (plant, buildings, land, etc.)
- variable assets (stock and debtors).

Now examine the Profit and Loss account for the two companies. The difference between sales revenue and costs is £250 000 for each company.

M. Tight Ltd has a return on turnover of 5 per cent, making £250 000 profit on the £5 million sales.

Slack has extra costs of £323 000 and thus makes a loss of £73 000. Good stock control has made the difference between profit for M. Tight Ltd and loss for The Slack Company.

The next year M. Tight Ltd made £500 000 profit on the same £5 million turnover, a 'return on sales' of 10 per cent. The Slack Company had a profit of £500 000 − £323 000 = £177 000 because of their annual stockholding costs.

Although M. Tight Ltd has a return on sales of 10 per cent, The Slack Company only get

$$\frac{177\,000}{5\,000\,000} = 3.5 \text{ per cent}$$

This shows how important it is to the profit of a company, that stockholding costs are minimized.

That is a surprisingly significant difference and, when we look at return on capital, the situation is even more dramatic.

Both companies have plant and buildings worth £2.5 million, with total assets consisting of the fixed element of plant and machinery, working capital tied up in inventory and creditors who are offset by outstanding debtors.

M. Tight Ltd also has £500 000 worth of stock, making a total capital investment of £3 million.

The return on assets for M. Tight Ltd is therefore

$$\frac{500\,000}{3\,000\,000} = 16.67 \text{ per cent}$$

The Slack Company has £2.5 million of stock making a total investment of £5 million. Their return on assets is therefore

$$\frac{177\,000}{5\,000\,000} = 3.5 \text{ per cent.}$$

Hold a minimum of the right items to be profitable.

By running the same business with two different inventory plans, we can earn anything from 17 per cent to 3 per cent for our trouble.

M. Tight Ltd has cash available for investment, wages, better working conditions, and to share out to the pension schemes, insurance companies and other investors in their company.

The Slack Company has little cash to spare other than to pay off the bank. Imagine the situation: poor profitability leading to disillusionment, poor working conditions and fears for the future.

Reasons for the current stock

It is easy to assess the effectiveness of inventory control in a company. All that is required is a visit to the stores. The stock level is a result of the effectiveness of the stock control, but what is there is probably not ideal! Instead of considering the detailed ordering history for each item (and there are some interesting causes for stockholdings), the inventory can be classified according to why it is now in stock. An examination of the factors which give rise to the stock in a typical store reveals causes that are detailed in the following sections.

Your current stock shows what is going wrong.

Purchase order quantity

There are several aspects which affect the ideal batch size. For example, most stores have items that are bought in bulk but sold in smaller quantities.

At any time there is a balance of receipts in stock – a half-empty box or pallet. This is usually an acceptable situation for keeping stocks at the right level, but that stock nevertheless still contributes to inventory value.

The size of logistic inventory is often determined by the transport methods chosen and the batch size of delivery which results. Changing the delivery method can change the stock levels in the stores.

Where items are being manufactured, the rate of production in one stage of the process may be greater than at other stages. The ideal solution to this would be Kanban (see the section on pull systems in Chapter 4), but often there is a stock of items manufactured slowly awaiting to fill a process load. Alternatively a bulk process, or one where items are produced at a fast rate, will create stocks which are subsequently used up at a much slower rate.

> *Frequent, small deliveries keep stock low.*

Safety stock

In many situations customers do not provide information of their demands far enough in advance. To compensate for this problem there are two tactics. The first is to organize the customer to give more forewarning. Is this impossible? Manufacturers of specialist, make to order or highly sought after products may make their customers wait until the product is ready. In more competitive situations, however, the customer has more power and may not be prepared to wait. This puts the pressure on suppliers to reduce the lead times and better to forecast demand. With better knowledge of the customer's demand, the unexpected peaks would no longer be a surprise and stocks could be varied to cover these occasions. The second option is to hold sufficient stock to cope with unexpected or excess demand. Safety stock (or buffer stock as it has been called) is there to cover our inability to predict demand. The inventory manager has an investment choice. Either invest in inventory or invest in information. Those who can accurately predict the customer requirements do not need safety stock.

> *Let the source of the risk hold the inventory.*

A secondary role for safety stock is to compensate for failure on the supply side, non-delivery of purchased goods, information failures, technical breakdown or, even, industrial action. You cannot be too careful! Generally the approach to these situations is to monitor the situation continually, and there are several standard ways of doing this (see lead times and schedules in Chapter 7).

Market change

Customers change their mind, contracts are lost, markets vary. These all cause excess or even obsolete stocks. But this should not be of a significant value. The changes that cause a build-up of inventory are either step changes or gradual changes. A gradual change is the most common, and leads to lots of minor excesses of stock. The cause is normally a forecasting method that is not reactive enough, and the simple cure is often to use a more advanced forecasting method. Methods discussed in Chapter 11 and used in the right way can avoid this situation.

A sudden reduction in demand can only be anticipated through understanding or information. Input from sales personnel is important in avoiding excesses as well as stockouts. Sales can warn of impending major changes, even if they are not good at identifying the size or the timing of changes. This is the best way to avoid having items remaining after demand is satisfied.

> *Detect trends and react automatically.*

Obsolescence

In addition to excess stock caused by customers, a significant amount of obsolete stock can be caused within the company itself. Consider several situations:

1 The marketing department put an item on promotion without informing the inventory controller. The demand increases as the promotion starts; inventory control see demand escalate and stocks run out. They therefore order more. The promotion ends just as the new stock arrives. The company now has three years' worth of stock cover at the normal usage rate, and the major customers have already bought extra during the promotion.

2 The design department has replaced the old version of a product with an improved one; the sales team were extremely pleased with the new product and started selling it immediately. However, there was a large amount of the old product left in stock which was not sold.

3 The marketing department has just launched the new season's catalogue. It contains some new items, and some old ones have been deleted. There are still good stocks of some of the deleted items.

4 The sales director has bought a job lot of items which are going to be sold as a new line.

All these are common occurrences in some businesses. They all cause self-inflicted obsolete stocks and can be avoided by improving communications between departments. The responsibility for organizing this communication revolution lies with stock control, since the other departments rarely understand the challenges of inventory management or the concept of lead time. (Refer to the excess and obsolete items section in Chapter 6.)

Poorly defined responsibility

When buyers procure to optimize purchasing costs and operations, and sales sell to suit customers, in the middle are the stores, accumulating stock and controlling the conflict between the two activities. Stock control in this case becomes simply recording and storing. Inventory management should have the responsibility for assessing the demand pattern, determining the appropriate stock levels to support this and then instructing purchasing what and when to reorder.

Planned inventories

Cyclic demand, on a monthly, quarterly or annual basis, may be supported most economically by constantly holding high stock levels. In the stores there may be stock which is building up to satisfy a demand event or which is stored to cover a definite requirement. The feature of these is that they cover situations where the demand events are certain and in excess of the supply capabilities, as, for example, in industries where there is an annual holiday shutdown. Stock has to be provided to cover customer demand over this period, and it is not possible to acquire the extra quantity at the last minute to cover the shutdown. The stock is therefore gradually increased over several weeks to maintain continuity. This effect can be complicated by different stages in a supply chain taking holidays at different times.

Layout and location of stores

The effect of a stock location must be analysed in the context of the whole business. In a company where the store is situated inconveniently or a long way from the users, the users tend to keep their own stock. In fact, they might keep some just in case the stores have run out when they want some. This lack of co-operation and control can result in satellite stores being set up and excess stock being held.

A worse situation results if different departments independently order the same items. In this case, there are totally independent stock systems, usually controlled by amateur stock controllers whose major interest is another aspect of the business and, as a consequence, stocks are excessive.

Fortunately where this happens the items are normally low-value consumables.

Where a company has more than one store (such as in distributed warehousing), a similar range of stock items are held in each, resulting in duplication of stock. If there are slow-moving items in one store (try to find a store without slow-moving items!) then the same item can be found in the other store – still moving slowly, and total stock across all the stores can exceed the annual demand. This is a natural result of distributed stock. The safety stock required will increase as the square root of the number of distributed stores increases; so managers had better think carefully about setting up local stock points. The increase of stock (by the square root) is based on the assumption that the local stores are part of an integrated, well-controlled inventory system. If this is not the case in practice, then the total stock can go through the roof (literally).

Company strategy

Some inventory controllers do not have complete authority to maximize availability and minimize costs. This may be because top management has imposed some constraints such as an extra stockholding imposed through lack of confidence in the ability of the stock system to provide adequate customer service. It may result from a commitment to customers to keep an agreed quantity locally or consignment stocks at customers' premises.

In the normal course of events the stock control system will provide the desired availability by adjusting the control parameters and should not need overriding by management edict. It is up to inventory management staff to instil confidence in their ability to control stock excesses.

System and control

The time taken to process information can substantially affect the level of stockholding. Consider a stores where an issue is being made. This causes the stock to fall below the ordering trigger level. How long does it take before the supplier gets to know that a new delivery is required? On a conventional order processing system the procedure is:

- Enter issue on system. (Delay one hour?)
- Print out list of items which need ordering. (Next morning?)
- Confirm that items do need ordering. (Rest of the day?)
- Print purchase orders. (Next day?)
- Get authorization for purchase. (When director is next available!)
- Send order to supplier. (Two days by second class mail?)

These processes could easily take a week, or much longer if the individual lines are batched up into a large order. If the supplier is a local stockist, then the item could arrive the same day, but the total supply lead time would be one week (internal) plus one day (external), and the stock level would have to be correspondingly large to compensate for this lead time.

For efficiency use real time procedures.

Slow systems cause extra stock

The solution is to use the computer systems properly, without paperwork, and to use communications technology, (fax or e-mail) for ordering.

If the stock is analysed according to the headings above, the causes of stock excess become obvious. Analysis will show that actual stock is not held in the right balance to meet marketing and financial targets in the most effective way. By establishing a reason for this, a plan can be devised and action taken to ensure the lowest stockholding levels necessary to meet company targets.

Initially, the reason for stockholding has to be determined. It could be to achieve good customer service, cheap purchase, bulk discount, security or a large number of other reasons. The stockholding strategy depends directly on the company objectives and policy. Company policy defines the service which it intends to provide to customers, what investment to make in stock, how orders are placed on manufacturers or other principles on which to base supply plans. This is the first ground rule of stock control – to have a policy worked out by top management to guide the operation. The more professional the stock control operation, the closer the policy needs to be defined so that targets are set against which the activity can be run. This is a generic, broad-brush approach, but useful for finding out where to put most effort in stock management.

Key points

- Consider the objectives and need for inventory.
- Target availability, inventory value and operating costs simultaneously.
- Create profits by avoiding non-essential inventory.
- Determine the causes of stock and use techniques to reduce it.

2

Customer service

- The key to successful business.
- Focus on customer delight.

Objectives

- Identify the purpose of customer service.
- Define the alternative measurements of availability.
- Analyse the components of administering customer service.
- Discuss the role of forecasting.

Meeting customer requirements

The focus on customer service has gradually changed over the years and suppliers are now becoming really interested in customer service instead of just talking about it. Customer service is a complex subject of its own, but there are two main aspects, namely customer relations and availability of service or items.

Please the customer and measure performance.

Customer relations

Customer relations is about keeping the customer happy. It requires interpersonal skills to ensure that the customers have the correct level of expectation of supply and that they are happy with their purchase so that there is potential for repeat business and wider sales.

This aspect of customer service has gradually emerged because it is the differentiating factor between many companies. The product can be similar from a wide variety of vendors, the vendor which is most successful is the one where the quality of customer relations is best.

Fix the customer. Don't just supply the item.

Customers may have a perception of suppliers which varies greatly from their actual performance. It does not matter whether a supplier is good or bad, what does matter is what the customer thinks about them.

Customers can look for a match between their own style of company and the style of the supplier. If there is a match, then relations are likely to be good; if not, then the co-ordination requires more work. The major factors affecting customer relations are outside the scope of inventory control, in sales and sales order processing. However, there has to be the same customer-oriented approach in stock control, which, like the company, needs to maintain credibility.

The second aspect of customer service – availability – is a key target for stock control. It requires technical management.

Measuring availability

Availability

The entire reason for stockholding is to have items available but, despite the key nature of customer service to the success of the business, some companies do not quantify proper stockholding. They simply take notice of complaints or other oblique assessments of service. To manage professionally there is a need for solid facts.

For each item in stock the risk of stockout can be reduced by increasing the stockholding. The larger the investment in inventory for an item the better the service. This is shown in Figure 2.1 for an individual stock item. The curve shows that it is relatively inexpensive to give a reasonable level of availability, but the value of stock increases very rapidly when we try to achieve very good service. A 100 per cent service is not possible because it needs an infinite amount of stock! However, 99.99 per cent availability is possible, but very expensive and no one can tell the difference from 100 per cent.

> *Availability: get the facts, then act.*

There is obviously a trade off between the investment and the availability of items from the shape of the curve in Figure 2.1. For each item, the stock policy can be adjusted to give the appropriate service.

The overall service to the customers, or other stores users, can be measured from success in meeting demands (on-time availability) on a daily, weekly or monthly basis. It can be calculated by adding the service

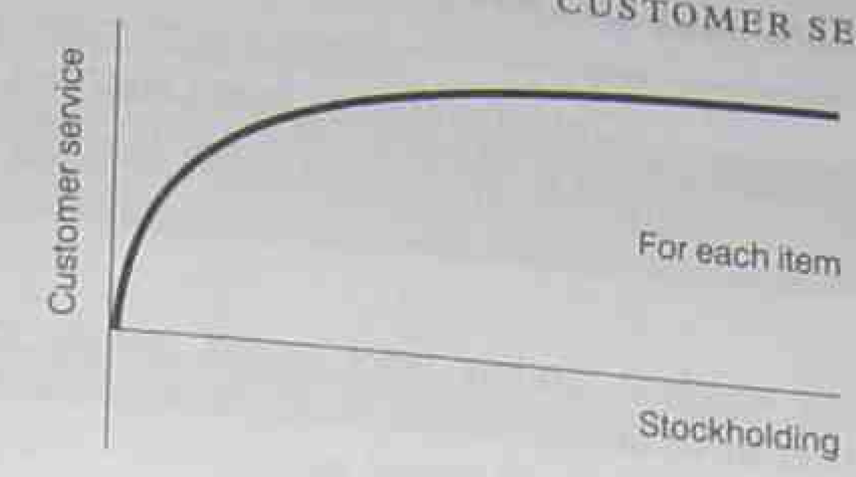


Figure 2.1 Availability of a stock item

provided on all the individual items, taking into account the issue frequencies for the total range of items. We can define availability as:

$$\text{Availability} = \frac{\text{Demand satisfied}}{\text{Total demand}}$$

This is the criterion which should be applied to all areas of inventory to monitor how well the investment in inventory is doing.

When the whole range of stock items is considered, the minimum stock levels result from all the items being under control at their optimum service level. Basically, each item should be on the same curve in Figure 2.1. The only way to reduce the stock beyond this balance, while maintaining service, is to improve the total control method. After all, the aim is toward the top left corner of the graph – high service with low stock across the whole of the inventory.

> *High stock goes with poor availability. They are caused by poor control.*

Measurement of availability in practice

Within this definition of availability there are a number of ways by which the service level can be measured, depending upon the type of demand.

For companies supplying single items to individual users as a result of individual telephone or mail orders, then the definition could be:

$$\text{Availability} = \frac{\text{Total number of items supplied}}{\text{Total number of items ordered}}$$

For a department which is sending large export orders, these are often covered by a 'letter of credit' or the whole shipment must be sent in one crate. In this case the service level would be:

$$\text{Availability} = \frac{\text{Total number of complete orders supplied}}{\text{Total number of orders}}$$

For businesses where the demand is not required to be satisfied immediately, care must be taken to measure availability at the time required for dispatch, not the time of receiving the order. This gives the inventory controller the opportunity to fill the demand from elsewhere rather than from stock.

Care must also be taken as to the definition of 'on-time' delivery. Delivery should really be measured as the time the customer receives the stock rather than when it leaves the stores. 'On time' is when the customer needs the stock. Where there is a history of poor delivery the customers are quite likely to request a delivery date well in advance of their real need, in order to have a chance of having the items when they need them.

Measure availability in line with typical customer needs.

The type of customer also affects the measurement of availability. For example, if a customer requires ten and the stock is eight what customer service can be provided? Is the answer:

- 0 per cent because the demand was not completed (non-stock/manufacturing answer)
- 80 per cent because eight out of ten can be supplied (normal answer)
- 100 per cent because the balance can be provided before the customer needs them (for stockholding customers).

Different situations have different priorities. To have an appropriate measure of availability, the service must be defined in one of the above ways

Table 2.1 Pitfalls of measuring availability

Part no.	Demand	Shortage	Service level (%)	No. of orders	Shortage	Service level (%)
A12	1000	10	99.0	5	1	80.0
B25	10	5	50.0	1	1	0
Average			74.5			40.0
Sum	1010	15	98.5	6	2	66.7

– the one which suits most customers. In Table 2.1 only two items are being considered. The demand for item A12 was ten orders totalling 1000 items of which 990 were sent on time and ten items were not available for one order. This means that 80 per cent of the orders (and 99 per cent of the items) were sent on time. The second item is a slow mover and there was only one order (for ten items) and only five were dispatched on time giving a 50 per cent service of parts (and complete failure in terms of complete orders).

How good is the service? Out of 1010 items requested only fifteen were short, so the number of items dispatched was 98.5 per cent of the total. However, in terms of item order fill, deliveries of item A12 were 99 per cent full but B25 was only 50 per cent full, giving an average per line of 74.5 per cent. If the customer is looking for complete orders, then out of the total of six orders two were short, corresponding to a 66.7 per cent success rate. As an order fill per line, this was 80 per cent for A12 and 0 for B25, so the average was 40 per cent. This illustrates that different assessment of levels of service which can often occur between customers and suppliers. In this example the supplier is claiming 98.5 per cent delivery on time, while the customer is measuring it at 40 per cent! It is important to agree a common goal between suppliers and customers, and to establish what the customers need in terms of availability. The appropriate measure can then be used for monitoring.

Availability policy

Customer service can be structured or focused on a particular group of customers or a market sector using alternative availability policies. Options which could be used are:

- the same availability across all the products
- minimizing the total cost of service

- concentrating on the most valuable customers
- enhancing the service on the most sensitive products
- greatest availability on the most profitable products, or
- better service on the major turnover items, reduced service on slow movers.

These alternatives will provide different availabilities for different products, with a management decision setting availability levels.

Whichever alternative is chosen, proper control of stock requires the inventory manager to regularly monitor availability.

> A better measure takes account of 'lateness'.

Back-order measurement

The measurement of availability discussed in the measurement of availability in practice section should always be used, but a further measurement is necessary for sophisticated stock control.

Once an item has missed its delivery date, it is out of the calculation. This introduces a further factor into the equation: the priority given to satisfying back orders. Consider a case where there are two demands for an item, one which is past the due date and one which is to be delivered on time. In such circumstances, the availability target calculation requires us to fulfil the current order, not the old one, since the current one improves the availability factor, but the other has already failed. In practice there is normally a need to give precedence to customers whose delivery is late. Therefore a more rigorous measure of availability would be to measure how many orders are late and how late they are.

An example of how this can be approached is developed from Table 2.2 (and Figure 2.2) which shows the dispatch position for two companies, A

Table 2.2 Comparison of dispatches for companies A and B (items dispatched in week 37: analysis of lateness on orders dispatched)

	Company A	Company B
Delivered on time	10	30
1 week late	60	15
1 month late	20	40
3 months late	10	15
Total dispatches	100	100

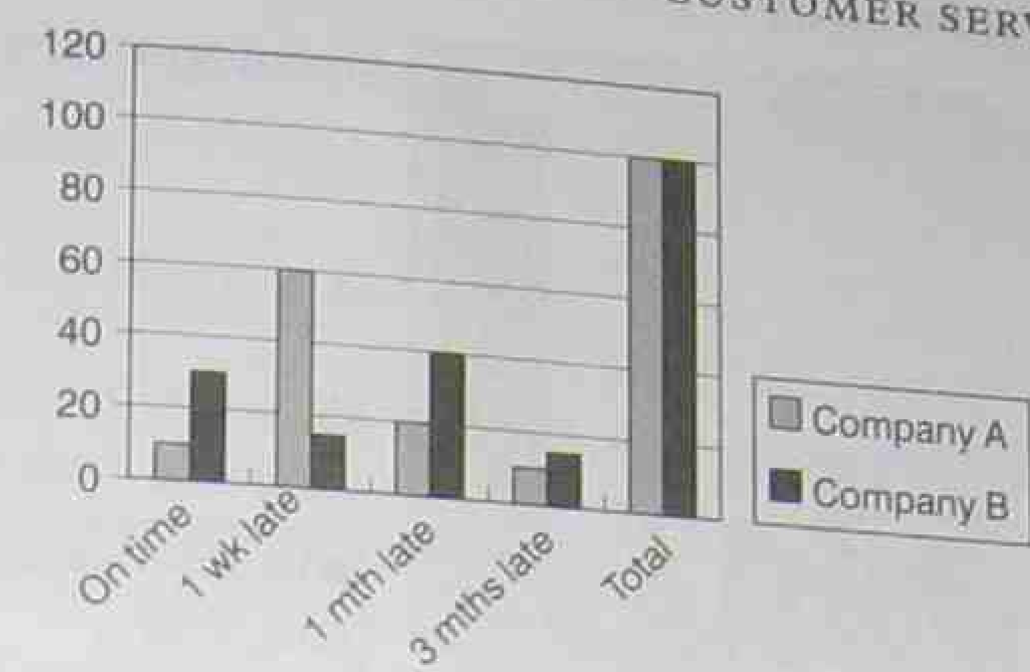


Figure 2.2 Comparison of dispatches for companies A and B

and B. Which company is providing the better service? From our service-level (level of fill – LOF) approach company B is the better because it gives 30 per cent LOF against 10 per cent LOF for A. However, there is a better chance of getting the items quickly from company A. Perhaps A is better after all.

A measure of customer service which takes into account the overall performance, not just the 'ex stock' part, can be developed by using a weighted average. In Table 2.3 (and Figure 2.3) the arrears (or 'back orders') have been analysed by categories of lateness. By multiplying the number of orders outstanding by the weeks of lateness, a back-order week figure has been calculated. This gives an overall figure of merit of 45 arrear weeks, corresponding to an average for arrears of 2.1 weeks per late item. By monitoring weeks of lateness against a target level, inventory controllers can assess the level of service achieved. Obviously, the best practice is to ensure that the bulk of demands are satisfied on time because those few items which are long overdue give a negative contribution to the figure of

Table 2.3 Distribution of arrears

Time overdue (T weeks)	No. of items (N items)	Weighted average (T × N)
1	8	8
2	6	12
3	4	12
4	2	8
5	1	5
Total	21	45

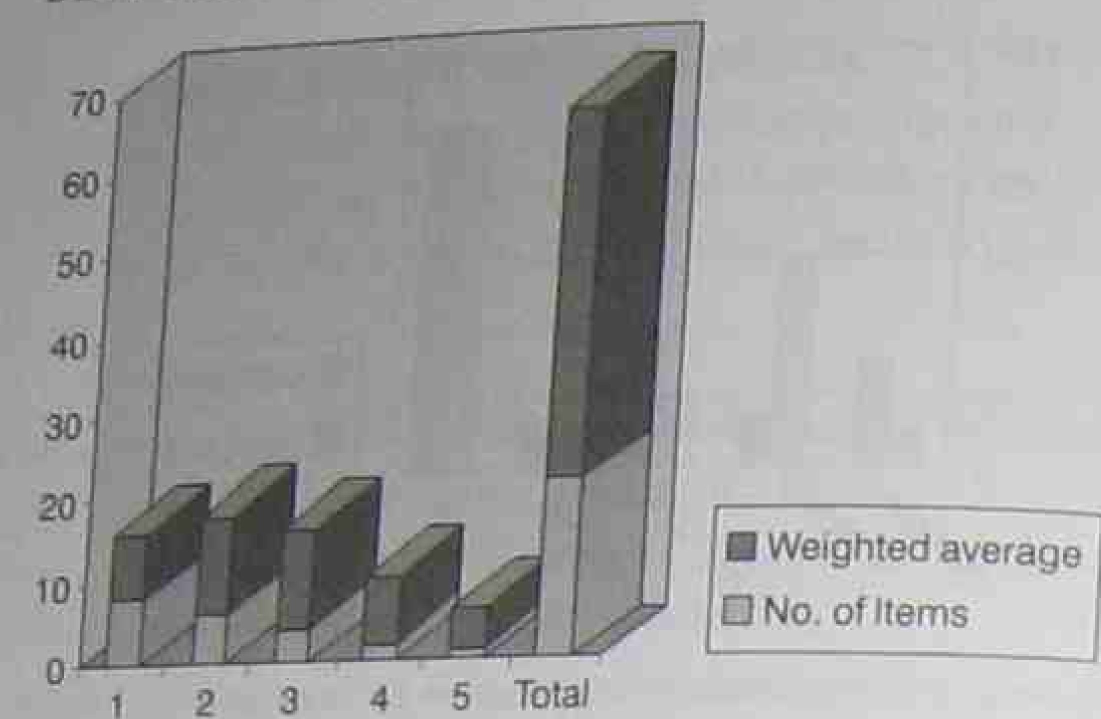


Figure 2.3 Distribution of arrears

merit. If the back orders are satisfied rapidly, the level of service can be measured in units of days or hours rather than weeks.

In practice the order backlog viewed on the system may not be all true demand. It often consists of:

- real arrears which are customer orders slightly overdue and a real priority
- virtual arrears which are often long-standing customer orders without any urgency to them; these virtual arrears are orders which the customer does not really want, or where they have bought an alternative and not cancelled the order for the original item.

Outstanding orders need regular review to ensure that virtual orders are either rescheduled or removed. If virtual arrears build up they have a distorting effect on the back-order measurement.

> Verify dates required before calculating arrears.

The extent of arrears can be measured in several ways. The number of items in arrears is the simplest measure, and the aim is to focus the control on the important items. Measurement in terms of value is not usually recommended since the aim is to provide customer service. Routine assessment of arrears should be carried out in a similar manner to the monitoring of availability.

Demand management

Order processing

> Entering the demand and making sure the customer is happy.

Stock should, of course, fill customer demand. In practice this includes both having stock available and dispatching it. The physical handling of the stock is a stores operation's responsibility, while in many companies the actual order processing forms part of the stock control function. Demand management consists of a series of steps:

1 Order receipt.
2 Order processing.
3 Estimating delivery times to customers.
4 Production of dispatch information.
5 Customer feedback.

Order receipt

This is the initial compilation of orders into a form which can be used to dispatch the goods at the right time to the customer. This information must be:

- complete
- specific
- accurate.

The information can arrive by post, telephone, electronically or by word of mouth. A first requirement is that the company has a good item identification system which enables sufficiently unique identification to satisfy a discerning customer. This has to be formalized into data for processing.

Stock control requires an identification code for each item, ideally provided by the customer. To assist this, some companies provide customers with catalogues and order forms to enable customers to specify precisely. It used to be customary for counter-sales customers of all types not to know code numbers, and to rely on stores personnel to remember them. This situation is no longer acceptable – if customers want the correct item, then they should expect to specify their requirement properly. An order form not only enables the correct data to be collected, it also presents it in a standard format, which can be arranged to match computer input or stores picking procedures.

Order processing

This is the entry of the order into the system. A second requirement is that the processing of orders translates, reliably, speedily and without creating errors, the customer order into company information. This process has to be fast, simple and, above all, accurate. For computer systems this should be done as a one-stage process with the source (phone, fax, letter, etc.) being input directly into the system. This avoids errors and delays. Where printed customer orders are required, these should be printed out from the data input to the system.

> *Ensure that the data input is accurate.*

For non-computer applications the source document (or a copy of it) should be used for stock picking. With technology, electronic data interchange (EDI) can create a demand in the system without any work. This is good as long as the demand is easy to fulfil. Stock controllers need the opportunity to vet orders, which means that all EDI orders should be formally accepted for dispatch. Application of the 'manage by exception' principle suggests that most orders are easy to satisfy and can be accepted without question – automatically. It is only those orders for fast delivery, for large quantities or for non-standard items which need control. Criteria can be set up to identify these so that the majority, the other standard orders, can be accepted without question. The stock control parameters for establishing these criteria are discussed later.

Estimating delivery times to customers

Estimating delivery times is the activity which causes a high level of customer confidence to be gained or lost. Promises have to reflect the real stock and replenishment situation or else customer service will be poor. (See the following section 'Estimating delivery times.)

Production of despatch information

The despatch information will consist of two stages:

- picking and delivery note production
- customer invoicing.

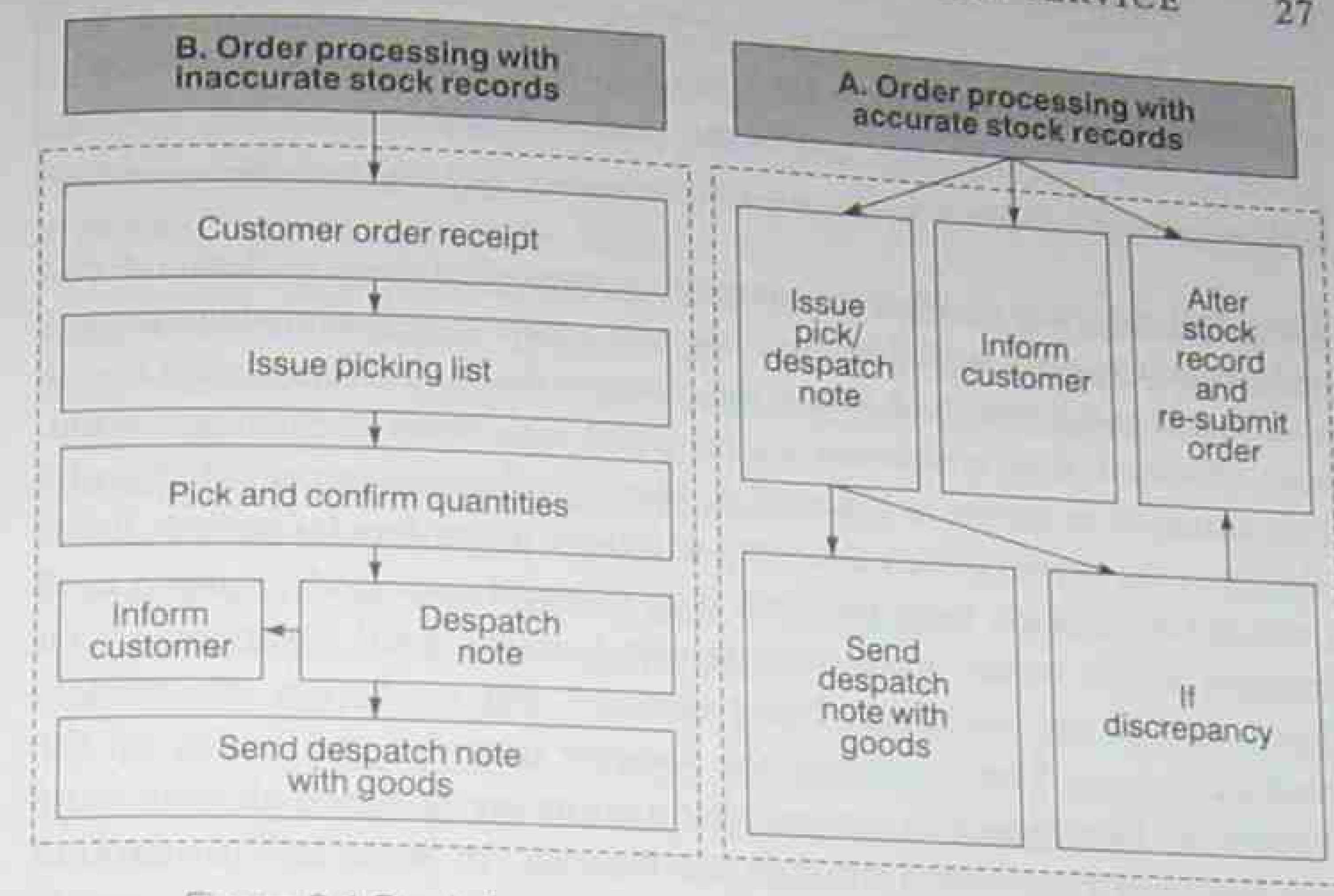


Figure 2.4 Stock item order processing

These stages should be carried out sequentially. The financial (invoicing) systems should not be allowed to interfere with the stock movement (dispatch) system, as invoicing is not part of the stock control system. Two alternative paperwork processes can be used. A delivery note can be prepared as soon as the order is processed, as long as the stock records are accurate as shown in Figure 2.4(B). Where records are somewhat suspect, a safer procedure is the one shown in Figure 2.4(A). This usually causes extra paperwork.

Customer feedback

Customer feedback to report on progress and to keep the customer feeling involved in the supply is very important from a customer relations standpoint (see previous section, 'Meeting customer requirements'). Customers often require acknowledgement of order and confirmation of price as well as a dispatch note and invoice. This confirmation should form part of the normal information processing procedures.

Consuming forecast demand

In almost all situations business works to a forecast until the real orders arrive. However, real orders rarely match the forecast and the question is raised as to whether an actual order received is really part of the forecast.

Consuming the forecast = take out a forecast demand and replace with the customer order.

The ideal situation is when a company accurately forecasts demand and makes that quantity available. As customer orders arrive, available quantity is allocated and dispatched. When the quantity for the current period has all been allocated, then customers are provided with items in the next period. This situation is shown in Figure 2.5. Here the forecast in periods 1 and 2 (shown as the curve) is fully allocated and there are orders for periods 3 and 4 which have already been received. The demand in periods 5 and 6 is all forecasts as no orders have been received which need satisfying so far ahead. Orders are constantly being received and product is allocated to these customers. This consumes the forecast quantity and builds up the allocations. This approach ensures that a company does not allocate more than has been provisioned through the forecast. In some fast-turnaround businesses with period length of a day or less, this situation is acceptable and almost inevitable. In other businesses more flexibility is required and there has to be:

- safety stock to meet high demand, or
- extra supply sources available to meet peaks, or
- resources reserved for emergency supply (e.g. allocation of only 90 per cent of capacity in manufacture leaving 10 per cent for last minute orders).

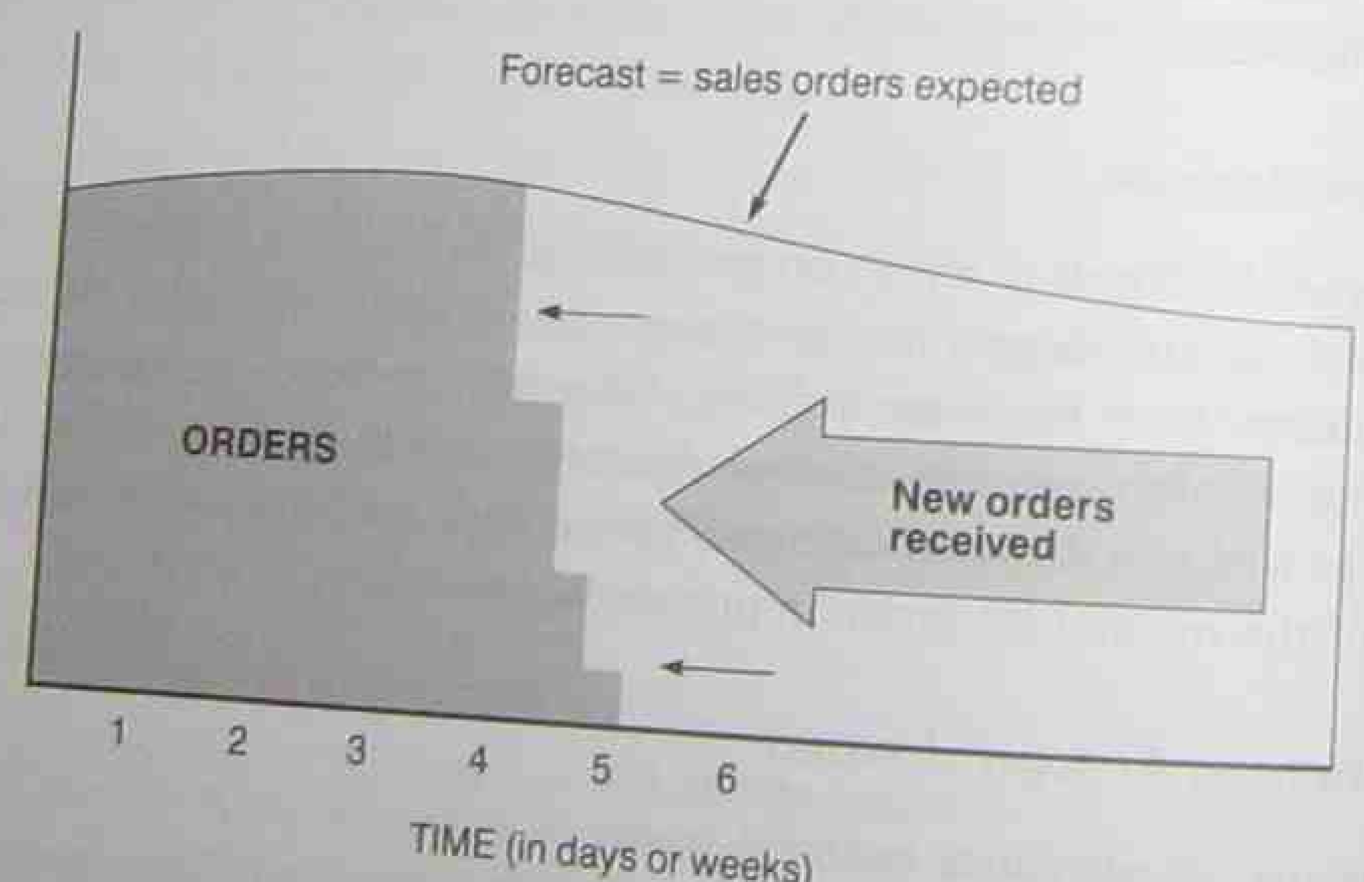


Figure 2.5 Consuming the forecast

Where a company agrees to fill every order from customers at a time to suit the customer, then it is important to forecast carefully, otherwise delivery achievement will be poor, much management time will be spent chasing problems, panic purchases will be made and extra deliveries will be required, causing costs to be high.

The differences between forecast and actual demand are measures of the quality of the forecast (see Chapter 6). Actions arising from order processing, besides accepting the customer's request, could be:

- arrange increase or decrease in stock levels
- bring forward or delay supply
- offer earlier or later delivery date to customer
- reject order or offer alternative item
- provide part shipment.

Limit order book to achievable capacity.

An example

A supplier of fasteners has trouble with widely fluctuating demand levels on some products. This is because the customer thinks 'I need a box of fixings for this job' and orders enough to do perhaps two months' work. Consequently the supplier gets hit by two-months' demand and then nothing. This is acceptable if there are many customers, but for the slower moving fasteners demand can be very erratic. By understanding the supply problem, the customer can be encouraged to order enough to start the job off with sufficient supplies, while the supplier, who understands the user's situation, can provide the range of fasteners but can get away with much less stock. Of course, the situation must be carefully monitored and the balance must arrive before the initial stock has been used up.

Large quantity orders can be filtered out using simple criteria (see Chapter 6) so that real demand priorities can be fulfilled. Demand can thus be managed through the order processing activity.

Estimating delivery times

For supply to order items, and for stockouts, an estimate of the delivery date is usually required. Sales are often not aware of the status of the supply and the order book, and so are badly placed to give more than a blanket estimate of delivery. The stock control or purchasing departments do have the information and can provide dates which are based on the actual current situation.

It is the responsibility of inventory management to provide delivery and availability information in support of customers' orders. The time taken to provide an item to a customer depends on the balance between stock currently held and on order and the quantity already allocated to customer requirements (see Chapter 12). If the order being accepted is allowed to use items already allocated to other customers, then the order can be fulfilled sooner. In general the priority sequence is 'first come first served', but there are often cases where certain customers gain priority. Sales may consider that promising the items regardless of ability to deliver will secure orders and keep inventory control people on their toes. This can lead to a mistrust of the optimistic delivery promises made and poorer customer service. The solution to this situation is for inventory management to provide customer delivery dates using priority rules specified in advance by the marketing department. Where there are exceptional circumstances, then priorities can be altered within the total availability profile. The request to 'ring the supplier and get a faster delivery this once' should be dismissed on all but the odd occasion, since it usually results from sales people not being able to negotiate successfully.

Monitoring delivery dates can become complex. Dates which can be considered as appropriate are:

1 When the customer placed the order.
2 When the customer requested delivery.
3 When the customer needed the item (usage date).
4 When delivery was originally promised.
5 When delivery is now promised.
6 When the item is dispatched.
7 When the item is received by the customer.

It is also useful to know:

8 Who changed the delivery date.
9 How many times has it been changed.
10 When the delivery date was changed.

> Decide carefully what is 'on-time' delivery.

The importance of these dates varies from one market to another and one company to another. There are so many dates and potential monitors, that companies have to choose which ones are important in their own circumstances. Providing too many management controls is confusing and dilutes the important ones. The prime controls would be delivery on time measured as '7–2' and '7–4'.

The difference between the customers' request for delivery and expected delivery dates (2–1) is their perception of lead time. This may be a general indication of market expectations. The difference between the customer order date and the date they received the goods (7–1) is the actual customer lead time. Unfortunately, most companies measure their dispatch dates (6) rather than the customer receipt dates, so there could be a difference between the two views of delivery performance. Many carriers now have sophisticated tracking systems that have the potential to provide suppliers with customer receipt information to use in their control systems.

Customers may leave an interval between their requested supply date (2) and the time they actually need the items (3), to compensate for delivery problems. If information is available to monitor this, it can be used to improve collaboration between supplier and customer. The monitoring of delivery performance, as with all other inventory monitoring systems, exists to provide the best factual information for management and control. Monitoring systems should therefore differentiate between when the customer wanted the goods (2), when the delivery was originally promised (4) and when it is now promised (5). It is useful to measure credibility (7–4), to ensure that the supply systems are working correctly. Measuring performance against current delivery promise (6–5) can fool a company into thinking falsely that it is performing well. The current delivery promise may bear no relation to when the customer requires the item, and may be the latest in a succession of promises as the delivery date slips back. The estimated delivery date can often be changed so that 100 per cent performance is achieved. Therefore, monitoring success against current delivery promise is not likely to be a reliable measure of effectiveness. Monitoring of slippage – the number of times the delivery promise has changed (8) – can also be revealing in showing whether the company is providing customers with good information and service. However, the reason for late delivery against the original plan could be simply that the customer does not require the goods until later. It is therefore useful to examine (9) what proportion of the changes are due to poor supply and how many are due to customers changing their mind.

and the order book, and so are badly placed to give more than a blanket estimate of delivery. The stock control or purchasing departments do have the information and can provide dates which are based on the actual current situation.

It is the responsibility of inventory management to provide delivery and availability information in support of customers' orders. The time taken to provide an item to a customer depends on the balance between stock currently held and on order and the quantity already allocated to customer requirements (see Chapter 12). If the order being accepted is allowed to use items already allocated to other customers, then the order can be fulfilled sooner. In general the priority sequence is 'first come first served', but there are often cases where certain customers gain priority. Sales may consider that promising the items regardless of ability to deliver will secure orders and keep inventory control people on their toes. This can lead to a mistrust of the optimistic delivery promises made and poorer customer service. The solution to this situation is for inventory management to provide customer delivery dates using priority rules specified in advance by the marketing department. Where there are exceptional circumstances, then priorities can be altered within the total availability profile. The request to 'ring the supplier and get a faster delivery this once' should be dismissed on all but the odd occasion, since it usually results from sales people not being able to negotiate successfully.

Monitoring delivery dates can become complex. Dates which can be considered as appropriate are:

1 When the customer placed the order.
2 When the customer requested delivery.
3 When the customer needed the item (usage date).
4 When delivery was originally promised.
5 When delivery is now promised.
6 When the item is dispatched.
7 When the item is received by the customer.

It is also useful to know:

8 Who changed the delivery date.
9 How many times has it been changed.
10 When the delivery date was changed.

> Decide carefully what is 'on-time' delivery.

The importance of these dates varies from one market to another and one company to another. There are so many dates and potential monitors, that companies have to choose which ones are important in their own circumstances. Providing too many management controls is confusing and dilutes the important ones. The prime controls would be delivery on time measured as '7–2' and '7–4'.

The difference between the customers' request for delivery and expected delivery dates (2–1) is their perception of lead time. This may be a general indication of market expectations. The difference between the customer order date and the date they received the goods (7–1) is the actual customer lead time. Unfortunately, most companies measure their dispatch dates (6) rather than the customer receipt dates, so there could be a difference between the two views of delivery performance. Many carriers now have sophisticated tracking systems that have the potential to provide suppliers with customer receipt information to use in their control systems.

Customers may leave an interval between their requested supply date (2) and the time they actually need the items (3), to compensate for delivery problems. If information is available to monitor this, it can be used to improve collaboration between supplier and customer. The monitoring of delivery performance, as with all other inventory monitoring systems, exists to provide the best factual information for management and control. Monitoring systems should therefore differentiate between when the customer wanted the goods (2), when the delivery was originally promised (4) and when it is now promised (5). It is useful to measure credibility (7–4), to ensure that the supply systems are working correctly. Measuring performance against current delivery promise (6–5) can fool a company into thinking falsely that it is performing well. The current delivery promise may bear no relation to when the customer requires the item, and may be the latest in a succession of promises as the delivery date slips back. The estimated delivery date can often be changed so that 100 per cent performance is achieved. Therefore, monitoring success against current delivery promise is not likely to be a reliable measure of effectiveness. Monitoring of slippage – the number of times the delivery promise has changed (8) – can also be revealing in showing whether the company is providing customers with good information and service. However, the reason for late delivery against the original plan could be simply that the customer does not require the goods until later. It is therefore useful to examine (9) what proportion of the changes are due to poor supply and how many are due to customers changing their mind.

Key points

- Find out how to delight the customer – go for added value.
- Measuring availability is essential for success (more important than financial reasons).
- Choose an availability measurement that suits the typical customer.
- Ensure that demand-processing procedures are effective.
- Keep the customer informed and involved.

3

Managing the inventory

- The use of Pareto analysis to save time and give control
- ABC for managing inventory
- Stock cover provides a monitoring tool
- Minimizing effort and inventory value
- A guide to reducing inventory levels

Using Pareto analysis for control

Applying effective control

In the stores there are a wide variety of items, with a stock record for each. Some have high value and others are very cheap. The high-value items are normally controlled tightly, whereas the low-value items are not treated as carefully and are issued in bulk in approximate quantities. Most effort should be put into managing the items which are most important for achieving the inventory targets. In inventory control the best results are gained by organizing effort correctly. There is not sufficient management time to maintain detailed control of all the individual items. If the immediate aim is to reduce stockholding costs then studying the stock of low-value items is unlikely to be the best place to start unless the sales volume is very large. If service is the aim, then attention to a few fast-moving lines often provides the bulk of the improvement required.

This simple principle is embodied in Pareto's Law, which is illustrated by the curve in Figure 3.1. It is also called the 80/20 rule because 80 per cent of the effect is provided by 20 per cent of the cause. In Figure 3.1 80 per cent of stock value is caused by 20 per cent of stock lines. The principle can be applied to many different areas of activity: 80 per cent of the purchased items come from 20 per cent of the suppliers, and 20 per cent of sales lines give 80 per cent of turnover. For a warehouse, 80 per cent of the space is occupied by 20 per cent of the lines. This simple fact makes it obvious which lines should be received little and often.

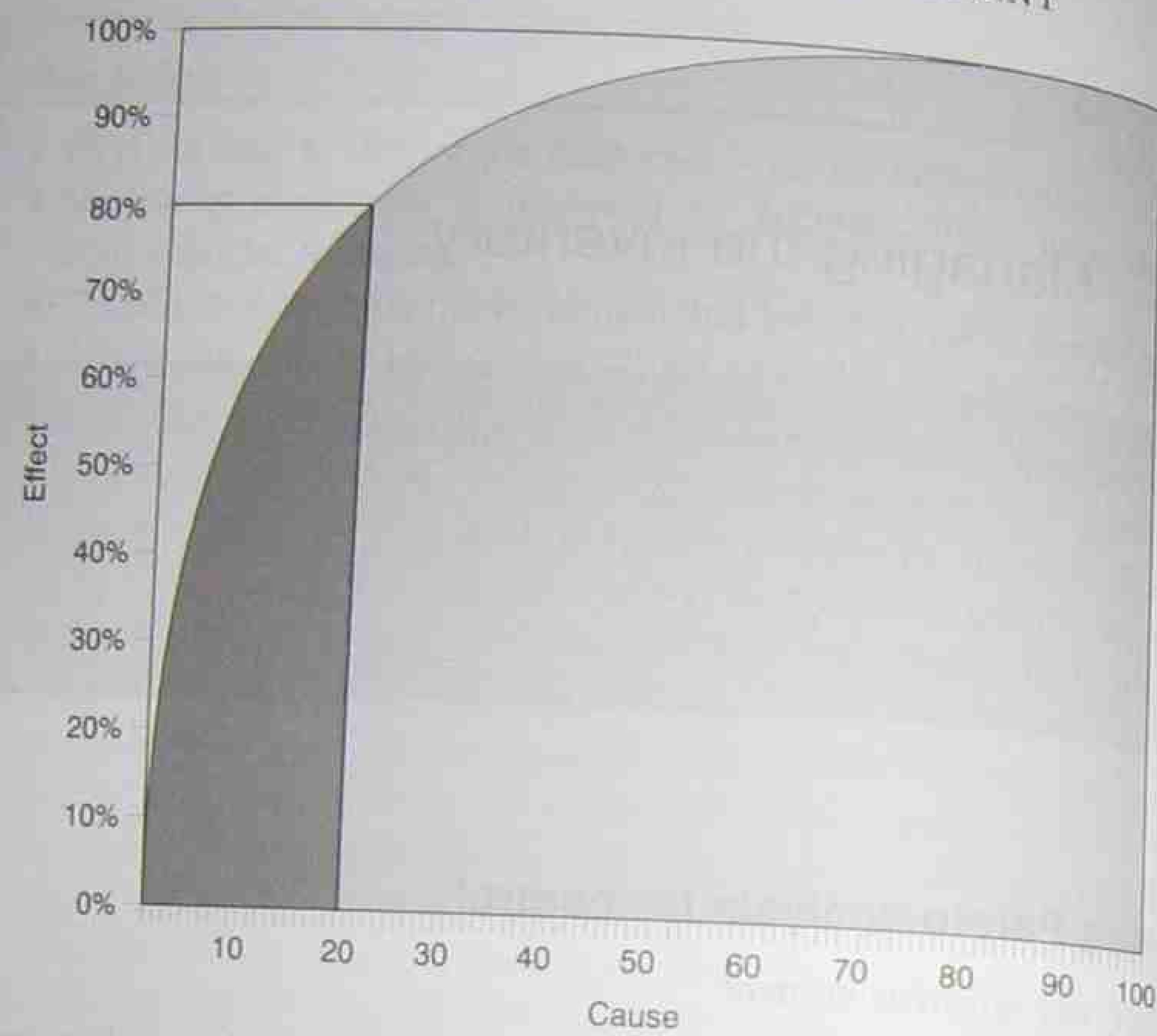


Figure 3.1 Pareto curve

Pareto analysis is the technique which forms the basis of inventory control thinking and is an important management principle which can be applied to minimize effort and to obtain best results. It can also be applied for time management, credit control and many other areas of control. To gain best control, effort has to be directed to the most important areas. The Pareto curve (see Figure 3.1) is often called the '80/20 rule', but the values can be read off at any convenient point. For example, the graph shows that 50 per cent of the product lines account for 97 per cent of the sales (or that the other 50 per cent only provide 3 per cent of the sales – a worrying thought). The shape of the Pareto curve arises from the range of volumes and values combined in a statistical distribution. The shape of the curve does not always give exactly an 80/20 relationship, but this does not affect the principles of applying Pareto analysis to inventory management.

Stores carry a wide variety of items, with a stock record for each. Some have high value and others are very cheap. The high-value items are normally controlled tightly, whereas the low-value items are not treated as carefully and are issued in bulk in approximate quantities. Most effort should be put into managing items which are most important for achieving

inventory targets. In inventory control the best results are gained by organizing effort correctly. There is not sufficient management time to maintain detailed control of all individual items. If the immediate aim is to reduce stockholding costs then studying the stock of low-value items is unlikely to be the best place to start unless the sales volume is very large. If service is the aim, then attention to a few fast-moving lines often provides the bulk of the improvement required.

> *Use Pareto analysis to save time and get results.*

Stores contain items ranging from main products to washers and labels, with a stock record for each. High stock value items need to be closely controlled, whereas minor items need not be treated as carefully. To control the resources of the company most effectively our effort and controls should be biased towards high cost areas. Pareto analysis formalizes our efforts to do this. It states that the majority of the effect is produced by a small proportion of the cause. (80 per cent of the effect is due to 20 per cent of the cause.)

The application of this to a stores stock control means that 80 per cent of the total stock value is made up from 20 per cent of the total stock items as stated. The other 80 per cent of stock items contribute only 20 per cent to the total inventory value (shown in Figure 3.1.) In a stock reduction exercise the majority of our cost saving will be gained by decreasing stocks of those few major items.

> *Eighty per cent of stock is in 20 per cent lines – so reduce the high stock values.*

Example

Consider a stock of 12 000 types of items in store. Pareto's Law shows that for a stock value of £800 000 we find that 2400 items account for £640 000 of inventory. The remaining 9600 items are worth only £160 000. Therefore, by concentrating on the 2400, control over the total value will be tight. If 2400 is rather too many to review individually, then the Pareto curve (Figure 3.1) shows that 5 per cent of items account for 55 per cent of cost so 600 items contribute £442 000 to the total stock costs. Again by working on these 600 items carefully the overall stock value can be controlled or decreased.

ABC analysis

Pareto analysis by the current stock level is good for reducing stock levels, but a more consistent classification is required when focusing on the management of inventory. The current stock does not necessarily show which items are important for the business. In fact there may be some important items where the current level of stock is low because Stores are awaiting an impending delivery. On the other hand some items may have a high stock value simply because no one is buying these. It is therefore usual to rank the items according to the annual turnover. The annual turnover is given by

Annual usage × Unit cost

It is not too important whether the unit cost is the standard cost, latest cost or an average as long as it is consistent across all the items. Annual usage has to be adjusted in the case of new or obsolescent items to reflect the future expected demand rate rather than the historical one.

Pareto analysis of this data shows that 80 per cent of the value of demand is for 20 per cent of the moving items. (There is often a number of items in stock for which the demand is zero, and therefore these items are not included in this turnover analysis.) For some businesses the 80/20 rule is not obeyed exactly, but the use of Pareto analysis is important for all inventory.

To use Pareto analysis properly requires the classification of stock by issue value, and the simplest way is to use ABC classes. These can be defined as:

- A = 10 per cent of stock numbers, giving 65 per cent of turnover.
- B = 20 per cent of stock numbers, giving 25 per cent of turnover.
- C = 70 per cent of stock numbers, giving 10 per cent of turnover.

> *ABC analysis is an excellent technique for achieving objectives.*

This is illustrated in Figure 3.2.

It is important to ensure that ABC analysis is based on turnover, but less important that the exact percentages are adhered to. In some instances a further classification (D) is useful to include a large number of very low turnover items. This enables the number of stock lines included in the A, B and C classes to be reduced to manageable numbers.

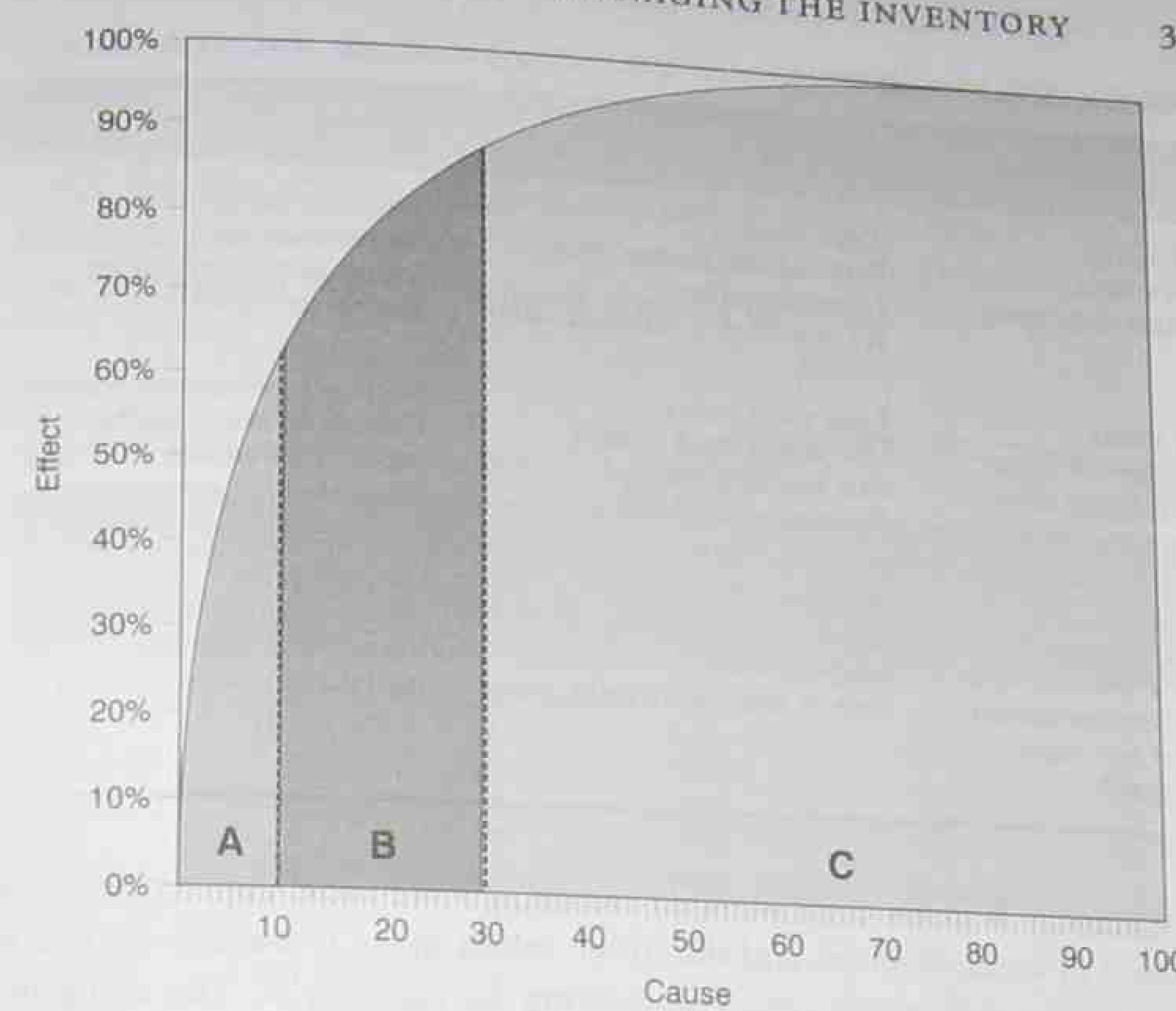


Figure 3.2 ABC analysis

The purpose of ABC analysis is not to provide different types of service but to provide service with the least amount of cost and effort. Different systems of control are used for the three categories of stock. As the A items carry the most value, accurate systems are required to control them. On the other hand, the C items are low turnover value but form the bulk of the inventory. For these the main requirement is to ensure that stock is available to meet demand.

> *A class – there aren't many, so control tightly.*

The control requirements for each category are shown in Table 3.1. Category A items have a disproportionate amount of time and effort used on them and have to be controlled tightly using systems in conjunction with market expertise and product knowledge to maintain stock at the lowest appropriate level.

> *B class – let the system manage these.*

Table 3.1 ABC inventory control

Characteristics	Policy	Method
A items Few items Most of turnover	Tight control Personal supervision Communication JIT approach – balanced safety stock	Frequent monitoring Accurate records Sophisticated forecasting Service-level policy
B items Important items Significant turnover	Lean stock policy Use classic stock control Fast appraisal method Manage by exception	Rely on sophisticated method Calculated safety stocks Limit order value Computerized management and exception reporting
C items Many items Low turnover value (Few movements or low value items)	Minimum supervision Supply to order where possible Large orders Zero or high safety stock policy	Simple system Avoid stockouts and excess Infrequent ordering Automatic system

For category B items computerized techniques are most appropriate. The number of items involved and the lower values make it a waste of time to use specialist skill which could be working on category A. The computer can maintain control through statistics and deal with the complex calculations using the computer forecasting models which are most important for B items. The use of management by exception is also important for B class.

The minor sales items, category C, should be controlled by a simple system which enables supply to be obtained with a minimum of administration. However, the control system for C items must be reliable and not result in stockouts or large excesses. An investment in extra stock of C class items is inexpensive but can greatly simplify the problems of controlling large numbers of stock lines. This is an appropriate policy for the faster moving C class items. For the very slow moving, higher value C class items a purchase to order policy should be adopted where possible, or if there is only one customer, they can hold the stock themselves and be responsible for reordering as required.

> C class – don't take risks, be lazy.
> High but not excessive inventory.

The most effective stock control systems are based upon ABC analysis combined in a common-sense manner with the other techniques we shall be discussing later.

ABC analysis is the basis for the total control of stock. It is also used as the basis for perpetual inventory stores control where annual stocktaking is avoided by routine counting of a few stock parts each week.

An example of Pareto analysis in action is given in Tables 3.2–3.4 and Figure 3.3. Table 3.2 shows a number of different stock items, their unit cost and annual usage in terms of quantity of value.

Table 3.2 Example of Pareto analysis

Item	Annual usage (units)	Unit cost (£)	Annual turnover (£)	Annual turnover (%)	Rank
A12	21	7	147	2.1	5
B23	105	11	1155	16.3	2
C34	2	15	30	0.4	10
D45	50	5	250	3.5	4
E56	9	14	126	1.8	6
F67	394	12	4728	66.6	1
G78	5	8	40	0.6	9
H89	500	1	500	7.0	3
I90	11	4	44	0.6	8
J01	3	25	75	1.1	7
Total			7095	100	

Re-arrange

Table 3.3 Classification by usage value

Item	Annual usage (units)	Unit cost (£)	Annual turnover (£)	Annual turnover (%)	Rank	Class	Cumulative percentage
F67	394	12	4728	66.6	1	A	66.6
B23	105	11	1155	16.3	2	B	82.9
H89	500	1	500	7.0	3	B	89.9
D45	50	5	250	3.5	4	C	93.4
A12	21	7	147	2.1	5	C	95.5
E56	9	14	126	1.8	6	C	97.3
J01	3	25	75	1.1	7	C	98.4
I90	11	4	44	0.6	8	C	99.0
G78	5	8	40	0.6	9	C	99.6
C34	2	15	30	0.4	10	C	100
Total			7095	100			

Classify

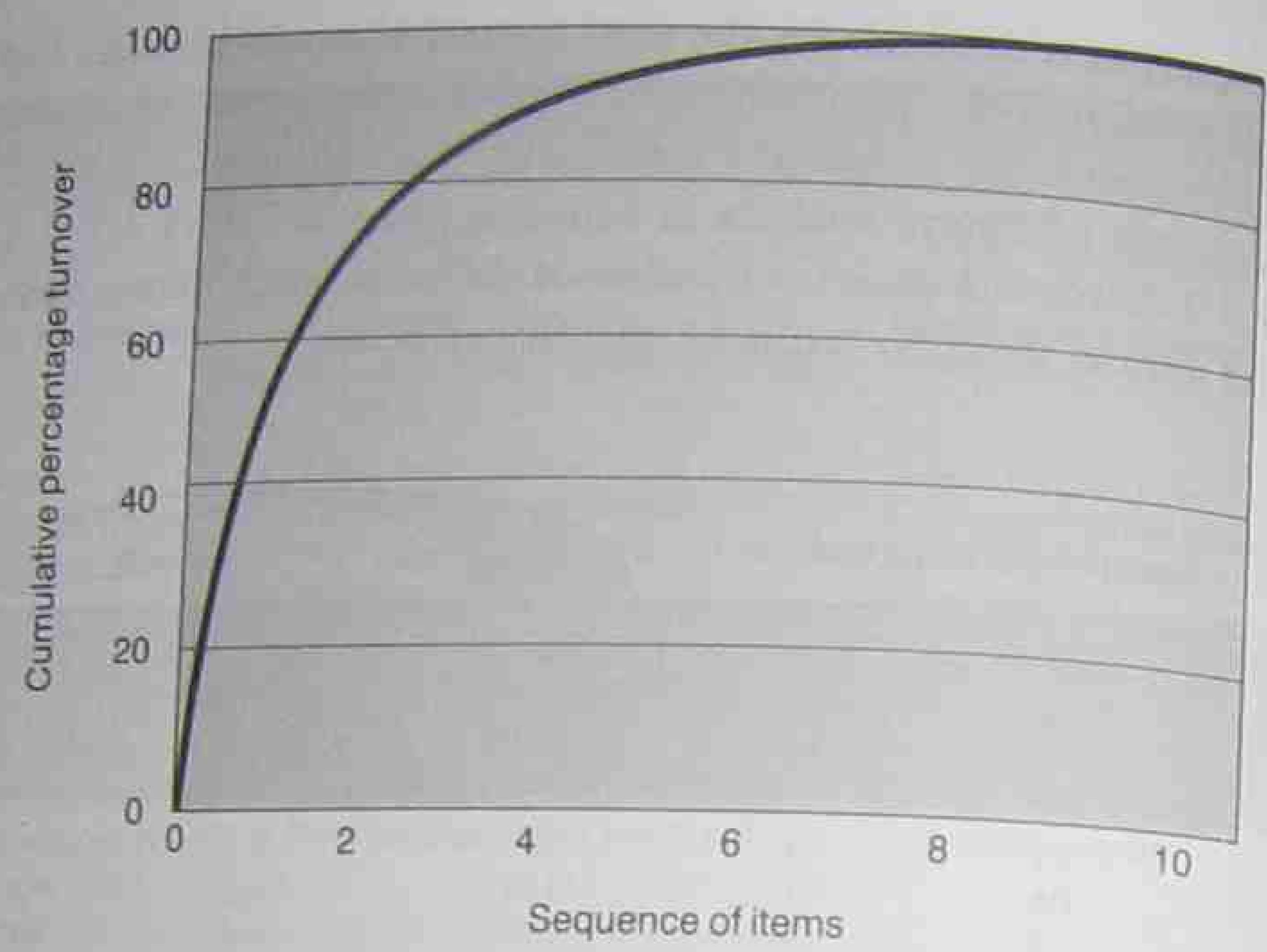


Figure 3.3 The cumulative Pareto curve

Items are then ranked in order of size of annual turnover value. This is displayed in Table 3.3 with the items in descending order. The cumulative annual usage value and percentage of turnover are also calculated. Items are then classified into A, B and C by assessing the number of items in each category and their percentage of total turnover value.

In the Table 3.4, classes A, B and C are compared as percentages of number of items and total value. The table shows that the seven items in class C have a turnover of £712, whereas the only item in class A has a turnover of £4764.

Table 3.4 Summary of ABC analysis

Classification	Percentage of items[1]	Percentage of value	Value per class
A	10.0	66.6	4728
B	20.0	23.3	1655
C	70.0	10.1	712
Total	100.0	100.0	7095

Note: [1]ABC analysis is carried out only for items with usage.

Saving time

Applying the Pareto principle is a way of balancing inventory, stock availability and critical resource spent on each item. How the law can be applied depends on what the critical resource is considered to be. The critical resource for all inventory controllers is time, because of the large amount of information required for tight control and the wide variety and quantity of stock held. The Pareto principle shows that 80 per cent of the time is spent doing 20 per cent of the jobs and a significant time saving can be made if a small reduction can be achieved in these jobs. They may be very frequent short jobs (such as keying stores issues into a computer) or more infrequent, long-winded jobs (such as writing a major management report).

Practical considerations in using ABC analysis

The ABC classification is a simple tool to enable the stock manager to control a large number of items in a limited amount of time. This simple approach is one of the most powerful tools employed to reduce stock value and to decrease the workload of busy inventory managers and purchasers. Experience with this technique over a large number of companies has suggested that there are some pitfalls and some practical ways to circumvent them. The situations which arise are typically:

1 Too many A class items.
2 Large numbers of lines (D class).
3 Non-moving lines (O class).
4 Fixed stock level items (F class).
5 Non-stock items.

Use ABC to save time.

Number of A class items

A class items are supposed to be reviewed on a daily basis, or a weekly basis at least. To be practical, there has to be an upper limit of, say, 300 A class items per inventory planner. Even where the planner manages many thousands of items, the number of A class has to be kept small.

D class items

For normal inventories of 3000 lines, the ABC classification works well. Where there are over 10 000 part numbers there should be a modification to the classification system so that the vast bulk of low-value turnover items are dumped into a further class, D. The D class contain the lowest turnover lines, say 50 per cent of the active item numbers which contribute only 2–3 per cent of the total turnover. The ABCD classification now is as follows:

- A class: 5 per cent of moving lines (300 per planner maximum).
- B class: 10 per cent of moving lines.
- C class: 35 per cent of moving lines.
- D class: 50 per cent of moving lines.

The classification can alternatively be carried out by turnover value, for example:

- A class: 45 per cent of turnover.
- B class: 30 per cent of turnover.
- C class: 22 per cent of turnover.
- D class: 3 per cent of turnover.

These classifications can be varied to suit the exact shape of the Pareto distribution for the inventory to be managed. The principles of use of the technique do not change even if the distribution is not 80/20.

Non-movers

So far in the discussion the moving inventory has been classified, and the non-movers have been ignored. Companies do sometimes have a need to keep non-moving items, or items where the movement is so slow that they appear as non-movers in the recording systems. These items will probably be subject to the stock cleansing discussed in the section on turnover of stock. In the interim they should be provided with a separate classification. Normally they are given class 'O' or 'X' which signifies that they should not be ordered again.

> Use ABCOFZ as a practical control method.

Fixed classification

As Pareto analysis is to be used for ordering, there are a few items where the stock level should not respond to usage rate. (Unfortunately in many older stock management systems, the stock control parameters are all like that!)

For example, the employment of two new maintenance fitters in a factory requires them to be kitted out with protective clothing, tools, tool boxes and a variety of items. The use of these items is likely to fall rather than rise as a result of that action since there is less likelihood of further recruitment. By putting these items in a separate classification (say F), they can be segregated and identified. The system can then identify that the stock level parameters from classification F are not updated.

Non-stock items

The decision as to which items are in the stock range is an arbitrary one and depends on the particular inventory policy and market conditions. Customers do not usually consider their requirements as 'stock' or 'non-stock' items and the difference to them is only in the lead time provided by the supplier. In fact, non-stock items are continuously being taken into the stock range and stock items being deleted. For this reason it is useful to include non-stock items in the ABC analysis. This can either be done by including them within the ABC classes or, more commonly, to have a separate classification (say, Z) for non-stock items. This separate classification then has the feature that no stock is ordered from suppliers unless the stock cover is negative (i.e. there is a customer requirement but no stock). It is very useful to include all goods in the inventory management system as it provides unified records identification and control over all goods, and makes management and analysis easy.

Stock cover

Turnover of stock

The current stock levels in the various stores throughout the company may not all be at the ideal levels, as we have seen. The purpose of controlling the inventory is to drive the stocks toward their proper level which is determined by the characteristics of supply and demand patterns. The major factors are:

- supply lead time
- average demand rate
- variability of demand
- supply frequency
- customer delivery time allowed.

There are also practical considerations such as:

- reliability of the supplier
- criticality of the item
- availability of item from other sources.

The concept of 'balance' is most important in ensuring that the maximum service is produced from a minimum of stockholding cost.

The best level of service will be provided if there is an equal chance of all items being available for the customer. High stocks of one item and no stocks of another will reduce the overall availability and increase the inventory cost. The percentage availability target (see Chapter 2) should therefore be the same for all items in the stock range unless there are specific marketing reasons for favouring some customers or products.

If the cost of managing the inventory is also considered, then the value of items being ordered and controlled becomes important. Valuable inventory management resources should be confined to the most cost-effective jobs. This means that time should be spent on high turnover value items and not wasted on items whose value is insignificant. The best balance of inventory leads to an optimization of costs and service over the full range of stock lines within the time available.

The inventory performance of each item can be monitored using a figure of merit for stock balance. This is the 'stock cover', which is defined as:

$$\text{Stock cover} = \frac{\text{Current stock} \times 52}{\text{Forecast annual usage}}$$

This gives the result in terms of weeks on hand. The same answer could be calculated for a month's predicted usage and multiplying the stock by 4.2 instead of 52. This is more convenient sometimes, but should only be used when demand is consistent. It is also convenient to use historical average usage in this equation. (See Chapters 10 and 11.)

A sample of stock items is shown in Table 3.5. Which of these items requires attention first?

Stock cover gives an insight into the priority for action. It is not an infallible guide, but it does indicate where review is required. The first instinct would be to look at 1P4 and 1P8 in Table 3.5, where the stock cover is small. However, there may not be a problem with these since a delivery

Table 3.5 Stock cover

Item code	Stock (units)	Annual usage (units)	Stock cover (weeks)
1P1	250	2000	6.50
1P2	700	1625	22.40
1P3	500	400	65.00
1P4	15	1000	0.78
1P5	20	25	41.60
1P6	40	250	8.32
1P7	500	200	130.00
1P8	8	400	1.04
1P9	6	40	7.80
1P10	65	20	169.00

may be arriving. At the other end of the scale 1P3, 1P7 and 1P10 have over a year's worth of stock, so means of reducing this level will have to be found. Stock cover shows whether the stock is 'in the right ballpark': if it is not correct, attention to the items which are obviously well outside acceptable stock levels will keep the shape of the inventory reasonable. Stock cover is used by inventory controllers because it is easily understood in terms of usage rates and lead times. Stock cover is the time in which the stock will run out at average usage rate.

Use stock cover for analysis but not control.

As well as being a crude analysis tool for each stock item, stock cover is also an important tool for measuring the total inventory. Overall stock cover is calculated from the total value of stock divided by the annual issue value and multiplied by 52 to change the answer to weeks. In many distribution stores the answer is between one and eight weeks. In fast-moving consumer goods, one to eight days is more appropriate. Financial managers are often more interested in the use of funds, and therefore measure the effectiveness of inventory management using 'stock turnover' or 'stockturn'. This is just the reciprocal of the stock cover, taken on a value basis for the complete stockholding.

$$\text{Stock turnover} = \frac{\text{Value of annual usage}}{\text{Value of stock}}$$

This calculation gives the number of times the stock would be used up per year.

Example of stockturn calculation

Value of stock in the stores is £150 000

Issues for the last twelve months amount to £900 000

Stockturn is therefore 900 000 ÷ 150 000 = 6.

This means that the stock value would be used up completely six times per year so that the stock cover for the total stock will be two months, or by the stock cover calculation as

Stock cover = 150 000 × 52/900 000 = 8.67 weeks.

Stockturn is based on historical data and is used for financial reporting. Stock cover is an inventory management tool for planning stockholding and can be based on known data and the forecast usage rate, so that the stock will meet the expected demand for the item. When the stock level is being assessed for accounting purposes, the ratio uses the historical usage rate, which enables a conservative view to be taken of the stock level. Although this sometimes leads to a divergence of views on the necessary stock-holding, the assessed future demand should always be used when controlling stock or placing orders.

Setting stock targets

For good stock balance the stock cover of all the items should have similar value. In practice differences in the variability of usage and the order cycles leads to a range of acceptable values for the items. Stock cover should not be used for working out reorder levels – there are proper accurate ways of doing this (see Chapter 7). Stock cover ratios can be used to calculate the broad ranges of weeks' cover which would be needed for inventory items. For instance it is unlikely that more than twenty-six weeks' worth of stock is planned for any inventory item, therefore a figure of more than twenty-six could be the boundary between 'OK.' and 'needs attention'.

As inventory control should be tightest for the A class items, these are the ones which can be controlled down to lower stock cover figures (leading to

the paradoxical situation of holding lowest stocks of the best sellers!) whereas extra stock of the minor C class items adds little to stock value and significantly reduces the work of controlling.

In practice the stocks could have control limits to avoid extremes of inventory, and an allowable stock cover range can therefore be set by the ABC inventory classes in a ratio which is theoretically 1:3:7. An illustration of the acceptable ranges for category A, B and C items is shown as:

- A class: between one and four weeks.
- B class: between two and eight weeks.
- C class: between three and twenty weeks.

A stockturn ratio (weeks of stock) for all the items in the category should lie in the ranges shown (see Figure 3.4.)

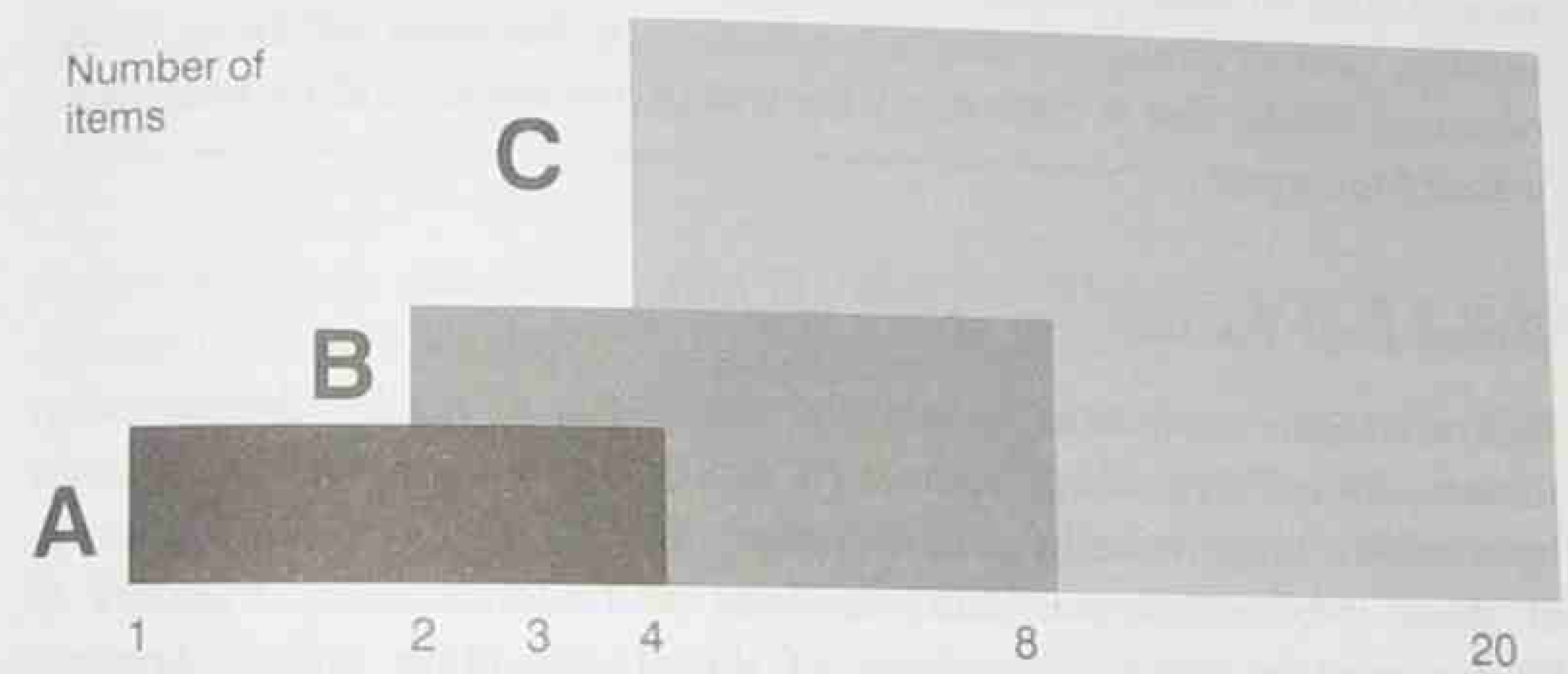


Figure 3.4 Ranges of stock cover

In theory the distribution of the population of items within a class is shown in Figure 3.4. Also shown in this figure is what the curve looks like in practice. (Note that the horizontal scale is logarithmic.) The theoretical curve allows for the variability of demand, the usage of the safety stock for some items and some slow movers, causing some of the items in a class to fall naturally outside the expected limits. The actual population curve when plotted against cover has an entirely different shape. Some items have high stock cover ratios, and since the stock value has to be limited, some of the other items have very low stocks to compensate. The shape of this curve is caused by the response systems initiated by low stock levels and the less effective action which normally is seen when the stock levels for some of the

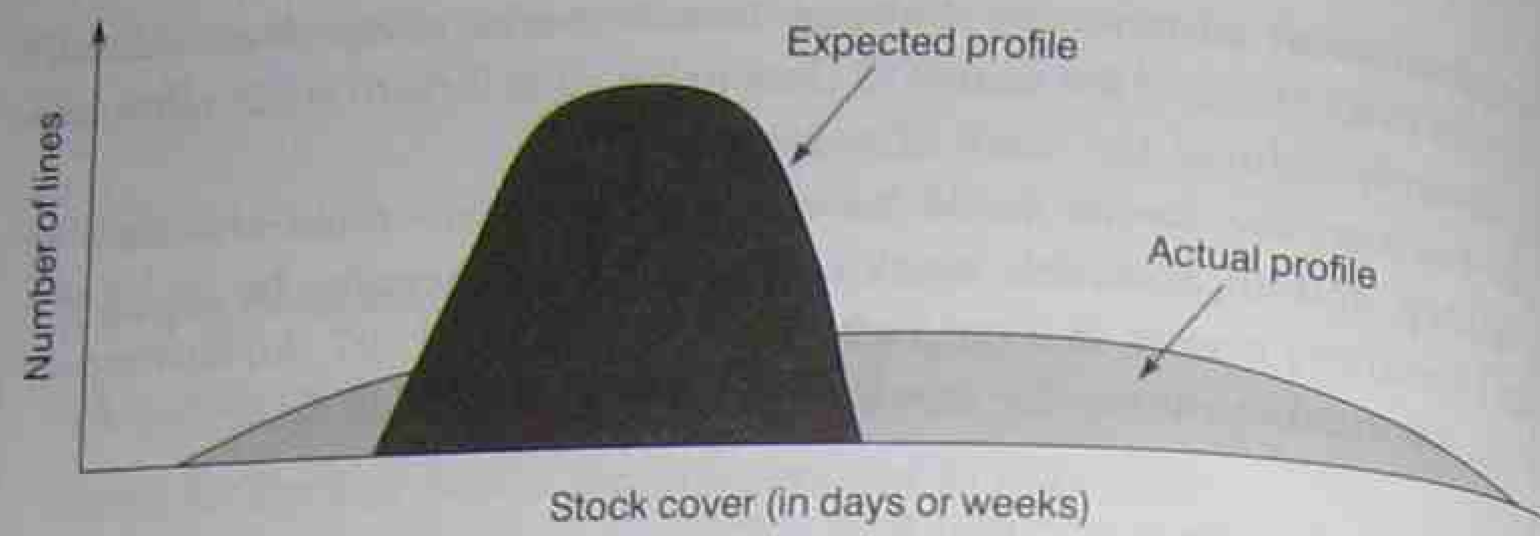


Figure 3.5 Expected and actual stock profiles

items rise over the maximum. The profiles of stock turn ratio in Figure 3.5 are typical, even where inventory management practices are good.

The weeks of stock cover shown in Figure 3.4 indicate how long it would take at best to reduce stock levels. If the stock reduction is to be made through C class items, the time taken to reduce the levels will be in excess of twenty weeks. For A class items the whole process should be completed within four weeks.

Pareto stock balance

Pareto analysis enables us to take the discussion of stock balance a step further. Each Pareto class, calculated according to annual turnover, can have its own target months of stock range, as shown in Figure 3.4. The use of these two techniques (ABC analysis and stock cover) together is fundamental in forming the basis of the stockholding policy. Another illustration of the power of the Pareto rule is shown when the overall value of stock has to be reduced to meet a new target. Say the inventory investment currently amounts to six weeks' of stock and this needs reducing by a minimum of 10 per cent, then ABC analysis of the inventory items would follow the pattern given in Figure 3.5.

In Table 3.6 the A, B and C classes have average stockturns of 2.0, 5.0 and 11.5 weeks, and therefore constitute a well controlled inventory. If these amounts of stock cover appear to be over generous, then the substitution of the word 'day' for the word 'week' in the heading will suit all but those involved in JIT supply (see Chapter 4). The classes contribute 65 per cent, 25 per cent and 10 per cent to the inventory value, respectively. This means that the total stock values arising from A, B and C classes are equal.

There are several options for decreasing the stock value. One way is to cut A items down to an average of 1.5 weeks of stock. This new situation is

Table 3.6 ABC stock cover

Class	Moving lines (%)	Stock cover (weeks)	Turnover (%)	Weeks of value
A	0.1	2.0	0.65	1.30
B	0.2	5.0	0.25	1.25
C	0.7	11.5	0.10	1.15
Total				3.70

Table 3.7 Options for reducing stock value

Class	Moving lines (%)	Stock cover (weeks)	Turnover (%)	Weeks of value	Reduction in weeks of value	Number of lines involved
A	0.1	**1.5**	0.65	0.975	0.325	100
B	0.2	5.0	0.25	1.250		
C	0.7	**8.0**	0.10	0.800	0.350	700
Total				3.025		

shown in the A class calculation in Table 3.7. The net result is a stock reduction of 0.33 weeks.

An alternative way to achieve the same result is to reduce the stock of C class parts down to 8 weeks (from 11.5 weeks). The new contribution to stock cover is 8×0.10 (= 0.8) weeks of value, a reduction of 0.35 weeks.

Tighter control of A items is the better alternative because:

- the change required is less (0.5 week as opposed to 3.5 weeks' cover)
- there are fewer lines involved, so less disruption of planning
- less work will be involved – 100 stock items instead of 700 (last column of Table 3.7)
- the effect will be more rapid.

This last point can be proved by looking at the stock cover for each of the ranges in Figure 3.5. The average taken for stock cover for A, B, and C, are 2.0, 5.0 and 11.5 weeks respectively. If the classes were perfectly balanced, and the supply was stopped, then the A class would run out after 2 weeks, the B Class after 5 weeks and the C Class after 11.5 weeks. Stock reduction through natural usage will take at least these respective times. As stocks are not often perfectly balanced, the true time to reduce the stock is normally over twice this long.

> Using ABC and stock cover together saves time and inventory.

Purchase order patterns based on Pareto analysis have been devised to save stock value and management time. This is well illustrated by the ordering process shown in Table 3.8. Consider an inventory of 1000 different lines. If each line is purchased each month from the suppliers, then there are 1000 lines ordered each month, or 12 000 per year, and a stock cover of just over two weeks plus safety stock. (Stock goes up by 4.2 weeks' worth on receipt and then reduces to nil, so the average is 2.1 weeks.) Safety stock can be omitted from the discussion since it is taken to be independent of the order pattern. The order pattern is A weekly, B fortnightly.

Table 3.8 Number of purchase orders with ABC inventories

Class	Moving lines (%)	Stock cover (weeks)	Turnover (%)	Weeks of value	Orders per year[1]
A	10	1	65	0.33	5 200
B	20	4	25	0.50	2 600
C	70	10	10	0.50	3 640
Total	100		100	1.33	11 440

Note: [1]Per 1000 lines.

C then weekly gives a reduction in both inventory and orders.

Further, if the A Class items were ordered twice as frequently and the C class half as frequently, what would be the effect?

The result is that:

- the total stock value is reduced from 2.1 weeks' cover down to 1.55 weeks' cover (excluding safety stock in each case)
- the number of line orders placed is also reduced from 12 000 per year down to 9750.

This shows that a reduction in effort and a decrease in stock are brought about at the same time. These results are typical and are borne out in practice. The effect on customer service from these changes is negligible because the safety stock remains the same in each instance. (The availability is based on the calculations shown in Chapter 6.) Stock discussed in Table 3.8 could be the same as overall inventory shown in Figure 3.4, where the total inventory includes the safety stock in the calculation.

Practical methods of reducing stockholding

The approach

Going back to the basic principle of

$$\text{Decrease in stock} = \text{Output} - \text{input}$$

shows that the way to reduce stock is to decrease input and increase output. A normal range of stock in stores comprises fast moving parts, a wealth of obsolescent and special items and fluctuating quantities of other parts, some in short supply, some overprovisioned. In this situation there is potential for decreasing stock by a significant amount, probably 30 per cent over eighteen months.

For many items there is little manoeuvrability to change the unit cost significantly. However this should be considered, since a 10 per cent reduction in price will cause a corresponding drop in inventory value in the long term. The average value of a stock line is given by multiplying

$$\text{Unit cost} \times \text{Average stock.}$$

The average stock quantity is typically halfway between maximum and minimum[1] stock. The minimum should be the safety stock and the maximum occurs immediately after a delivery. If the stock falls to safety stock level before delivery then:

$$\text{Average stock} = \frac{(\text{max.} + \text{min.})}{2}$$

$$= \frac{(\text{Safety stock} + \text{Order quantity}) + \text{Safety stock}}{2}$$

$$= \text{Safety stock} + \frac{(\text{Order quantity})}{2}$$

See Figure 3.6.

[1] The 'minimum' on some systems is set to the review level for triggering resupply even though the stock continues to fall below this level until the delivery arrives. Here, the 'minimum' is the lowest average stock level immediately before delivery.

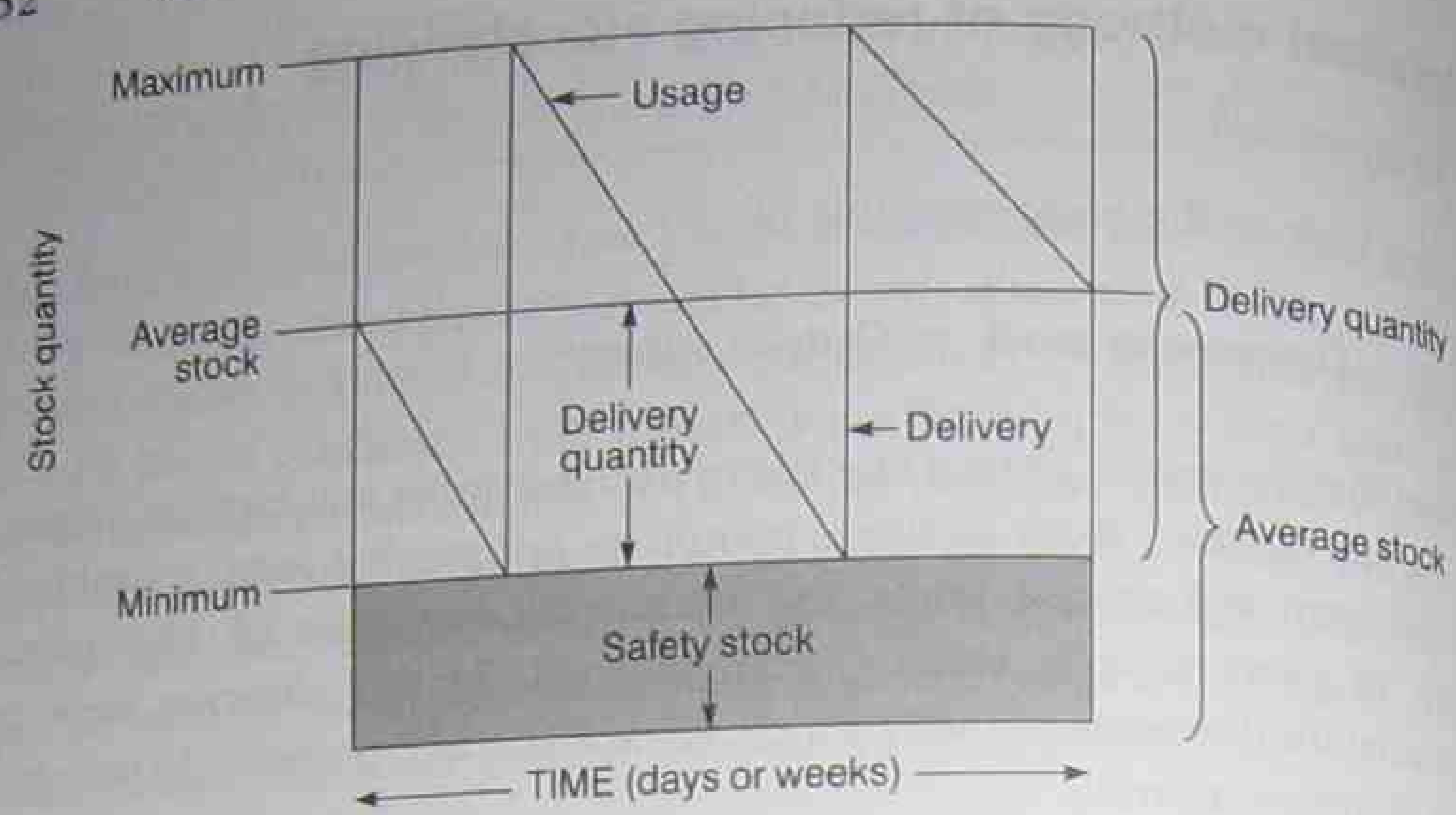


Figure 3.6

In order to reduce the average stock, therefore, there needs to be a reduction in either

- order quantity, or
- safety stock.

Batch quantity

The lowest inventory will result from receiving the same order size into stock as the issues to customers and, of course, this will be at a minimum if the order size is small. For those items which are individual, then the minimum order size is one. The parameter for order quantity on this basis is therefore the simple message 'buy one'. This is also the obvious answer for the majority of slow moving stock items. It is not so obvious when stated in the form: 'hold the least stock cover for the most important stock items', a consequence of good stock control and also of JIT supply. Minimizing the delivery quantities of the A class items (or those contributing to the highest stock value) is normally the most fruitful way of reducing stock value.

Supplier delivery quantities are more important than lead times.

Safety stock

Minimizing the safety stock could result in lower customer service. Because the balance of stock is never perfect and the levels constantly need to be

changed, availability provided by the stock management needs to be monitored continuously. Therefore, the first step in controlling the high-value stocks is to ensure that the safety stock is sufficient to provide the service required and not more. Achieving a higher availability level than necessary is very expensive. The availability for each of the high-value items has to be matched to the requirements of the market. Balancing these stocks is most important. As the safety stock depends on the availability required, the lead time and the variability of demand, the next topic to consider is lead time.

Little and often for A class.

Lead times

By negotiating shorter lead times, the safety stock can be reduced. Modern purchasing practice should eliminate random demand 'over the counter' but it does require collaboration, forecasting and supply chain management so that the supplier can arrange to make items available by the demand date with a much shorter lead time. Lead time can be reduced by negotiation, and in the context of reducing stock investment, it is the few major stock items which will offer the most opportunity to negotiate.

The third factor is demand pattern. If the demand were ten every week consistently, then there would be no need for safety stock. (There are people from some areas of business who would question this simple statement, since the holding of extra stock would enable extra orders to be serviced should they arise. This extra stock is not safety stock but an investment in opportunity – or folly – and should be considered separately.) The question to be answered is 'what opportunity is there to smooth out the demand pattern?'

Example: a case of a pan

The stock records for a range of materials sold by a major supplier of aluminium were typically as follows:

Week	1	2	3	4	5	6	7	8	9	10	11	12
Demand	225	2625	75	100	200	50	125	100	0	2050	300	125

The items were aluminium circles (flat round sheets) in various sizes used for making pans. The conventional stock level calculations suggested very high stock levels in order to cope with the demand peaks in weeks 2 and 10. This was introducing a high level of stock for these items, and inflating the inventory value.

Investigation of this order pattern showed that across the range of products, these peaks were caused by one major customer who had an order cycle of two or three months and this was causing the large demands. There are several solutions to this problem:

- ***Understand demand*** – *greater understanding of the demand pattern improves the forecast. Since safety stock is required to compensate for forecast inaccuracy, the improvement of forecasting reduces safety stock. There is, therefore, the opportunity to try to smooth customer demand by getting the customer to buy little and often.*
- ***Flatten demand*** – *remove sporadic demand from the stock by buying to order and giving the customer a longer lead time.*
- ***Agree scheduled supply*** – *It would be better to organize a continuous supply by scheduling delivery.*

The reduction project

Order quantities and safety stocks should be examined to ensure that they are synchronized, starting with the item with the most stock value and continuing down the stock value Pareto curve. A target should be set so that the stock reduction for each item can be judged against it. For example, an overall stock value reduction of 20 per cent can be achieved through a 25 per cent reduction in stock of the top 20 per cent of lines. Starting from the highest value stock line, a percentage reduction should be attempted for all lines. This is unlikely to be achieved for some items, but it is essential to concentrate on achieving the target for the highest value lines or the overall target will not be met. The value under-achievement on some items has to be compensated by extra value savings on other lines.

The approach – how to reduce stock successfully

Stock reduction is usually undertaken as a project, and concentration should be given to achieving the objective in a short time. By examining

the first few lines in the Pareto curve, major savings can be achieved in a short timescale. The next step within each area of activity is to find the major cost items. By Pareto Analysis it can be assumed that 80 per cent of the value of stock exists in 20 per cent of the item types. Items with the most stock value are likely to be a mixture of fast movers with reasonably high unit value and high-value items where there are relatively few demands per year. Pareto analysis shows that the way to reduce the stock value is to concentrate on the high stock value items, whether they are slow moving or fast moving. The value of inventory is the same for each and the financial investment is the same. In fact it is easier to vary the inventory levels for the faster moving items. As far as basic Pareto classification is concerned, there is no difference between the stock for fast-moving and slow-moving items with the same stock value. The differences will be managed through the individual stock calculations for the items.

> High stock value fast movers reduce the fastest.

The stock of a fast-moving item is likely to be consumed quickly and consistently though servicing a wide variety of demands from a large customer base. The inventory profile for a fast mover is therefore large numbers of demands taking a small proportion of the inventory. Demand is relatively stable. For high-value items, there are relatively few issues to a smaller number of customers. This means that the demand is more unstable for two reasons:

- As the stock lasts longer, the risk of demand changing within the stock cover period is greater and therefore there is an enhanced risk of the stock remaining as obsolete or excess.
- The demand comes from only a few customers, if any of them reduces their offtake, this will have a much more significant effect than if there were more customers.

It is therefore especially important to manage risks on the slow-moving, A class items by forecasting, ordering little and often, and collaborating with customers and suppliers. The A items are selected for special control, applying a scheduling approach where possible and negotiating closely with the supplier. The supplier may be keeping a stock of these items specially or can be persuaded to do so if desirable. The delivery quantity for A class must be cut to a minimum and reviewed each time an order is placed.

The stock reduction project

Specific stock reduction can be achieved through a dedicated project. Continuous stock reduction is a background job for all inventory managers, but since insufficient focus is given to it, arranging a project tends to be more effective. The following steps may be taken:

1 Identify the potential for reducing stock.

(a) Find the total stock value.
(b) Find the annual material usage.
(c) Calculate the total stock cover for the various stores or categories of stock.

The results can then be tabulated, to identify which areas of stock have the best potential for significant value reduction, under the following categories: stock type, stock value, calculated stockturn, target stockturn.

The analysis of stock type can be generalized for distributors. Typical classifications are: market sector, product type, supplier, or other obvious classes. For manufacturers classifications can be into raw material, work in progress, bought-out parts, finished goods, consumables, etc.

All that is required is to reduce average stock value of significant stock lines. With this analysis it is easier to see which areas to concentrate on in a stock balancing programme. By assessing targets based on experience an initial plan can be launched, to be followed by a detailed analysis of appropriate targets at a later time.

2 The next step is to classify stock items into ABC.

3 For the few A class items, it is important that the minimum stock is held for the availability required. This is where the major effort in stock reduction should be focused. A items have to be reviewed individually and the supply arrangements reconsidered. Significant reductions are to be sought, through supply practice and better demand forecasting.

4 The reorder level and safety stock depend largely on the lead time. It is, therefore, important that lead times reflect current trading conditions and are kept low through contact with suppliers (especially for A class items).

5 For the medium price items, i.e. B class items, computer monitoring of safety stocks and order sizes will enable smaller batches to be ordered and safety stock to be reduced.

6 Avoid ordering more than (say) three months' supply of anything. Setting a limit ensures that the amount of slow-moving stock is reduced. Very long lead time items will have several purchase orders outstanding simultaneously.

7 Low value items purchased in category C are often standard products which can be obtained off the shelf. By proper planning, these need not be kept in stock. They can be purchased to meet demand or provided by a supplier on consignment. However, an efficient reorder procedure is necessary or else the works van will be touring the neighbourhood continuously picking up odds and ends.

8 From the stock turnover figures it becomes obvious that some items have too much stock cover, and are simply contributing to stock value. They are either obsolete or have very low sales. From our introduction to this section the formula shows that this stock has to be disposed of in order to have a low inventory presence.

The first step is to decide which stock is obsolete and to remove any chance of it being reordered by marking the records accordingly. The most profitable way to dispose of stock is through the servicing department. They can often sell obsolete lines to customers with old machines at full sales value or at an offer price near to it.

Stores are also likely to hold proprietary products such as components, fittings, motors, consumables, bearings and packaging which can often be sold back to the supplier at a low price. With high inflation over the past few years this return value can equal the original purchase price and so there may be no loss of assets involved.

By concentrating on major cost items the value of obsolete stock can be reduced. It generally takes time to negotiate the sale of items and therefore can be viewed as an ongoing project. Where these approaches still leave a significant number of useless items in stock, then disposal for scrap is the simplest course, and there should be an approved budget for this. The loss of assets through scrapping stock reflects directly on profit margins, and overzealous disposal projects can mean a period of poor financial results.

Armed with a budget for scrapping, items can be thrown out of stores and written off until the budget is used up. As stores are often short of space and control of slow movers is tedious, it is convenient to scrap a large number of the lowest value items in stock, thus reducing the number of stock lines and consequently freeing most space.

9 Design and planning departments can help to reduce stock requirements by very large amounts. All they need to do is to use the same items widely rather than marginally different ones for each application.

A stock rationalization exercise is normally fruitful for standard items such as electronic components, motors, fasteners, raw materials, packaging items, gearwheels and tools. This can be carried out initially by inventory control and then passed on to technical people to continue.

10 The order is the cause of increased stock. Therefore it is necessary when undertaking a stock reduction exercise to look at every order placed to ensure that:

(a) The item is required at the time purchased.
(b) A minimum is ordered.
(c) There is no stock existing which can be used instead.

11 To maintain overall control of the stock reduction project reliable management information must be available on a monthly or weekly basis for such things as the value of items on order, stock cover and availability. By plotting them on a graph the effect of these actions can be measured continuously.

Key points

- ABC analysis provides the best tool for saving time and structuring inventory.
- Stock cover enables inventory managers to look out for likely problems.
- Stock cover is not used for setting safety stock levels.
- Inventory is reduced chiefly by using Pareto analysis and controlling supply.

4

Just-in-time management

- The opportunity for low inventory
- Improving the operating conditions
- How to use just-in-time supply
- Identifying the benefits

The zero inventory philosophy

Conventional and JIT approaches

Stock exists because items have been bought before they are required. It is normally uncertainty or overcaution that causes inventory. The principle of JIT is simply that we have items when they are needed and none when they are not needed.

JIT supply is a result of high-quality supply.

The idea may be simple but the application of JIT has given the opportunity to decimate stockholding without affecting customers. Instead of trading availability and stockholding as discussed in Chapter 2, the trade-off is between organization and stockholding. The better organized and controlled the supply chain, the less inventory is required.

Companies which are considering how JIT can work in their business, or avoiding it in a Canute-like manner, should realize that JIT is an outcome of other techniques, not a technique of its own. It is the logical aim of tight inventory control, effective process planning and plant design, workforce motivation, cost reduction, logistics and even material requirements planning (MRP). The optimization of these together inevitably leads to the JIT approach. The elements of JIT are the techniques to be developed, for example:

- supply what is required
- supply the quality required

- reduce lead times
- organize effectiveness
- use all the expertise available (i.e. people who do the jobs plus technical backup).

These are the fundamental changes which lead to JIT – all good inventory management techniques. The most important of these improvements has been in quality, particularly in management processes – quality of records, supply, delivery, forecasting and target setting.

> Holding inventory shows we don't have control.

Stock controllers have been sceptical about the efficiency of frequent deliveries of smaller batches without investigating all the options and the potential. There are many instances where the accepted delivery quantity is now much smaller than it was a few years ago, and is destined to continue to be reduced. Some of the contrasting features of managing conventional stock control and JIT control are shown in Table 4.1.

Table 4.1 Contrasts between conventional and JIT inventory control

Conventional	Just in time
Satisfied with the status quo	Continuous improvement
Lead time is fixed	Reducing lead time is a continuing challenge
Product range is a sales issue	Product range reduction is an inventory issue
Management provide methods	Operators are responsible for practices
Stock in case of customer demand	Purchased to meet demand rate
Convenient purchase batch size	Buy singly or small quantities

Just in time works as a pull system whereas conventional stock control and material requirements planning are essentially push systems. In the push system the stock is provided for the next stage of supply, e.g. buying items for sales to sell or starting manufacture without having the total production path clear. The philosophy with push systems could often be thought of as packing so much into the warehouse or production plant that they will have to send some items out because there is only limited space!

With a pull system the first action in the chain is that the item is demanded. To satisfy this demand there is an item in stock. As soon as this stock is used up, then another item is supplied, either from outside or from

a production process. The supply is organized so that the effective supply lead time is very short (under one day) and the quantity supplied is small – a few minutes up to a week's worth of demand. In production the demand filters backwards through the processes creating a demand down the bill of materials, and ending up with issue of one product's worth of raw material.

Continuous improvement processes are associated with JIT including product quality, process efficiency, information systems and operating value-added activities more effectively while eliminating non-value-added activities.

Who can use JIT?

Quality process improvement is usually thought of as a continuous journey of improvement, with no definite ending, since there is always more potential. From the point of view of material flow the principle of JIT is obviously ideal but it is often difficult to implement in practical situations unless the conditions are right. Of course, the right conditions do not happen by accident and anyone looking for the benefits has to work hard to create the appropriate situation. Just-in-time supply should be considered as a quality process, although most of the objections to JIT are based on lack of quality in some aspect of supply or demand. If stock levels are incorrect, this is often the result of complacency or lack of understanding (usually both). There is no perfect solution to stockholding but, like any other quality improvement process, JIT operations gradually develop an existing unsatisfactory situation into an improved one. A decrease in lead times and simplification of processes should be the aim of all inventory managers. From the viewpoint of JIT, time is a value-added commodity and wasting it is unprofitable. The more time saved the better, and continuous improvement means reducing the timescales.

> JIT results from best practice in inventory control.

The definition of just in time presented so far can apply to any material management process which actively minimizes timescales. The purist would think that there is more to JIT than this simple concept and there are some specific concepts for achieving this reduction in timescales, particularly:

- desire to improve
- simplification
- demand-led supply (pull)

- quality conformance
- devolution of responsibility.

If these concepts are put into practice, then an operation has a JIT philosophy (most of these concepts are fundamental to good inventory management anyway). Supported by improvements in communications and driven by the need for better service and lower costs, the influence of JIT has been felt in all types of business and has fuelled change.

JIT in manufacture and warehousing

Just in time was taken up by some sectors of manufacturing originally as a space efficient method of production. For some supply chains, the application of JIT is natural because of the nature of the products or the processes, e.g. in fast-cycle manufacture, where a stage in the production process takes very little time. In a high-volume manufacture, delays between processes would cause large stockholding (or, more specifically, work in progress). In other businesses JIT techniques are less easy to apply. Retail shops find it important to hold stock so that the customer can have a choice of several options, so JIT supply would not be appropriate for this stage of supply without changing the nature of the relationship. If the customer were buying the same item by mail order, then there is a very good opportunity to use JIT techniques.

Although the original impetus for JIT was for space saving, the greatest benefits are financial. In Chapter 1 (section on profit through inventory management) the benefits of lower stock on profits were illustrated. This magnitude of improvement is likely to occur for a second time as a result of applying JIT.

Consider the example used in Chapter 1 with each of the companies having a £5 million turnover. Often companies find that half their costs are in purchased materials; let us assume that the same is true for these companies.

If these companies are distributors, then the stock cover for The Slack Company is 52 weeks and for M. Tight Ltd is 10.4 weeks. If the application of JIT on selected lines (product types) could reduce the overall stock cover to 1 week, then this would result in an inventory holding of only £48 000, a saving of £202 000 even for M. Tight Ltd. The stock cover and relative inventory investment are shown in Table 4.2 for a turnover of £5 million. The decrease in inventory values is large when the initial stock cover is high. However, as stock cover becomes low the potential savings are lower and

Table 4.2 Warehouse stockholding

Stock cover	Inventory investment (£)
52 weeks	2500
10.4 weeks	500
6 weeks	288
4.2 weeks	202
1 week	48
1 day	10
1 hour	1.3

the risks greater. It is then not the value of inventory that is the driving factor toward zero inventory, but the operational cost savings.

Applying the same logic to a manufacturing situation shows additional benefits through the savings in planning resource. The operational situation is more complex in manufacturing. The material cost starts out as, say, 50 per cent of turnover but then collects added value until it is worth typically 70 per cent of sales value. For simplicity let us assume that there is an equal balance of inventory throughout manufacture, so that the average inventory value is 60 per cent of sales value.

> *Lead time should be reduced – it results from a planning problem.*

A further aspect of manufacturing is the number of jobs in progress. It is typical for a company with a turnover of £5 million to issue about fifty new jobs to production each week. (The logic of the following discussion works just as well for companies with different numbers of jobs and turnover values.)

Using this figure a manufacturer's performance table can be created as shown in Table 4.3. Stock cover in manufacturing is required for the duration of average throughput time. For The Slack Company the £2.5 million inventory equates to 43.3 weeks' cover and a total of 2167 jobs in work (assuming that all the inventory value is in work in progress – WIP). For Tight Ltd the WIP equates to 8.7 weeks' cover and 433 jobs.

The advantage of reduced lead times in manufacturing is twofold, as can be seen from Table 4.3. Inventory investment is reduced as well as the number of jobs and the complexity of the materials management. It takes a great deal of effort to control 433 jobs, but much less to track fifty.

Table 4.3 Effect of throughput time on manufacturing inventory

Throughput time (weeks)	Inventory investment (£)	Number of jobs in progress
43.34 weeks	2500.0	2167.00
8.66 weeks	416.0	433.00
6 weeks	288.0	300.00
4.2 weeks	202.0	210.00
1 week	48.0	50.00
1 day	10.0	10.00
1 hour	1.3	1.35

Table 4.3 shows that major cost savings go hand in hand with reductions in manufacturing complexity. Inventory cost savings in particular can be achieved by cutting the lead time down to a few weeks, days or hours – and then benefits flow from other operating costs, such as less shop-floor space required, improved flexibility, less expediting, more control and less material handling. There is a similar simplification in warehousing where lower stocks means less to count and less to control.

JIT environment

The ideal situation

Over the years the perception of lead time has changed in most businesses. Lead times have shortened due to the combined influence of customers and accountants, the former wanting faster service and the latter lower inventory. This has been a gradual culture change, but the move to true JIT requires a further leap in thinking. An attitude of 'we're doing alright, don't risk it!', has to be replaced by 'this is what we need to do, now how do we make it happen?' – i.e. continuous development. Just-in-time individuals have to create an environment where things do not go wrong and their positive attitude is the basis of JIT philosophy, whose essence is reliability and consistency. Ideally this reliability must pervade the whole of the material control operation, not just one stage. Performance is dependent upon a number of factors, for example suppliers, the market, the company itself and the individuals within it, and these factors will always throw up a problem of some sort. (These should be considered as the challenges of everyday stock control.)

> Business needs to use JIT to remain competitive.

In high-volume manufacture it is easy to see how items can be transferred from operation to operation, and lead times and cycle times can be kept very low. In distribution operations where the demand is variable and uncertain, then the application of JIT is rather more difficult, but not impossible.

Typical features of the ideal company where JIT concepts could be applied would be:

- narrow product range
- manufacturer
- high volume
- stable market
- influential
- good quality management
- local, reliable suppliers
- dependent suppliers
- fast-cycle processes
- affluent.

The one major element which was left out of this ideal list is personal commitment, and although comparison with one's own current business profile might prompt many to discard JIT as inappropriate, a positive attitude will prompt others to consider how close their business is to these ideals and how JIT can work in reality. This is a major factor because successful JIT depends on commitment by a core of people at operational level and equal dedication of managers and directors. As in all good management, it is useful to have monitors for success. These are local targets decided upon by the local team to set and manage their own performance and criteria for success. Quality and throughput will be part of these measures.

Repetitive processes

Just in time can be seen to work best in repetitive processes, whether in warehousing or in manufacturing. Of course, almost all of business is based on continuity of supply, so this is not really a major constraint. For a usage of twenty per month, a traditional approach could be to buy twenty at a time, whereas JIT would require one per day. Such a small order rate may not appear economic for one item, but it would be when dealing with a large number of items.

Consider a packing or manufacturing line that each week completes

- 2000 of item a
- 3000 of item b
- 4000 of item c
- 2500 of item d
- 1000 of item e.

Instead of working on each type of item in turn, it would be better to provide each item at the rate at which they are required. This can be achieved in two ways:

- produce at a balanced hourly rate across the whole range
- produce to order and adjust the rate to suit demand.

For the first option the rate of production is consistent. The supply rate is shown in Table 4.4. For this option to work, there has to be a very consistent demand or a buffer stock somewhere in the system. This is generally the way that JIT operates in warehousing and manufacture where internal planning is on a fast-cycle basis, but customer demand is on a slower-cycle basis so that demand fluctuations can be smoothed out over the customer supply cycle.

Examples of this situation are the automotive industry and high technology equipment manufacture. In these cases the supply can be geared to a constant rate of output and the resources organized to efficient and cost effective supply (and production where appropriate). In Table 4.4 the options for the stores or production unit are to have a single throughput channel which completes an item every 11 seconds, or to have a multiple of channels which deal with a single or a limited range of items.

Table 4.4 Continuous supply cycle times

Item	Throughput per week	Throughput per day	Cycle time (secs)
A	2 000	400	68
B	3 000	600	45
C	4 000	800	34
D	2 500	500	54
E	1 000	200	135
Total	12 500	2500	11

> By making manufacturing repetitive planning capacity and materials is easy.

Example

In the automotive industry there is generally one production line making the variety of processes at the same time but on one car at a time (with a parallel line for flexibility and offsetting the effects of changeover or plant failure). In a mail-order warehouse the capital investment per line is not so great, and stock items are more diverse, so that generally there is a number of specialized packing lines.

In warehousing it is important to apply JIT where it is easiest to use, and the conditions discussed earlier in this chapter point toward the A class items with high usage, more consistent demand patterns and significant benefits from inventory reduction. The items sold can be split into two types:

- those controlled conventionally
- those using JIT.

The major items will be JIT controlled and the wide variety of low turnover value items controlled conventionally.

When we have achieved full control through conventional means (see Chapter 7 for methods) the control for the second option (JIT) becomes an extension of this. Here the customer requires the items on a fast-cycle basis. This means that fluctuations in demand have to be taken up by varying the capacity available. This is less efficient for the use of resources, but may avoid inventory build-up elsewhere in the supply chain.

JIT supply

Processes can be considered to be suitable for JIT if the timescale is in hours rather than days. For most suppliers this requires special arrangements for feedback of demand and issue of items within a short timescale. Effective JIT supply requires the same approach as that used for single source suppliers (see Chapter 8): there has to be a longer-term agreement on anticipated throughput levels, quality management and commitment by the supplier to provide an infallible supply of specified items. The supplier should be treated like an integral part of the organization.

One of the changes in attitude that this brings is the approach to transport. Often the suppliers organize delivery so that they incur the cost. With JIT supply there is an advantage if the customer collects the items at a time convenient to them (the pull system being extended in this way through the transport system), since the transport cost would be incurred in the final product cost whichever company pays for it.

For a fast pull-type relationship to work, then the supplier has to be local to the customer, otherwise the reactivity is not good enough. Location is of significant importance, and some suppliers have moved to the vicinity of a major customer in order to ensure close cooperation.

JIT and distribution

Just in time requires more frequent deliveries from suppliers but this does not automatically lead to higher transport costs, although it could do so if no changes are made to mode and arrangement of supply. Simple examples of where there is definitely no extra transport cost are:

- when the freight is charged simply by weight
- when a wider variety of items is put into each shipment and these compensate for the reduced amount of bulk orders
- where freight is charged at a flat contract rate.

Costs might increase where:

- the same delivery method is used more frequently
- batches of items are bought separately from different suppliers
- a flat rate per consignment is charged
- transport is arranged one journey at a time.

> *Change transport method and get frequent supply.*

So these situations are to be avoided.

Just-in-time philosophy requires close co-operation with suppliers of distribution services as well as of goods. By moving toward longer-term, higher-volume commitments with suppliers, prices can be reduced based on the total traffic volume over an extended period. The distributor requires a continuity of demand and the opportunity to work out routes, loads and schedules well ahead and this will reduce the operating costs as long as the business is significant for the distribution partner.

Many people are sceptical about increasing the number of deliveries without increasing the overall costs. Others who have actually tried increasing the delivery frequency have found distribution partners who can provide the service without increasing financial disadvantage, and there is a growing number of companies working this way.

The general cost savings produced by JIT are discussed below. The economics of frequent deliveries can be added to these savings for delivery processes. Some of the typical benefits are:

1. Smaller delivery quantity means smaller loads, which enable different transport methods to be used (parcel post instead of lorry load). Small bulk transport often costs considerably less, since the operating costs are reduced.
2. Frequent fixed route deliveries enable a carrier to have a base load upon which to build other business.
3. Frequent deliveries do not require planning and management effort once set up.
4. Discounts on contracts can be very significant.
5. Standard pack quantities and containers are smaller, less expensive and recycle faster.

> *JIT saves costs if it is organized properly.*

The secret of success in saving costs on frequent JIT deliveries is to:

- plan carefully so that delivery schedules are not altered at the last minute (this is easier with JIT since the planning horizon is relatively short)
- be bold – radical changes in delivery methods are required if costs are to be minimized (this often means a complete change in transport method)
- negotiate carefully – transport costs are adjustable because they contain a large fixed element of cost. The apportionment of this cost (e.g. running a lorry from A to B) can be carried out in any way. Transport suppliers may be persuaded to apportion the costs away from routine major deliveries
- be open – the size frequency and costs of delivery required should be shared with the transporter. Their relevant costs should also be known in order to ensure that their service is viable.

Quality management

Concurrent with the development of JIT has been a great move to improve quality. This has taken two forms.

First there was the move toward consistency through the development of quality standards (including ISO 9000) which were instrumental in providing reliable performance not only for the products provided, but also for procedures and paperwork. These standards have changed attitudes within industry and there is now an expectation that items provided shall be correct to specification.

The second was the development of the TQM philosophy, which does not accept current quality and strives always to provide gradually improving quality. When this philosophy is applied to the operating processes of the organization, the key concerns are the elimination of waste and operating effectiveness. Total quality management then becomes a practical approach to productivity improvement, standardization of procedures and packaging, improvement of the quality of delivery on time, record accuracy information and operating effectiveness, and the complete operating processes of the company.

> *Suppliers should pay a penalty for poor quality.*

Quality management principles integrate well into stock control methods and operations because it is a basic assumption of stock control systems that the items supplied are fit for their purpose. If a delivery batch is not acceptable, then it has to be replaced immediately for the stock control calculations to be correct. Delivery in smaller batch quantities often reduces the overall effect of rejected supply since it impacts on fewer customers' orders and leaves less time before the next delivery. Nonconformances will cause extra work, especially:

- creation and holding of extra records as a cross-check
- quality inspection on goods inward (also causing booking delays)
- sales personnel requiring physical stock checks rather than relying on the recorded quantities
- time spent investigating and checking
- arranging replacements and priority deliveries.

Formal specification of inventory management procedures enables them to be reviewed and identified for quality management processes. As time is costly, the allocation of time to administration has to be done sparingly. The cost and benefit accrued by maintaining full records and checking procedures has to be balanced against their added value benefits. Nonessential activities have to be avoided (such as paperwork and manual systems where there are computer records). Procedures have to be

improved and rationalized continuously, to maintain the impetus of the quality improvement.

Customer support

Supplier quality is not only measured by the acceptability of the product, but also by the customer support provided. Rating of suppliers should include:

- delivery performance
- packaging and labelling
- information systems and communication
- support
- style
- flexibility.

Supplier quality issues are discussed fully in Chapter 8.

Advantages of JIT

Operational benefits

How can it be more efficient to deliver in small quantities, manufacture in small batches and increase the amount of administration? These are often the queries from those brought up in the traditional school of thinking. Equally, someone who only knows JIT would ask:

- Why do you buy things when you don't need them?
- How do you know what the demand is so far ahead?
- What use is a warehouse?

The accounts of a conventional company show a large investment in inventory. In the case of the two companies in Chapter 1, the profits were increased by a low stock philosophy. If M. Tight Ltd were to adopt a JIT philosophy this could bring the stockholding down even further to one or two days' worth, say £20 000 total stock value. This would release a further £480 000 cash for the business and increase the return on assets from 16.7 per cent to 20 per cent as a minimum. (As a move to JIT will eliminate the stores and some overheads, the real improvement will be greater than calculated here.) In this case each company had a large investment in fixed assets. For a distributor with rented warehousing the increases in profitability are very large, even if the average stock value turns out to be weeks' worth rather than days' worth.

The operational benefits arising from JIT are:

- inventory investment
- 'supply to order' instead of 'provision for stock' (see Chapter 7, section on managing lead times)
- easier forecasting so less slow moving stock
- better flexibility
- simplified administration
- waste elimination
- less scrap should there be a problem.

For each of these operational benefits there is a corresponding cost benefit which can be offset against any additional costs which arise. These additional costs usually occur because methods have not been changed to suit JIT. For instance, if an item is delivered in a batch once per month, it can be invoiced, delivery documentation produced and payment made. If through a change to JIT the item is delivered every day, then it would not be sensible to place a purchase order for each load, to raise delivery paperwork and to arrange separate payments for each load. Information is still required for control, but the information system has to be re-created to meet the new conditions. In the short term this may not be possible and, so, extra costs can arise. As JIT embodies the process of continuous improvement, the inefficiency will eventually be eliminated.

Efficiency benefits

Just in time supports the continuous output of a variety of different items, which in turn creates a variation of items going through the processes (picking, manufacture, packing, etc.) Flexibility is essential, but conventional thinking is that changeovers are to be avoided since they take up time and use resources. The JIT approach is to find the output rates as illustrated in Table 4.4 and then determine how to achieve the throughput and mix. Changeover time is therefore considered as a variable and this is a basic difference in the JIT approach. Normally the changeover time is associated with a particular product.

Example

Conventional logic is that if it takes ten minutes to change over for a particular item type, there is no point in making just one (which takes fifteen seconds); better to make at least eighty to be efficient. Accountants usually back up this type of simplistic logic with financial break-even analysis (see for example economic order quantity –EOQ – in Chapter 8). This results in large batches being produced and high stocks being held. The real situation is that changeover time is a variable depending on what the change is from and to.

Fast changeovers give the opportunity for lower stockholding. On a packing line, the change from one type of item can be fast if they are packed in the same type of box, but slow if the box size has to be changed. The secret of success is therefore to always run the right job sequence so that the number of short changeovers will be large and long changeovers small. This requires planning and load balancing which is difficult when forecasting for weeks ahead, but easier when looking ahead a few hours as required by JIT.

Cyclic sequencing reduces overheads.

Assuming that the items in Table 4.4 are all processed on the same process line, then one item needs to be completed every 11 seconds (5.6 per minute). If the process is carried out on a cyclic basis as shown in Table 4.5, then the cycle is repeated every 4.5 minutes. The number of each item in the cycle is shown, together with the output rate required to cope with the average demand. If the 4.5-minute cycle is to be carried out most efficiently, then the changeover times between the items need to be reduced. There are two ways in which the changeover time can be reduced, namely:

- Improve the planning of production sequences.
- Reduce the time per changeover.

Considering these two options, the improvement of product planning of sequences can make most changeovers minor. Within a short repetitive cycle it is much easier to sequence the production to optimize the setup time than it would be with conventional priority planning. The quantity of each item can be varied to suit the demand required if this varies, and items

Table 4.5 Cyclic production

Line	Throughput per minute	Number per cycle
Item A	0.9	4
Item B	1.3	6
Item C	1.8	8
Item D	1.1	5
Item E	0.4	2
Total	5.6	25

Times per complete cycle = 4.5 mins

with the same setup can be substituted should this be required by the programme.

Changing the production philosophy from weekly batch quantities (see Table 4.4) to cyclic JIT production (Table 4.5) will reduce stock levels by a factor of twenty, and also save overall cost. Just in time organizations often have small management teams with few management levels. This is a result of the JIT process itself. The planning operation is simple and the management is largely a repetitive activity which does not require a great deal of managing, especially as the direct process control is carried out by the direct operators.

From an accountancy viewpoint, if the cost of supply is split under the headings of materials, labour and overhead, the major impact of JIT can be in the lower overheads resulting from the simpler operations. The proportion of overheads applied to the direct costs are much lower in JIT operations.

Stock control using JIT

Lead time reduction

Inventory control has to work under the conditions imposed by the market. To get the stock levels right calculations are based on the real information which is available and not on idealized values. This situation can be improved because JIT gives potential to improve, particularly the potential to shorten supply lead times. In working with a dedicated supplier effective lead times can be changed. The supplier and customer can work together and share the risk.

Lead times are a result of queuing, distribution and sometimes process time (see Chapter 7.) For JIT suppliers, the demand is assured by the

agreement with the customers, but the demand timing and quantity varies. As the JIT demand is frequent, the effects of these variations cause only minor stockholding and delays – days' worth not weeks' worth. Small stocks enable the supplier to meet the small daily fluctuations of the customer very rapidly. Lead times can therefore be reduced.

To set up an effective JIT supply:

- the customer must *smooth the forecast*
- the supplier needs a regular demand for the item in order *to estimate the average demand rate*
- the supplier also needs a good estimate of the long-term demand for their planning in order to *understand the supplier's constraints*.

> *JIT results from making other things reliable.*

Suppliers can extend some aspects of their service and struggle to meet other aspects:

- Greater understanding improves the match between the needs of the two companies and *builds confidence*.
- Trust is the major ingredient for effective collaboration in order to *maximize supplier resources*.
- If suppliers know the volume of demand then they can react much more quickly to individual requirements. For instance a supplier of painted items could often paint and deliver within a few hours, as long as the demand for the item is expected and the choice of paint colours has been agreed previously. This enables them to offer *flexibility in demand changes*.
- As short-term demand is always variable, an agreement is needed to ensure the extent to which demand can be raised or reduced. This avoids failures in the JIT supply. It is therefore necessary to *communicate and monitor*.

Close liaison between the two parties involved is essential. It involves regular meetings for assessment and review between people at operational level.

Reducing the lead time is a continuous process starting with the A class items. These have the greatest cost benefit and are also amongst the simplest to change. They have most importance to the businesses and are among the easiest to forecast.

Pull systems

The essential difference between the JIT approach to material control and the stock planning approach is in the mechanism for triggering replenishment. Conventional inventory control predicts demand, works out how much stock is required to supply it and then procures that quantity. Just-in-time operations only acquire stock as a result of demand. This can be thought of as a 'pull' process, with the issue of an item triggering the demand for another – the number of triggers depends upon the rate at which demands are received. Stock-level systems purchase items to cover forecast demand (in larger quantities) and then try to supply the items to the customers (a 'push ' system). The advantage of the pull system is that only sufficient stock is held to meet the immediate demand for a few hours or days. The effective supply lead time is arranged to be short (JIT) and the deliveries frequent, therefore the stock can be very low.

In inventory management simplicity of operation is the key to success. Kanban is an operational method, consisting simply of a ticket, which achieves a pull system from within the company or from a supplier. When an item is demanded a request must be passed back to the source to provide another. This information can be a kanban card or a simple signalling system. The card will identify the item, the quantity required (ideally one, but possibly more) and where it is required. It will also state times so that performance can be monitored. This ticket or traveller has to be provided to the source very rapidly if the supply is to be prompt.

The ticket is held with the physical goods and is an easily maintained and simple method of informing the previous stage of supply. Using the principle of the ticket, there are other trigger mechanisms which can be used instead, such as coloured lights or sound. Fax or network communications can be used for external supply. The way information is transferred is immaterial so long as the process is instantaneous, completely reliable and accurate.

An obvious application of JIT is for manufacturers producing large quantities of similar products. They can set up a flow line and have in-line stocks of a single item or a small batch between each process stage. The trigger for manufacture is simply 'use one so make one' starting from the last process and feeding back gradually to the first. The restructuring of production into a flow line is a most effective way to manufacture. Where the product range is wider and the process routes differ, cellular production is more applicable, and the pull system is used between the cells and by the cell operators for processes within the cells.

For purchased items the supply is triggered in the same way but with larger batches. The delivery batch would be expected to meet the demands

for a few hours or a day. As the lead time is normally over an hour, a supplier kanban does not run out before signalling. It has a residual stock to cover the demand during the supply lead time.

Just in time has made a major impact on manufacturing and the reduction of lead times is changing the whole of inventory control. The changes will provide competitive advantage for those companies with the best ongoing developments of JIT.

Key points

- JIT supply is the aim of best practice in inventory control.
- JIT is a result of improving the quality of everything and organizing 'flow'.
- JIT business is fundamentally different from the traditional. The key concepts are:
  - pull – only buy when needed (kanban)
  - one – only make or buy small quantities
  - flow – balance movement throughout supply chain
  - simplicity – operating controls should be simple and visible
  - quality – high reliability for supply, products and control processes.
- Apply JIT to where it is most beneficial – A class.

5

Organization and management

- Set the direction
- Identify objectives and targets
- Manage stock value
- How to value and depreciate inventory

Where stock control fits into the organization

Nowadays the inventory controller has extra responsibilities. The job has grown out from storekeeping to become a management discipline of maintaining the inventory to meet company policy. Gradually this responsibility has broadened to include all stock whether inside or outside the stores. The place of inventory management in the organization has altered as responsibilities have changed. This is illustrated in Figure 5.1.

The original role of stock control was part of the physical stores control shown in the structure in Figure 5.1A and its position in an organization was lowly, working for either purchasing, finance or, in manufacturing

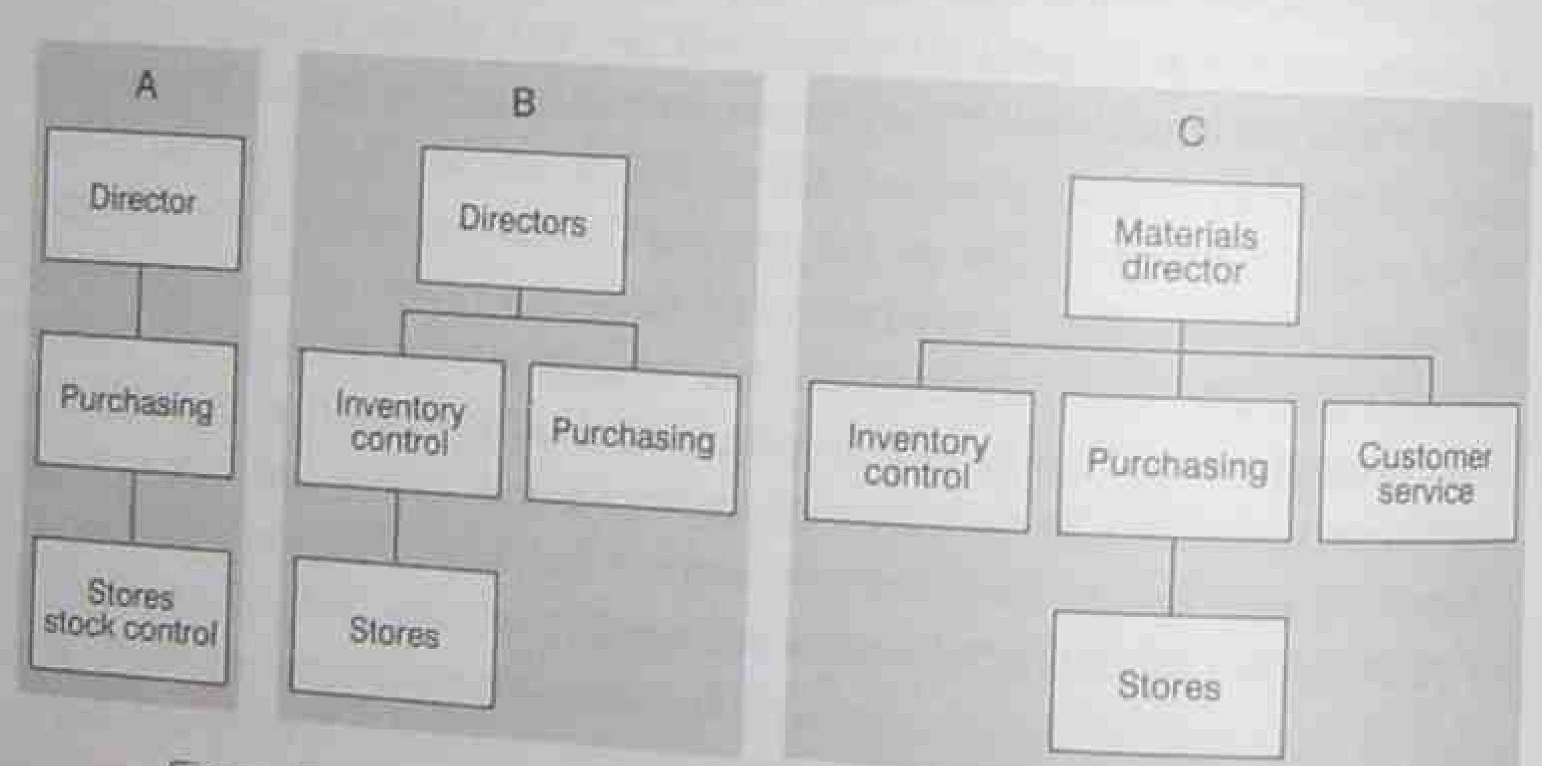


Figure 5.1 Changing inventory management structures

companies, for production. In this structure stock acquisition was initiated by buying and sales pressure, and there was no clear responsibility for, or focus on, inventory control. The tightening of company objectives changed the approach and separated stores and stock control activities. Inventory management is now accepted as an operational activity and works alongside purchasing, as illustrated in Figure 5.1B.

> *Inventory management is a senior management activity.*

This arrangement has gradually evolved into structure C as pressure increases for better performance. The central role of inventory management has been recognized in many businesses, both in distribution and manufacturing, and material management or logistics directors have been appointed to co-ordinate various aspects of inventory management into one supply activity. Stock control encompasses:

- purchasing – especially reordering (but see Chapter 8)
- forecasting demand
- planning inventory against company targets
- stock allocation and delivery promising
- monitoring and controlling service and inventory
- production planning and control (for manufacturing).

These are organizational activities. The physical management of inventory, including stores, dispatch and movement then becomes treated as a separate problem to be managed by a parallel organization. A similar relationship can arise in manufacturing between the production control department, who carry out the planning, and production supervision itself, which seeks to work to the plans provided by the production control department.

Where there is a separation of the stock and stock control activities, the warehouse become responsible not only for managing the movement and holding of stock, but also for providing information to the rest of the company. It is particularly important for inventory management that the information provided is accurate. Checking stock physically is very time-consuming, and the recording systems should be sufficiently accurate to make this unnecessary apart from stock auditing.

> *Stores control takes care of inventory.*
> *Inventory control sets the amount to keep.*

Stock control in a professional organization can be carried out at a different location from the stores. Thus stock controllers can manage site stocks or multi-location storage as long as there are real-time communications with the stores. If these activities are split, warehousing can be located at the most economic location for transport and customers and inventory control can be located independently at the most convenient place for management.

Inventory management is in the process of evolution. Targets are being expanded to include extra aspects of operations, cost and service and there is much potential for development (see Chapter 13).

Responsibilities and targets

Rationalizing the problem

Inventory controllers have to take a simplified view of the stock items and treat the majority of them in the same way. In driving a car there is a distinction between the driver (who may not be able to repair the car) and the mechanic (who may not be able to drive). In stock control, most people are the drivers but the stock control vehicle is not well designed, mechanics are few and, so, the driver is often forced to make improvements. The driver who understands the mechanics can always get better performance out of a vehicle.

> *Inventory management has the same principles and techniques – only the situation is different.*

Every item has its own characteristics, often with a unique story – 'do you remember when the supplier's warehouse burned down?' 'The customer was frozen up for five weeks?' 'It was put on a boat to Antarctica by mistake?' In the past individuals managed stock through their memory and understanding but this is not enough to meet the tighter targets of modern inventory management even though such understanding is an important additional tool to use for A class items (and a waste of time for C class).

The stock controller should largely ignore the details of the item and consider the characteristics of the demand pattern and the supply, despite the widely different types, sizes, shapes, usage, origins and characteristics of the lines, but whatever the item is, inventory control remains the same. An arguable point? What about liquids, explosives, perishable foodstuffs, pine trees? They are all very different. True but, despite the differences, their supply and demand characteristics are not that different: they depend on market conditions and not on the product. Just as financial controls are applicable across different industries, so is inventory management. The characteristics of the stock determine the mix of inventory techniques to use, but the components are the same:

- demand forecasting
- determining safety stocks
- negotiating supply patterns
- avoiding slow-moving stock.

Inventory techniques are common for all businesses, and their skilful mixing for aggregate and individual item inventory control allows many different types of stock to be managed. Simplicity is the key, but this is not the simplicity of systems which give superficial answers. It is the simplicity of operating processes resulting from rationalization of procedure, systems and approach. It arises from, say, using the same system for placing all orders and controlling all the stock; in having the same item identification throughout the business; in treating priorities through an organized system. The elements of the simple procedures have to be supported by complex tools such as software packages for forecasting, modelling and communicating, but these need not concern the daily operation as long as they work correctly. The actual logic of such tools has to be intricate to cope with the complicated nature of inventory planning.

Simplification should be considered as an ongoing challenge for development of good stock-keeping. The following areas can usually be improved:

- single stores philosophy
- inventory system applied to everything
- single operating procedure
- paperless information systems
- purchasing as part of inventory management
- real-time systems.

By rationalizing these activities, the professional stock controller can free time to work on the most beneficial topics, and improve the level of control.

External objectives

Discussion earlier in this book was about setting inventory targets for availability, inventory value and operating costs. The balance between these targets is a matter of inventory policy. The management of inventory relies upon routine and regular monitoring of these values. The objectives set for inventory management should be chosen to meet the immediate demands of the business, and specifically:

- customer service
  - measured as stock availability or delivery on time (percentage against own and customers' targets)
  - analysis of age distribution of lateness
  - assessment of customer satisfaction – survey results
- inventory investment
  - value
  - stock cover ratio
- operating costs
  - warehousing and inventory operations costs
  - cost per transaction, movement and purchase.

> Key objectives – 100 per cent availability, zero inventory, no operating costs.

The process of setting targets usually begins by measuring current success. Current levels of achievement are assessed to see whether they are good or bad. (If inventory control staff consider that they have the correct stock level, this is often because of their complacency and not because of adequate achievement assessment.) Targets can then be set through negotiation with those who are affected by them and with those who have to achieve them. The three target dimensions are:

- quantity – how much the monitored values will change
- date – the stages by which the quantities will be changed
- responsibility – who will be accountable for achieving the results?

The prime aim is to meet all the inventory targets simultaneously, and not to sacrifice one to meet another. Measurement methods should meet the prime business needs (in Chapter 2 alternative ways of measuring customer service were discussed) and should be chosen to give the most satisfactory results.

> Identify all your real objectives.

Routine monitoring of these items is required on a weekly or monthly basis. The results, both operational and financial, should be available within three

days of the period end, otherwise the information cannot be used for control (although it is commonly used for apportioning blame).

Managers should also be aware that targets are set as part of company policy and the art of correct management is to set targets which are both achievable and challenging. They should reflect an expectation of performance, not an achievement under ideal conditions. This can be difficult when trying to co-ordinate planning with other departments. A good method of improving the accuracy of planning is to give two targets, one optimistic and one pessimistic, and to ensure that performance lies between the two. By adjusting the two limits, a target range can be identified which is both attainable consistently, and which is acceptable for planning and co-ordination with other activities.

Internal targets

External targets have to be supported by more detailed objectives within an inventory management department and these should include personal objectives for individuals which all contribute to departmental targets.

Customer service can be structured by setting availability targets for different types of items. Some service is likely to be ex-stock with different availability levels, while other service will be supplied on a range of lead times. Information required to give fast direct performance measures include:

- number of lines out of stock
- number of customers and orders affected by shortages or late deliveries
- analysis of shortage according to cause
- monitoring of customer complaints
- details of customer-care initiatives.

> Set quantitative targets – or you won't know when you've achieved them.

Stock levels will differ for these product ranges and have different expectations in terms of stockturns. Items with a JIT-type of operation have very good stockturns while slow-moving items such as service parts can inevitably achieve only relatively poor results. In addition to the inventory value, the investment can be controlled by managing purchase commitment and write-off value as discussed below.

We need to aim for several targets at the same time.

Operating costs are increasingly becoming the responsibility of individual inventory controllers. The elements of cost are controlled through cost analyses and budgets, and monitoring operational activity levels often provides the basis for cost reduction. These controls include:

- value of transactions or number of transactions per day (e.g. customer orders processed, stock reconciliations done, adjustments to stock levels, etc.)
- processing delays or queue lengths (e.g. allocated customer orders)
- cost per transaction by type
- performance against budget for each activity.

Other targets are required to complete the management of inventory activities. Prime external targets do not necessarily cover all the requirements for good inventory management, and there are additional aspects which are important as well: requirements for individuals to manage change and control projects; other business objectives to be met, including quality management, credibility, legal, motivational, health and safety, communication and process development. The real requirements for inventory control are numerous and, where possible, targets are needed to evaluate and monitor these objectives. Inventory management must also meet general business objectives. Targets need to be set in the areas of:

- personnel development planning
- improvements in systems and procedures
- quality development for products, service and administration
- management of change processes
- supplier audit
- environmental issues
- presentation, style and image of the department.

For well-managed inventory control, all the requirements of the market, the operating environment and the supply chain have to be considered. Greater understanding of the real objectives will improve the effectiveness of the business as a whole and the worth of inventory management within the business.

To reduce stock, keep value of receipts less than issues each week.

Purchase commitment

One way of maintaining control over inventory value is to balance the values of input and output. If the value of input is kept to below the value of output, then stock reduces (as discussed in Chapter 3, in the section on practical methods of reducing stockholding). The value of supplier deliveries should be less than the forecast value of customer dispatches in each time period. Supplier deliveries are a result of purchase order and schedule commitment, so it is the ordering process which has to be kept under budgetary control. The control process relies on collecting information on:

- the forecast value of demand for each time period (week or month)
- the amount of schedule or order commitment already made for delivery in each time period
- the typical value of last minute (emergency) orders which are committed each time period.

A policy on stock reduction, or increase, has also to be defined in terms of what value change in inventory level is considered practical or desirable each time period. From this information it is easy to calculate:

1. The maximum delivery value for each time period (= Forecast customer demand – Value of stock reduction target per period).
2. The maximum value of order commitment for delivery per period (= Maximum delivery book value – Typical value of emergency orders).
3. The maximum value available for new orders placed for delivery during that period (= Maximum value of order book commitment – Amount of order book commitment already made).

This maximum value available for new orders is the ceiling for total order value for delivery in the time period. If this value is exceeded by placing a further supply order, then that order should be rescheduled into a later time

Table 5.1 Order book values

Period	Supplier deliveries already ordered	Value of order book available to place
Next month	185 000	5 000
1 month ahead	130 000	60 000
2 months ahead	60 000	130 000

period (or another order delayed) or else the resulting stock value will increase.

Example

The current order book value with suppliers for delivery over the next three months is as shown in Table 5.1, and the maximum order book value for delivery in each time period is calculated in the final column. The task for the inventory controller in M. Tight Ltd is to ensure that further purchases are placed so that the planned deliveries next month are not more than £5000 and then £60 000 for one month ahead, and £130 000 for the month after that. By monitoring delivery book value in real time, an inventory manager can maintain the stock within the required budget. The control of stock reduction using purchase commitment is illustrated in Table 5.2. (See also Tables, 5.3, 5.4 and 5.5.)

Table 5.2 Inventory management targets – stock value

Location	This week/month			Last week/month			Target	
	Inventory value	Slow movers value	Weeks on hand	Inventory value	Slow movers value	Weeks on hand	Inventory value	Weeks on hand
Central								
NW store								
SW store								
NE store								
SE store								

Table 5.3 Inventory management targets – availability

Location	This week/month	Last week/month	Target	
	Availability (%)	Availability (%)	Availability (%)	Date
Central				
NW store				
SW store				
NE store				
SE store				

Table 5.4 Inventory management targets – supply

Purchasing activity	This week/month	Last week/month	Target	
	Value	Value	Value	% success
Orders placed				
Goods received				
Late deliveries				
Emergency orders				
No. of orders	Qty	Qty		

Table 5.5 Inventory management targets – outstanding demand

Location	This week/month		Last week/month		Target	
	Overdue	More than x days	Overdue	More than x days	Overdue	More than x days
Central						
NW store						
SW store						
NE store						
SE store						

Inventory valuation

Unit cost of stock lines

Accounting procedures for stock are usually organized by the accounting functions of a company . The costing options which are commonly used are explained below. It is important to have a consistent costing system which values all items. There are several ways of updating inventory cost when new items are purchased by changing the unit cost. The preferred costing method is average value, although standard costing provides better budgetary control. The costing method should not influence the actual items being picked, in almost all circumstances picking should be organized as a first in first out (FIFO) process.

Don't associate costing methods with individual stock items.

First in first out

Stock is valued at its purchase value. The oldest stock is assumed to be used first (as is required by good inventory practice) and the stock value is therefore the total of the most recent purchases. First in first out is best used as an accounting procedure, but not for identifying which items in stores to pick. Stores stock rotation should be arranged through the warehousing control system.

Last in first out (LIFO)

The issues are valued at the most recent purchase price, leaving the remaining inventory valued at a previous (generally lower) value. The effect is to minimize the profit on the stock being sold and to minimize the value of the remaining inventory. This method of valuation can be used to reduce the profit reported by a company and decrease the valuation of the stock. It is, however, only a valuation technique and not appropriate for organizing stock movements. For example, there is a stock of ten of an item valued at £80 each, and a new batch of five has been bought at £120. The stock value is therefore £1400. If a customer buys two, then the stock will be reduced to thirteen. By LIFO these two issues will be valued at £120 each, and the new stock value for the thirteen items will be £1160. If the items sell at £150 each, the recorded gross profit is £60. (Using FIFO, the remaining stock of thirteen would be valued at £1240 and the gross profit would be £140.)

Replacement value

Stock is valued at the current price for buying replenishments. This enables a simple rule to be applied to sales price, (such as 'sell at purchase price plus 50 per cent'). It increases the value of stockholding but requires market information to manage.

Standard cost

Fixed item value is calculated financially and held the same for a long time, usually a year, giving a good stable valuation. Differences from the standard cost are considered as good or bad variances. Standard costs are used widely in manufacturing, as they can account for typical costs and overheads.

Average value

A running average is the safest stock valuation method. The cost of new deliveries is added to the total stock value and the total value spread over

the new total stock. For example, there is a stock of twenty of an item with a unit cost of £10. Ten more are purchased at a cost of £13 each. The total value is now 20 × £10 + 10 × £13 = £330. The stock is now thirty, so the average cost is £330/30 = £11. This is the new average cost and any demand will reduce the stock value by £11 per item.

Stock-keeping unit (SKU)

Stock-keeping units are especially useful in distributed warehousing. These SKUs acknowledge the fact that the stock value can depend on where the item is held. If a customer requires an item which can be supplied either from the main stores or from a satellite store and the cost of distribution is £15 from either source, where should the supply be made from? Since the stock in the satellite store has already been transported once (probably at a cost of £15), then it would be cheaper to supply from the main store since the satellite stock has already cost more. Stock-keeping units are stock at a specified location taking into account the transport and purchase costs. It is normal to talk about a 3000 SKU warehouse, which is equivalent to saying that there are 3000 lines in that remote location. The actual stock value element of cost in an SKU could be calculated by any of the preceding methods.

> SKUs take into account logistic costs.

These alternative valuation methods are illustrated by an example in Table 5.6. The 'cost of the next item to be issued' is the unit value of the item remaining in stock which will be issued next to fulfil customer demand. Each of the valuations give slightly different unit costs. For LIFO and FIFO methods, allocation of the costs depends on working out from which supplier delivery the next sales item should be taken.

In Table 5.6 the standard cost set by the accountants is £24, but purchase costs are higher. There will be a purchase 'variance' when the outstanding orders are received:

Order 1	10 × £2	= £20
Order 2	10 × £4	= £40
Order 3	10 × £0	= £0

At a demand of ten the total (cumulative) variance = £60.

Table 5.6 Alternative methods of determining unit stock value

Date (week)	Transaction: Type		Quantity	Unit cost	Stock balance (quantity)	Cost of next item to be issued: FIFO	LIFO	Standard cost	Average cost	Replacement value
1	Stock check[1]		10	20	10	20	20	24	20.0	20
		Customer demand	3		7	20	20	24	20.0	20
3	*Supply order 1*		10	26						
		Customer demand	2		5	20	20	24	20.0	26
4	*Supplier delivery*		10	26	15	20	26	24	24.0	26
5		Customer demand	4		11	20	26	24	24.0	26
		Customer demand	1		10	20	26	24	24.0	26
6	*Supply order 2*		10	28						
7		Customer demand	6		4	26	26	24	24.0	28
		Customer demand	3		1	26	26	24	24.0	28
9	*Supplier delivery*		10	28	11	26	28	24	27.6	28
		Customer demand	4		7	26	28	24	27.6	28
10	*Supply order 3*		10	24						28
		Customer demand	2		5	28	28	24	27.6	24

Note: [1]Existing stock of ten is valued at £20 each.

Aggregate stock valuation

Stores stock can be valued either at full value given by costing at one of the above costing methods, or by a written down value to acknowledge the fact that some stock is no longer usable or saleable. There must be an amount of money deducted from the stock value to account for this. There are two ways of achieving this: depreciation and provision.

Depreciation

The first is to depreciate the value of stock of each item. The longer an item is in stock the less it must be saleable. The rate at which depreciation is allocated will depend on the type of stock. Fashion goods can lose value very fast. Aircraft spare parts depreciate very slowly. The simplest and most common way of calculating depreciation is to deduct a portion of the item value each year it remains in stock. Taking 25 per cent off the original value each year would mean that items have no value after four years, and would call forth acknowledgement by the accountants that inventory control have made the expected mistakes and have squandered company profit on useless stock. It is not a healthy situation when inventory managers are forced to minimize the value of stockholding because of poor control.

> Depreciation steals profit to cover for mistakes.

Writing down the value of non-moving stock is unfortunate but essential The accounting code of practice requires stock to be valued at 'the lesser of its purchase cost and its realizable value'. This means that items bought at a discount have to be valued at a discount, and items which have no likely use have to be valued at zero value.

Provision

A second way of reducing the value of stock is to allow a provision for the total value likely to be lost in a year. This provision can be based on a detailed depreciation calculation or a simple assessment. Whereas depreciation is applied to individual items, provisions are applied at a global level. This enables the provisions 'reserve' to be used for whatever accidents happen to cause stock obsolescence (sudden changes in legislation, contract loss, etc.).

Provisions are financial judgements. They can therefore be used in situations where there is an impending change which will affect inventory, but which is not predictable from historical information.

Professional stock controllers do not use write down as a method of reducing stock value because this only transfers funds from company profits to pay for the write down. (The proper methods for reducing inventory value were discussed in Chapter 3 in the section on how to reduce stocks successfully.) For stock control purposes it is therefore good inventory management practice to value stock at its full value and to apply the depreciation reserve or provisions later for the financial accounts. This ensures that the slow-moving, high-value stock receive some priority in the control activities.

Skills and systems

Systems should not only be designed to perform a task, but also to enable the user to achieve the objectives most conveniently. The term 'user-friendly' is often applied but seldom true, and the usability of a system should be geared to suit the type of person expected to use it, and the knowledge, skill and time they are likely to devote to it. Where a complex system is in use, all operatives must be fully trained to understand it, otherwise they will not use it efficiently.

The more professional the user, the more complex the computer system can be and the better the results achieved by its operatives. This is true from the user options point of view. A routine operator may only require an uncluttered transaction screen with the system set up for fast input or enquiry, showing the relevant information and nothing else. However, if information is to be accessed by less frequent users, they will still need to have a thorough understanding of what the system can be made to do. If they do not understand it, they will not be able to use it effectively. Further, although systems should be straightforward and user-friendly, their underlying logic should be complex enough to cope with the many variables which can occur. For instance, a user with a seasonal demand pattern will have to disregard a system only providing a simple minimum stock level to help in reordering. What they require is a seasonal forecasting system using percentage weighting, base series or Fourier analysis. Just as it is not necessary to understand the circuit for electronic ignition if you want to drive a car, so it is not necessary for the user to understand underlying computer logic as long as the method can be applied, the answer is good and the forecast type is easy to switch on and off.

Two distinct features for choosing a successful system are:

- apparent simplicity geared to user skill
- system sophistication geared to the importance of the application.

Successful implementation develops the right balance between systems, procedures and the user's knowledge of a particular application. As an example, take a computerized parts list. A customer requires an item but does not know the part number. The old hands in stock control with many years experience will just quote the item number and look at the stock record, because they know it. A new starter requires support in finding the number. How can this be achieved? There are basically three methods, namely:

- training
- support procedures
- system features.

These different approaches are illustrated in Figure 5.2.

Training

If the user has an understanding of the application and memory of the codes, the system will work effectively. Training is essential for providing the understanding and knowledge. Inventory personnel need specific job

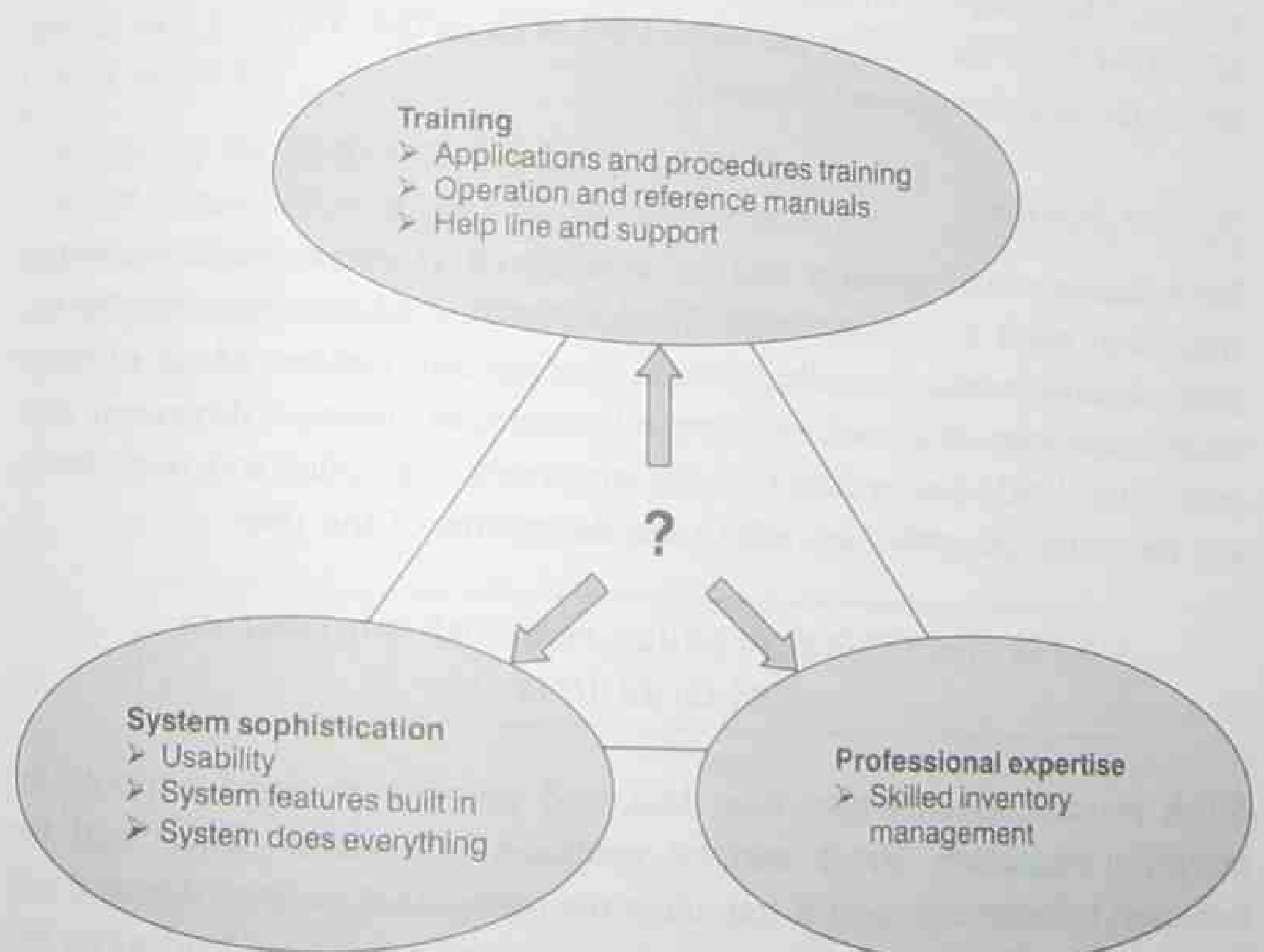


Figure 5.2 Inventory control options

skills which enable them to gain the full benefits of their systems and contribute to operating effectiveness. Investment in training is minuscule compared with the direct inventory savings.

Support procedures

Where the user of the system is unlikely to have knowledge of the application (e.g. technical knowledge by sales clerks), the company can provide the necessary information through supporting procedures. These can be paperwork or other aids which enable the user to follow a logical sequence with which to provide or gain the necessary information.

An example of this could be the person who often receives enquiries from customers needing advice on which stock item to buy. This person has either to gain sufficient technical knowledge of the stock to give advice or pass the item on to an expert, or to have the information provided for them in an appropriate form. A simple method of supporting this person is to provide a 'decision tree' for routine queries consisting of a set of questions. The answer to one question directs the user to the next question until the query is resolved. A person without detailed knowledge of a subject can still be effective using such support procedures.

Development of these procedures has to be overseen by a highly skilled person. Thus there is an initial cost, but the procedure can then run and advice can be given without the same level of expertise. This is a cost saving where the situation occurs regularly.

System features

Stock management systems can be considered as either stock-recording systems or stock control systems. Stock-recording systems concentrate on physical movements, recording issues, receipts and orders. Many of these systems also value the stock and keep a transaction history. 'Maximum' and 'minimum' levels can be input by the inventory controller and these levels and the order quantities are left to the discretion of the user.

> *Make the system produce relevant information, not large lists.*

Stock control systems take this data and provide decision support for inventory managers. Stock can be analysed into ABC classes and the historical information used to calculate the (weighted) average demand and the variability. Once policy on customer service and supply lead times has been decided, the system calculates the safety stocks and order quantities.

The outputs of a stock control system are:

- purchase schedules and orders
- expediting notes for suppliers
- management control information
- exceptions – unfulfilled customer orders
  - forecast predicting the wrong demand level
  - wrong supplier lead times
  - excess and obsolete stock analyses
  - options for special investigation identified by the user.

A successful system will produce information which can be used without amendment by the inventory controller. Some data may need modifying because the full information was not provided to the system, but this should be rare if the system is operated properly. In practice, most systems fail to provide this full specification resulting in time being wasted on routine analysis and interpretation. This time could be much better employed on customer and supplier liaison, and on ensuring that the system parameters are accurate. It is therefore important to use a system with stock control features, and to use these features to manage the majority of inventory.

Key points

- Identify the responsibility for stock management.
- Set quantitative targets (cover all aspects).
- Value all inventory.
- Design systems and information for the user.

6

Safety stocks

- Analysis of historical data
- Demand distribution
- Safety stock calculations – MAD and SD
- Customer service factors to set availability levels
- Determining obsolete and excess stocks

Learning from history

Collecting historical information

Stock is held either because it is convenient to buy in bulk, or because the item is required faster than the supply can provide it. In the latter case there is some uncertainty as to the quantity required, and so some safety stock is needed. In Chapter 2 the balance between the conflicting requirements of good service, low inventory costs and small operating costs were discussed. Now the means of achieving the controls can be examined.

The amount of safety stock held in an organization depends upon three main factors:

- the variability of demand
- the reliability of supply
- the dependability of transport.

The general approach to this situation is to set stock levels to cover the normal variability of demand and to adjust the other two variables so that they are relatively insignificant. Quality initiatives have improved the supply situation but it is usually found that the major uncertainty is caused by customers and their unpredictable requirements.

Store records show the movement of stock in and out of the warehouse and such historical information is essential for evaluating what level of stock to hold (unless the customer provides firm orders). However, the best guide

to the appropriate stockholding is the amount of demand, rather than the number of issues.

Note: When using demand history in forecasting, it is conventional to neglect the demand information from the current, incomplete period unless the demand is extremely large.

Individual stock movements logged on the stock records have to be analysed into movements per time period, normally weekly, so that weekly usage statistics are produced for each stock item. In cases where the demand is small or very variable, then monthly periods (time buckets) can be used. For fast-moving and rapidly varying demand (such as chart music), daily time buckets are required.

Collect data into time buckets.

The first step in analysing demand history is to collect the total demand for each period, as shown in Table 6.1. Sometimes such a table is used to estimate the stock levels and order requirements, whereas a little more understanding would enable accurate stock levels to be established.

Table 6.1 Demand history

Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Demand	23	35	28	19	34	25	41	15	39	33	28	48	31	38	28

The demand pattern can be displayed on a frequency distribution graph or histogram which shows how many periods the demand was at a certain rate. Figure 6.1 shows that the demand was between twenty-five and thirty at three times in Table 6.1 and it was between forty and forty-five once. This histogram gives a good idea of the usual demand rate and the spread of rates. In Figure 6.1 rate intervals of five per period were chosen (11–15, 16–20, 21–25, etc.) and the figures on the horizontal axis represent the maximum figures in the range). As can be seen the most common demand rate is 31–35.

If the demand comes from a similar type of customer, then the variations in orders will generally follow the same characteristics, although they are unpredictable. These demands form part of a population. If this assumption is correct, then the properties of this whole population can be used to predict the demand pattern. The most common demand pattern is the

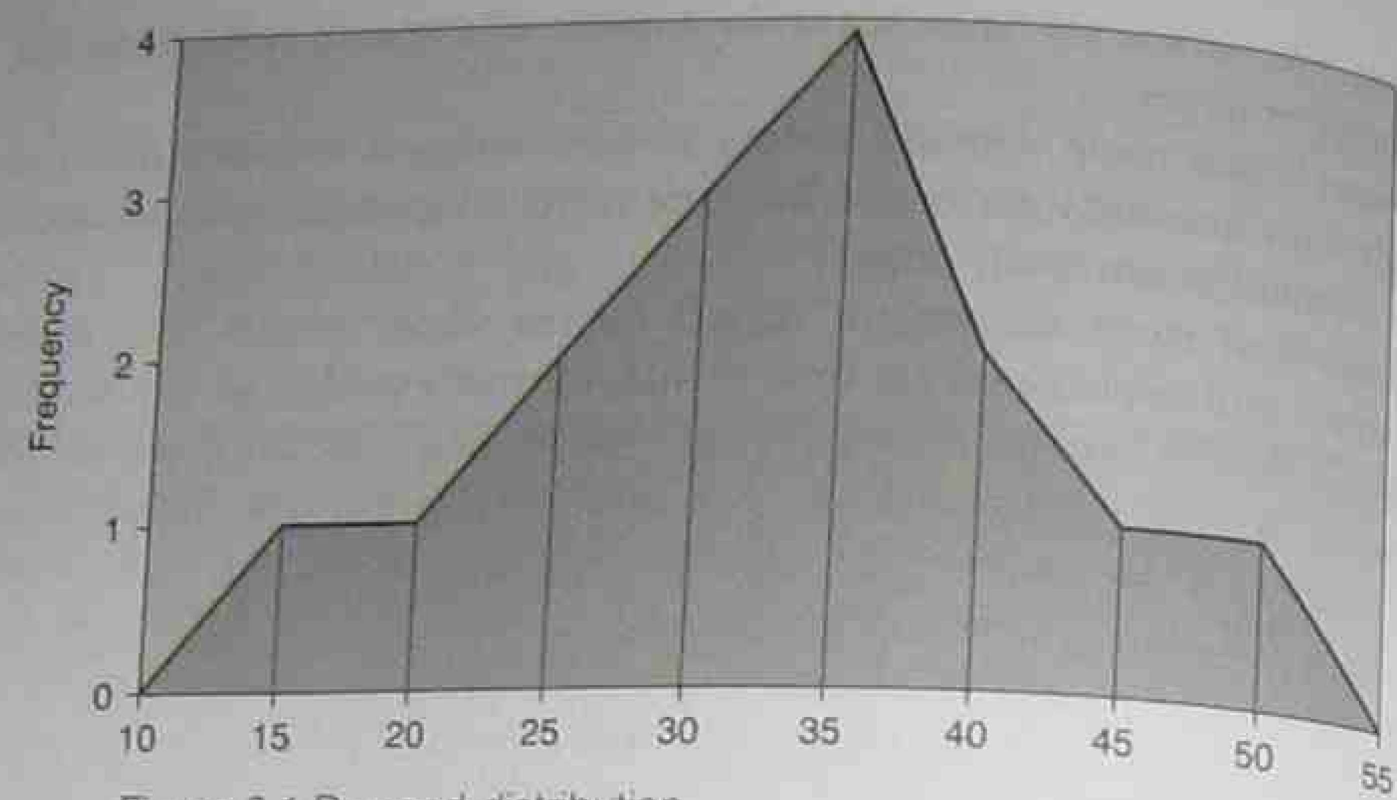


Figure 6.1 Demand distribution

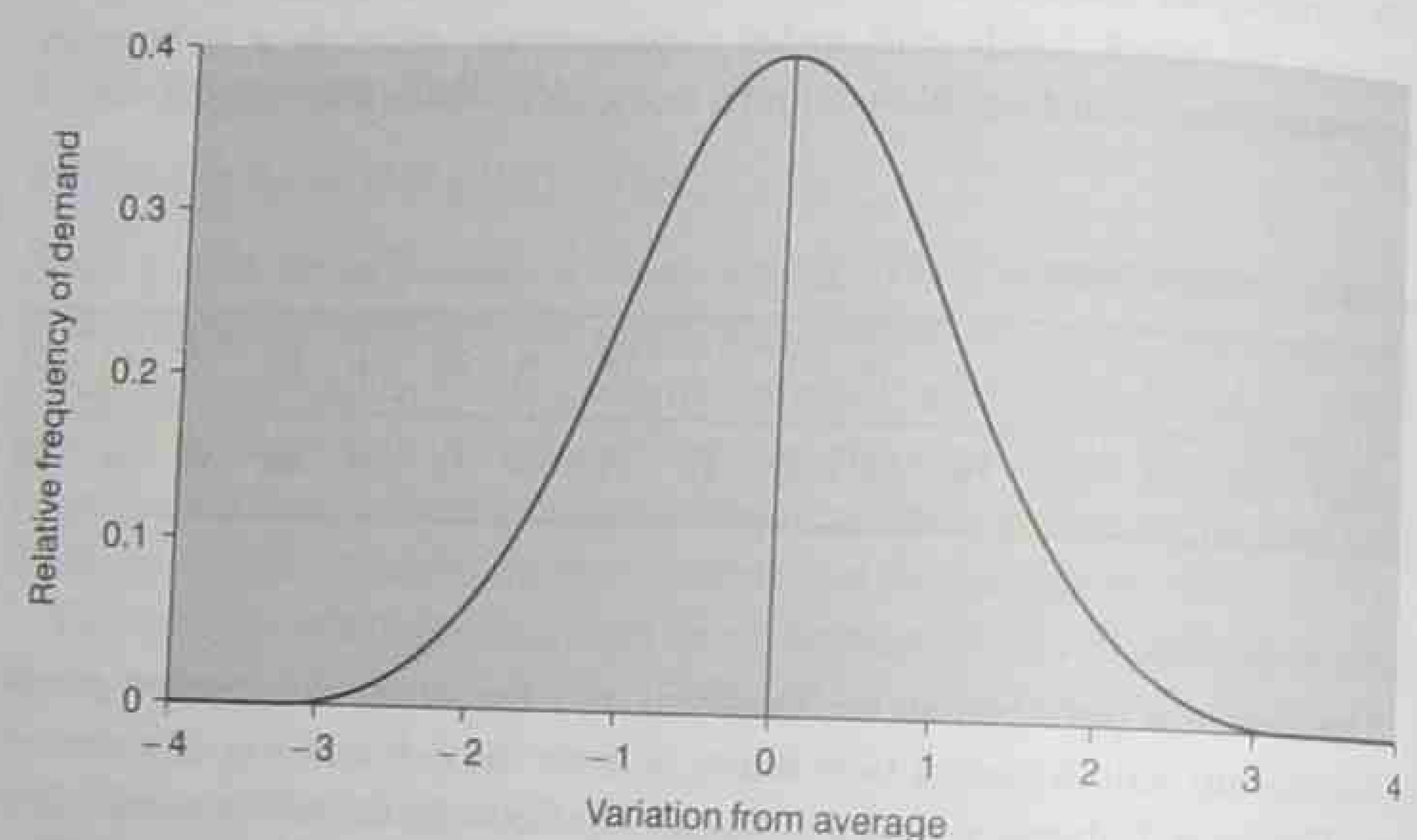


Figure 6.2 Normal distribution

Gaussian or 'normal' distribution shown in Figure 6.2. This is an idealized form of Figure 6.1 for a very large number of periods of demand. Using the small amount of information given in a table such as Table 6.1, the shape of the curve in Figure 6.2 can be calculated quite accurately.

Demand patterns can be assumed to be 'normal'. (Even if they are not, taking them to be normal is much better than simpler methods of control.) As can be seen from the distribution graphs, most of the time the demand is near to the average. Sometimes, but less often, it is much more or much

less. The vertical scale of the graph is frequency – how often that value occurs. The horizontal scale is really measured from the middle (the mean) and shows the values at which the populations occur. In practice, careful batching of the data is needed to produce a distribution like the one shown. In general, there is just a small amount of data to work with and this is used to fit a normal distribution.

A key property of the normal distribution is that it can be described using only two parameters, the average and the average width. Calculation of the average value is relatively simple. Measuring the width is more difficult.

> *Analyse the demand into true average and variability.*

It is no use taking the highest and lowest figures, as these are usually the result of atypical situations (stockouts, one-off demands, holidays, etc.) It is better to take values which are, say, a quarter of the way in from the extremes of the sample (quartiles) or apply a more rigorous approach which uses all the available data. In general the measure of width used is standard deviation (SD), although in stock control a simpler approach is used, namely mean absolute deviation, MAD (see section on standard deviation below).

Of course more complicated demand patterns do arise. Occasionally there is more than one Normal distribution pattern mixed in the demand for an item. For example, a product could be sold in a local market in single units, whereas for export it is boxed in fifties. The demand pattern will then consist of two normal distributions, one with small fluctuations in the demand pattern, and the other with a more sporadic pattern.

Reasons for safety stock

The normal amount of stockholding can be determined by statistical methods which rely on history to predict the future, and assume no change in circumstances during periods ahead. Safety stock is primarily to cover random variations in demand, but it can also cover many other situations such as:

- supply failure
- production shortfall
- transport failure
- slow, unreliable or inaccurate information
- and any other source of disruption of service.

Safety stock is the buffer between supply and demand. It decouples customer service from manufacturing, and enables each to operate independently and more effectively.

Normal demand patterns

Stock availability

There is the need first to identify the major demand and supply variables discussed in this chapter and to quantify their variability. The need to cover for changeability of demand is a central task of inventory control. Unexpected demand requires a level of safety stock to cover for it. In fact using the normal distribution, the unexpected demand can be anticipated, or at least taken into account.

The more safety stock the better the customer service. Going back to the sample data in Table 6.1, a stock of 100 would probably be far too many and a stock of thirty-five too few. With a stock of forty-five it looks as though there will only be a shortage once in fifteen periods. But a surer way of finding this out would be to use all fifteen sets of data to give an accurate normal distribution, and then to find the risk of running out of stock.

Consider a supplier who arrives at the start of each period and tops up the stock to the required amount. One of the items supplied has the usage history shown in Table 6.1 and a unit cost of £1000. If the average usage amount, thirty-one, is held in stock at the start of each time period, then the expectation would be that the stock would run out during half the time periods, since the demand exceeds the average half the time. (This happens seven out of the fifteen times in the sample data.) If there were no stock at all then the stock would run out every time, and of course, if there were an infinite number in stock, then there would be 100 per cent availability. As the stock increases, the availability increases, and this relationship is shown in Figure 6.3. As safety stock is added above the average demand, the availability increases rapidly, but once the availability becomes high, then it requires a lot of stock to improve the situation.

Standard deviation

There are several ways of finding the typical variability, but there are two measurements which are generally used, namely standard deviation and mean absolute deviation. These are different ways of measuring the same thing, just as wealth can be measured in dollars or pounds, but the total value is the same. There are relative merits for both measurements.

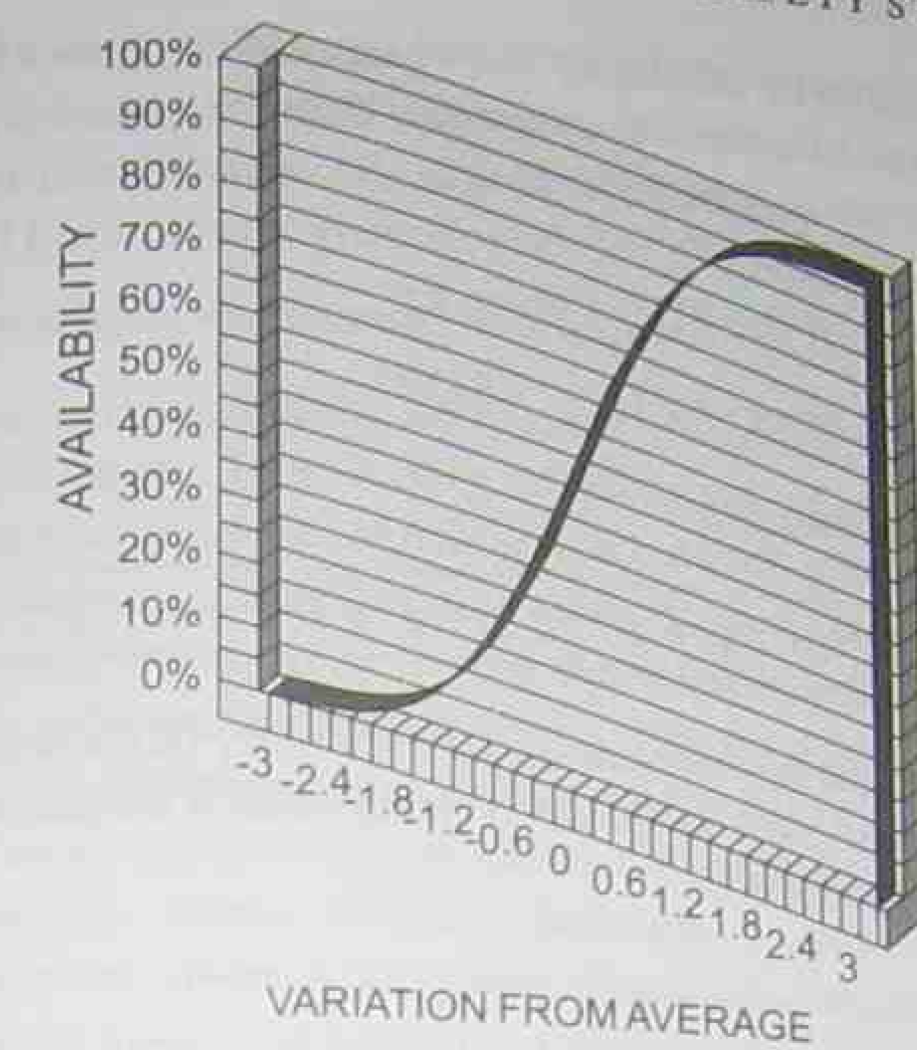


Figure 6.3 Stock and availability

Standard deviation is theoretically the correct measurement, but people find it more difficult to understand and work with. Standard deviations should be built into computer systems. Mean absolute deviations are much more user-friendly. They are easy to understand and, arguably, easier to bias. (The reason for wanting to bias them will be discussed in Chapter 10.) Mean absolute deviation is a good enough measure in practice. The standard deviation will be described first.

> Use SDs if they are in the software
> or if you like mathematics.

The basic premise for safety stock is that the differences between the forecast and the actual demand are random. These random variations in demand usually follow a 'normal' distribution (see Figure 6.2). This is symmetrical about the average value. The size of the variations differ from one stock item to another and, consequently, the width of the normal distribution differs also. This width, measured by the SD, is calculated from the sum of the squares of x, where x is the error, the difference between the actual value and the average (or the variations of the individual demand figures from the average). Mathematically this is:

$$SD = \sqrt{[(1/N) \times \Sigma xi^2]}$$

The SD is calculated by adding up the squares of the errors from a sample of N values; in our case periods of history. Using the history given in Table 6.1, this gives an SD value of 8.4. If there is only a little history, then N is low and the value of the SD will be relatively inaccurate. This inaccuracy (the standard error of the SD) is given by

A standard error of SD = $\sqrt{\text{SD}/\text{N}}$.

Mean absolute deviation

Use MADs for ease or to weight for more accuracy.

The MAD is a simple assessment of the variability of the demand pattern. Like SD, MAD is the basis from which safety stock is calculated. It is easiest to explain using an example.

Example

For a demand during successive periods of 5, 14, 6, 15 the average demand is 10.

The swings about that average are –5, +4, –4, +5.

Adding up these values gives 0, which shows that 10 is the true average. For measuring the variability, it is the size of the errors not the signs which matter, so forget the signs and add up the numbers. This gives 18 in total, or an average of 4.5 for the four periods, and is called adding the 'absolute deviations' or 'moduli'.

A second demand pattern has the form 9, 12, 8, 11 for the four periods of history, and an average of 10 as in the first example.

By the same analysis, the absolute deviations are now 1, 2, 2, 1 and the MAD 1.5 (6/4). Obviously the safety stock required in this case will be considerably smaller than in the previous example, in the ratio 4.5 to 1.5.

The MAD is calculated by this method, or by taking the last value of MAD and updating it with the latest absolute deviation. The formula for MAD is:

$$\text{MAD} = \frac{\text{Sum of absolute deviations from mean}}{\text{Number of periods included in the sum}}$$

For a normal distribution, MAD = 0.8 SD.

The MAD is easily calculated but to avoid the nuisance of recalculating each time, and to give a better value, an exponentially smoothed MAD, is normally used (see Chapter 10, section on weighted averages).

Choice of deviation measuring technique

With any statistical distribution, such as the difference between forecast and actual demand, there is a variety of standard ways for assessing the dispersion, such as quartiles, MAD and SD. Approaches using the greatest and least demands are unreliable since the extreme demands usually result from an abnormal occurrence. Basing the variability (and safety stock) on a stock cover calculation is a common, but incorrect assessment. This assumption gives completely the wrong stock balance.

There is a more serious debate about the relative merits of MAD and standard deviation. Mean absolute deviation is easy to understand, to calculate and to set up on a computer. Standard deviation is theoretically more correct, is used throughout statistics and is available as a pre-programmed function on spreadsheets. It is more difficult to calculate and understand.

For improved safety stock evaluation, the assessment of variability should be biased towards using the most recent data and the value should be exponentially weighted (see Chapter 10, section on improved values for MAD). This can be done very effectively with MAD, whereas the SD has to be converted to the variance (= SD^2) in order to use the weighting technique. Exponentially weighted MADs are therefore preferable to SD. Their use results in improved safety stocks, and offers a better selection of forecasting techniques to meet future demands (see 'Choosing the best forecast' section in Chapter 10).

The smoothing factor (see Chapter 10, 'Selection of the smoothing constant, α) most commonly used for MADs is 0.1, since there is normally no rapid change in the demand variability. In practice this may prove to be a little insensitive and 0.15 is often found to work better.

For seasonal demands, the variations in size of the demands are likely to be greater at the maximum and smaller at the minimum. The MAD can be considered as a percentage of the demand level. Using this 'percentage MAD' approach enables an appropriate and up-to-date variability to be maintained and provides the appropriate safety stock levels. The use of exponentially weighted variances should give better values in theory, but is seldom used in practice. Table 6.2 shows a simple summary of the options. There is an arbitrary quality scale included to illustrate the relative merits of the techniques.

Table 6.2 Summary of deviation measurement techniques

Historical evaluation technique	Comment	Relative quality rating
Stock cover	Only for the amateur A major improvement on continuous inspection	1
MAD	Simple to understand and calculate Good for most situations	10
SD	Theoretically more accurate Available on spreadsheets	13
Smoothed MAD	Gives better results Also suitable for seasonal demand Easy to put into spreadsheets and software	25
Smoothed variance	Should give best values	30

For appropriate safety stocks, interpret the historical data, measure demand not sales, and respond to changing variability in demand.

Evaluation of safety stocks

Risk measurement

The reason for having to use SDs and MADs is that they are the only sensible ways to work out stock levels. The shape of the normal distribution curve measured in terms of SDs (or MADs) is always the same. This is the shape of the curves shown in Figures 6.2 and 6.3. These graphs indicate that when stocks are held at average usage, 50 per cent service is given. The other values behind Figure 6.3 are given by the customer service factors shown in Table 6.3. The customer service factors are shown in the table for both SDs and MADs. Either measure is acceptable and will give similar answers.

Adding one MAD's worth of safety stock to the average requirement means that customer service increases from 50 per cent to 79 per cent; or when stocks are increased by one SD, the service increases to 84 per cent, so there is a 16 per cent chance of stockout. By increasing the buffer stock to twice the SD, service is increased to 97 per cent. Additional stockholding increases the service toward 100 per cent but the cost in terms of inventory level is very high. Three SDs' worth gives 99.86 per cent and four SDs 99.99 per cent, But these are increases of only 2.1 per cent and 0.13 per cent respectively for the same investment in stock that

Table 6.3 Customer service factors (assuming a normal distribution)

Desired service level (% periods without stockout)	Multiply SD by	Multiply MAD by
50.00	0.00	0.00
75.00	0.67	0.84
79.00	0.80	1.00
80.00	0.84	1.05
84.13	1.00	1.25
85.00	1.04	1.30
89.44	1.25	1.56
90.00	1.28	1.60
93.32	1.50	1.88
94.00	1.56	1.95
94.52	1.60	2.00
95.00	1.65	2.06
96.00	1.75	2.19
97.00	1.88	2.35
97.72	2.00	2.50
98.00	2.05	2.56
99.00	2.33	2.91
99.18	2.40	3.00
99.50	2.57	3.20
99.70	2.75	3.44
99.86	3.00	3.75
99.90	3.09	3.85
99.93	3.20	4.00
99.99	4.00	5.00

gave 34 per cent for the first SD and 13.6 per cent for the second. This raises the question of which service will be too expensive to give, and whether the customer can tell the difference between 99.86 per cent and 99.99 per cent availability.

The availability level is measured by how often the designed stock level will meet all the demands during the period. (Other measures would be 'how many items were short during a year?' or 'how many customer orders were not completed?')

In Table 6.4 the stock is unable to meet demand in three periods. With review level at 5.0, average stock is 3.7 units (ignoring negatives).

In Table 6.5 all demand is fulfilled with safety stock of two MADs (four items). With the review level at 9.5, average stock is 6.5 units.

An example using stock records

Taking an example, Table 6.4 shows a conventional stores movement record for one item, with receipts, dispatches and supply orders for a stores. The average usage per period is 2.5. If the policy is to purchase sufficient stock to cover three periods, then the order quantity will be 7.5, rounded to 8. If it takes two periods' lead time for the stock to be replenished, then replenishment orders need to be placed when the stock falls below 5 items. This is called the 'review level'[1] because when the stock falls to this level the decision should be taken on whether to order. All too frequently an order is raised at some arbitrary 'minimum stock level' which bears no relationship to the current usage rate. As a result of assessing the demand, a purchaser will change the review level, place an order, or both.

The stock levels resulting from this are shown in the 'stock' column in Table 6.4. The orders are raised each time the stock falls below 5. (This review level is taken as a constant over the short time shown.) Notice that there is a shortage in periods 6, 7 and 10, resulting from the large number of sales in period 6 which coincided with the stock being below the stock review level. It is prudent to hold safety stocks to cover for these occurrences.

The amount of stock can be assessed from a statistical analysis of the demand pattern as discussed in the previous sections. The total sum of the variations from the mean is 20 for the 10 periods (demand variability in Table 6.4.) This gives a MAD of 2. If the safety stock is now increased by 4 from 5 to 9, this should give added protection of 2 MADs. (2 × 2 + 4). This corresponds to a 94.5 per cent safety stock. The stock review level is also increased from 5 to 9 and the orders have been placed immediately at this level. The new pattern is shown in Table 6.5. The extra two stock has reduced the number of depleted weeks from 2 down to nil. It also meant that it was possible to supply all of the 25 items needed. With no safety stock it was possible to issue only 21. This calculation will be refined below.

Of course, from the statistics, this level of safety stock will cause a shortage in 5.5 per cent of the order periods and this may not be good enough. Additional safety stock can be added to give the service level required. The higher the safety stock, the more costly is the stockholding, not only because money is invested in the stock, but also because an investment in stock may prove fruitless if the item is superseded or becomes obsolete.

Note: [1]Up to this point it has been referred to as 'reorder level'.

Table 6.4 Stock record

Period	Demand	Stock	Purchase orders	Demand variability
0		8		
1	1	7		1.5
2	5	2	Place order for 8 ↓	2.5
3	0	2		2.5
4	3	7	Delivery of this order	0.5
5	2	5		0.5
6	6	−1	Place order for 8 ↓	3.5
7	1	−2		1.5
8	0	6	Delivery of this order	2.5
9	6	0	Place order for 8 ↓	3.5
10	1	−1		1.5
Average	2.5	37	MAD	2

Table 6.5 Stock record with safety stock

Period	Demand	Stock	Purchase orders		Demand variability
0		8			
1	1	7	Place order for 8 ↓		1.5
2	5	2			2.5
3	0	10	Delivery of this order		2.5
4	3	7		Place order for 8 ↓	0.5
5	2	5			0.5
6	6	7		Delivery of this order	3.5
7	1	6	Place order for 8 ↓		1.5
8	0	6			2.5
9	6	8	Delivery of this order		3.5
10	1	7			1.5
Average	2.5	6.5			2

Additional safety stock

Safety stock is normally calculated on the basis of unknown demand. The quality of supply from the manufacturers may also prompt an increase in safety stock, either because the lead time for an item may vary between placing one order and the next, necessitating extra stockholding, or because the supplier's delivery performance against the lead time is unreliable and delivery is more often late than early. An allowance in stock has to be made for this factor. If stock has to be held against poor supply, then theoretically a supply performance MAD could be calculated and a safety stock worked out. This supply safety stock would not be additional to the demand safety stock, since the same safety stock would cater for either high demand or late delivery, but not both. It is found that the addition of MAD from different causes can be added using Pythagoras' theorem (in a similar way to adding variances).

> Make the supply effective. Don't hold stock to compensate.

In dealing with the supplier, the supply pattern should be a result of compromise for both organizations. Whoever is the cause of the extra safety stock should incur the cost of it. Minimum stock is held when each party can trust and accept the information provided by the order.

Slow-moving items

For a large number of stock items the usage rate is low. This includes obsolescent items, non-standard lines and spares. For these the 'normal' distribution is squashed against zero usage and is replaced by a skewed distribution. The chance of an item being required is governed by Poisson's distribution law. Compared with the normal distribution, the Poisson distribution has a longer tail where there is an enhanced (but small) chance of high demand during a period. To cover for this, a higher stock would be required where high levels of availability are needed.

> Demand for slow movers is a Poisson distribution.

An important feature of the Poisson distribution is that

$$\text{Standard deviation} = \sqrt{\text{Average demand}}$$

This relationship simplifies the review level formula for slow-moving items. When using the Poisson approximation, the usage events have to be independent (i.e. issues in ones to unconnected demand). If this is not the case, then

$$\text{Standard deviation of sales orders} = \sqrt{\text{Average number of sales orders.}}$$

Using the calculations developed in this chapter and applying the lead times discussed in the next provides the basic method for inventory management for the majority of stock.

Excess and obsolete items

In practice, stocks are never balanced ideally and some items become excessively slow moving. These items can be described as 'excess' or 'obsolete'. One of the areas which generates little interest and lots of cost is obsolescent items. The definition of obsolete should be that there is no use for the item. This information is normally only available when the customer changes to a direct and improved substitute. In many instances, goods are classed as obsolete as a result of a long period with no demand: so the practical definition is usually that an item is obsolete if it has not moved for 'x' months.

> Have the system identify obsolete items, then remove them.

The selection of 'x' varies with the type of item and market, and the position in the supply chain. It is relatively easy in fast-moving consumer goods to get rid of items which have not moved for three months and this may well be the definition in this environment. Suppliers of heavy machinery and spare parts, where world demand has been zero for two years, may not regard demand to have finished and they will still desire to keep stock (only one!). In this case failure to supply quickly may have important strategic or economic effects, and supply lead times are long. A mechanical part on a machine which fails every seven years or so is not obsolete, but has a very low usage rate. The decision to purchase such slow-moving stocks is one problem, the decision to throw away existing stock is entirely another. Each company has to decide on a proper obsolescence policy for disposing of non-moving items.

Excess items

The 'not moved since' analysis can cope with non-movers, but there are other useless items, namely the excess items. If the stock of an item is 1000 and the usage is one per month, the item is certainly not obsolete, but may be before the stock is used up. Therefore the definition of excess could be 'Excess items have stock of more than y months' usage'. Again the value of 'y' depends on product, market, position in the supply chain, and also on the Pareto class of the item, since the stockturn definition is a limit on the stock investment.

Figure 6.4, with cut off values 'x' and 'y', shows simple control of excess and obsolete stock. It allows accountants to identify how much money to put aside to cover likely stock disposal. Writing down inventory value is, of course, an admission by the stock controllers that they failed!

In theory the values of x and y could be different for each item. For example, a stock of helicopter rotor assemblies could be held in stock for a long time before use, but there is a mandatory requirement to change and repair these units, so the stock is not obsolete, just slow moving. Whether stock is excess depends on whether the items are now units (bought too early) or repaired units (which would be in inventory anyway). In practice, therefore, the cut-offs shown in Figure 6.4 are used, and some allowance is made for items with stock of one.

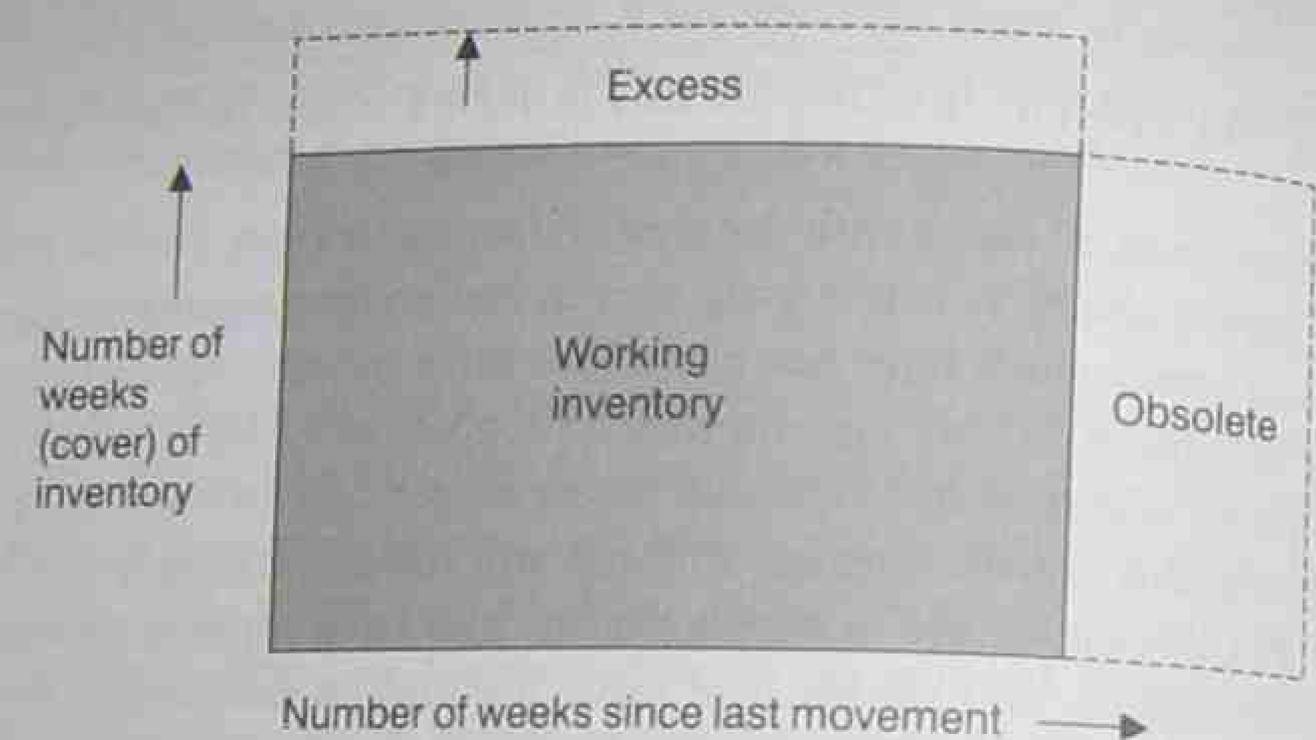


Figure 6.4 Definition of excess and obsolete stock

The first step in dealing with the phase-out of stock lines is to identify early the lines in this terminal condition. These lines can then be included in the ABC classification as 'O' or other suitable flag to inhibit normal ordering. The disposal of excess and obsolete stock has to be carried out carefully and professionally to avoid a large loss in profits. Before selling a company, the value of obsolete stock items has to be reduced since it is illegal to overvalue the company and the stock can be a major part of this value.

Key points

- Safety inventory compensates for unknown demand.
- Supply problems should be removed, not buffered.
- Demand can be split into average and variability.
- Variability (safety stock) should be calculated using MAD (or SD) but not estimated.
- Customer service factors should be used to decide the level of availability necessary.
- Identification of obsolete and excess items will enable old inventory to be used up more effectively.

7

Setting the right stock levels

- Order patterns with safety stock
- Successful calculation of review level
- When to expedite
- The effect of supply lead times
- Co-ordinating supply with target stock levels

Simple assessment of review levels

The basic calculation

Our aim in stock control is to maintain a balanced inventory so that customer service for each stock item is maintained within its proper limits. The time at which we can influence the stock levels most effectively is when we order, and it is at this point that the major opportunity occurs for ensuring a balanced inventory. This is the major control mechanism for ensuring a balanced inventory. It is also the point at which customer satisfaction or excess inventories are created. The technique for calculating the right time to order inventory is crucial to balancing these conflicting pressures.

Originally stock-recording systems used to include stock control levels as 'minimum' and 'maximum' stock levels. These basic concepts are still applied today as a simplistic approach where tight control over stock is not necessary. The use of a minimum stock level for order control is not sensible, since the minimum occurs immediately before delivery. Items have to be ordered well in advance of this, so control is through a 'reorder level' not a minimum stock. However, a minimum stock level is vital for ensuring that there is warning of low stocks. In a stock control system, therefore, there is a need for both reorder level and minimum stock level. These are called the 'review level' and 'safety stock' (see Figure 7.1).

The review level is triggered by the information from the system that stock is low (i.e. it has fallen below the 'reorder level') and makes a decision

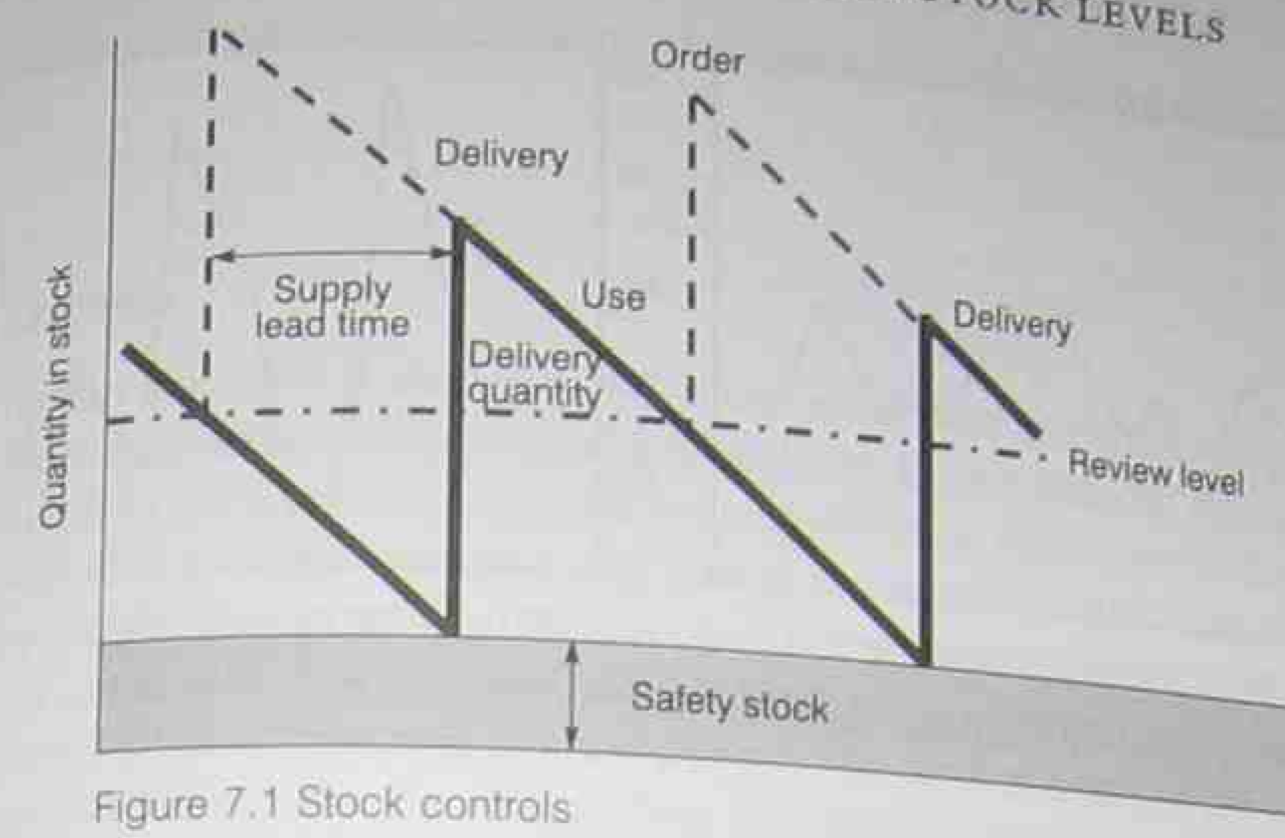


Figure 7.1 Stock controls

as to whether to order. Where the demand has increased or lead times increased, orders are necessary; where the lead time is reduced or demand has diminished, then there is no need to purchase. Since demand can go up or down, the trigger level is treated in practice as a review level.

Minimum stock has also been replaced by the safety stock which is calculated as discussed in Chapter 6, and allowing for lead time (the analysis is shown below). Maximum stock is a result of negotiation with suppliers on delivery quantities. The maximum stock level should be a result, rather than a direct control. It is therefore necessary to control the delivery quantities directly by using three parameters to carry out simple and effective stock control, as shown in Figure 7.1, namely:

- review level
- safety stock
- delivery quantity.

The starting point is the calculation of the safety stocks. A quick look at the historical demand for the two items illustrated in Figure 7.2 shows that the second will require a great deal more safety stock than the first if the customers are going to be equally satisfied. The average demand rate is the same in each case. The more the variability of demand, the more safety stock is required. Therefore the safety stock does not depend on the demand rate. (This means that safety stock calculated as number of weeks' demand is simply wrong.) The safety stock for periodic delivery items is

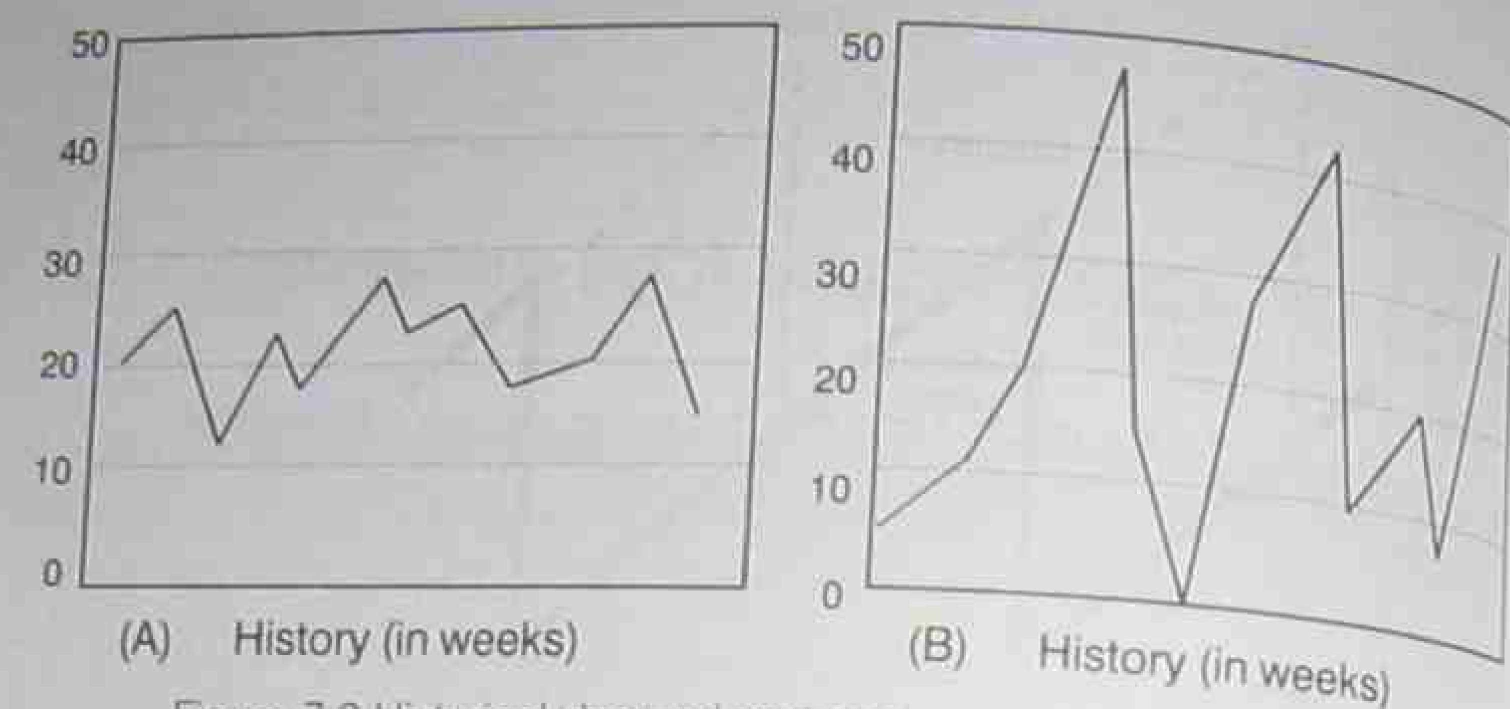


Figure 7.2 Historical demand patterns

given by multiplying the customer service factor by the standard deviation (or MAD) of the demand during the period.

> *Safety stock depends on demand variability and not demand rate.*

If the demand is thirty per month and the supply lead time is two weeks, then there had better be fifteen in stock or on order or else there will be a shortage. This is because the time to order is given by:

Review level = Usage rate × Supply lead time

and to be safe:

Review level = Usage rate × Supply lead time + Safety stock

The safety stock calculation discussed in Chapter 6 showed that the stock depends on the variability of demand (MAD or SD) and the availability level required (customer service factor). Although the effect of supply lead

Table 7.1 Variation of safety stock with lead time

	Per day	Per week	Per month	Per year
SD scaled proportional to lead time	0.4	2	8	100
SD scaled as the square root of lead time	0.9	2	4	14

time was not discussed, the safety stock also depends on the supply lead time.

Example

For an item with a demand of five per week plus or minus two (i.e. MAD = 2) what is likely to be the variation in demand per day, per month or per year? Two options are suggested for the answer in Table 7.1. The average demand per period increases linearly with time period. The uncertainty of the demand, however, does not increase at the same rate. The average is more likely to be achieved over a long period of time than over a short period. The annual sales of an item are often similar to the year before, but from day to day the demand will vary widely. The weekly sales will also be quite widely dispersed but the monthly results are quite consistent. If the SD (or MAD) were to increase in proportion to the time period (see Table 7.1), then the inaccuracy in the long-term forecast would be great. However, because of the 'compensating errors' over longer time periods, the deviations are relatively small. In fact it is found (and can probably be proved) that the SD only increases as the square root of the time period. This is shown as the last line of Table 7.1.

This same effect can be illustrated in Figure 7.2B. A short sample could give a wide variety of answers to the average use, whereas a longer sample will give a more reliable answer. In other words, the longer the sample period (in the case under discussion – the lead time) the less the relative safety stock need be. From Chapter 6 this would be a square root relationship. As the variation in demand is largely a statistical effect, the square root relationship occurs very frequently in stock control.

The full formula for safety stock is therefore:

$$\text{Safety stock} = \text{Customer service factor} \times \text{SD} \times \sqrt{\text{lead time}}$$

For values of customer service factor, see Table 6.3. This equation can then be used in the review level calculation above.

The full calculation of review level can be carried out for the item with the demand history shown at the start of Chapter 6 (Table 6.1). The average was thirty-one and the MAD was 6.8. Some practical information is now required, namely, the supplier's lead time and the availability level to be offered to customers.

Assuming that the lead time is three periods and the required customer service is 90 per cent, the customer service factor for 90 per cent customer service is 1.6 MAD from Table 6.3. First the safety stock has to be calculated:

$$\text{Safety stock} = \text{Customer service factor} \times \text{MAD} \times \sqrt{\text{lead time}}$$
$$= 1.6 \times 6.8 \times \sqrt{3}$$
$$= 18.84$$

From this the review level can be calculated:

$$\text{Review level} = (\text{Usage rate} \times \text{Lead time}) + \text{Safety stock}$$
$$= (31 \times 3) + 18.84$$
$$= 111.84$$

This means that if the stock falls to 111, then more need ordering (assuming there are not enough on order already to cover the 111). If it is decided that 90 per cent service is not good enough and 98 per cent is more appropriate, then the customer service factor becomes 2.56, the safety stock 30.15 and the review level 123.15.

In reality, all the items do not need to be in stock to provide availability for customers as long as they are on order and due for delivery before the current stock runs out. In the example above, if the stock is currently sixty then there is a problem. However if there are already 100 on order, there is no need to raise another order. Therefore the condition for ordering is when:

$$\text{Current stock} + \text{Supply orders outstanding} < \text{Review level}$$

(To be exact, the supply orders which count here are only those which are due for delivery within the supply lead time.) The purpose of the review level is to trigger replenishment orders at the right time. Of course, triggering the order at the right time does not necessarily ensure that there is stock in the stores. The review level calculation just described is the most important piece of stock control theory.

> Reordering triggers orders but does not eliminate stockouts.

When to expedite supplies

From the purchasing point of view the supplier has to deliver on time and should be chased if the delivery is late. From a stock control point of view, we rely on the supplier delivering at a fixed lead time. However, if the supply is late, it does not necessarily cause a problem if the rate of issue has not been as fast as expected (and the stock is still relatively high). In other circumstances the demand may be higher than average and it would be convenient to receive delivery early. In the example discussed before, a 90 per cent availability means that once in every ten time periods a stockout is to be expected. When this happens it would be useful to try to enhance availability by rushing in supplies. Therefore, a warning is needed for when the physical stock is getting low, not when the supply is late. As the supply should, on average, arrive when the stock gets down to the safety stock level (see Figure 7.1), then the safety stock level is a suitable trigger level to remind the supplier that goods should be arriving. Sometimes this reminder will be after the proper lead time (if delivery is late), or it could be before the item is due when the demand is above average. In either case there is still the safety stock to use before there is crisis. The condition for chasing supplier is therefore when current stock is less than safety stock. This should be a standard routine printout from the system used for automatic expediting.

> Chase suppliers when you start using safety stock.

Managing lead times

Lead time

Lead time is the time between a shortage occurring and items being available to maintain supply. The lead time for supply is a balance between what the supplier wants and what the customer wants. If the suppliers are powerful, they will specify a long lead time (such as a one year firm programme). If the customers have the upper hand, they will expect to be supplied immediately, without any commitment. Both of these extreme cases increase the total operating costs. The compromise is a lead time which gives the suppliers sufficient supply time but do not force the customers to commit themselves further than they can predict. The accuracy of their forecasts of requirements reduces rapidly the further ahead they look. Long lead times and large batch supply produce an order pattern which is irregular and subject to change. Stocks are needed where the sales lead time is shorter than the supply lead time. If this can be altered so that supply lead time becomes less than sales lead time, the stockholding can be avoided.

Lead times for purchasing are governed by production and transport. Lead times for manufacturing result from capacity, material availability and production planning. Lead times required by customers depend upon availability and supply policy (make to order, assemble to order, stockhold).

Components of lead time

Lead time can be analysed into several components, some of which are essential and some avoidable. The supply lead time comprises:

- order review time – the intervals at which the low stock situation is reviewed (e.g. printout every day, hourly, weekly, etc.)
- order processing time purchasing and communication) – the time it takes to:
  - review the order
  - decide to buy
  - transmit to the ordering system
  - raise an order
  - gain the appropriate authority
  - inform the supplier (post, fax, e-mail)
- supplier lead time (vendors' manufacturing, buying and dispatch)
- transport time – transfer of the item from supplier to receiving bay
- receiving time – time taken for goods inwards and updating the stores records.

From this analysis, the actual supply lead time contains some elements of administration in addition to the lead time quoted by the supplier.

In stock control equations, the supply lead time is assumed to be constant. It is very important that the estimate of lead time in the system reflects the true lead time. If the supply lead changes, then the lead time in the calculation should be changed to reflect this. It is a continuous task and part of the maintenance of the stock systems. Changes can normally be made by groups of items from the same source, or by using the actual lead time to update an (exponentially weighted) standard lead time.

> Keep the lead time realistic.

Setting the lead times

If the lead time on the control system is set too short, this will lead to stockouts, since the stock level will be kept too low. If the lead time is set overcautiously (long), excess stock will result. The compromise for setting the lead time is a fine judgement for the inventory controller.

Minimizing the lead time

Longer lead times make forecasting more difficult and increase the safety stock. Reducing the lead time is therefore desirable and can lead to stockless supply if the supply lead time is short enough. It is part of inventory management, therefore, to work on reducing the overall lead time.

One of the simplest ways of reducing the lead time is to specify it more precisely. 'Delivery in the month of June' as a request gives the supplier plenty of leeway. 'Delivery on 12 June at 11.18 a.m.' suggests an entirely different expectation, and is likely to gain a much better delivery performance (during the day required rather than the month). Numerous experiments attempting this have shown that a majority of suppliers will improve their delivery precision greatly, simply by being asked.

Companies also find that supplier reliability is often affected by lead time. With a lead time of one day, if the supplier of an A class item is a day late, then this can be a major issue. If the lead time is twelve weeks and the supplier is a day late, this is often expected, or even considered as on time. One day's worth of stock is held to offset this. The perception of delivery performance is as a proportion of lead time. However, from an inventory perspective, a day's worth of stock is the same value, whether it is added to an extra six weeks' worth or to one day's worth of total inventory.

From the analysis of lead time above, the total lead time should be minimized by:

- inputting information to stores and computer systems without delay (removing administration leadtime). It is important to minimize the time between:
  - the stock getting low in the stores and the supplier receiving the purchase order
  - receiving the delivery and the event being displayed on the company computer system

- partnership with major suppliers (giving schedule flexibility or kanbans).

Once these elements of lead time are small enough, (see also Chapter 8, section on purchasing procedures) the suppliers' lead times should also be examined. Supply lead time consists of actual physical transport time, manufacturing time in some instances and organizational time. Lead time should be a negotiated time, not something that happens without manipulation.

> **Example**
>
> *A supplier quotes six-week delivery. What does this mean? It could be:*
>
> - *all the existing stock is allocated to someone else*
> - *there is a stockout because the safety stock was not high enough or the reordering procedure broke down*
> - *it takes six weeks to get these from their supplier.*
>
> *In the first two cases, the six weeks are caused by priorities, first, because the order is not jumping the queue and, second, because the supplier does not consider the order important enough to put in higher safety stock or improved systems. This is a matter for the supplier and the customer to negotiate. The third case could be argued as similar. If their supplier is a manufacturer, then the six weeks could reflect the manufacturing time. However, only a very little part of manufacturing time is taken up with production. The manufacturing time per unit assessed by work measurement (time being worked on) is much smaller than the production lead time taken for the work to pass through the manufacturing plant. This is usually more than ten times the measured work content. There is, therefore, potential for manufacturing lead time to be collapsed to a much smaller timescale. However, this cannot be done on a one-off basis; it is has to be done gradually for repetitive products because of its effect on capacity balance and replanning. The amount of replanning manufacture has to be minimized in the short term or else the plant becomes inefficient and delivery performance suffers. Again, collaboration with suppliers may enable them to reduce their own lead time.*

Because of the large numbers of items in inventory, reduction of lead time has to be carried out for selected types of items. These are initially:

- A class items
- items where a small change in lead time can make them into supply-to-order items.

Effects of lead time

Lead time does not govern stockholding. Order quantity is not affected by lead time. It is possible to order twenty weeks' worth of an item with one week's lead time and one week's worth of an item with a twenty-week lead time (as long as further orders are placed each week). True the review level is calculated using the lead time, but the review level determines only what order cover is required and not how much stock is held.

Lead time does affect the amount of safety stock to hold. The longer the lead time, the wider the absolute variation in demand and therefore the higher the safety stock for a given service level. Fortunately the amount of extra stock only increases as the square root of the lead time.

In summary, the effects of long lead time are generally:

- forecasting demand to cover the supply period is more difficult
- the reliability of the supplier may be poorer.

Errors in forecasting cause higher stocks to be held. If demand is certain over the lead time, then stocks can be very low. A short lead time means that:

- the forecast is more likely to be correct
- errors will be smaller.

Dealing with inconsistent lead times

With closer partnerships and higher expectations, inconsistent delivery is not the problem that it used to be. However, customers are now more sensitive to supply reliability, so strategies should be considered for any supplier where delivery is inconsistent and changing supplier is undesirable.

It is difficult to set reasonable stock levels or to provide good customer service if lead times are not known accurately. The policy on inconsistent lead times should be determined by the impact of the problem, and on whether the financial risk is high or low. For high turnover value items (A class) the inconsistency has to be resolved through scheduling and working more closely with suppliers to smooth out the inconsistencies. The objective should be to smooth out the variations rather than to improve the updating, since each time the lead time is updated, the stock parameters will be changed and the order book modified.

For low turnover value items (C class) the situation is not so important unless there are very many inconsistent suppliers. The strategy for avoiding the problem for routine supplies is to use pessimistic lead time in calculations. This will cause higher stocks, which may be necessary if the item is important. However, as the purchase quantity is probably large, there are only a few occasions on which there is a chance of running out so C class items may not be worth the investment in extra stock to provide the extra availability.

In some instances lead times are cyclic. This is the result of the supplier having a fixed plan and not responding to customer demand.

> *Examples of this are where there is a logical sequence of production which cannot be broken such as size (steel-mill rolling plan – thickest to thinnest over several weeks), contamination (pure to impure in process industries, light to dark in injection moulding) or other cycles (food crops etc.).*

In this case supply orders should ideally be placed when the lead time is minimum. Using target stock levels (described at the end of this chapter) the timing of orders can be arranged to achieve this. However, the process can only work if any changes to the supply cycle is communicated to the customer.

Effect of order frequency on safety stock

The safety stock calculated by the methods described earlier in this chapter are adequate for most practical situations. However the calculations can be improved because the risk of stockout is not constant all the time. Availability is good when there has just been a delivery. The risk of running out are greatest when the stock is lowest, i.e. while the goods are below the review level and a delivery is expected. The main risk during a year is, therefore, the risk during the lead times. This depends upon the order

frequency. The more replenishment orders for an item that are placed per year, the higher the risk and the greater should be the safety stock. This factor is usually small compared to the benefits of ordering more frequently and reducing the inventory costs and risks of obsolescence. A simple addition to the calculation above suffices for control, as shown in the following example.

> More refined calculation adjusts for batch size.

Example

Assume that the historical data for an item is as follows:

- *Annual demand = 10 000 (200 per week)*
- *MAD = 100 (weekly)*
- *Service level required = 94 per cent (2 MADs) (measured as number of items short per year)*
- *Lead time = 4 weeks*
- *Order frequency = 6 weeks*

So the total stock shortfall allowed = I – (94 per cent of 10 000) = 600 per year.

If there are 50 ÷ 6 orders per year (8.3 orders per year) then the allowed stock deficit per stockout is:

$$\frac{600}{50} \times 6 = 72 \text{ (assuming 50 sales weeks per year)}$$

$$\text{Usage in the lead time} = \frac{10\,000}{50} \times 4 = 800$$

The allowed stockout is therefore 72 in the lead time (LT) demand of 800 so

$$\text{Service level required} = 1 - \frac{72}{800} = 91 \text{ per cent}$$

91 per cent = 1.5 (MADs) but this is for a 4-week MAD since the period is the lead time, so

$$\text{Safety Stock (SS)} = 1.5 \times 100 \times \sqrt{4} = 300$$

and

Review level = 300 + 200 × 4 = 1100

The average stock level is SS + 1/2 order quantity

= 300 + 1/2 × 6/50 × 10 000

= 300 + 600 = 900

Now if the order frequency is changed to 10 weeks, then there are now 5 orders per year

$$\text{Allowed stock deficit per stockout} = \frac{600}{5} = 120$$

The usage in the LT = 800 (as before), but because there are less risk periods

$$\text{Service level required} = 1 - \frac{120}{800} = 85 \text{ per cent}$$

Eighty-five per cent is 1.2 MADs (i.e. 4-week MADs). Therefore:

Safety stock = 1.2 × 100 × $\sqrt{4}$ (in weekly MADs)

Safety stock = 240

Review level = SS + usage in LT = 240 + 800 = 1040

Average stock level = 240 + 1/2 × 10/50 × 10 000 = 1240

The review level and safety stocks are less, but the average stock is more because of the larger order quantities. If the order frequency is changed to 4 weeks, the service level per stockout becomes 94 per cent, the same as the simple calculation. This is to be expected since there is continuous exposure to stockout risk. The safety stock becomes 400, the review level 1200 and the average stock 800.

Target stock levels

Application of target stock levels

The target stock level (TSL) is a maximum stock level which can be used to calculate the order quantities. Target stock levels marginally differ from

the review level approach to ordering: Review levels fix the order quantity and vary the order frequency, whereas TSLs fix the order frequency and vary the order quantities.

When ordering regularly from a supplier, this enables the stock controller to have a cyclic work routine, (such as always ordering from supplier X on Thursdays) and the TSL approach is therefore good to use for the regular demand items, and is recommended for A class products.

> *Target stock levels support efficient purchasing.*

The TSL is a level to which the stock is topped up (theoretically) when the cycle time for ordering arrives. The process for setting up the ordering procedure is, first, to establish a routine for placing orders on the various suppliers by specifying the day each week or month on which the order is going to be dealt with and, second, on the allotted day, review the stock of all the goods supplied by that supplier. The amount to be ordered is given by the formula:

Order quantity = TSL – Free stock – Supply orders outstanding

Here the 'orders outstanding' are, of course, only those orders which are due within the current supply lead time. Orders booked further ahead are not included in the calculation.

Example

TSL has been calculated as	*34*
Free stock is currently	*16*
Outstanding order due tomorrow	*5*

Therefore

Order quantity = 34 – 16 – 5

= 13

Calculation of target stock levels

The calculation of TSLs is very similar to that for review levels. An extra factor is included in the calculation of the TSL. The effective difference is that stock replenishment now needs to cover the period until the review cycle comes round again. If the review cycle were a month and the supply lead time only one day, then the TSL would have to cover a month's extra usage (plus safety stock) because the next order would be placed only in a month's time. Thus the formula for the TSL is given by:

TSL = [Usage rate × (Lead time + Review period)]
+ Safety stock

And there is a similar adjustment to increase the safety stock level:

TSL = [Usage rate × (Lead time + Review period)]
+ Safety stock

Safety stock = Customer service factor × MAD
× √(Lead time + Review period)

The demand 'in review period' is included because the stock level may not be checked continuously, or deliveries may only be made once per month. If we do not include the extra time in these processes then the risk of running out of stock is increased.

Example

Review period for an A class item = *1 week.*

Supply lead time = *5 weeks.*

Average demand rate = *4 per week.*

MAD = *2.5 measured on a week basis.*

Required customer service = *90 per cent (this gives a customer service factor of 1.6).*

Therefore,

Safety stock = *1.6 × 2.5 × √(5 + 1)*
= *9.8*

and

TSL = *4 × (5 + 1) + 9.8*
= *33.8*

Applying target stock levels

The TSLs are levels of stock to which the free stock never rises in a stable situation (see Figure 7.3.) The actual stock will rise to a maximum level which is

TSL − (Usage rate × Lead time)

The TSLs can be used for long review periods (e.g. one month) with short lead times or, of course, with short review periods (one week) and long supply lead times. In this latter case, there will always be one or more orders outstanding on the suppliers. This does not affect the validity of the calculation.

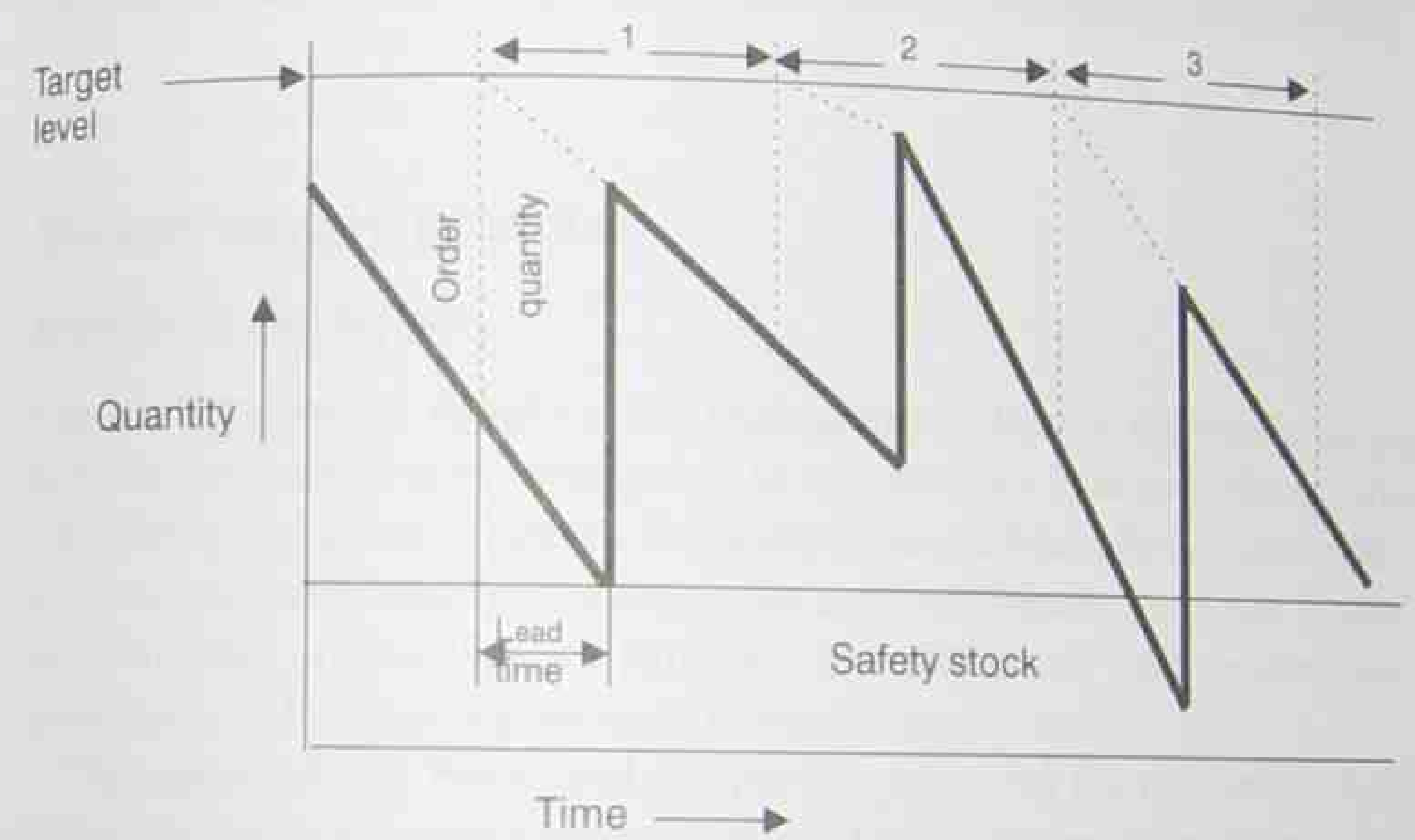


Figure 7.3 Target stock levels

It is important to keep on adjusting the TSLs to match the usage rates and lead times which are experienced, or else this approach can lead to last minute purchases and extra work.

It is also important to ensure that the stock does not run out without warning during the order cycle. This is why it is essential to monitor continuously to find those items where the stock is falling below the safety stock level, calculated in this chapter.

Key points

- Review levels should be continuously updated in line with demand.
- Review levels do not avoid stockouts; to do this requires expediting when physical stocks get low.
- Batch sizes are more important than lead time in keeping stocks low.
- Using target stock levels is good when buying many items from one supplier.

8

The changing role of purchasing

- The role of procurement
- Single sourcing
- Supply partnerships
- Vendor evaluation and rating
- Types of orders
- Calculating delivery quantities
- Effective purchasing administration

Modern supply practice

Improving the supply

The development of supply chain management changes the relationships between companies providing goods to one another. This has affected the role of the purchasing function. Purchasing has traditionally been an order-by-order process based on obtaining the best combination of price, quality and delivery (availability). Estimates are requested, best options are negotiated, orders are placed and delivery is monitored.

This approach has been eroded over the years and the time-consuming task of obtaining estimates is not now usual in routine material supply. The development of supplier relationships has introduced longer-term arrangements involving contracts, call-offs and schedules. This has changed the role of the purchasing function. Instead of individual orders, supply is negotiated less frequently. This has meant a different style of purchasing activity because low level routine negotiation is being replaced by high level contracts. The number of suppliers is also reduced.

Purchasing people are no longer order processors.

These changes have a major effect on the type of expertise required in purchasing. The traditional type of 'buying' is being superseded by a combination of high-level contract management and supply scheduling as part of the material control activity. This means that materials management will absorb the purchasing operations as part of control and scheduling. Why is this now practical and possible? The answer lies in the current supply environment:

- total quality – supply to specification is expected
- supplier partnerships – work closely with supplier
- logistic control – consider production and transfer costs
- linear production – manufacture products through same process route
- single sourcing – only one supplier for each item
- supply chain management – maximize effectiveness and minimize total supply cost.

The new role of purchasing in a company is in negotiation – a skill which inventory management has been practising internally for a long time. Purchasing no longer requires the wheeler-dealer, buying from a favoured source, hard-hitting on price and always with a good idea of where to go for a cheap deal or a quick supply. Purchasers must now be contract managers, working in partnership and shrewd enough to balance the well-being of their company with the aspirations of their supply partners. They will have fewer suppliers and maintain multilevel collaboration within the supplying and user company. In fact, in some companies the purchasing agreements will be carried out by top executives, as they were in the early days of business enterprise. This situation arises because of the improvement in communication between companies and the better organization of the supply chain.

The purchasing environment

The environment in most businesses has changed from a supplier-led to a customer-led operation. This has resulted from the increase in international trading, multinational competition and the improvements in transport systems.

The development of the customer-led market has caused major changes in supply structures. There have been significant developments:

- Supplier relations are much closer.
- Quality management works on audit, not inspection.
- Stock control philosophy is to keep low inventory.
- Supply chain structure is fixed.

- Number and type of suppliers – there are fewer suppliers and these are comparable sized companies.
- Distribution has become reliable and a direct cost, not an overhead.
- Administration has less people with tighter targets.

There has been a gradual change of power from supplier to customer leading to changes from a product-based supplier to a service-based supply (see Table 8.1).

Table 8.1 Changing attitudes to service

Service was supplier led	Service is customer driven
Manufacturer provides products	Customer requires solutions
Supplier provides narrow range of options	Customer sources wider range from each supplier
Non-standard individual items	Items are commodities, system components
Generic service	Different service for each customer
Delivery when available	Fast reliable service
'Provide products' philosophy	Packaged items ready for customers to use or sell

This means that the requirements for suppliers and customers have changed (see Table 8.2).

Table 8.2 Effects of changing attitudes to service

	Issues for service providers	Issues for users
Supplier-led service	Supply lead times Single logistic chain High demand rates Good margins	Variety of suppliers Many orders to control
Customer-led service	Good availability Wider type of demand Stock and supply risks Commodity type supply	Convenience Good control and price Single source Supplier partnerships

Single sourcing

Work closely with fewer suppliers.

A major development in purchasing has been in the area of supplier relations, with a pressure to decrease the number of suppliers and to develop partnerships with them. This stems from the need for improved supply, and has led to tighter control and higher quality standards. One solution to the supply problem is seen to be a single supplier for each item, providing the supply risks can be minimized. The approach toward single sourcing can be made in the same way as any other quality improvement process:

- Improve process until a problem is encountered.
- Identify the cause of the problem.
- Find means of solving or avoiding the issue.
- Advance to the next problem.

Managers often consider that their own situation is not appropriate for single-source supply. Traditionally, companies purchased items from a variety of suppliers, choosing a supplier at any time on the grounds of cost or availability or, less often, quality. Such a formal, arm's length relationship with the customer giving an 'order' to the most fitting supplier, was more costly and time-consuming than a single-sourcing arrangement, and did not necessarily give the customer the best service.

A supplier who is not confident of the customer will:

- present a price from which to negotiate
- only meet the customer's specification and no more
- offer improvements only under duress
- provide no add-on services
- consider only the short-term profit
- fear losing the next order.

This negative attitude stops the supply chain from being co-ordinated properly. The exception to this situation is where the customer is so dominant that the supplier is treated like a part of the customer's company and is dictated to by the customer. (This is the situation in many Japanese companies.)

The normal objections to single sourcing are:

- the supplier has too much power
- inflated prices

- risk of poor delivery performance
- lack of flexibility
- supplier may cease trading
- divulging sensitive information
- risk of quality problems.

Focus on least total cost operating, not lowest unit price.

These are real fears for those not accustomed to single sourcing and present the first barriers to be crossed.

There must be compensating benefits to encourage companies to adopt single sourcing. These benefits can be considered as short-term operational advantages, and as long-term improvements. The long-term benefits are:

- supplier is responsible for the item
- closer relationship with suppliers
- joint improvement programmes to reduce cost
- buyers' leverage is enhanced
- focus on quality and efficiency
- fewer vendor contracts
- more stable schedules.

Looking at single sourcing from the supplier's point of view, the advantages are:

- higher share of business
- better opportunity to plan resources
- confidence in continuing demand
- opportunity for product and process development.

The disadvantages are:

- more dependence on one customer
- need to divulge financial information.

In the long term, the development of mutual confidence and a close working relationship will enable two partners in the supply chain to plan facilities to match the ongoing demand, focus on quality and delivery (just in time?), eliminate formalities (time and paperwork), and provide a service rather than simply a supply of items. The aim is to optimize added value throughout the supply chain to the benefits of both partners. This requires a commitment by both parties to collaboration over several years.

The short-term benefits of single sourcing can be classified as:

- reduction in buying and paperwork
- less vendor management
- easier traceability for non-conformancies
- single shipping route
- consolidated deliveries
- faster replenishment cycle.

To make single sourcing work in practice, management have to be convinced that the cost savings from these benefits far outweighs the potential costs of the risks identified.

> Take the supplier point of view.

What really is single sourcing? The obvious answer is 'only one supplier for each item'. This does not mean putting all your eggs in one basket, but it does mean relying on a supplier for a particular size of egg. This always gives the back-up that the size 3 egg supplier could be persuaded to provide size 2 if there is a big supply problem. Single sourcing also means empowering the supplier to provide the quantity, quality and delivery without fail. Strategies to support single sourcing are therefore:

- ensure supplier profitability
- ensure that demand is important to supplier (significant turnover)
- partnership relationship
- management and monitoring of performance.

Vendor appraisal

The traditional purchasing targets of good quality, price and delivery should be interpreted to mean total quality, long-term total cost and consistent or JIT supply.

Quality is the responsibility of the supplier. Quality issues at the supplier interface are concerned with the provision of usable goods, delivered safely and on time, and accurate information including the customer's identity code, about what is sent, shipment by shipment, box by box.

It has been normal practice in some industries, to inspect goods arriving, on a 100 per cent basis in some instances, and on a sample basis in more advanced companies. It is also customary for manufacturers to inspect goods being dispatched. Items are therefore inspected on dispatch from one

company and inspected on arrival at the next, either to ensure the integrity of the transport or because there is a degree of mistrust. The integrity problem is relatively easy to solve, at least within one country. The mistrust is more difficult.

There are always stories of suppliers sending wrong quantities or wrong items, even though the consignment may have been inspected on dispatch, particularly if the supplier is also the manufacturer. The integrity of the method of supply may also be questioned, more so if it involves international or long-distance transportation. If a customer inspects a delivery and finds it defective in some way, he or she will complain to the supplier, and the questions which then arise are:

- Who paid for a full inspection and when?
- Who separates the defective parts?
- Who pays for the return transport?
- Who owns the problem?
- Who pays for a vendor rating and quality control system?

In most cases the answers to these questions is the customer. If the supplier is not worried about this aspect of customer service (especially for a single source or accredited supplier), then getting the customer to carry out the quality checks seems to be a good cost-saving idea. Only where the supplier's business is jeopardized by the quality situation is the matter resolved. The options for the customer are:

- Send the whole shipment back.
- Establish and circulate a league table for supplier rating.
- Charge the suppliers for the cost of inspecting their work.
- Communicate serious problems to their chief executive.

> Make suppliers pay for cost of their quality.

The ability of a business to meet commitments to its customer depends on the quality of support from its suppliers. If that support is missing, the business must look elsewhere for better service.

Appraisal criteria

Audit is necessary as part of vendor assessment, to ensure that the excellence of supply is maintained. The purchaser's appraisal consists of three major areas:

- Technical – is the item fit for function?
- Supply – is the quality, quantity, timing, identification and cost acceptable?
- Support – is the supplier assisting in developments, queries, re-engineering?

The appraisal criteria in each of these areas can be listed and ranked according to importance – some features will be essential and others merely desirable, and suppliers can be classified by their past achievements against the set criteria. Results can then be used to rate the suppliers, as shown in Table 8.3 for the type of inspection required.

Table 8.3 Classification of suppliers

Class	Status	Inspection
Certified vendor list	First choice Close relationship	No inspection
Qualified vendor list	Select group of suppliers Business quality audited Offered expanding business	Minor audit inspection
Approved vendor list	Used on occasions Items not available elsewhere Prototypes, non-inventory items, office supplies	Guarantees or full inspection

Vendor rating systems are operated by most major companies to ensure that their suppliers meet specified criteria. The final result is usually to categorize the suppliers into these classes:

- *Certified vendor list* – these are the small band of good suppliers who need little supervision and minimal quality checks. They should include the major suppliers.
- *Qualified vendor list* – medium supplier performance, but causing some non-conformance costs. Acceptable but needs work to improve.
- *Approved vendor list* – this includes all the other suppliers. Their performance history is either low quality or unproven, so they require a quality assessment project to either upgrade their achievement or dismiss them.

Operating vendor rating

To establish and maintain vendor rating, a programme of vendor liaison has to be created, which should include not only a rating system, but also: frequency of inspections; inspection topics, focus and method; other visits and liaisons; the duration of the visits and methods of recording and relaying the results.

For those suppliers not given a top classification, the purchaser should report back on areas of concern, publicize to the suppliers the selection criteria and objectives, and agree an improvement programme.

When a rating system has been specified, it has to be actioned. It is, therefore, useful to specify in advance the conditions under which suppliers would be abandoned or supported. It may not be possible in all cases to apply these conditions because of restrictions on supply, but it is useful to have prime criteria defined for the best working partnership.

The objectives of vendor selection are primarily to ensure adequate supplier performance, and an agreement on financial rewards and penalties is very useful to make it work in practice. The supplier gets a premium if the rating is excellent and the purchaser gets a discounted price if the rating is poor. The overall effect of this can be to reduce the total quality cost between the two companies. As a spin-off from vendor selection further benefits accrue, namely:

- fewer suppliers
- certified suppliers
- objective performance data
- long-term business relationships
- better understanding and dialogue.

Supply partnerships

Organizing a partnership

The modern relationship of supplier and customer is considered in terms of:

- long-term collaboration
- open relationships
- mutual development
- supply chain management
- quality conformance.

There are also some practical considerations including geographic location and flexibility.

The supplier base consists of a wide variety of different types of suppliers providing individual blends of services. These services should provide what the purchaser requires but, of course, there is always room to improve in one way or another. Some suppliers are retained because they are essential (particularly where one part of an organization is purchasing from another); others are retained because they are excellent. There is a continuing need to develop supplier relationships, to secure improvements and, sometimes, to seek alternative sources where suppliers are unable to meet the requirements.

When selecting possible vendors, both long-term and short-term considerations should be used. What information is there on the supplier, and do they meet, or are they capable of meeting, the supply quality standards? It is important to ensure that the proposed demand level will not overload the suppliers, causing them to fail. If the style, objectives and business ethic of the supplier match the customer's, then a partnership is more likely to be successful.

> Partnerships give long-term success.

The main reason for supply partnerships is to improve effectiveness for each of the participants through better inventory management. Improved mutual trust means better forecasting, less safety stock, availability of stock records to each party, shared risk and, therefore, low inventory. Successful partnerships agree on attainable targets which stretch both partners to give an equal share. The effects go outside inventory management into sharing development skills, operational and financial data, and providing support to each other.

The process for developing a new supply partner is usually:

- Select vendor.
- Assess expertise.
- Accredit.
- Agree supply type.
- Test performance.
- Set up agreement.
- Monitor performance.

This monitoring is then fed back to the supplier and maintained on a continuous basis.

Pricing method

Because of the huge variation in market and trading conditions, the purchase price can be fixed in a number of ways:

- market price – variable from day to day
- fixed price – optional supply channel, mandatory supply channel
- market adjusted price – using an agreed formula
- fixed term contract – with discounts.

Each of these has advantages for supplier and customer in different markets, and the choice should be one which both partners consider to be fair.

The ordering process

Planning orders

There are two separate aspects of procurement to be considered, namely:

- purchasing
- reordering.

A large proportion of inventory control in a mature business will be the reordering of supplies from an agreed vendor, and begins with a communication that a further delivery is required. This situation is more about material control and scheduling than traditional purchasing.

If the purchasers can separate out those items where there are routine orders from those which are not likely to be repeated, then a different focus can be made in the initial negotiations.

The strategy for supply should take into consideration the overall commercial, logistic, inventory and added-value services which are to be provided by the supply chain, and a variety of purchasing options can be considered, as appropriate. Supply can be obtained from:

- the open market
- a manufacturer
- a dedicated provider (specialist or distributor).

Each of these supply sources has different characteristics which should be examined to ensure their use is right for the items to which they are applied (see section on order types below). Current practice is to set up a supply pipeline which goes through as few stages as possible. However, the nearer the original source (manufacturer) of the item, the less flexibility there is likely to be, the longer the lead times and the higher the supply volumes.

Order types

With the increased pressure for effective supply, simple one-off purchasing processes are largely irrelevant. No longer does a buyer act as a dealer, seeking a number of suppliers, negotiating price and delivery, and ordering a batch. This is far too time-consuming and ineffective for the bulk of stock items.

> Scheduling suppliers is part of inventory control.

Ordering practices separate the purchasing aspects from the reordering aspects. Purchasing is strategic and carried out at intervals to set the structure, volume and price of supply. Reordering is carried out as a result of stock control and forecasting processes as necessary for managing replenishment. The two aspects require different skills and can be carried out by different individuals. Supply orders can be broadly classified into:

- one-off orders
- bulk purchases
- annual contracts
- split delivery purchases
- rolling schedules.

An analysis of these types of orders and their application is shown in Table 8.4. The more organized companies use JIT or scheduled supplies. Unfortunately, much of the software used in purchasing systems assume one-off orders will be placed, and so schedulers are forced to carry out extra work and make schedules look like split delivery orders. Rolling schedules and JIT supply are the only options which meet the real needs of both the supplier and the customer.

Table 8.4 Analysis and application of supply orders

Order type	Advantages	Risks	Application
One-off orders	No commitment	No continuity of supply	Capital purchase
Bulk purchases	Traditional method for low prices	High inventory; high risk	C class items
Annual contracts	Assured supply from awkward supplier	Need to forecast accurately one year ahead	Seasonal purchases (food)
Split delivery purchases	Lower inventory	Large stock cover or commitment	'Order' based supplier; C or B class items
Rolling schedules	Assured supply, low inventory, low commitment	Non-adherence to schedule, poor forecast adjustment	A class items, delivery more frequent than fortnightly
Just in time	'Zero' inventory pull system; short effective lead time	Delivery failure; unreliable forecast of demand	High volume limited range; responsible suppliers

Order quantities

If the aim is to minimize overall cost, by keeping order quantities low, efficient stock monitoring and goods received systems are required. If this necessitates more frequent ordering, then the purchasing workload is increased. However, an increase in inventory actually results from the delivery, not from the ordering, and the prime task must be to reduce the delivery quantity. Sometimes this is easy where reducing the order quantity is more difficult. There are several approaches, including:

- supply scheduling (see next section)
- multiple item deliveries
- source from suppliers who sell in small quantities.

A reduction in delivery quantities can be achieved in two ways, namely:

- increase the variety of items per delivery
- increase the delivery frequency.

The stock value is decreased in direct proportion to the number of lines delivered.

Consider the simplistic situation of a weekly delivery from one supplier who delivers one item one week and another the alternate week. They deliver a fortnight's worth of each item each time, which causes an average of one week's stock of each. Instead, it is agreed that half the quantity of each will be delivered every week The delivery quantity is now enough for one week and the average stock is then half a week's worth.

> Go for less orders and more frequent deliveries.

The stock is halved by simply splitting the order quantity between the items supplied.

As paperwork would be increased by the smaller deliveries, it is important to minimize the effect of the extra time spent by limiting this approach to A class items.

Pareto-based order quantities

The key decisions in stock control are when to order (discussed in Chapter 7) and how much to order because the larger the order quantity, the higher the risk of excess stockholding. The order quantity depends on the value and usage rate of the item, and purchase order quantities should be related to the predicted requirements. The use of the Pareto technique described in Chapter 3 provides the best balance between usage rate, costs and ordering effort.

The simple logic is to order A class items frequently to avoid stock value and order C class infrequently to avoid too many orders and deliveries.

A typical order pattern would be:

- A class: order enough for one week.
- B class: order enough for one month.
- C class: order enough for ten weeks.

This, of course, is not dependent on the lead time taken by the supplier. Deliveries can be arranged every day even if the lead time is four weeks, it just means that there are about twenty supply orders outstanding all the time. The lead time does affect the safety stocks and therefore the overall stock level.

> Get A class items delivered weekly or daily.

It is a gross assumption that a week's worth of stock is the appropriate order quantity for A class, but it appears to be generally correct across a wide range of businesses and markets. The main exceptions are fast moving goods where there is daily delivery and the equivalent balance in days is used as the order level. Companies where A items are delivered less frequently than fortnightly have large stores and unbalanced stocks, or they have several months' lead times for slow moving A items. They have the wrong idea.

Table 8.5 Pareto order patterns for 1000 stock lines

Class	Moving lines (%)	Order cover (weeks)	Turnover (%)	Average weeks of value	Orders per year[1]
A	10	1	65	0.325	5200
B	20	4	25	0.500	2600
C	70	10	10	0.500	3640
Total	100		100	1.325	11 440

Note: [1]For 1000 moving stock lines.

The effects of ordering pattern on the number of line orders placed per year has already been discussed in Chapter 3, a reduction from 12 000 orders per year to 11 440 per year. If the above order pattern is applied to the example provided in Table 3.8, then the values shown in Table 8.5 result.

The result, compared with monthly purchase for everything, is a reduction of stock from 2 weeks' worth to 1.325 weeks' worth (plus safety stock in each case) and a decrease in line orders of just under 5 per cent. This gives more orders but less stock than shown in Table 3.8, and is better balanced.

The ABC order quantities reflect the natural ordering practice for many inventory controllers, but the Pareto technique provides rules for controlling all the stock in an appropriate manner and makes it work better.

Scheduling supply

> Rolling schedules reserve supply and reduce risks.

The best arrangement for ensuring reliable supply and low stocks is the rolling schedule. It is commonly thought that parts which take a long time to arrive should be ordered in big batches. This is not necessary. Large savings in inventory cost can be made by receiving regular small amounts of the high turnover value items which have long lead times. For example an imported part may take two months from the order to receipt into stores. If there are 420 used per month, then an order could be placed every month for about 420. This would lead to a stock of 210 on average (plus safety stock). Alternatively, an order can be placed each week for delivery in two months' time from the date of each order. This would mean a delivery quantity of 100 each time and average batch stock of fifty. However it might be possible to place orders every day, in which case the batch stock would be reduced to an average of ten. (Distribution methods have to be altered to deal with more frequent transfers without incurring extra costs.) This is a significant reduction in stock and cost without any effect on customer service. The number of orders per year may increase, but the workload can be reduced instead by using supply schedules rather than single orders.

Rolling schedules are not fixed on an annual basis because this may lead to a problem at the end of the year where there is either a shortfall or a large unwanted amount still to be delivered. The best schedule is an agreement between supplier and customer to the supply of an agreed quantity over a number of weeks. Outside this period the customer indicates the expected requirement so that the supplier can organize capacity and materials to cover it. Although the supplier would prefer a long fixed schedule, while the customer would like very short commitment, the scheduling technique has many advantages when operated properly with the full commitment of both supplier and customer to keep to the schedule. They include:

- less risk for both customer and supplier
- reduced stocks for the customer
- reduced stocks for the supplier
- the opportunity to plan supplies
- regular flow of materials.

A further advantage of scheduling arises from a regular flow of orders where, instead of placing an order or a new schedule, the old schedule can be used and extended. This enables a continuous supply of items at the average rate without the need for any purchasing action. There should be a simple agreement between supplier and customer that the schedule quantity rolls forward unless there is specific instruction otherwise, so that supplies are ensured. The only reason for changing schedules should be to adjust rather than create.

Not only does the rolling schedule embody the principle of 'no change-no action' (the classic management by exception principle) but it improves the stability of supply. Even if the supplier's lead time varies, the scheduled quantity is not affected, nor should the regularity of deliveries be affected either. A schedule is essentially a sequence of delivery times, not order placement times, so the equivalent of lead time for schedules is the inflexible front end of the schedule. The only change should be in the length of the fixed part of the schedule. In consequence, as the lead time increases, the customer has a longer time ahead before the programme can be changed. Schedules solve many of the problems of ordering and minimize the risks.

To use schedules successfully, both supplier and customer have to be committed to the quantities agreed. There should also be no sudden changes

in schedule outside the fixed part of the schedule. These are caused by late or overcorrection to demand or stock levels and, if not avoided, will give a stop-go ordering pattern. When usage changes the schedules should be modified gradually. Minor action early prevents a major problem later on.

Accurate systems are required to control scheduled orders. Running orders and continual deliveries are included on the schedule, and the cumulative delivery quantities have to be correct with a check to show which scheduled delivery is actually being received. A computerized technique is an advantage where there are many items to schedule.

The economic order quantity approximation

Where there are no other guidelines to determine the order quantity a theoretical model can be used, the 'economic order quantity' (EOQ). This model considers batch size, Q, which results from balancing the cost of holding stock and the cost of ordering it. All costs are assumed either to vary with the batch quantity or to be invariant with the order cost. For an item with a cost of P and annual sales of A, the cost of holding stock is given by $\frac{1}{2} \times r \times P \times Q$, where r is the financing factor. The half arises through averaging the usage rate to get stock levels. r is usually 20 per cent to 30 per cent, because it takes into account the cost of borrowing the money for the stock and the running cost of the warehouse. The order cost, S, contains the cost of running the order department, goods receiving, and supply progressing and can be typically US$100 per order.

Consider an item with a usage of 420 per month (5040 per year) and a cost of $35. Using typical values of $100 for the cost(s) of raising an order and 25 per cent for the return expected on the stock investment, the ordering options are shown in Table 8.6. The costs for various order patterns show a minimum for fifteen orders per year, so this is said to be the EOQ.

> *EOQ does not work in reality.*

At the bottom of Table 8.6 is the actual minimum value, fifteen orders per year. This has been calculated using the formula:

$$Q = \sqrt{\frac{2 \times A \times S}{r \times P}}$$

Substituting the values in this equation gives the economic order quantity, Q, of 339, which suggests that the item should be ordered fifteen times per year.

Table 8.6 Balancing order and stockholding costs

Orders per year (N)	Weeks between orders	Order quantity (OQ) (A/N)	Ordering cost (OC) ($N \times S$)	Stock cost (SC) ($0.5 \times OQ \times P$)	Total cost (OC + SC)
1	52.0	5040	100	22 050	22 150
2	26.0	2520	200	11 050	11 225
4	13.0	1260	400	5 513	5 913
15	4.0	388	1 300	1 696	2 996
26	2.0	194	2 600	848	3 448
35	1.5	144	3 500	630	4 130
52	1.0	97	5 200	424	5 624
104	0.5	48	10 400	212	10 612
Balance					
15	3.5	339	1485	1 485	2 970

Although the EOQ theory gives simple answers, it generally gives too high a stock level and should only be used where there is no alternative. There are many reasons why EOQ gives the wrong answer in practice. These include:

- Order and stockholding costs are assumed to be fixed, but these costs should vary according to the stock and ordering situation.
- The valuation will give different results when estimating by full or marginal costs.
- The demand is assumed to be regular, and ignores batching and timing of issues which have a major impact on order costs in reality.
- Split deliveries and schedules do not fit well into the cost equations.
- The EOQ ignores the balancing of stocks which is the most important factor in manufacturing and many other cases.
- There is no allowance for co-ordination of orders for similar items where the order cost can be shared.

There are other more minor problems which arise from the impractical assumptions made by the model. The application of this EOQ technique is therefore in the quick assessment of the stock and ordering situation. It is not appropriate for use in manufacturing, nor should it be used as part of a stock control system.

As can be seen in the example, the total cost is insensitive to the number of orders per year near the minimum cost. An increase of order frequency from fifteen times per year to twenty times per year only increases the total cost by 4 per cent but decreases the inventory by 26 per cent. Even ordering every two weeks rather than every 3.5 weeks will increase the total cost by 14 per cent but decrease the inventory by 43 per cent as evaluated from the table.

Good techniques can reduce the order quantities below those given by the economic order calculation without increasing the costs. In many situations, batch sizes in the region of 60 per cent of EOQ minimize the actual costs. The best answers are given by considering the practical situation and keeping the delivery quantities as small as can be sensibly achieved.

Other order size considerations

There are situations where the supplier imposes a minimum batch quantity, or a penalty for ordering small quantities. This suggests that there is a mismatch between the supplier's perception of a 'reasonable amount' and that of the customer's requirements, and is a sign that the supplier is not interested in the amount of business which is available. Good liaison between purchaser and supplier could set up a rolling schedule to assure the supplier of continued custom, albeit for small quantities.

Minimum order quantities, as set by suppliers, are generally arbitrary round numbers. It is possible to erode them by a few per cent in many instances, and gradually move to the required quantity.

Minimum order constraints on A class items should be studied energetically to avoid expensive stockholding. Where the minima are imposed on C class, then the solution may be to accept the constraint as long as goods are likely to be used in under a year. The minimum order quantity problem can be resolved in a number of ways offering different degrees of flexibility as shown in Table 8.7.

Table 8.7 Minimum order constraints

Constraint	Initial policy	Long-term changes
Minimum order value	Place order for phased delivery	Negotiate on volume of annual business
Minimum shipment size	Co-ordinate orders for widest mix of purchases	Change method of shipment
Minimum order quantity per line	Order for phased delivery	Schedule, or arrange payment when stock is used
Minimum delivery quantity per line	Share item with others (locations or competitors)	Renegotiate or look for alternative supply

Where the supplier is a sole source, then more effort has to be made to resolve the situation:

- Customers may be persuaded to accept an alternative.
- A second source may be found.
- The item can be sourced through a third party (avoiding significant cost increase).
- The item could be destocked and the minimum quantity supplied direct to the customer.

Negotiation and an understanding of the situation can sometimes lead to a sensible accommodation on both sides.

Apart from minimum order constraints there are other situations requiring specific policies. In many situations, the forecast will be considered to be acceptably accurate for a number of weeks ahead, up to a limit, commonly called the 'planning horizon' illustrated here for each. (*Note*: planning horizon is the length of forward planning. In Table 12.4 it is ten weeks.)

> Inventory should flow through the business.

Different situations give rise to different order policies:

- *Single warehouse* – forward cover to horizon fixed by Pareto class.
- *Standard product* – cover to horizon against forecast.
- *Distributed stockholding* – co-ordinate supply to all stores.
- *Manufacture to order* – buy and make required quantity plus scrap allowance.
- *Parts for assemblies* – make balanced sets.

In manufacturing, batch quantities should be the same for purchase and production where possible. The actual batch sizes are often determined by the practical limits of capacity on a process, the overall process throughput time and the physical need to make a batch a more easily handled quantity. By co-ordinating the manufacture of like products the batch size of each can be reduced because the set-up costs are reduced. Thus a slow-moving item should be produced in small batches along with similar fast-moving items.

Purchasing processes

The stages of purchasing for a reorder are usually as follows:

1 Usage causes requirement.
2 System identifies need to purchase.

3 Need is assessed and actioned if necessary.
4 Purchase is initiated.
5 Order is produced.
6 Order is authorized (if high value).
7 Order is sent to supplier.
8 Supplier initiates supply.

If the workload in purchasing is too high to allow proper vendor development and liaison, companies must find ways to reduce the ordering workload by attending to the eight stages and rationalizing them, as follows:

1 *Usage causes requirement.* This is the initial step in the reorder process and the whole supply process is dependent on the information being accurate. The first issue is to ensure that the stock and demand information is correct.

2 *System identifies need to purchase.* Much time can be wasted searching through printouts, screens or even stores, to see what is really required. Systems should identify by exception reporting those items considered to be ready for reordering. The reporting of this should err on the safe side by selecting all items which need ordering plus extra items whose demand patterns may lead to a high usage, a decision on which can then be taken by the inventory controller.

It is always useful to have a communication from stores which reports what they consider to be at risk. This avoids orders getting overlooked and adds weight to the supplier-expediting process.

3 *Need is assessed and actioned if necessary.* The information provided by the system should initiate orders for a minimum of 85 per cent of the items where orders are recommended. If this is not the case, then the stock control system is inadequate and requires improving. The development of good stock control systems through forecasting or MRP is not a complex computer problem. Solutions are available for most situations, including sporadic demand and seasonal sales.

4 *Purchase is initiated.* The actual purchasing process can be a simple confirmation at the press of a computer key, or a time-consuming discussion with one or more suppliers. Some companies even require a quotation for routine supplies! For fast-moving items, the ordering process should be done in a flash once the order quantity has been decided. (For most items the order quantities should be determined using the formulae discussed previously.)

5 *Order is produced.* Order production is part of the normal computerized purchasing system and is printed automatically. Manual orders are to be avoided since they complicate the system.

> *Organize routines minimizing purchasing workload.*

There is a good argument for a petty cash purchase system where over-the-counter purchases of low value are carried out by an authorized person outside the normal purchasing system. This will avoid a large bulk of formal purchases and immediate access to the items required (often maintenance parts). Control over these purchases should be through the monitoring of monthly purchase values.

6 *Order is authorized (if high value).* For capital goods and non-routine purchases, there has to be executive involvement in the commitment of company money. For routine supplies, this process only serves to delay the purchase, and is a great disadvantage. The signing of reams of purchase orders by a senior person wastes their time and usually the orders are not investigated during this process. There should therefore be generous spend limits for buyers. Executives need to control the purchases by:

(a) monitoring the value of order placed and outstanding order commitment weekly or monthly
(b) signing the major value orders (say twenty per week)
(c) auditing the general order level by random checks each week
(d) setting targets for reduction of value of purchases relative to demand.

7 *Order is sent to supplier.* By post an order can take two days to reach the supplier, longer for overseas suppliers. The use of fax or more modern methods enable the supplier to receive the information immediately. The use of e-mail, wide area networks (WANs) or the data superhighway is important for ensuring fast ordering. This saves stockholding as well as improving communication. The well-organized purchaser will not have to use the telephone for normal routine orders. The telephone is time-consuming, and should be reserved for supplier liaison and negotiation.

8 *Supplier initiates supply.* The information sent to the supplier should then be processed into their supply system. Any delivery problems should be relayed to the customer without them having to prompt. In situations where the customer is apt to alter the demand, there is a tendency not to report difficulties.

Key points

- Buying consists of creating strategic contracts and supply scheduling.
- Close co-operation with suppliers pays dividends.
- Single sourcing is best, but risks have to be identified and ameliorated.
- Quantitative assessment of suppliers will identify potential problems.
- Organize ordering to minimize delivery quantities and effort.
- Use Pareto ABC as a basis.
- Avoid the use of EOQ.
- Review purchasing control and administration to reduce work.

9

Forecasting demand

- The role of forecasting
- Types of forecasts
- Avoiding forecast data error
- Managing data for forecasts
- Benefits of group forecasting

Options for assessing demand

Buying safety stock, or purchasing a batch quantity, means making a prediction that the items will eventually be sold. This is a crude type of forecast. In order to improve the stockholding (increase availability and reduce stock) the quality of the forecast has to be improved. If forecasts were excellent and the supplier could predict when the customers would want items, they could be obtained and delivered just when they were needed: Good forecasting means low stock. Poor forecasting means high stock.

Forecasting should be based on data which is accurate and appropriate for the purpose. This is often a problem, and a forecast frequently has to rely more on balanced estimates of future sales rather than history. The data used to calculate the forecast include established demand patterns but ignore irregular demands for which stock is not normally held, such as scheduled customer orders (for which specific purchases are required to cover the demands). For example:

- one-off promotions
- sales campaigns
- one-off very large orders (generally orders where the demand is more than four MADs).

Good forecasting results in low stocks.

The selection of the forecasting technique is a difficult decision and either the inventory controller selects the most appropriate one or employs a focus

forecasting technique. In general, it pays to use the most sophisticated technique available because it is a better model for the demand pattern and gives better results. Forecasting techniques can be divided in many ways. One simple classification is between forecasts which use prediction and those which use history, another is between qualitative and quantitative figures, and another between intrinsic forecasting and extrinsic forecasting.

Considering these concepts in turn, historical forecasts rely on looking back to predict the future. Historical forecasting forms the basis of most inventory forecasting because it has sufficient detail, and demand for products usually changes continuously. However, it is obvious that looking back to go forward could lead to disaster and this is where predictive forecasting plays a role. People will predict an event based on understanding of future changes, rather than extending what has already happened. The problem with predictive forecasting is that it is often qualitative. A sales campaign will increase demand, but by how much? The qualitative prediction is marginally useful, but for stock control a quantitative assessment is necessary.

The difference between intrinsic and extrinsic is that the intrinsic forecasting is based on information which is available within the system. Extrinsic forecasting uses information which comes from external sources such as trade statistics or market information. It is convenient to split extrinsic information into leading indicators, concurrent indicators and lagging indicators, according to whether the information collected shows future, current or past activity. For instance, government statistics on housing starts is likely to be a useful leading indicator to a national roof tile manufacturer. The same information would be a concurrent indicator, showing what is happening to the general demand this month, to a company hiring digging equipment for house foundations, and for architects it would explain the level of activity which they had already experienced. Obviously, leading indicators are the most useful, even though the information provided is often qualitative.

The manipulation of forecasts is part of the inventory manager's art, since computerized systems are unable to provide the right interpretations when conditions change.

Types of inventory forecasts

Some types of forecasting are used only for specialized purposes. Of those used regularly by inventory managers, the following five approaches are the most useful and form the basis of stock control:

1 *Market research.* There is nothing like knowing for sure what customers are going to buy. Firm orders are the ideal! Sometimes it is possible to obtain firm order commitment, but more often it is possible to gain only estimates from customers. Good liaison with the biggest customers, or the biggest users of A class items, can have a major impact on forecasting accuracy. If customers trust a supplier, they will provide accurate information, since they should be in the best position to know their market and their likely order requirements.

An alternative would be to carry out surveys to discover trends in the market. These can be particularly useful in one-off investigations for new products or for determining the effect of different sales or promotional policies. It is a skilled and complex job to devise market surveys which give factual results for calculating the demand level. A market survey is only cost-effective as an ongoing exercise for a narrow item range where the volume is high.

Another method of assessing customer requirements is to ask the sales people. The risk with this is that sales staff sometimes confuse the difference between maximum potential sales and expected average sales. Consequently, their estimates often have to be interpreted before use!

2 *Market demand models.* These are generally based on knowledge of the customer's markets. If the major factors which cause the demand can be identified, then a model can be created. The influence of each factor can be assessed and put into an equation, or the factors can be input into a simulation which can develop its own parameters.

> Choose the right forecasting technique for each product.

Models are very effective where there are leading indicators, and where there are simple deterministic relationships causing demand. Some models can be financial (expenditure on promotion versus increased turnover), others technical (item needs replacing every × years) or based on commercial factors.

3 *Historical techniques.* Forecasts should be based on historical data, and then modified by the influences. The only exceptions to this rule are for new products, and those which have a causal link to another forecast (via a model or MRP).

There is a wide variety of historical techniques to choose from. Some are simple and give adequate results, others are much more sophisticated and give excellent results in the right circumstances. In general, the more sophisticated the technique, the more data is required to realize its potential, so established products can be forecast best using these. Also,

a weighted average technique will be better than a simple average. There are many approaches to historical forecasting, favoured by a variety of authorities. The basic models for these will be discussed in the following chapters.

4 *Minimum stock levels.* Many systems, both manual and computer have a 'minimum stock level' field displayed. This field is usually used for the 'reorder level' instead of for the minimum stock, and is either a user input field or fixed by some useless rule (such as cover for 'x' weeks).

With a large range of stock items, it is a very big task to maintain stock levels up to date. In practice, where the minimum is too low, the stock controller will increase the minimum level, but where the stock is too high, there is no availability problem and so fewer adjustments are made. As a result, the stockholding can easily increase gradually.

Minimum stock levels should be avoided and proper safety stock levels should be recalculated on a regular bases by a computer system. An area of development of stock control has been the improvement in forecasting techniques used by the professional stock controller. These should continue to be used to the maximum.

5 *Demand patterns.* The demand for stock items often follows a pattern which results from a variety of customer requirements for each item. There are two general approaches predicting demand patterns:

(a) treat the usage as a pattern and analyse the demand profile (an intrinsic approach)

(b) investigate the causes of the demand and forecast the requirements separately (an extrinsic approach).

The first alternative has the advantage that it is based on accessible information and is quick and easy to apply. This type of modelling does not normally predict rapid changes to the usage rate because it is essentially based on the historic demand pattern.

If a change in demand is understood, predictions can be made of likely usage and total likely requirements. This method should provide a more accurate answer, but requires thorough communication, knowledge and management.

> Predicting for a group is more accurate than for each item.

In practice option 1 based on historical data is used for the majority of items. High-value items and critical items are usually treated by the second method, often through an informal system, although such systems should be avoided if possible in favour of extrinsic forecasting which will manage stock more efficiently.

Product life cycle

For a company's marketing and manufacturing departments, a product has a life cycle of three phases – design, development and manufacture. The end of manufacture is determined by marketing strategy and the introduction of a new more competitive product. This product life cycle is shown in Figure 9.1. It should be noted that product life cycles are getting much shorter with new technology. The product life cycle for the mechanical typewriter was more than twenty-five years; for the modern word processor it is frequently less than one year.

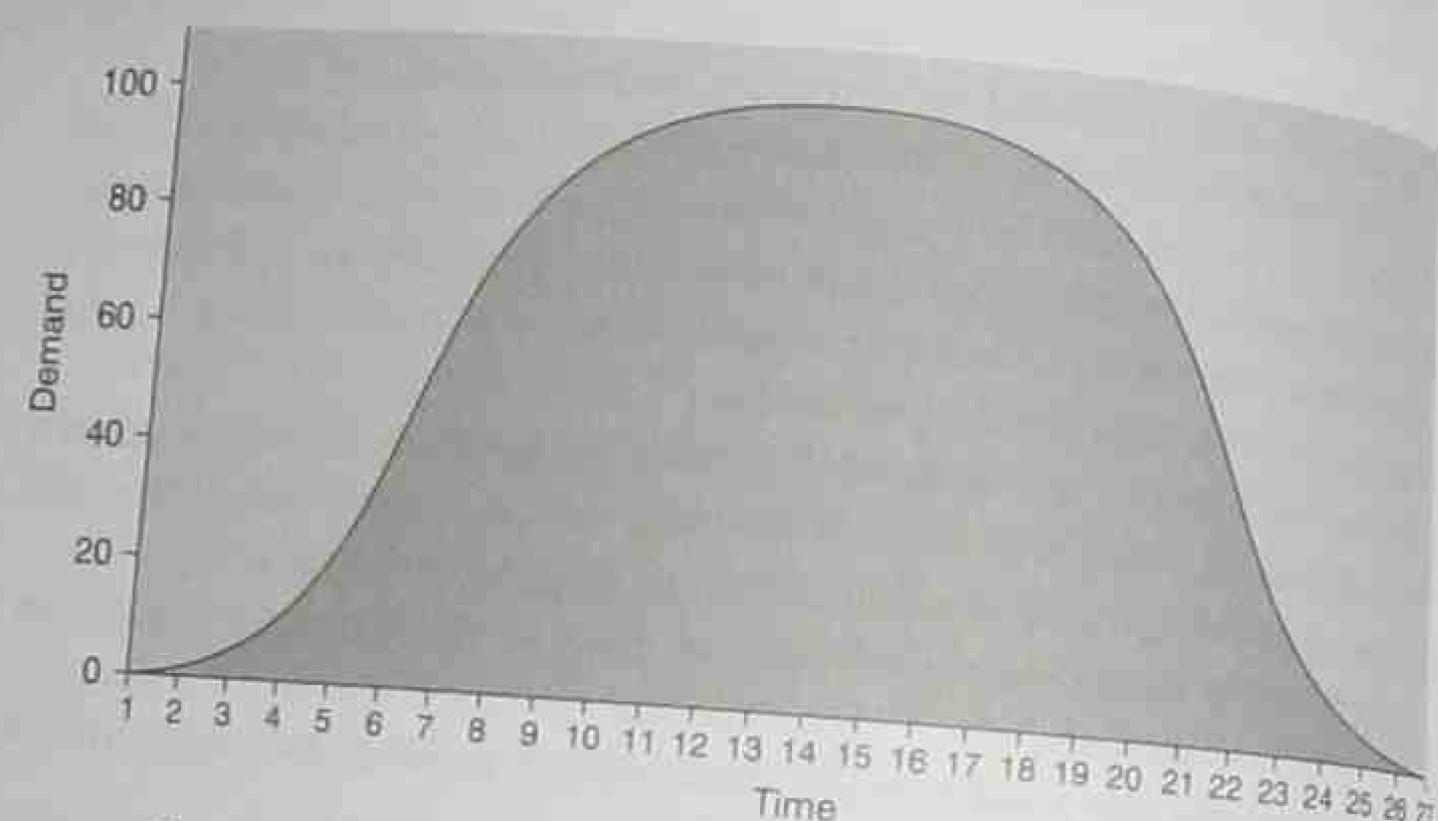


Figure 9.1 Product life cycle

For a service organization, the need for stockholding may extend beyond the product manufacturing cycle since customers will retain products for many years after the end of manufacture and still expect spares to be available to support a service for their equipment. For some companies, products manufactured more than twenty years ago are still in use and being repaired as required. In some cases, service organizations have announced stop dates for service in conjunction with sales initiatives to sell customers new products. However, this has frequently failed since third party maintenance organizations have seen an opportunity and taken over a segment of the business which is very lucrative.

The product life cycle for the service organization is different from the marketing and manufacturing life cycle and can be divided into three phases:

- introductory phase
- supply phase
- obsolescence phase.

Introductory phase

During this period, which should coincide with the design and development phases, the inventory planner will be working with the product project team, when the customer service strategy for the product should be decided. To determine the stock requirement, the inventory planner will want to know the forecast demand and the possible geographical distribution of customers, which will enable the planner to determine the requirement and place time-phased orders for the items necessary to support the market strategy for the product ahead of the launch date.

Product launch is a critical period in the life of the item and it is essential to have sufficient to support the first customers. However, if too much is bought initially, the profit can be seriously impacted because much of the initial purchase may never be used if too much 'just in case' inventory is bought. The planner must monitor carefully early life usage as the sales build up and adjust his ordering accordingly.

Supply phase

Established products have a steady or slowly varying demand pattern, and during this part of the product life cycle major demand of stock items occurs. Inventory control is concerned at this time with maintaining availability. The ongoing requirement can be determined from historical usage data. Standard stock control systems with provisioning facilities can be used for reordering replenishment stock. At some point during this period, manufacturing may cease which will create a 'last time buy' opportunity for distributors. Discretion is needed to ensure excessive 'just in case' inventory is not bought at this time.

Obsolescence phase

It is important that the inventory planner is involved as the requirement for the product declines. The task now is to prevent a build-up of surplus stock, and the stock control system should order fewer replenishment parts as the usage fails. However, a large proportion of inventory will still be in place to meet the requirements of customer service. When the product demand ceases then all the inventory should be withdrawn promptly. Companies find this a valuable exercise since it provides an opportunity to redeploy inventory in other countries where the product life cycle is in its introductory or supply phase, and not its obsolescence phase.

During this obsolescence phase, consideration should be given to the relative importance of customer support and the high risk of having to write off unsold stock.

Causes of forecasting inaccuracy

People sometimes say 'you can't forecast this' – a statement which is untrue. Everything is capable of being forecast, but some things can be forecast well and others are very difficult to forecast. To improve the quality of forecasts, and therefore of stock control, it is useful to review what risks there are and what causes forecast inaccuracy. Typical pitfalls for the forecaster are as follows:

- inaccurate data
- sales information rather than demand statistics
- bias
- speed of response to change
- poor assessment of supply capability
- inclusion of extra demand in the forecast
- shortage of data.

Inaccurate data

The whole of inventory management depends upon knowing how much stock is available so that the appropriate controls are maintained and actions taken. Inaccurate data leads to poor availability, unreliable customer service and excess stocks. It is therefore imperative for stores to maintain records accurately, and not just to maintain records by a weekly or monthly physical stock check. Records should be maintained accurately through a variety of techniques including:

- automatic recording and checking
- batch control
- stores empowerment
- perpetual inventory checking
- bar coding.

If records are not accurate, then it is not worth improving inventory management processes until inaccuracies are small enough to be a minor influence on customer service.

Sales information rather than demand statistics

When considering inventory statistics, it is normal to discuss 'demand'. However, most companies' records focus on 'sales'. The difference is only important where the actual amount issued does not correspond with the amount ordered. This could happen in a number of ways, including sending boxed quantities, or when there is a stockout. In the case of a stockout, the lack of sales could lead to the conclusion that no stock is required if no notice were taken of the actual demand.

Professional stock records should therefore monitor 'demand' (order requirements at their delivery dates) rather than just sales.

Example

Taking an extreme example: if an item has a demand in three months of thirty, forty-five and forty, but there is no stock of the item in the first two months and then a delivery during the next month, the sales records would give monthly sales of 0, 0 and 115. The actual fluctuations in demand are only about ten whereas the sales statistics would give a much higher value to month 3.

Bias

Sales tend to use a forecast either as a motivator to the sales team by exaggerated optimism, or to ensure that stock is available. As a consequence, projections from sales departments are notoriously high. Stock controllers have to resolve this basic issue either by interpreting sales data during forecasting, or by making sales (or marketing) departments more responsible for the accuracy of their forecasts. This can be done by apportioning financial responsibility for excesses.

> Get the data right, then you can forecast.

Speed of response to change

It is difficult to get the right compromise between overreaction to events and lack of response. One or two periods of high demand could signify an increase in market demand or simply a statistical variation. Either assumption will cause risk. Similarly when considering what time intervals to use for collecting data, if the time period is long (e.g. a month), there could have been major changes in customer demand profile within that time or, for short periods, demand can be erratic and make it difficult to interpret trends.

Poor assessment of supply capability

Fulfilment of increasing demand may be restricted by availability from suppliers. This can happen even with good suppliers who are used increasingly until they become overloaded and cease to supply successfully. This will result in forecasts which are not achieved. Supplier restrictions have to be understood by the forecaster. If a supplier cannot cope with peaks in demand, then there is no point in including them in forecasts for peak rates.

Inclusion of extra demands in the forecast

Special offers, campaigns and fixed scheduled orders should not be included in the routine forecasts. They add variation to the demand rate which is not true random demand. Instead, there should be planned supply to meet the requirement which is assessed separately. For example, a demand pattern may be:

20 510 30 5 25 520 15 35 25 515

and the reason for this may be one major customer who requires 500 every four weeks plus a variety of customers who are taking about fifteen per week. A sensible policy is to stock for the many small demands and to order in specially for the 500 to arrive just before it is required. The 500 would therefore not be included in the data for forecasting but as a separate type of demand in the item stock record.

The ability to forecast closely to the actual demand rate will depend on the selection of the method of forecasting. This can be done automatically by the computer from the historical models, but staff may also possess additional information which could be significant in forecasting future demand.

One-off abnormal demands can be excluded from forecasts by using a filter set at, say four MADs so that demands outside this range are treated as non-stock sales.

Shortage of data

For a 'good' historical forecast, sufficient data is required. It would be sensible to go back into history until the demand became different but it is often difficult to assess this from the data. For seasonal sales, it is essential to have data for more than one year.

The smaller the amount of information available, the less accurate will be the forecast. For example, the statement 'last two months' sales were 500' gives very little information on which to base stock levels. If the information available is 'the sales were 300 for month 1 and 200 for month 2' then this starts to give an indication of the demand pattern. If weekly demand data were available, then it would be possible to see whether the demand was:

- only two customer orders or sporadic demand
- regular orders each week or each day
- decreasing demand.

This suggests that it is better to use more smaller time periods rather than large ones, unless the demands in the time periods become non-standard (e.g. if demands on Fridays are different from the rest of the week, daily time periods would give cyclic demand with a peak each Friday). How much data is needed to make a decision? The answer is indicated in Figure 9.2. The size of the forecast error depends on how erratic the demand is –

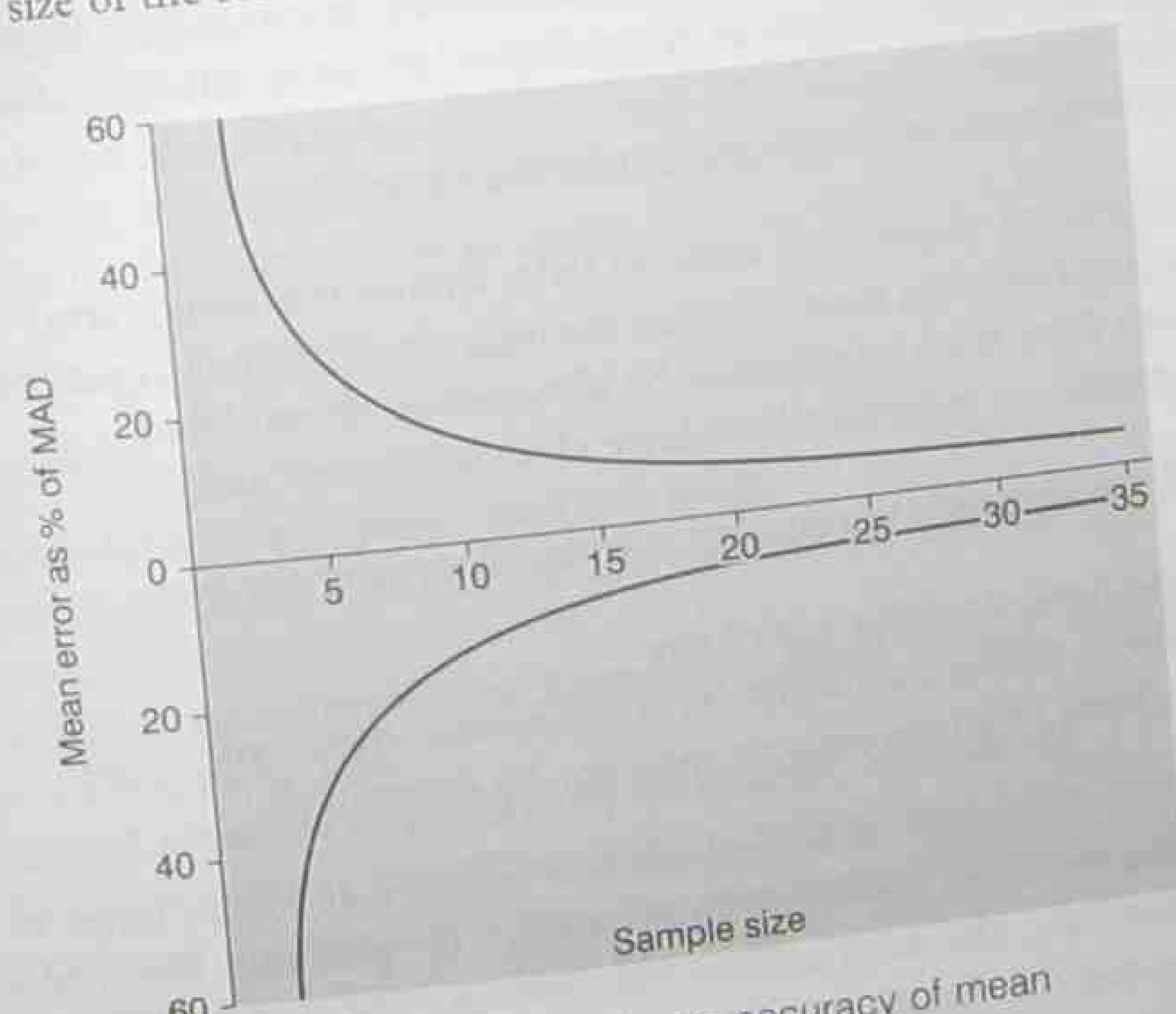


Figure 9.2 Effect of sample size on accuracy of mean

as indicated by the vertical scale. Figure 9.2 suggests that there is a reasonable degree of accuracy afforded by twelve periods of demand history.

> Work with the data available – it's the best you have!

Methods of improving forecasting

Feedback and monitoring

The simplest improvement to forecasting comes from ensuring that the forecasts are modified to reflect actual demand. This simply requires the re-evaluation of forecasts at the end of each period. New forecasts have then to be applied to inventory calculations to optimize the stockholding. Problem items can easily be identified in each period using a stock cover exception report (see Chapter 3).

Feedback from the forecast error can show where the forecast technique is lacking and can identify when the forecasting method is not coping and needs changing.

Data aggregation

Statistically there is a major benefit from using as much data as possible by going back further in time or spreading wider geographically, locally and globally.

The obvious risks of using old information are:

- undetected population changes have altered the usage rates
- demand patterns have changed through modifications in user practice
- items have been substituted by alternatives or modified
- contracts or customer base have changed significantly.

If a wider geographic population is used, then the risks are:

- usage patterns differ regionally
- different demand patterns (or modification mix) exist locally
- data from different sources may not be compatible.

It is therefore a matter of assessment and balance as to the range of data which can be used for individual items. In general, the wider the information base, the more reliable the forecast. The accuracy improves as the square root of the amount of data included.

Comparative items

For a high usage item, random fluctuations are small compared with the forecast demand. Where demand is small, fluctuations can be large compared to the forecast. Alternatives to aggregation are, first, to find a similar item which is expected to have the same demand profile and, second, to vary the forecast for a minor item by scaling the forecast for a major item.

> **Example**
>
> *Consider two items in the same product range, one with a demand of 100 per month and the other with a demand of 2 per month. The forecast for the major item shows that the forecast has drifted from 100 per month to 110 per month. Assuming that the profiles are similar, the demand for the minor item should increase by 10 per cent as well – to 2.2. It is much more difficult to detect confidently a change from 2 to 2.2 than from 100 to 110. It does not matter whether the two items change in exactly the same way. This technique works if the minor item is more likely to follow the major item than to remain static.*

If there are a large number of items in a product range with low usages, comparison with the major items can be a very good way of allowing for changes in the underlying level of demand. The forecast for each item can be based on individual historical demand, and the resultant forecast multiplied by the relative change factor used on the major item.

> **Key points**
>
> - Better forecasting gives low stock levels.
> - There are many techniques and options so several should be used.
> - Forecast can be predictive, based on product or sales information, or on historical data.
> - Record accuracy is paramount for realistic forecasting.
> - An ongoing effort should be made to keep the records accurate.
> - Extending the amount of data (by aggregation or copying profiles) increases the reliability of the forecasts.

10

Historical forecasting techniques

- Historical averages
- Better forecasts with weighted averages
- Choosing the weighting factors
- Choosing the best forecast
- Warning when forecasts are bad

Basic forecasting techniques

Forecasting

Minimum stocks are possible when a good knowledge of demand exists. If demand were known accurately enough and far enough in advance, then no stock would be necessary. On the other hand, if there is very little market information, high stocks are needed to ensure reasonable customer service. Forecasting is therefore one of the important aspects of stock control. It is a complex art which embraces many factors.

There are two basic approaches to forecasting:

- assessing future market requirements
- using demand history.

Assessment of demand

Assessment of demand requires the knowledge of customers, products and background conditions. That comes from discussion and evaluation with users further down the supply chain and through techniques such as market research, questionnaires and customer surveys. The strength of these techniques lies in providing general information on the levels of business and on the likely uptake of a new product, or the phase-out of an old one. This information is invaluable where large changes in demand level are likely to occur.

These techniques need to be better managed to make them more acceptable as forecasting tools for each line item, and inventory management need to work with sales departments to ensure that they do not draw optimistic conclusions – higher demand – which then has to be second guessed for planning stocks. However, assessment of the market demand is an essential tool for the professional stock controller, and the technique should be used for A class items and those at the start and end of the product life cycles.

One method of determining customer demand is to assess potential sales. Many companies produce estimates, enquiry responses or have contracts and formal agreements. Using a proven conversion factor on this information to calculate an expected average demand by item number can be the best way of estimating demand. Of course, a customer order could be considered as a market prediction, since there is always the possibility of them changing the date or quantity.

Demand history

The second approach to forecasting, using demand history, provides a convenient basis for predicting demand. For most items it is:

- readily available
- detailed by item number
- generally reliable
- easy to use.

However, using demand history is rather like rowing a boat. To go forward means sitting looking backwards and working hard. This is fine if there is no obstruction in the way, but it is advisable to look round from time to time to see what is in front. In a similar manner, inventory controllers use history to determine the future. They have the problem that they are steering many hundreds of items at the same time, and it is not practical to look round for each one to see if it is heading for collision. What is required is a general warning mechanism to detect impending disaster, and a forward look for the more sensitive and important items.

Historical forecasting works for the vast majority of items and is therefore a basic tool of inventory control. The more sophisticated the forecast, the better the results, but there is always the need to keep abreast of real changes in the market so that forecasts can be overridden if necessary.

Base forecasts on history.

In practice, a variety of simplified models, based on major features of the demand pattern, can be used to meet stock control requirements for standard items. More sophisticated forecasting methods are applied to the non-standard items.

Historical forecasting methods are based on the mathematical manipulation of historical data. This approach is fine for most items, and can lead to excellent forecasts where the demand pattern is consistent. It is preferable to off-the-cuff estimates because it is:

- consistent
- rapidly calculated
- frequently updated
- based on facts not reaction to pressures.

However it should be stressed again that it is desirable to modify forecasts in the light of additional information.

Moving average

The most straightforward method of forecasting the next period's sales is to take the average of sales for each preceding period, add them up and divide by the number of months. The moving average technique is a formal way of calculating this for each period in turn, taking into account a fixed number of preceding periods (usually weeks or months). Financial accounts have a fixed annual cycle in which history builds up from the start of the year. This data is unsuitable for stock control, since a fixed number of periods of history is required. The moving average can be calculated from:

$$\text{Average demand} = \frac{\text{Demand for N periods}}{\text{N}}$$

where N is the number of periods over which the average is taken. The selection of N is of prime importance in getting the best from the moving average technique. It is a compromise between the inaccuracy caused either by too little data (see Figure 9.2) or by using old irrelevant data. Often the average is taken over twelve months (or fifty-two weeks) to avoid complications due to holidays or seasonality. However, it has to be remembered that by averaging fifty-two weeks of history, the data used is on average twenty-six weeks old: a lot may have changed since then! Therefore it could be safer to use a shorter moving average.

Example

Take, for example, the demand history for three items over a twelve-week period, as shown in Figure 10.1. The average demand for all of them is 20.5 but the reliability of the figure 20.5 as a forecast for the next week could be different in each case. In Figure 10.1A the demand is in between 11 and 28, and 20.5 is quite a reliable forecast. For Figure 10.1B demand is more sporadic and 20.5 is not such a good estimate – perhaps more data would be useful. In Figure 10.1C there is a gradual increase in demand and 20.5 is rather low, so taking a shorter average would give a better answer in this case (i.e. a smaller number of periods).

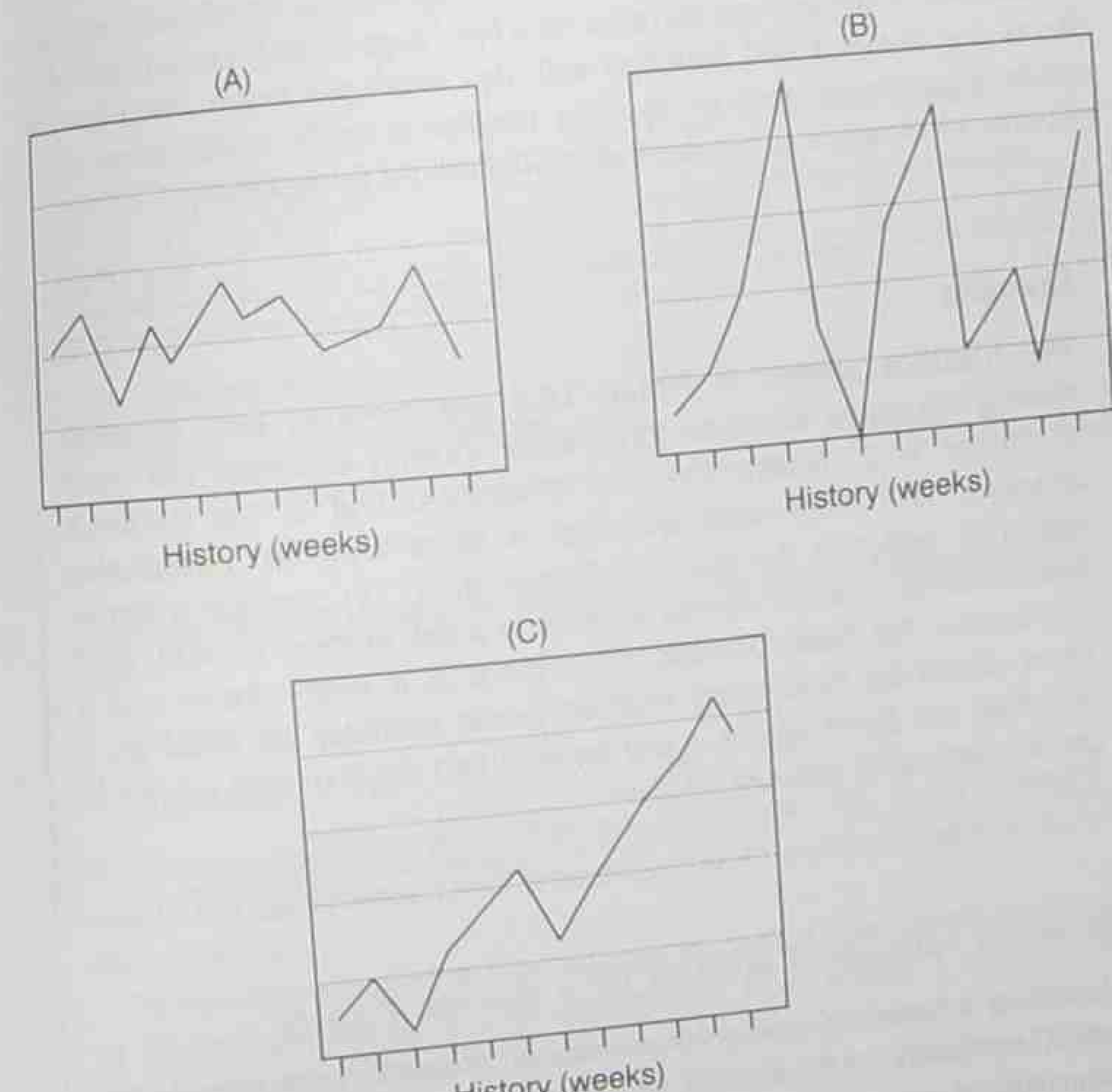


Figure 10.1 Historical demand patterns

In general a 'suitable' number of periods is chosen for a whole range of stock items. This is not the best option but a practical solution. Calculations including many time periods are best where underlying demand is static, where there are large variations between demand in each period (Figure 10.1B), where stock level is not too critical and the forecasting is allowed to run itself without much supervision. Reducing the number of periods in the average enables the forecast to cope faster with a changing demand level (Figure 10.1C). This is useful where the demand is smooth or where the sales in a period are dependent upon those in the previous period (autocorrelation).

> Each inventory line needs its own averaging technique.

Taking an average is really a method of segregating trends and random fluctuations. If the average includes data for a long period then any spurious effects are damped out, but this will also make the result insensitive to trends. Conversely a short moving average with N small, gives good response to changes in market but is affected by fluctuation.

Example

The histories plotted in Figure 10.1 are provided with six-week moving averages whereas three-week moving averages are used in Figure 10.2. Notice that the smoothest line is the six-week average. The three-week average is more variable, but it also responds better to the trend in Figure 10.2C. In practice a three-week moving average contains rather a small amount of data, so if it provides the best forecast, then there is a major trend and a more advanced forecasting method would normally be used or, if not, then the three weeks could be split into days or half weeks to get a reasonable amount of data.

Alternative calculation methods for a moving average

Calculating the moving average by summation month by month can be long winded, especially when dealing with twelve-month averages. There is a shorter way of generating the new moving averages.

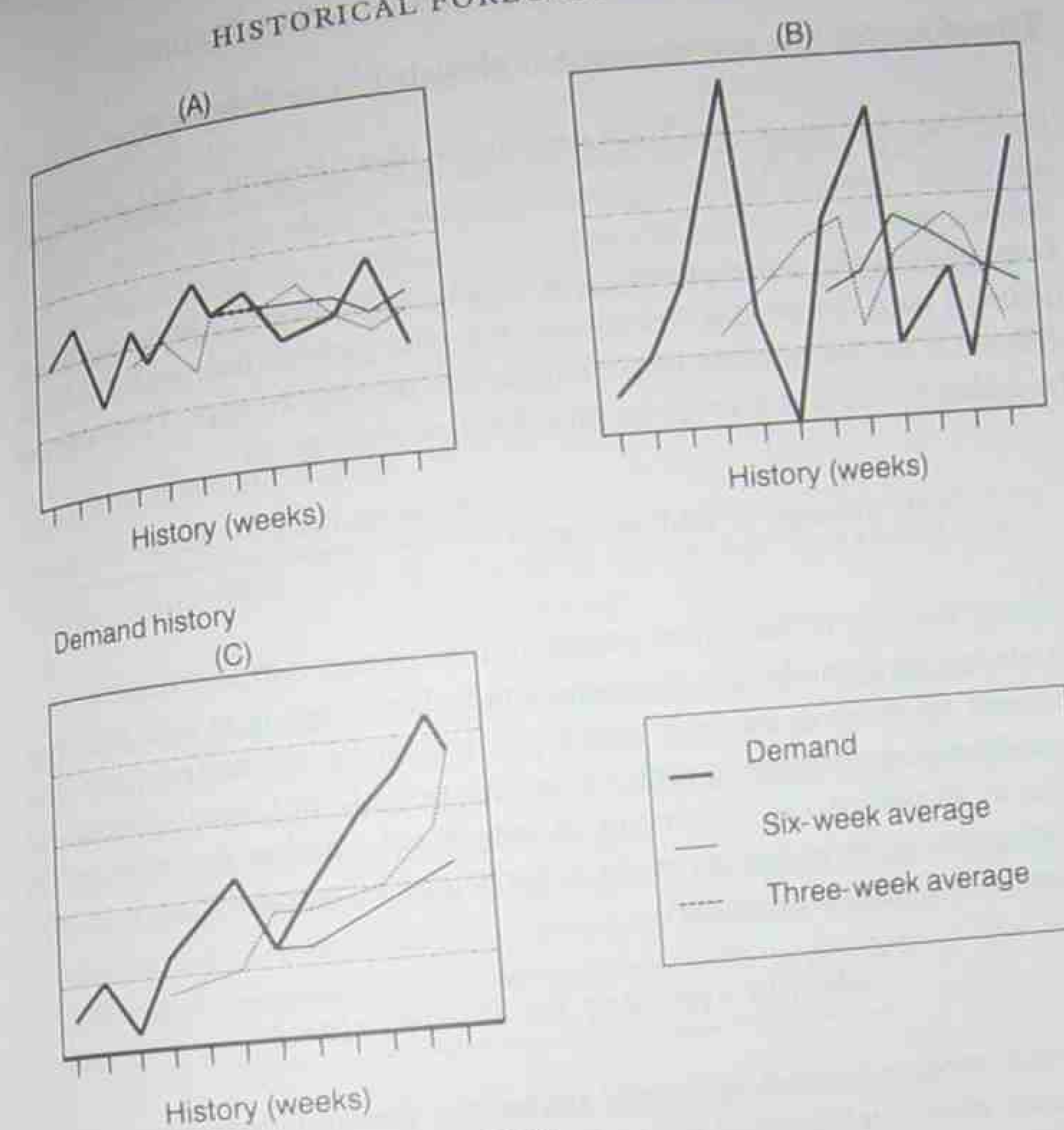


Figure 10.2 Moving averages

Example

Consider that the demand for an item over the previous seven months was:

Month	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Demand	*17*	*11*	*15*	*12*	*14*	*15*	*?*

The average demand, A_o = 14 for the first six months. In the following month the sales were 11. The new six-month average demand A_N is therefore 13.

To calculate the old moving average, A_o the total six-month demand of 84 was divided by 6. The demand for 17 in month 1 is not now included in the calculation, since it is older than six months. However a simple way of calculating that A_N is 78 was 84 − 17 + 11 = 78.

Therefore the new average can be calculated by working out:

$$\text{New average} = \frac{\text{Total demand} + \text{New demand} - \text{Oldest demand}}{\text{N}}$$

Even easier, and more practical, is to find the difference between the old demand data (to be discarded) and the new period data, and divide the difference by the number of periods in the moving average. Then this can be added to the old average to give the new average, or:

$$\text{New average} = \text{Old average} + \frac{(\text{New demand} - \text{Oldest demand})}{\text{N}}$$

Moving average is the most common forecasting method because it is simple to apply and easy to understand. However, it is not the most favoured forecasting method because it has to be calculated from data extending back over the month by month period, and equal weights are given to all periods. It is a matter of judgement whether the most recent information is more important in a particular instance. Often it is, and better techniques are required.

> *Moving averages for C class items.*

Moving average is most successful where the demand fluctuates widely, because more sophisticated methods cannot identify demand trends and profiles in the presence of sporadic demand. Moving average is often used for C class items because the demand for these is often very variable compared with the mean, and it is more important to have a forecast which is generally reliable rather than one which requires some management.

Weighted averages

Benefits of weighted averages

The moving average technique suffers from two difficulties. First, it does not respond well to changes, as it gives equal importance to all periods and, second, a large amount of information is required to recalculate the average each time.

An average which takes more notice of recent history and less notice of older data has major benefits. The use of a weighted average improves historical forecasting. There are many options for weighting the averages

but there is one option which is better than the others and is used almost exclusively.

The technique of exponential smoothing avoids both the problems outlined above. It produces a weighted average which is based on all the information available (see Figure 10.4). The calculation of a new average only requires the old average, the new demand and a weighting factor.

Exponential smoothing

> *The simplest proper forecasting technique is exponential smoothing.*

The name 'exponential smoothing' comes from the fact that contributions from history for an exponential curve go backwards in time. There is theoretically a contribution from many years of data, although this will become increasingly insignificant as it ages.

Forecasting with exponential weighting is the simplest of the professional forecasting tools. A new forecast demand, or historical average, is obtained by mixing a portion of the old average with a portion of the new demand.

$$\text{New forecast} = \alpha \times \text{Demand in Last Period} + (1 - \alpha) \times \text{Forecast for last period}$$

The mixing portion, α, is the smoothing factor.

A new forecast has to be calculated as soon as the data for the previous week or month (i.e. last period) has been collected. The new forecast then applies to the week or month starting at that time.

The weighted forecast is easier to use when written down mathematically:

$$A = \alpha D + (1 - \alpha B)$$

where

A is new forecast
B is the old forecast
D is the demand in the period just completed
α is the smoothing factor.

Table 10.1 Exponential forecasting

Month	Sales	Forecast ($\alpha = 0.1$)	Error ($\alpha = 0.1$)	Forecast ($\alpha = 0.3$)	Error ($\alpha = 0.3$)	Cumulative Error
1	10	10	0.00	0.00	10	
2	14	10.00	4.00	4.00	10.00	0.00
3	6	10.40	–4.40	4.40	11.20	4.00
4	8	9.96	–1.96	1.96	9.64	–5.20
5	16	9.76	6.24	6.24	9.15	–1.64
6	12	10.39	1.61	1.61	11.20	6.85
7	7	10.55	–3.55	3.55	11.44	0.80
8	11	10.19	0.81	0.81	10.11	–4.44
9	13	10.27	2.73	2.73	10.38	0.89
10	5	10.55	–5.55	5.55	11.16	2.62
11	10	9.99	0.01	0.01	9.31	–6.16
12	8	9.99	–1.99	1.99	9.52	0.69
		MAD = 2.74		MAD = 2.90		–1.52

'A' can be described as the 'average', which it is, although it is being used as a 'forecast'. The value of α is between 0.1 and 0.4 as discussed below.

The smoothing effect of exponential smoothing is illustrated in Table 10.1 and Figure 10.3. Like moving averages, this calculation simply analyses the demand to give an average plus the remainder which is treated as statistical fluctuation. The smoothing factor, α, defines the amount of notice which is taken of each period's demand.

Selection of the smoothing constant, α

The nature of the forecast can be changed just as for moving averages, by altering the value of the smoothing constant, α. Low values of α make the forecast consistent; high values make it reactive to change. Making it too high would render the result unreliable because it would depend too heavily upon demand in the last period. For effective exponential forecasting, the range of values which can be chosen for α are 0.1 to 0.4.

If the demand varies slowly then a low value of α, say 0.1, is suitable (see Figure 10.3). If there are changes in average demand level to which the forecast should adapt then a high value of α, say 0.4, is chosen.

If the variability of the demand is high from month to month, then a low α is required to show up a stable average. Where the demand is smoother a larger value of α can be used. This responds more rapidly to trends in the average. If little is known about the item, or a value of α is required for a range of items, then a median value of α = 0.2 can be used initially. The

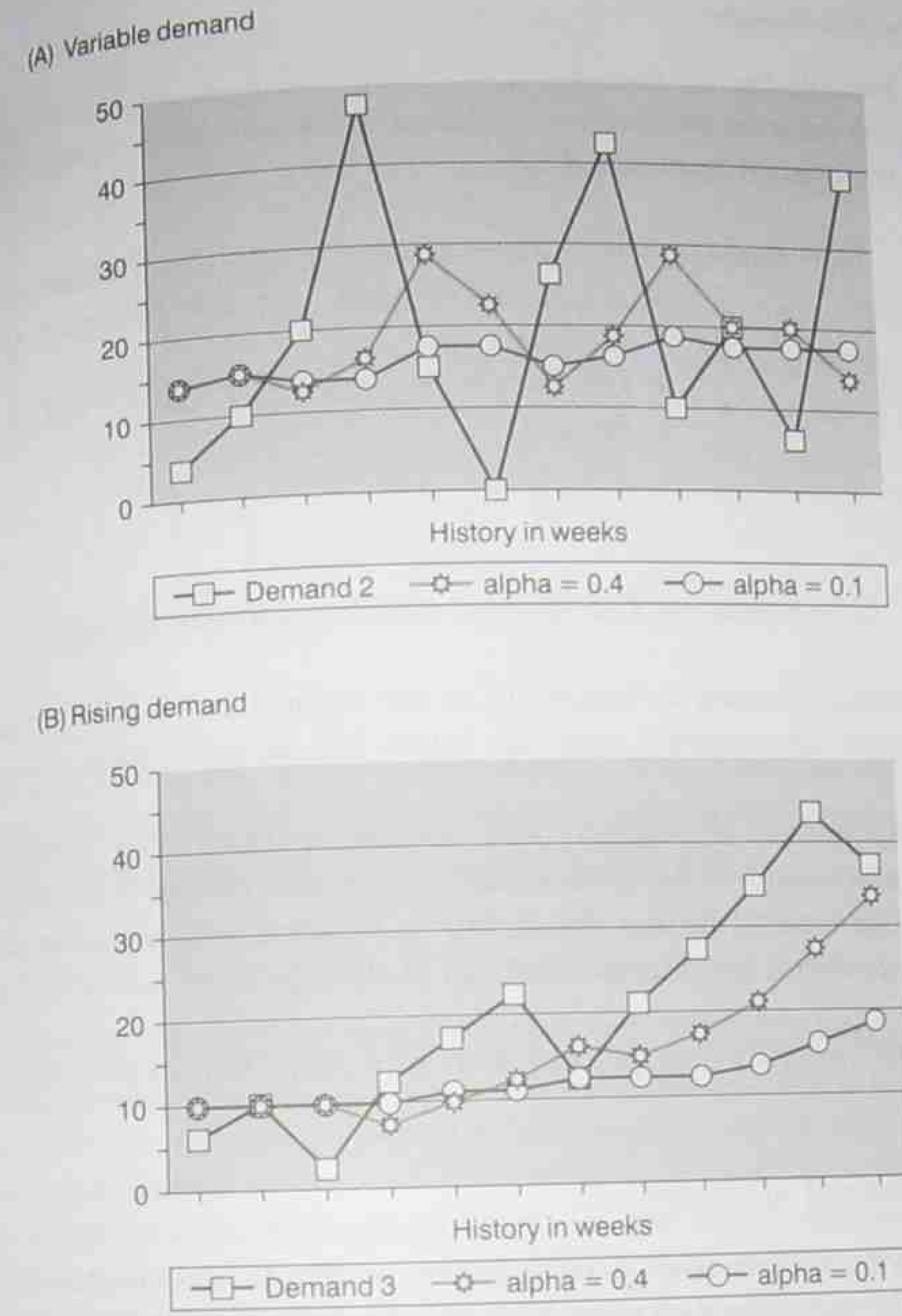


Figure 10.3 Effect of different values of α

value of α is then reviewed at regular intervals and altered where necessary.

In Figure 10.1 a sales pattern has been predicted each month using exponential smoothing with α = 0.1 and α = 0.3. The forecast for the next month has been worked out using the old average and the error. The success of the forecasting has been gauged by calculating the MAD in each case. The result indicates that α = 0.1 gives better answers in this particular case.

Alternative formula

This formula has an alternative, more convenient, method of calculation using the error on the forecast instead of the actual demand, which differs from the original form as follows:

$$\begin{aligned}\text{New forecast} &= \alpha \times \text{Actual demand} + (1-\alpha) \times \text{Old forecast} \\ &= \alpha \times \text{Actual demand} - \alpha \times \text{Old forecast} + \text{Old forecast} \\ &= \text{Old forecast} + \alpha\ (\text{Actual demand} - \text{Old forecast})\end{aligned}$$

so

$$\text{New forecast} = \text{Old forecast} + \alpha\ (\text{Error in old forecast})$$

This formula is used as frequently as the original but it is important to ensure that any negative signs are taken into account when using this version.

In exponential smoothing (and moving average), the forecast for subsequent periods is the same as that for the next period. It is possible to project this forecast further ahead, but its accuracy will be poorer since it will not predict a trend in demand (up or down).

Let the computer do the calculations and keep them updated

Contributions of history

The moving average takes data at full value until it becomes too old and falls off the end of the data table. With exponential smoothing, history never gets discarded, it just becomes less significant. The forecast for this week with $\alpha = 0.2$ is 20 per cent of last week's demand and 80 per cent of history. But that 80 per cent was calculated from 20 per cent of the demand of the week before and 80 per cent of the previous history, i.e. last week is 20 per cent of the forecast, the week before 16 per cent (20 per cent of the 80 per cent) and previous forecasts are 64 per cent (80 per cent of 80 per cent). Taking it back further, the contribution from three weeks ago is 12.8 per cent and from four weeks ago is 10.24 per cent.

The contribution from each period to the forecast is shown in Figure 10.4 going back twelve periods for $\alpha = 0.4$, $\alpha = 0.2$ and $\alpha = 0.1$. The forecast using $\alpha = 0.4$ takes little heed of demands which are more than

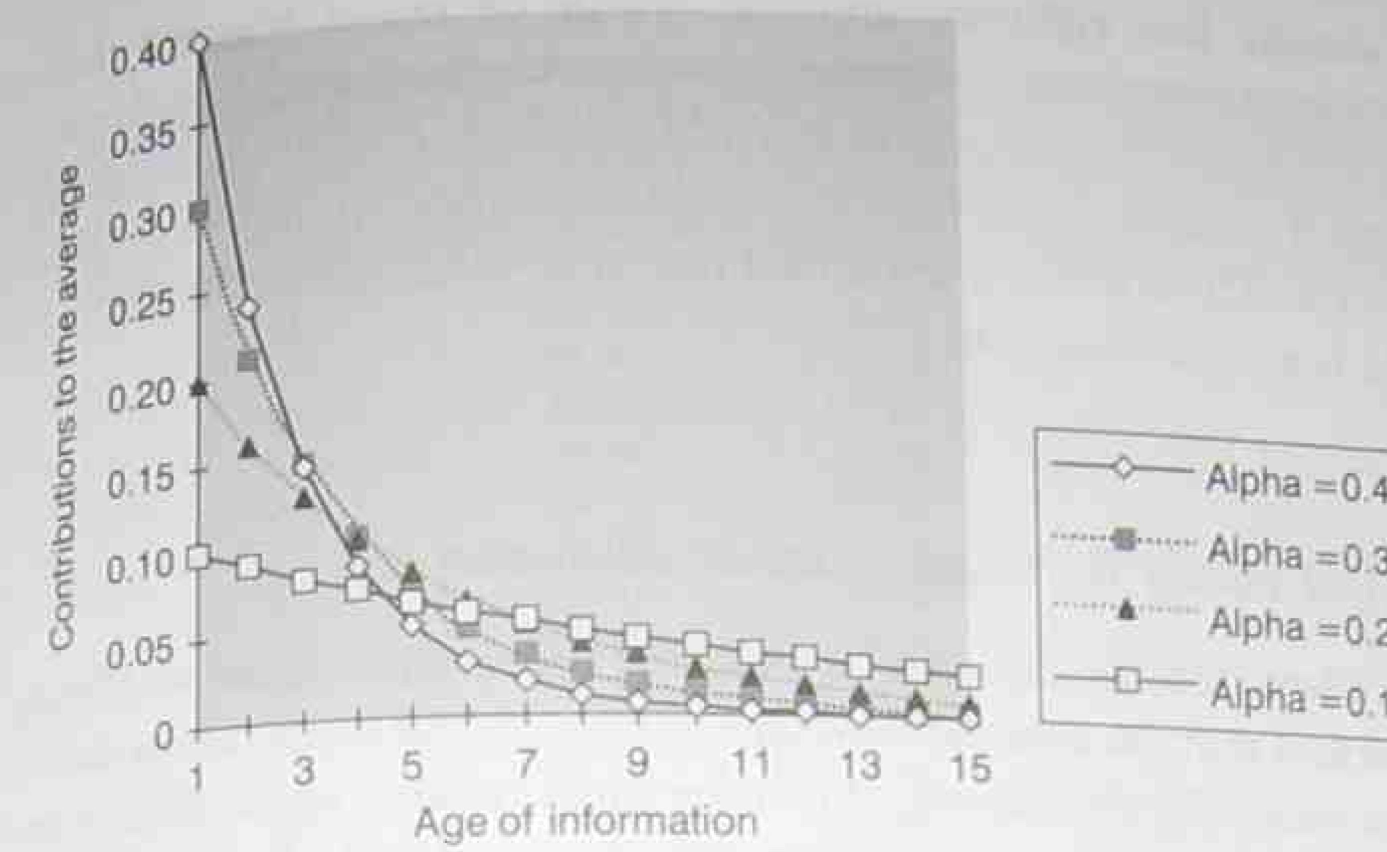


Figure 10.4 Contribution from each period to the average

about six periods old. The forecast using $\alpha = 0.1$ still has significant contributions from twelve-month-old data, and demands from as long ago as eighteen months can affect the answer significantly.

Initializing the forecast

The fact that exponential smoothing depends on old history gives a problem of when to start the forecast. What should be taken as the 'old forecast' in the equation? There are two ways of solving this, namely, estimate and history.

For a new product, or for the addition of a new item to the stock range, it is best to make an estimate of expected demand to start off the forecasting. This can be done by:

- inspired estimate
- model or replicate a similar item
- an arithmetic average of such data as is available
- historical forecast.

The effect of this estimate will become reduced at the rate shown in Figure 10.4. It is not a good practice to start with zero as the 'old forecast' since it takes too long to settle down, especially with low values of α.

Where there is some usage information, a large number of items to forecast or inexperienced staff, then a historical forecast is better. If the item already has eighteen months' worth of usage history, then the technique is to work out a six-month average for the demand up to one year before. This can then be used as the 'old forecast' to start twelve months of calculation using the exponential formulae to bring the forecast up to the present day. Once the initial forecast has been established, the exponential average is self-perpetuating.

Improved values for mean absolute deviation

Any tabulated information can be weighted using exponential smoothing. It is important that the forecast is weighted towards the most recent data, and that the variability data incorporates the most reliable up-to-date information to set the safety stocks. The calculation of MAD or SD should therefore be exponentially weighted to allow their wide application in assessing forecasts and calculating safety stocks with exponentially weighted MAD and forecast error functions. Where standard deviations are used, the weighting has to be applied to the variance $(SD)^2$. Updating of the MAD can be done most conveniently using the second version of the exponential smoothing formula:

$$\text{New MAD} = \text{old MAD} + \beta\ (\text{Error in MAD})$$

where

$$\text{Error in MAD} = \text{Actual absolute deviation} - \text{Old MAD}$$

and β is the MAD smoothing constant and is exactly the same as α but using β shows that it is the MAD that is being smoothed rather than the forecast.

It is normal to take β = 0.1 when smoothing the MAD. The reason for this is that the MAD reflects the characteristics of the demand pattern, and this would only be expected to alter gradually. If there is a sudden change in the variability of demand it is likely to be detected first through a change in demand level and forecast bias.

If the forecast for period 1 in the example (Table 10.2) is 36.2 and the MAD 8.5 from a previous assessment, the new MAD can be calculated either by taking the errors from the last six periods and averaging them to get a simple average, or by taking an exponentially weighted average. The last two columns in Table 10.2 show the difference between the simple and exponentially weighted average.

Table 10.2 Exponentially weighted MAD

Period (wk/mth)	Demand	Exponential forecast (α = 0.2)	Error	Simple MAD (6 period)	Smoothed MAD (β = 0.1)
1	28	36.20	8.20	8.50	8.50
2	45	34.56	10.44		8.47
3	33	36.65	3.65		8.67
4	39	35.92	3.08		8.17
5	19	36.53	17.53		7.66
6	27	33.03	6.03	8.57	8.64
7	22	31.82	9.82	8.16	8.38
8	48	29.86	18.14	8.43	8.53
9	36	33.49	2.51	9.71	9.49
10	25	33.99	8.99	9.52	8.79
11	18	32.19	14.19	10.50	8.81
12	41	29.35	11.65	9.95	9.35
Forecast for period 13		31.68		10.88	9.58

Examples

1 MAD Forecast = 4, actual absolute deviation = 6
New MAD = 4 + 0.1 × (6 − 4) = 4.2
2 MAD forecast (old MAD) = 4
Demand forecast = 17
Actual demand for period = 15
so actual absolute deviation = 2
new MAD = 4 + 0.1 (2 − 4) = 3.8

This calculation would be the same if the actual demand was 19. Although MAD always has a positive value, the error in the MAD can be positive or negative.

Exponentially weighed MADs, which give a better assessment of current variability, should be used wherever possible. An example using exponentially weighted averages and MADs is shown in Figure 10.5 together with the exponentially weighted errors of the mean. The exponentially weighted standard deviation is insensitive to variances during each time period because a low value of β is used. This is a good feature for ensuring a stable calculation of safety stock, but it does make the task of setting the initial SD more critical. The initial estimate used for MAD or SD will affect the

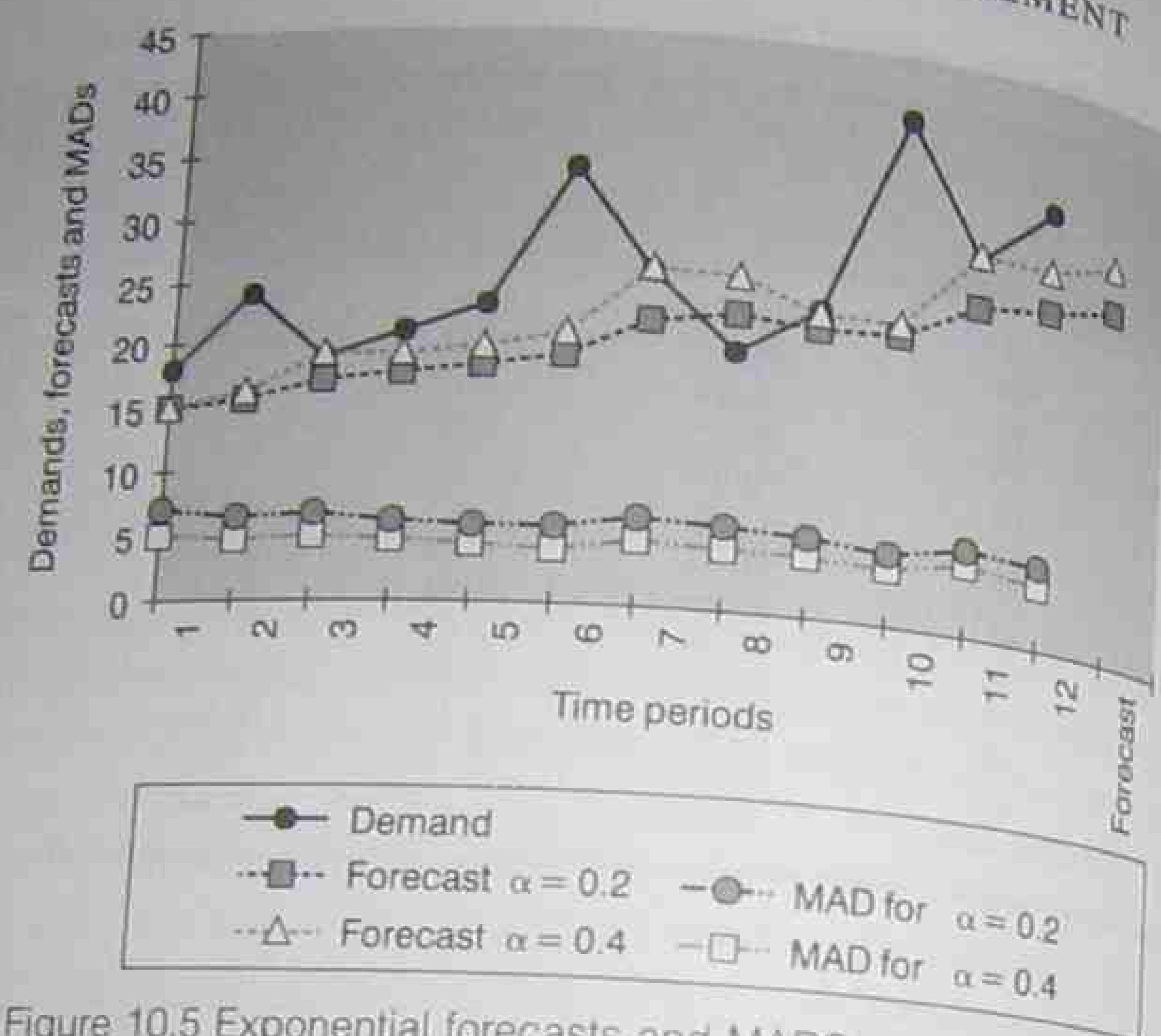


Figure 10.5 Exponential forecasts and MADS

calculation for many cycles, and can give misleading results if not chosen carefully. (An initial assumption of zero for MAD or SD is not good enough.)

In order to compensate for a change in variance or to reduce the effect of initial conditions, the weighting factor can be increased for a few periods but β should be kept below 0.2.

Choosing the best forecast

Application of forecast types

The forecast models discussed previously give a variety of estimates for the future, some of which will turn out to be better than others. The problem is how to choose the best option. At one time, it was necessary to decide in advance what the characteristics of the demand pattern were, and to choose the best model to use. Nowadays, technology can choose the best forecast, so long as the rules are defined.

The general principles which have been used to choose between the simple forecasting models are shown in Table 10.3. (The forecasting models are described in this chapter and the next.)

The exponentially weighted models have the advantage that they are more sensitive to recent events and, therefore, are generally to be preferred

Table 10.3 Uses of simple forecasting models

Forecasting model	Demand characteristics
Moving average	Static demand level, irregular pattern
Regression analysis	Continuous trend, irregular pattern
Exponential smoothing	Variable demand, gradual changing level, and slow moving spares items
Base series	Seasonal demand, mature products with low variability
Double exponential	Varying trend in demand, low variability, product phase-in and obsolescence
Fourier	Product with long history, cyclic demand
More sophisticated techniques	Much history often needed, for longer-term forecasting
Specific models	Based on market knowledge, potentially most accurate

over the simple averaging techniques such as moving average, regression, or trend measurement.

Table 10.3 can be taken as an indication of which method is likely to give best results in any circumstances, or it could be used to set the model for each item for forecasting. With the large number of lines in many stores, some gross assumptions have to be applied, and rules can be suggested such as those shown in Table 10.4.

Focus forecasting

With the technology now available it is not necessary to make these assumptions. The approach is to calculate a range of options and then to select the best average.

An assumption in making this selection is that the forecast which has been most successful recently will be the best one for the period being forecast. This assumption is open to debate, but is the best working option.

Table 10.4 Simplistic selection of forecasting models

Pareto class	Model	Method of use
A	Double exponential	Monitor and use knowledge
B	Single exponential	Exemption report and control
C	Moving average	Accept forecast
Other	Product specific	Supply to order

In inventory control the forecasts are usually for a short period ahead, making this concept more acceptable.

> *Do many forecasts and choose the smallest MAD.*

The question still remains as to which forecast can be considered the best. The best forecast is the most accurate one, of course. The most accurate forecast is the one which has the least errors (errors are the difference between the actual demand and the forecast). Therefore, the choice of the best forecast can be made by finding the one which recently has demonstrably given the least errors.

Now the basis of a safety stock calculation is that stock is required to cover a divergence between the forecast and the actual demand (perfect forecasts mean no safety stock). The calculation of the safety stock is based on finding an average divergence (SD or MAD) as one does in working out the accuracy of the forecast. The MAD is a measure of the forecast error – this is focus forecasting.

If there are, say, three different forecasts for the same item, created in different ways, then the way to choose the best one is to calculate the MAD for each (see Table 10.5). The best forecast is the one with the smallest MAD.

The use of a simple MAD is reasonable but, as in the case of safety stocks, a better measure is readily available. This is because a forecast which was brilliant a few months ago, could now be rather poor and it would be preferable to choose a forecast which has given reasonable results in the last couple of periods. For this reason a weighted MAD is used – in fact an exponentially weighted MAD (see the section on improved values for mean

Table 10.5 Three different forecasts for the same item

Month	Demand
1	28
2	45
3	33
4	39
5	17
6	27
7	22
8	24
9	15
10	18
11	28
12	25

Forecasts	Model	Forecast demand	Error (MAD)
Moving averages	12 period	26.8	5.9
	6 period	22.0	7.4
Exponential smoothing	$\alpha = 0.1$	28.3	9.5
	$\alpha = 0.2$	24.7	6.2
	$\alpha = 0.3$	23.8	5.7
	$\alpha = 0.4$	23.9	5.2
Double exponential smoothing	$\alpha = 0.2$	21.9	5.4
	$\alpha = 0.3$	21.8	5.0
	$\alpha = 0.4$	21.4	5.7

absolute deviation, earlier in this chapter). This provides a more reliable method of assessing the potential of alternative forecasts.

Monitoring forecasts

Forecast tracking using cumulative sum of errors

The forecasting methods discussed will cope with most demands, but they need adjustment from time to time when the method is inappropriate to the conditions. We therefore need a monitoring process to enable prompt action to be taken.

Tracking signals provide a monitor of how well the forecast is performing. When the errors fall outside acceptable limits, the tracking signal identifies the problem or modifies the controls.

> **Example**
>
> *The demand pattern for an item over the past twelve periods is shown in the two left-hand columns of Table 10.5. As a result the forecasts in the three right-hand columns were calculated. By looking at the MAD column, it can be seen that the best forecast is double exponential smoothing with α = 0.3, and the worst forecast is exponential smoothing with α = 0.1. Therefore the forecast to choose is 21.8. However at the next period end a different forecast might well prove to be even more accurate.*

The Cusum technique provides a warning signal when there is a significant asymmetric shift of actual results from the forecast. It is very simply obtained by cumulating the errors in forecast over successive periods (see the last column of Table 10.1). When the forecast is correct, the actual demand is distributed equally above and below the forecast and cumulative error is small. Once the forecast diverts from the average demand, the errors build up period by period.

If the Cusum reaches an unacceptably high level, an error signal can be triggered and the forecast adjusted. The size of the error in the mean depends on the variability of the demand pattern and so the limit on acceptable forecast errors is set in terms of MAD – six times the MAD. However, for important items, values above four times the MAD should be investigated.

When the Cusum has signalled a problem by a high error value, the Cusum should be reset to zero. Otherwise the controlled forecast will give a level which is consistently high.

> Monitor forecast performance.

Improving control – Trigg's tracking signal

When the Cusum indicates that there is a significant discrepancy, the forecast needs to be changed rapidly. For exponentially weighted averages this means increasing the smoothing constant, α (see section on selection of the smoothing constant, α). Where there is a large variety of items this will result in many changes to be made. Trigg's method is a way of feeding back the Cusum information automatically and using it to modify α. Instead of generating the cumulative SUM of error, the tracking signal first works out the average error. This is done using the exponential smoothing formula on the forecast errors and is called the 'smoothed error', calculated from the actual errors, in the normal way:

$$S_t = \delta e_t + (1 - \delta)S_{t-1}$$

Where S is the standard error, e is the actual error and δ is the smoothing factor for the errors. The subscripts on the standard errors, t and t–1, stand for this period just completed and last period. This is similar to the usual exponential smoothing of the MAD:

$$MAD_t = \beta e_{t-1} + (1 - \beta)MAD_{t-1}$$

δ and β are set to the same value (0.1 or 0.15).

The tracking signal provides a figure of merit for the quality of forecast, and is calculated by the ratio

$$\text{Tracking signal} = \frac{S_t}{MAD_t}$$

The tracking signal shows whether the forecast is good or not. It is considered good if it is not biased (i.e. not consistently high or consistently low) where S and MAD are the exponentially smoothed values. The difference between the calculation of S and MAD is simply that S takes into account whether the forecast errors are up or down, and the MAD only uses the size (absolute value) of the forecast errors. Hence the biggest value

that the ratio, the tracking signal, can have is 1, when all the errors are either positive or negative. The smallest value is zero where the smoothed errors compensate and are very small compared with the individual fluctuations in demand.

Trigg used this to give direct feedback into the α factor in the forecast. The argument is that an accurate forecast gives a low value for tracking signal, and it needs little modification, so a low value of α is best. If the tracking signal is large, there is a significant error in the forecast so a high α factor is required to alter the forecast. The tracking signal can therefore be used as the smoothing constant for the forecast, and this will enable an exponentially weighted forecast to be more self-compensating each period. For this purpose, the tracking signal has to be a positive number and so any negative tracking signals (TS) are changed to positive.

The forecast F_t is then determined by:

$$F_{t+1} = F_t + |TS_t| \times e_t$$

where e is the forecast error for the latest period, and the value of TS is always positive. The tracking signal can be used for feedback into exponential smoothing or double exponential smoothing. It is also possible, but less beneficial, to calculate a tracking signal using normal averages. To make the tracking signal work properly the value of δ has to be the same as the value of β. Since both of these are usually set to 0.1 this does not present a problem.

There is a danger when using Trigg's method that the α factors become very high (above 0.4) when it tends to overcompensate for small changes. In the discussion of exponential forecasting, it was stated that the value of α should be below 0.5. It is therefore advisable in practice to modify the formula so that values above 0.5 are not created.

The tracking signal is very useful and allows a computer-calculated forecast to monitor itself and warn the inventory manager when it is producing bad forecasts. The tracking signal can be used simply as a warning flag. If the tracking signal is high, say above 0.4, then a manual adjustment to the forecast is required, and an exception report can be triggered.

> Trigg's forecasts can be a more accurate method.

The tracking signal can be combined with focus forecasting, the optimum forecast being selected by choosing the one with the smallest smoothed MAD as normal. If this forecast has a tracking signal below, say 0.4, then the forecast is acceptable. If the tracking signal is greater than the unreliability level (0.4), then the forecast with the next smallest MAD can be chosen. This is a simple, mechanical and effective way of getting good forecasts across the range of stock items.

Example

In Table 10.6 the characteristics of the demand pattern change after period 12.

The forecast gradually drifts up and the MAD increases. The smoothed error which was very small up to period 12 starts to increase. This is because the change in forecast lags behind the actual demand level. Consequently the tracking signal increases. It reaches 0.5 in period 11, which is quite quick. At that stage the forecast can be readjusted or an alternative model used, or it continues to give poor results for a whole (periods 12 to 14).

Table 10.7 shows what happens when the tracking signal is fed back to change the α as proposed by Trigg. The α factors are much smaller at the beginning. For the first few periods this is a result of the initial condition that the smoothed error is set to zero. The tracking signal stabilizes and responds less to the low demand in period 7, which is sensible given the consistently larger demand in the previous periods. When the demand level changes, the tracking signal increases and enables the forecast to rise faster to cope with the problem. However, this does not fully compensate and the tracking signal still rises above 0.4 . The forecast rose quickly to the new level as the α factor increased. It did this without the inventory controller having to intervene. The effect is shown graphically in Figure 10.6, which contrasts the demand, the forecast with α = 0.2 and the forecast using Trigg's method. The data are taken from Tables 10.6 and 10.7 and represents a typical demand level change. The only problem in period 19 is that the tracking signal is still rather large. As the smoothed error continues to decrease because the forecast is again accurate, this tracking signal will reduce gradually.

Although it is not the ideal answer, since it lags behind the demand pattern significantly in a period of change, it is an improvement where there are many stock items and little time to manage them individually. However, it would be expedient to limit the maximum value of tracking signal fed back to 0.4 and to manually readjust when it exceeds this value.

Table 10.6 Forecast tracking

Period	Demand	Exponential forecast ($\alpha = 0.2$)	Error in forecast	Smoothed MAD $\beta = 0.1$	Smoothed error $\delta = 0.1$	Tracking signal
1	13	16.00	–3.00	4.00	0.00	0.00
2	1	15.40	–14.40	5.04	–1.44	–0.29
3	10	12.52	–2.52	4.79	–1.55	–0.32
4	17	12.02	4.98	4.81	–0.89	–0.19
5	11	13.01	–2.01	4.53	–1.01	–0.22
6	19	12.61	6.39	4.71	–0.27	–0.06
7	16	13.89	2.11	4.45	–0.03	–0.01
8	15	14.31	0.69	4.08	0.04	0.01
9	26	14.45	11.55	4.82	1.19	0.25
10	38	16.76	21.24	6.47	3.20	0.49
11	31	21.01	9.99	6.82	3.88	0.57
12	40	23.01	16.99	7.84	5.19	0.66
13	35	26.40	8.60	7.91	5.53	0.70
14	32	28.12	3.88	7.51	5.36	0.71
15	Forecast	28.90				

Table 10.7 Forecast using Trigg's tracking signal

Period	Demand	Trigg's exponential forecast	Error in forecast	Smoothed MAD $\beta = 0.1$	Smoothed error $\delta = 0.1$	Tracking signal
1	13	16.00	–3.00	4.00	0.00	0.00
2	1	16.00	–15.00	5.10	–1.50	–0.29
3	10	11.59	–1.59	4.75	–1.51	–0.32
4	17	11.08	5.92	4.87	–0.77	–0.16
5	11	12.02	–1.02	4.48	–0.79	–0.18
6	19	11.84	7.16	4.75	0.00	0.00
7	16	11.84	4.16	4.69	0.42	0.09
8	15	12.21	2.79	4.50	0.66	0.15
9	26	12.62	13.38	5.39	1.93	0.36
10	38	17.41	20.59	6.91	3.79	0.55
11	31	28.72	2.28	6.44	3.64	0.57
12	40	30.01	9.99	6.80	4.28	0.63
13	35	36.30	–1.30	6.25	3.72	0.60
14	32	35.52	–3.52	5.98	3.00	0.50
15	Forecast	33.76				

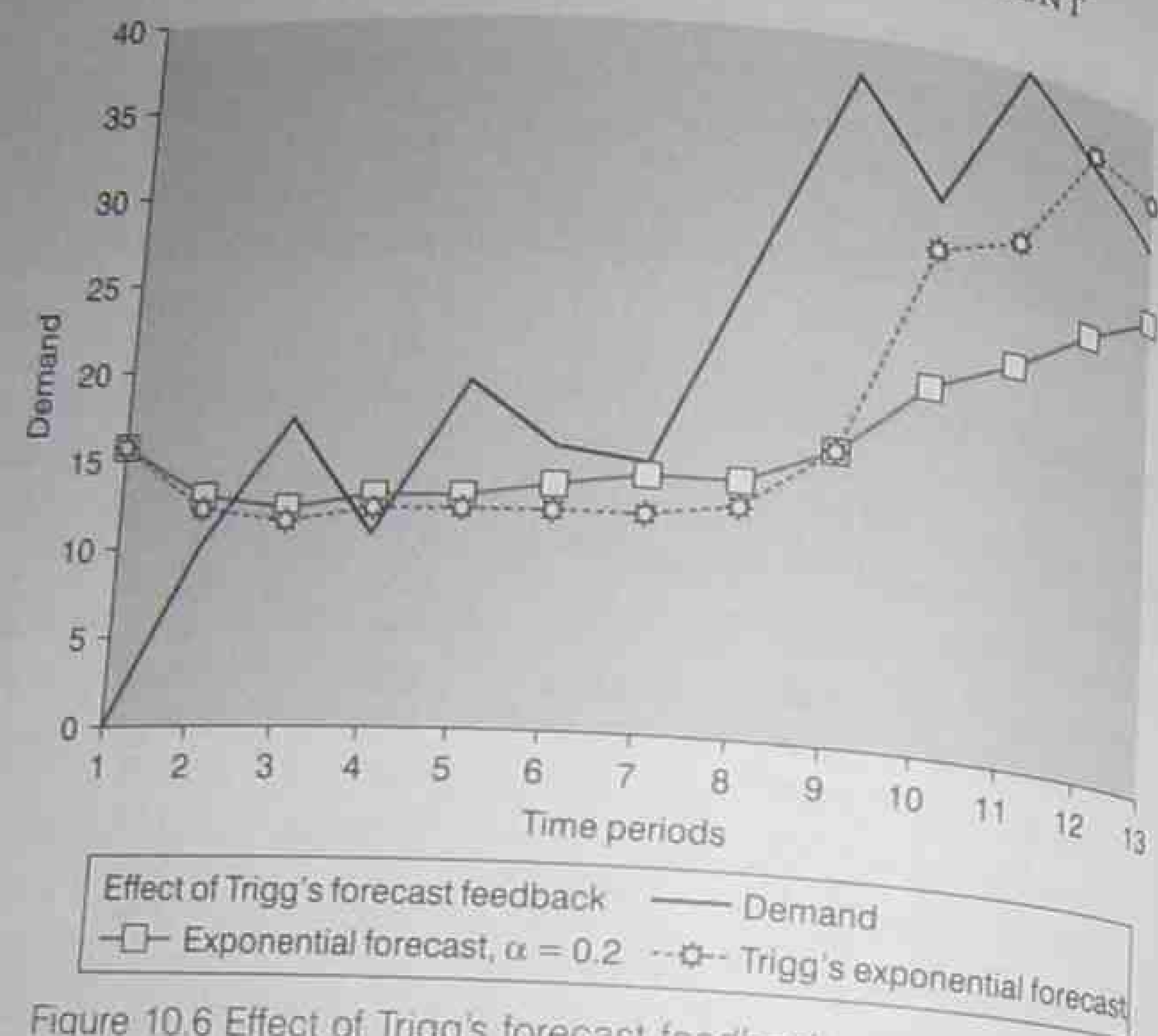


Figure 10.6 Effect of Trigg's forecast feedback

Key points

- Historical forecasting is essential for basic stock control.
- It is necessary to add information about future events.
- A rolling average gives a basic forecast.
- This can be greatly improved using exponential smoothing.
- Carry out several forecasts, then select that with the smallest MAD.
- A tracking signal shows when the forecast is wrong.

11

Advanced forecasting methods

- Better forecasting for rising or falling demand
- Coping with seasonal demand
- Other techniques

More forecasting tools

Improved forecasting techniques

Exponential smoothing is the basic model, which can, of course, be improved. It splits demand into two components:

- average, which is recalculated each period
- random fluctuations about the averages.

The next stage of complexity is to add either:

- a constant increase or decrease in demand, using double exponential smoothing, or
- a seasonal variation of demand, using base series forecasting.

Alternatively, entirely different approaches to intrinsic forecasts can be used giving:

- unweighted average trends (regression analysis)
- profitability of changes in demand (Baysian forecasting)
- curve fitting (Fourier analysis)
- or more complex methods. (These are noted but not discussed as exponential methods are a better option.)

The mathematics can become quite complex and are usually left to the computer. In general, the more complex the model the more history is required which is why simple forecasting has been acceptable until recently. Intrinsic forecasts are best for data collected under similar conditions, even when the demand is erratic. Then it is quite possible that exponential forecasting with $\alpha = 0.1$, or even long moving averages are best. If a change from single to double exponential smoothing gives a significant improvement in the forecast for an item, then perhaps more sophisticated methods would work better, but there are fewer and fewer items benefiting from the better forecasting techniques. As a general rule forecasting can benefit more by improving the customer communications to detect changes in the market than by going to highly sophisticated models.

Double exponential smoothing

As stated in the last section, single exponential smoothing splits the demand into 'average + fluctuations'. This is improved in double exponential smoothing by analysis into 'average + trend + fluctuations.' In Figure 11.1

Figure 11.1 Improved forecasting through double exponential smoothing

the difference in forecast can be seen. The exponential forecast for the next period will show a weighted average, whereas by adding the average trend, the predicted rate of demand will be higher, and the forecast will be better.

To understand double exponential smoothing it is best to go back to simple mathematics. Figure 11.2 shows a normal straight-line graph given by the formula $y = bx + a$, with slope m and cutting the y axis at point c. The double exponential model is equivalent to this straight line, with y being the demand and x being time. Single exponential smoothing is like the line 'a' above the x axis.

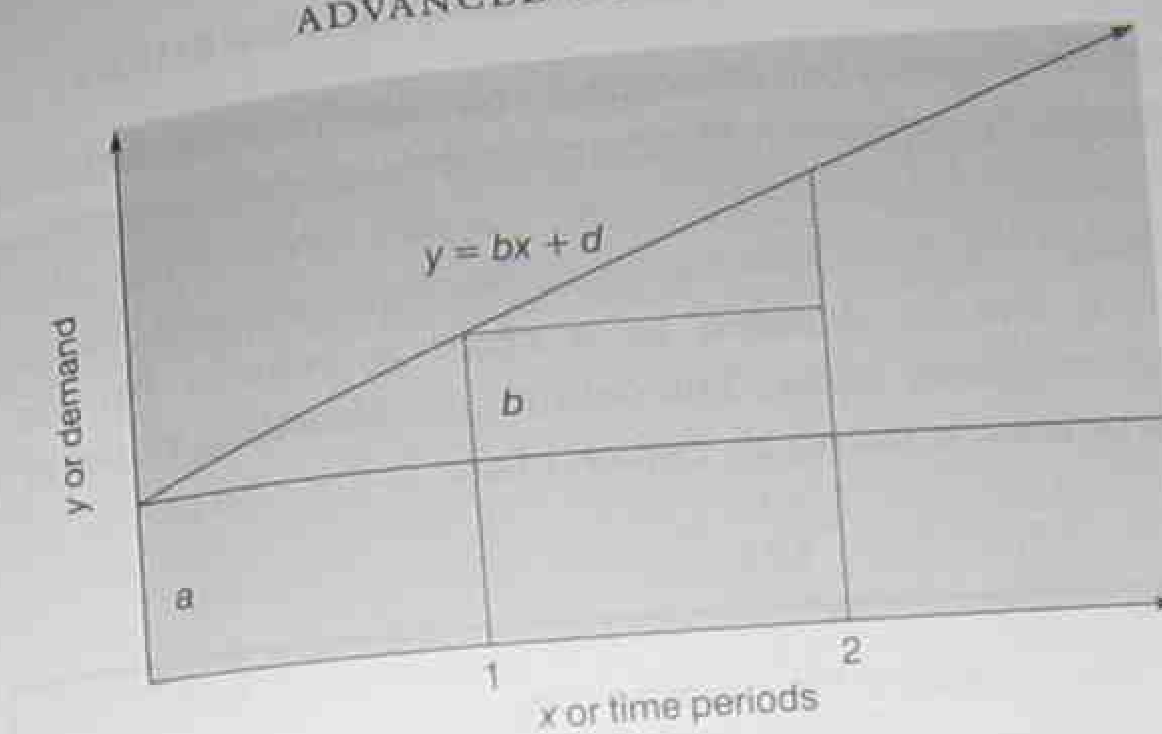


Figure 11.2 Double exponential model

In double exponential smoothing the forecast of demand, F_t, is therefore composed of two factors. $F_{t+1} = a_t + b_t$, with one factor arising from the movement of the average due to random fluctuations, similar to single exponential smoothing. The second factor calculates the trend of variations and adds a 'slope' factor. This has important advantages because the extrapolation to future periods continues, the increase or decrease being observed rather than assuming a static demand level.

In addition to this, double exponential smoothing can also compensate for lag in the forecast. It is obvious in graph (C) of Figure 10.1 that the forecasts are lagging behind an increasing demand level (and will signal a tracking error). Double exponential smoothing can bring the forecast in line with the demand once a linear trend has been established. This can be very useful when there is a gradually increasing or decreasing average demand.

The lag in an exponentially weighted average is simply given by the expression

$$\text{Lag} = \left(\frac{1-\alpha}{\alpha}\right) \times \text{Trend}$$

where α is the smoothing factor.

From Figure 11.2 it can be seen that the double exponential forecast for one time period ahead is $a + b$ (demand + trend). The forecast for two periods ahead will be $a + 2b$ (demand + twice trend).

There are several versions of the double exponential formula. The simpler version is Brown's model. A more comprehensive model is Holt's model, but this requires two smoothing constants (one for the base forecast '*a*', and one for the trend '*b*'). Holt's method does allow optimization of the smoothing factors for selected items by iterative programming.

The simpler formula is adequate normally and uses only one smoothing factor, which can be denoted as α since it behaves like the single exponential smoothing factor. The method of calculating values of '*a*' and '*b*' and the forecast is shown in Figure 11.3.

1. *Find the forecasting errors*
   Error = ForecastActual demand
2. *Update the demand level*
   Demand level = Actual demand + $(1-\alpha)^2$ error
3. *Update the trend*
   New trend = Old trend a^2 error
4. *Add the components*
   New forecast = Demand level + New trend

To set the forecast requires:

old trend
old forecast
actual demand

Figure 11.3 Double exponential smoothing

The general forecast for *n* periods ahead is:

$$F_{t+n} = \text{'}a\text{'} + \text{'}b\text{'}n$$

By this linear extrapolation of trend, a much more accurate forecast of demand in future periods can be made, if a linearly increasing or decreasing sales pattern is expected.

For items with widely varying demand patterns, the estimated trend can be misleading. A high or low value for the previous month's demand reflects too heavily on the calculated trend. If α is reduced to compensate, then the model becomes unresponsive. In this case it is better to take a longer sample period to even out the statistical variations, and this is often done by

Table 11.1 Double exponential forecast

Time period	Forecast F	Demand D	Error e	Demand level a	Trend b
				65	−5.00
1	66.0	55	5.0	58.2	−5.20
2	53.0	50	3.0	51.9	−5.32
3	46.6	45	1.6	46.0	−5.38
4	40.6	55	−14.4	45.8	−4.81
5	41.0	80	−39.0	55.0	−3.25
6	51.8	105	−53.2	70.9	−1.12
7	69.8	130	−60.2	91.5	1.29
8	92.8	155	−62.2	115.2	3.77
9	118.9	180	−61.1	140.9	6.22
10	147.1	205	−57.9	168.0	8.53
11	176.5	230	−53.5	195.8	10.67
12	206.4	255	−48.6	223.9	12.61
13	236.5				

Note: α = 0.2

calculating the average demand over the last three periods and using the averaged error to modify '*a*' and '*b*'.

The trend calculation is smoothed through the demands over preceding periods. Therefore, the greater the smoothing the slower is the trend to respond to change in trend. This causes lag – the amount of lag is shown in Table 11.1.

Changing forecast methods to double exponential smoothing means that the initial values of '*a*' and b_{t-1} have to be estimated. The smoothing constant behaves like the smoothing constant for single exponential smoothing. The larger the value, the faster it responds to change, but the more unreliable is the value.

As the constant is squared for double exponential smoothing, selection of the correct value is more important. Lower values of α are chosen. In Figure 11.4 the equivalent future demand is shown for single and double exponential smoothing.

The method of estimating '*a*' and '*b*' depends upon how much information is available. If the demand for one period only is known or an estimate of it, then there is little alternative to setting '*a*' equal to that value, '*b*' to zero, and using a high value of smoothing constant so that the forecast will respond rapidly to the actual trend. Thereafter, the smoothing constant can be reduced to a more stable value.

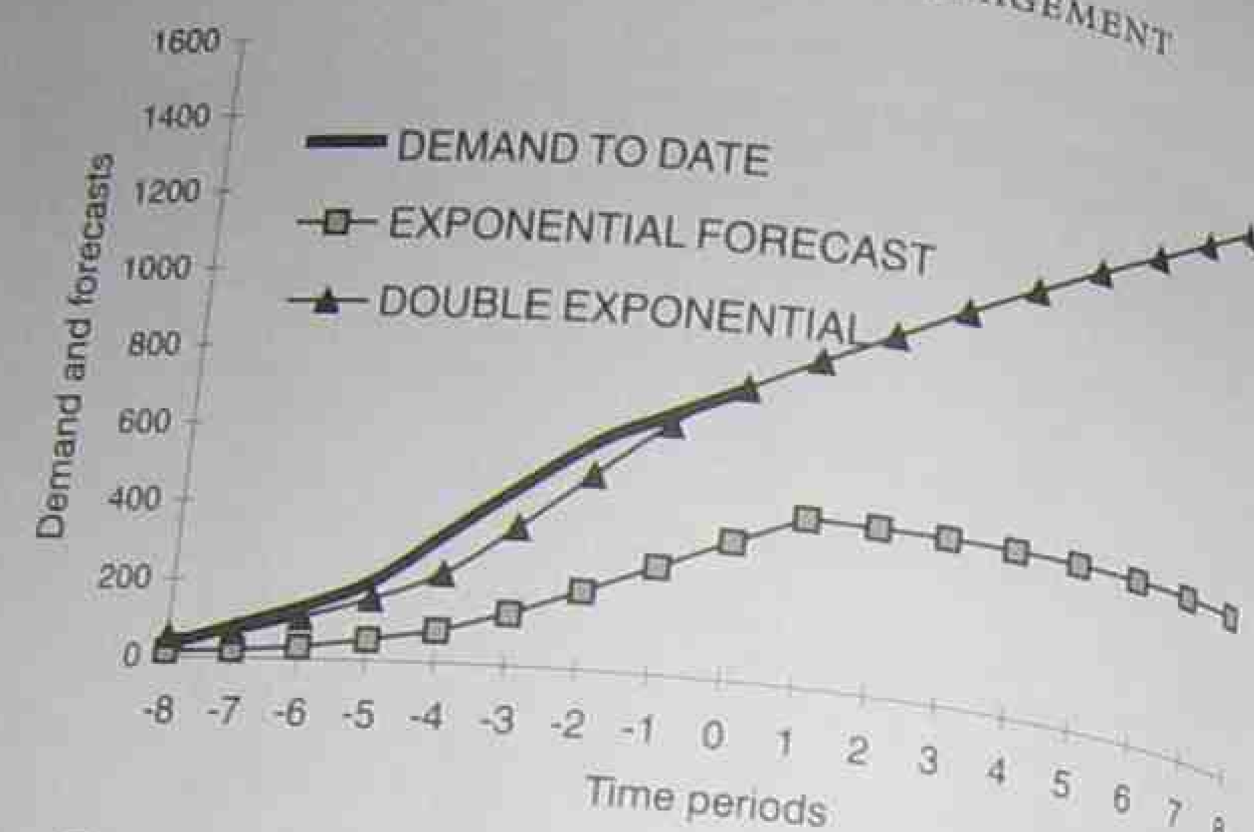


Figure 11.4 Forecasting several periods ahead

For an established product, the history is known and the values of 'a' and 'b' can be estimated accurately. In this case it is best to plot a graph and estimate the value of demand, F_t, and a trend 'b_t.' Then

'a_t' = $F_t - b_t$

These values of 'a' and 'b' can then be used as a starting condition for double exponential forecasting.

An example of double exponential smoothing is given in Table 11.1 and the effect of single period forecasting in Figures 11.4 and 11.5.

Forecasting for seasonal sales

> Forecasts do not cope with seasonal demand without using base series.

For many products seasonal or cyclic demand patterns exist. This can be caused by natural seasonal factors, for example temperature, rainfall or the probability of thunder storms, or by other seasonal factors like holiday patterns, Christmas or other religious festivals, or financial year ends. These factors can be the overriding cause of demands, or a small reason for fine tuning the forecast. Seasonal demands can be masked by the sporadic

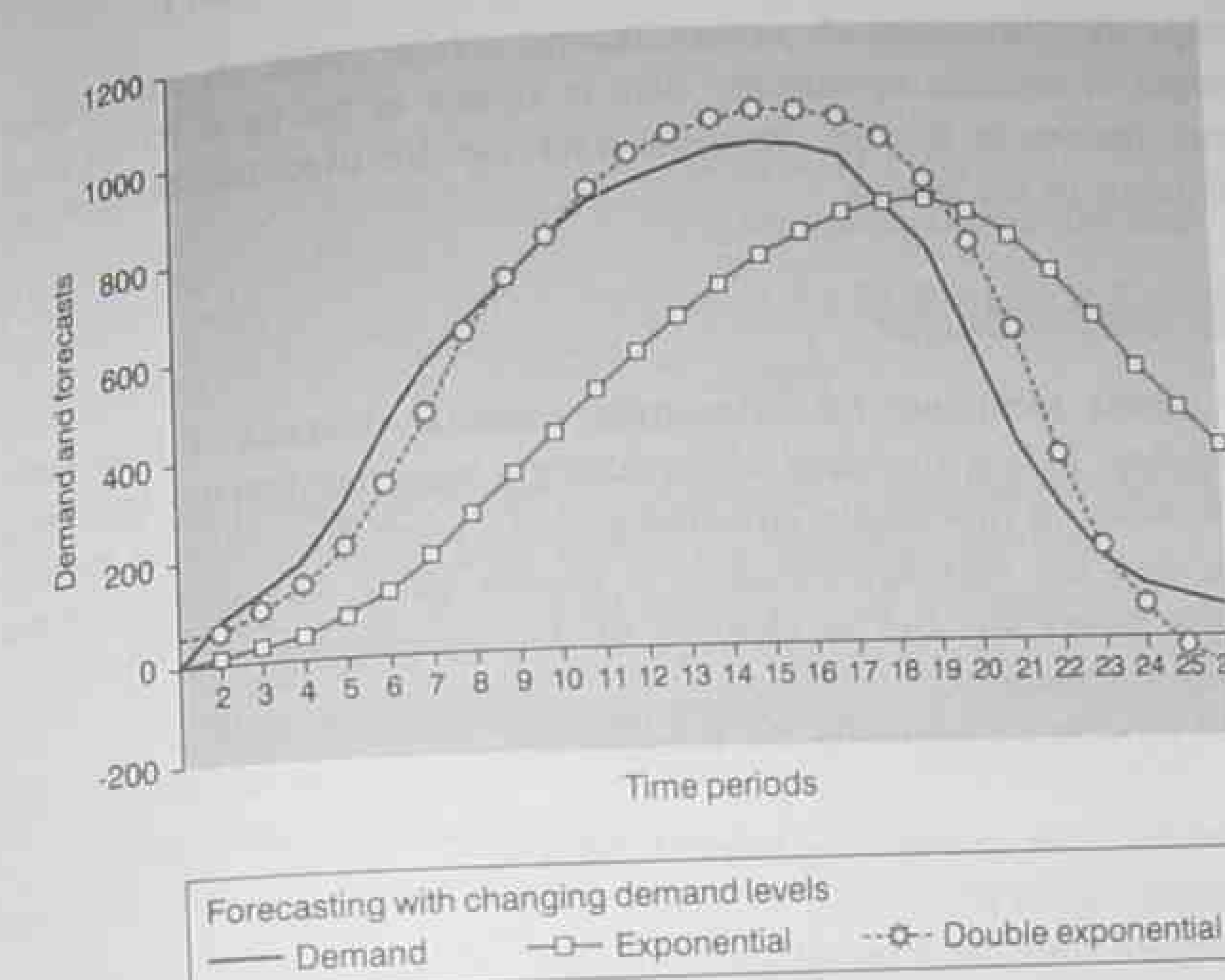


Figure 11.5 Forecasting with changing demand levels

nature of the demand for an item. Where this happens, as is often the case for C class items, it may well be expedient to ignore the seasonal variation and treat this as a random fluctuation too. Here a twelve-month moving average is a simple solution, although the safety stock will, of course, inflate the service outside the faster moving season and cause more stockouts during the season.

Historical data

To use history to forecast seasonal demand requires at least a year's worth of data! In fact the more years for which data are available, the more reliable the forecast, unless there has been a change in the pattern of requirements for the item. This is usually a matter of interpretation, since the data is often distorted by specific, non-repeating factors to do with the market, price changes at irregular intervals or large contracts. The historical data often has to be interpreted to get the best forecast.

Generally seasonal data is collected in monthly buckets, although where Christmas dominates the sales, weekly information around November and December is often needed. The average demand for each corresponding month for three years or more is averaged so that an average monthly demand pattern emerges. It is useful to use weighted averages of demand to bias the data in favour of the most recent years. Where history is shorter, seasonal data can be used but there is less confidence in the results.

Base series technique

The classic technique for calculating seasonal demand is base series forecasting. This is a method of producing a seasonal demand forecast as an extension to exponential smoothing, and makes the following assumptions: that the seasonal variation is known (from the data) and that forecast errors are due to changes in the average demand level or to random fluctuations.

The process for updating the forecast is:

1. Take the new month's (or week's) demand.
2. De-seasonalize it.
3. Compare with the forecast.
4. Update the forecast for the next months (exponentially).
5. Re-seasonalize the forecast.

Before the forecasting can start, seasonal demand factors have to be calculated, and an initial forecast made for the average demand. The processes 1 to 5 are then carried out.

Example

Demand for July is forecast as thirty-six. This is because the average monthly demand is thirty and history has shown that the demand next month is typically 20 per cent above average. At the end of the month, the records showed that thirty-three were sold. What is the likely demand for August, which has typically proved to be 10 per cent above average?

The demand in July should have been $30 \times 1.20 = 36$, but instead it was 33, which is 27.5×1.2. (The assumption is that the seasonal factor is correct.) The de-seasonalized forecast should have been 30 but was 27.5. Therefore the forecast was too high and can be exponentially smoothed downwards (assuming α is 0.2 here) as follows:

$$\begin{aligned}\text{New de-seasonalized forecast} &= \alpha \times \text{de-seasonalized demand} \\ &\quad + (1-\alpha) \times \text{old de-seasonalized forecast} \\ &= 0.2 \times 27.5 + (1-0.2) \times 30 \\ &\quad + 29.5\end{aligned}$$

If the de-seasonalized forecast is 29.5, and the demand for August is 10 per cent above average, then the new expected sales for August are 29.5 + 10 per cent of 29.5, i.e. 32.45.

The de-seasonalized demand forecast of 29.5 can be used to forecast all future months by multiplying by their respective seasonal factors.

Other methods

From the wide variety of forecasting methods available, only the simplest have been described. Others such as regression analysis, Fourier analysis and Baysian forecasting are based on different assumptions about the demand pattern.

More comprehensive approaches such as Holt Winters and Wagner Within, which extend the methods to greater sophistication, can be used to give more accurate forecasts, but require more historical information.

As the quality of sales information is often inconsistent, the opportunity to use refined techniques is limited but the development of improved forecasting will make it easier for the stock controller to use better techniques in future. Meanwhile, application of exponential techniques is adequate for many purposes although changes will be required in other areas including:

- the measurement of demand not sales
- the segregation of stock from non-stock demand for an item
- time bucket control.

The application of refined forecasting will then provide useful benefits.

Key points

- Double exponential smoothing gives tangible benefits where there are
  - significant trends in demand
  - long lead times for A class.
- Double exponential works best where demand is consistent.
- Seasonal demand has to be forecast using base series.
- More sophisticated techniques need better data and may not make a significant improvement in the forecasts.

12

Material requirements planning – an alternative to forecasting

- Identifying dependent demand
- The MRP process
- Benefits of using MRP
- Structured demand and bills of materials
- Triggering orders – projected available
- Meeting demand – available to promise

Avoiding uncertainty

When there is a clear picture of the demand for a particular item then stock can be minimized. No stockholding would be needed if demand were known precisely and far enough in advance because supply could be matched exactly with demand.

For items outside the range stocked, the customer orders and receives the supply after a lead time. There may be some transient stock because the supplier is asked to deliver in advance of the time that the customer requires it. The greater the confidence in supply, the shorter time items can spend in stores.

Treat demand as dependent to reduce inventory.

The situation with non-stock supply is mirrored for raw materials in manufacturing and for production components in assembly. Stock is required to feed the process, and firm plans exist for usage rate just as for scheduled customer orders and non-stock supplies. The approach to these types of inventory is to treat them as dependent demand rather than the independent demand as discussed in the previous chapters. The contrasts between dependent and independent demand are shown in Table 12.1. The essential feature of dependent demand is that it is calculated from the

Table 12.1 Dependent and independent demand

Independent	Dependent
Inventory-level system	MRP system
Forecast	Calculated
Keeps stock	Supply as required
All lines separate	Lines co-ordinated
Reactive	Proactive
Good customer service	Very good customer service
High inventory	Low inventory

demand of the next item up the supply chain; the aim is to have stock when it is to be used and no stock the rest of the time.

The inventory profile for dependent demand is shown in Figure 12.1. When the need is identified an order is placed. After the supply lead time, the delivery arrives. The items are then transferred or transformed for the customer and, when this is complete (which takes the process lead time), then the item can be sent out. The stock is therefore only on site during the process lead time and the investment in stock is very low. There is potentially some safety stock included in case the customer order quantity turns out to be greater than that originally agreed. The safety stock here can be calculated from the forecast errors; in this case the 'forecast' is the originally agreed delivery quantity. Ideally the safety stock should be zero.

A basic concept used in planning dependent demand is backward scheduling. This is simply the process of taking the delivery date required by

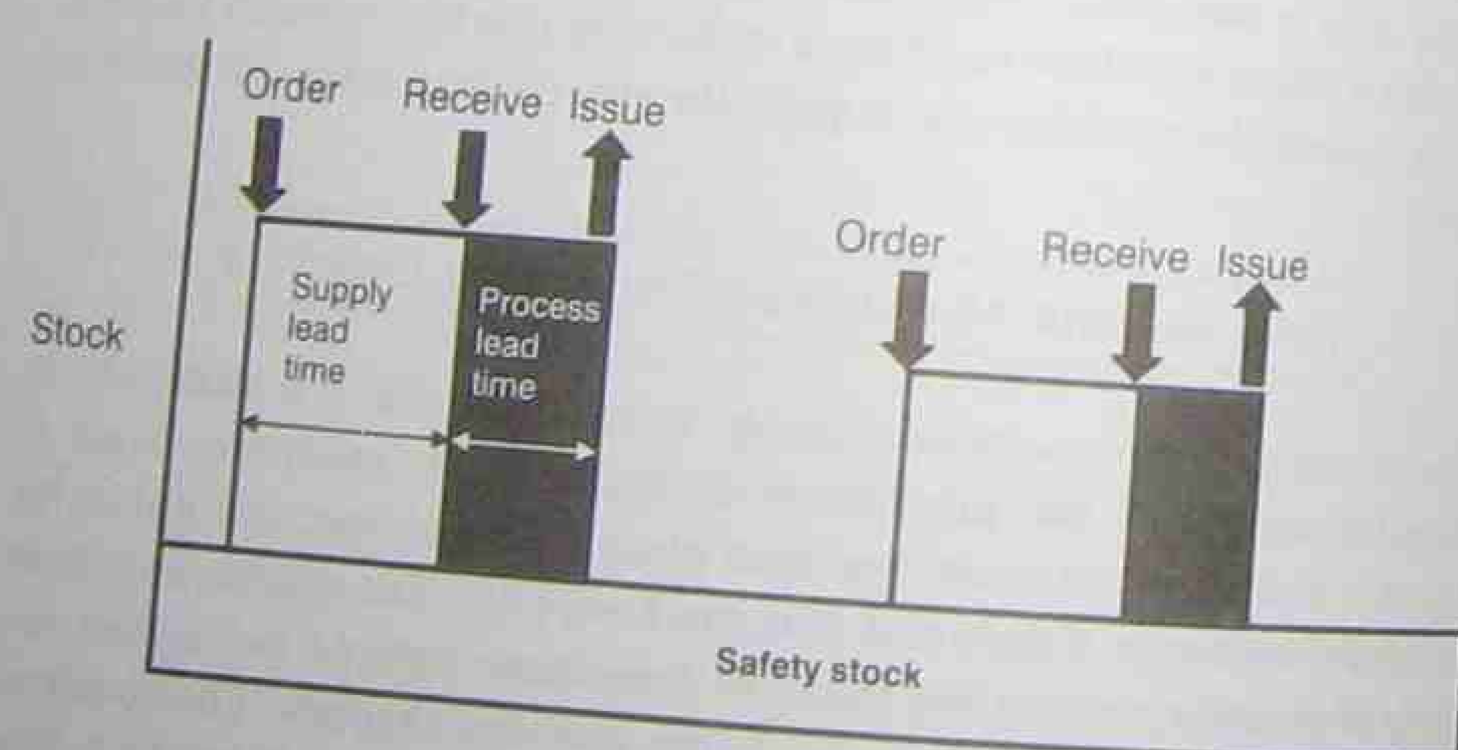


Figure 12.1 Inventory levels for dependent demand

the customer, and working back to identify when the item needs ordering from the supplier. 'Offsetting by the lead time' is the jargon used for this method. In Figure 12.1 the 'issues' should be made at the time requested by the customer. Moving back by the 'process lead time' gives the time to start on that order. If the system is working properly, this should happen just as the supplier delivered (marked 'receive' in the Figure 12.1). Going back further by the supply lead time identifies when to place the purchase order.

If this calculation results in a date which has already passed, the options are either to find a way of reducing the lead time or to forward schedule from today's date. This forward schedule will then estimate a later customer delivery date, so that the customer can be informed and the plan delayed.

Material requirements planning

The structure

There is a well-developed technique for planning dependent demand called material requirements planning. This is a generally applicable technique for all types of dependent demand. The basic concept is to have stock when it is needed and to have none the rest of the time. With dependent demand, the size and timing of the requirements are known from the next level down the supply chain. Material requirements planning can therefore give good control. It is used extensively in manufacturing because there can be several levels of dependent demand in a product (assemblies made from sub-assemblies made from components made from purchased materials.) Here MRP is well developed and complex. In many inventory situations MRP is a much simpler process.

The MRP process

Material requirements planning is a useful planning tool outside manufacturing. Normal stock management treats each item as completely independent, unaffected by usage rates of any other item and the demand is assumed to happen at random. The MRP approach is entirely different, based on:

- interdependent usage rates between stock lines
- acquiring stock for when it will be used.

Many customers tend to use items together, so demand is not truly independent. For example, there must be a connection between the demand for hinges and doors, paint and paint brushes, nuts and bolts, although, in general business, there is no certainty that the sale of one item will be matched by the sale of the linked item. If a customer buys 100 bolts, there will be no automatic order for 100 nuts. They may already have a stock of nuts, or use some in threaded holes or, even, buy from elsewhere. The situation will become clear after a while. For example, it may work out in the long term that for every 100 bolts bought by a customer, they require forty nuts of one type and twenty of another. This relationship could be consistent, because the customer's use is consistent, sometimes without identifying the reason. The demand for these items could therefore be interlinked, and could vary as their level of business changes. It is important, therefore, for the items to be balanced in stock.

Material requirements planning links in the inventory systems with the sales and financial plans of the company. It is a professional technique for a modern company and it requires adequate information, particularly:

- a master schedule of planned supply to customers for each product, projected far enough ahead to permit ordering of bought-out items, raw materials and components. This 'planning horizon' must be long enough to cover the procurement lead time for bought-out items plus the total manufacturing time
- a well defined 'bill of materials' showing purchased items and manufactured components to identify what to make and buy
- lead times for the purchase or manufacture of all parts, realistic times taken
- accurate records of stock, work in progress and on-order parts.

Material requirements planning enables parts to be scheduled for the day that they are required. The advantages over the traditional inventory level systems are shown in Table 12.2.

Table 12.2 Advantages of MRP over traditional inventory levels

Inventory-level system	Material requirements planning
Treats each part individually	Deals with structure
Depends on demand history	Looks at future plans
Assumes average	Handles erratic demands
Aims to keep stock levels up	Holds stock only to cover demand
Priorities inflexible	Sensitive to priority changes
Mainly runs itself	Needs managing

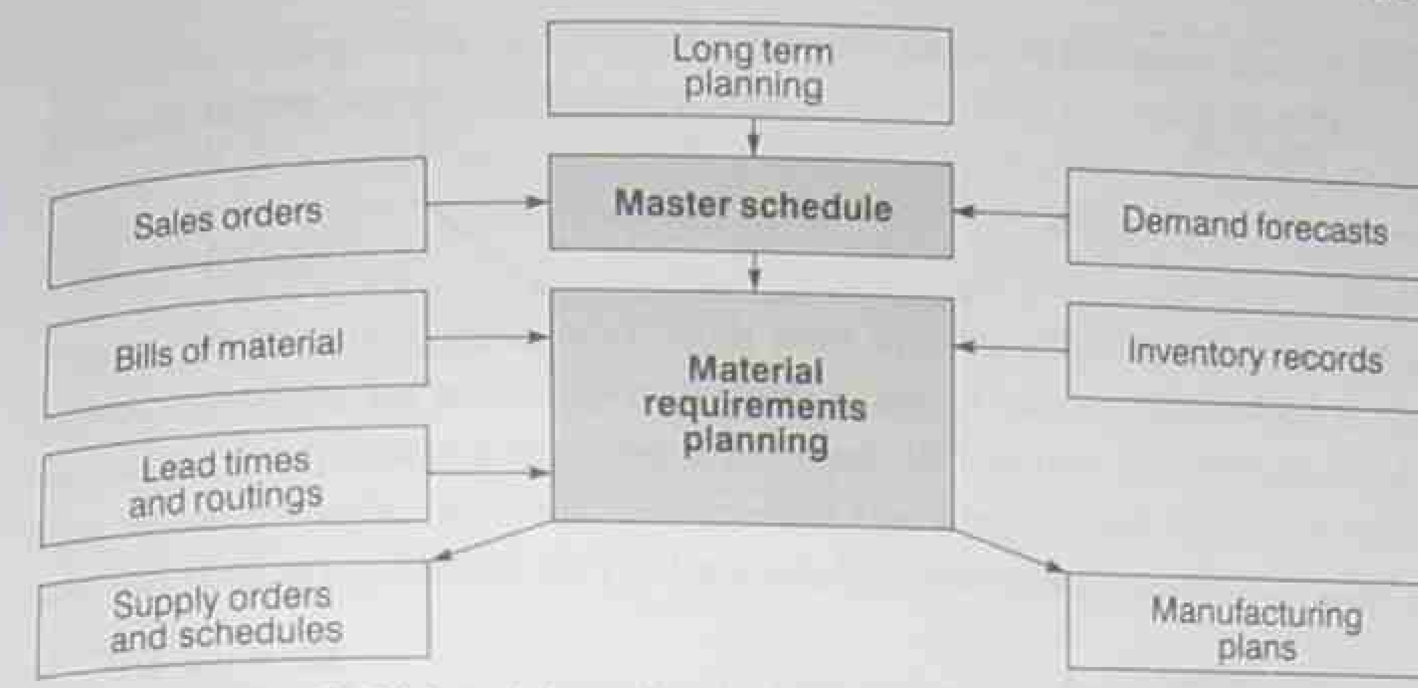


Figure 12.2 Material requirements planning

The two main attributes of MRP, namely 'time-phasing' and 'structure' are now being applied in a wide variety of stocking situations outside manufacturing. If it is known, or can be forecast, when major customers want their supplies, any business can create a master schedule, with the consequential reduction in stocks.

Material requirements planning is a business planning tool as well as a material supply calculation. The planning of MRP is shown in Figure 12.2. The process, known as manufacturing resource planning (or MRP II), starts with a long-term plan, which identifies what range of items are offered, how fast, with what market focus, and all the policy decisions to enable inventory control to operate successfully. As a result of this long-term process, actions should have been taken to ensure that the resources (people, plant, systems, logistics, organization) are available to make the policies work successfully. This long-term plan stretches ahead at least six months and, typically, by two to three years, depending how fast-changing the market is. Long-term planning is usually carried out by product group rather than by product. A master schedule is generated from the long-term plan, and is simply a list of what the business is going to send out, allowing for the real constraints (see section below).

Material requirements planning logic

The MRP backward scheduling process starts by identifying the customer requirement in the 'master schedule'. Once the output to customers has been decided, then the focus is on achieving it. If there is already stock, how far ahead on the schedule will it last? If there is no inventory, then what has to be purchased – the item itself or are there components which are needed?

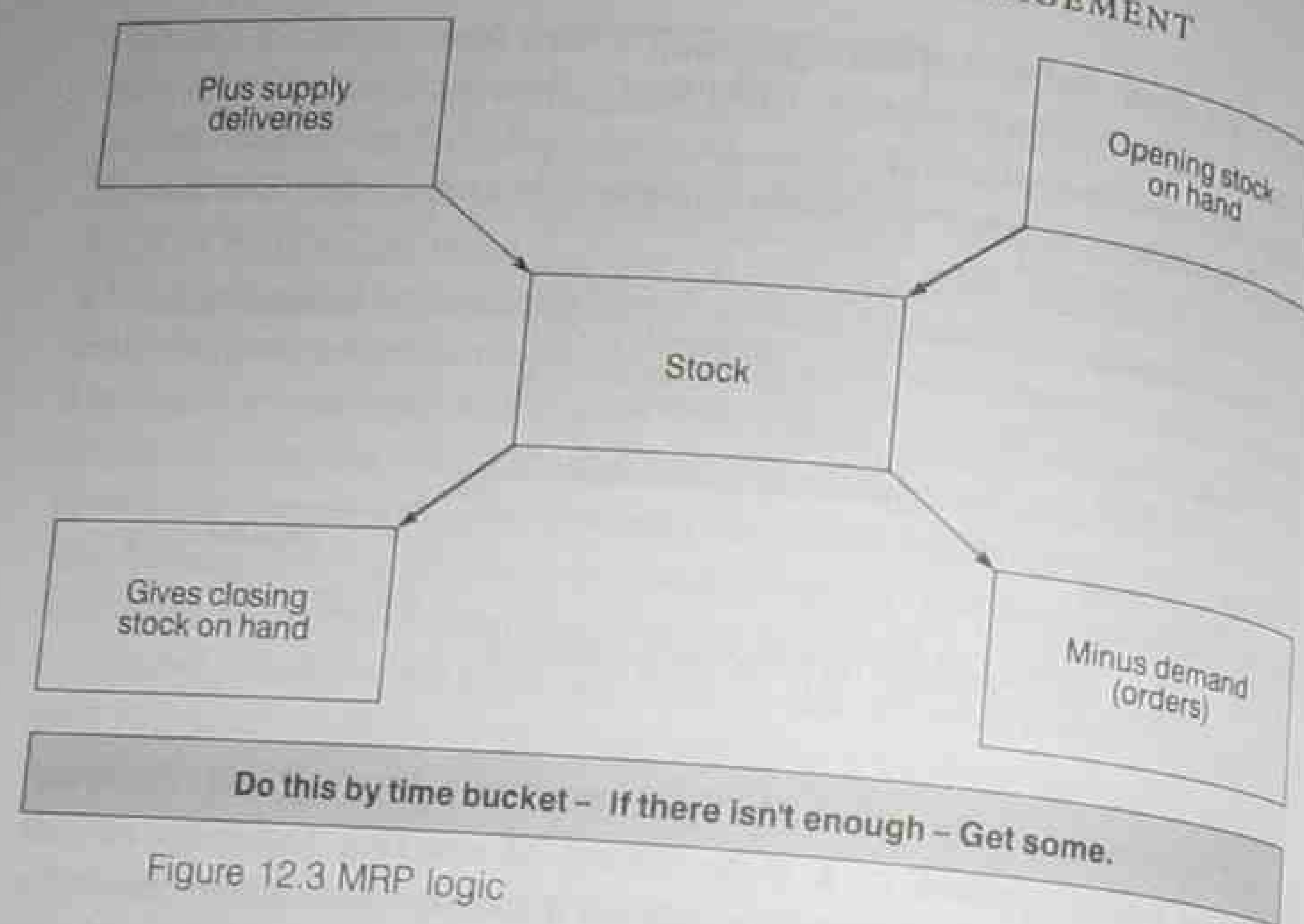


Figure 12.3 MRP logic

A bill of materials is required to cover this. At the same time, information on how long the supply will take and where it will be done – the lead times and routings – is needed. All this information is fed into the MRP calculation.

The actual MRP calculation is very simple. First it batches up all the requirements within each time bucket (week or day) so that there is no differentiation between different customer orders. Inventory is deducted from this total requirement. If the inventory is sufficient to cover the demand, then the same calculation is carried out for the next period. If the inventory is insufficient then delivery (order receipt) is required.

The process is simply shown in Figure 12.3.

$$\text{Initial stock} + \text{Receipts} - \text{Demand} = \text{Closing stock}$$

If closing stock is above zero (or safety stock) repeat calculation for next 'time bucket'. If closing stock is below zero (or safety stock) then create more receipts by ordering more. This then causes extra supply for orders and demand for that component.

This netting off process is carried out for all periods of the final process first, and then it is repeated for the supplying level (components or assemblies) if there is one.

The result is a list (or a series of lists) of what has to be done by when; a list of when to place and receive purchase orders and, where appropriate,

a manufacturing 'work to' programme for each production area (called 'manufacturing plans' in Figure 12.2).

Bill of materials

Classic stock control deals with each item individually, independent of all others. However, for most inventory, the usage of one item is linked to that of others, e.g. printer cartridges and paper. Inventory controllers need a way to acknowledge this link and to save the work of two forecasts where one will suffice (see Table 12.1). This can be done using an inventory bill of materials.

Traditionally, a bill of materials (BOM) is the list of parts, ingredients or materials needed to produce or assemble the required end product. The bill of materials for a product is not simply a parts list. It contains more information. Most products are packaged and the box, packaging material, pallet and even the documentation is part of the full BOM. The BOM can also include essential tooling for production. This is a useful way of planning for all the resources necessary for the process.

In manufacturing, an assembly is made up of components. If they are all bought out and given to an operator to assemble and pack, then the BOM

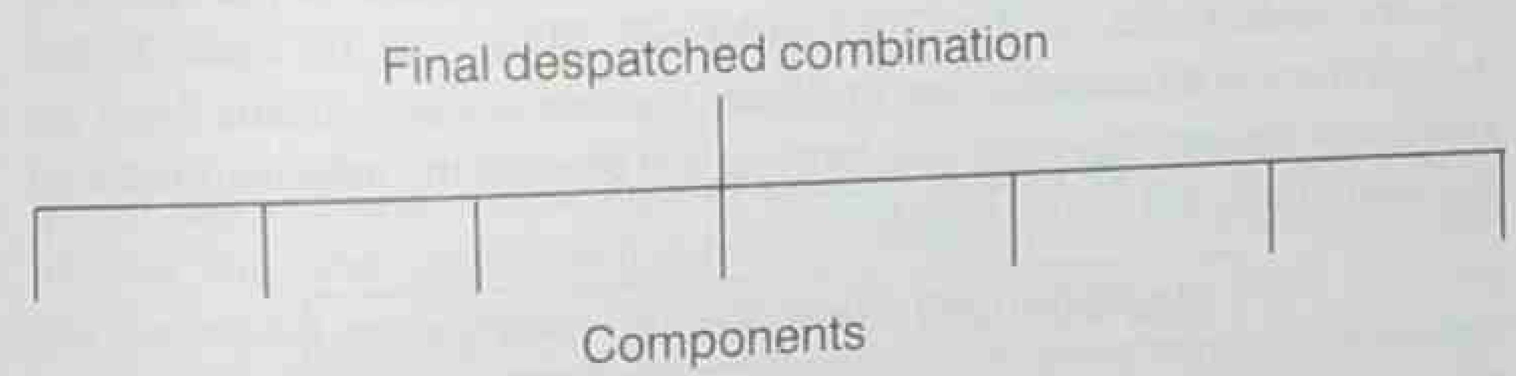


Figure 12.4 Balanced demand – bill of materials

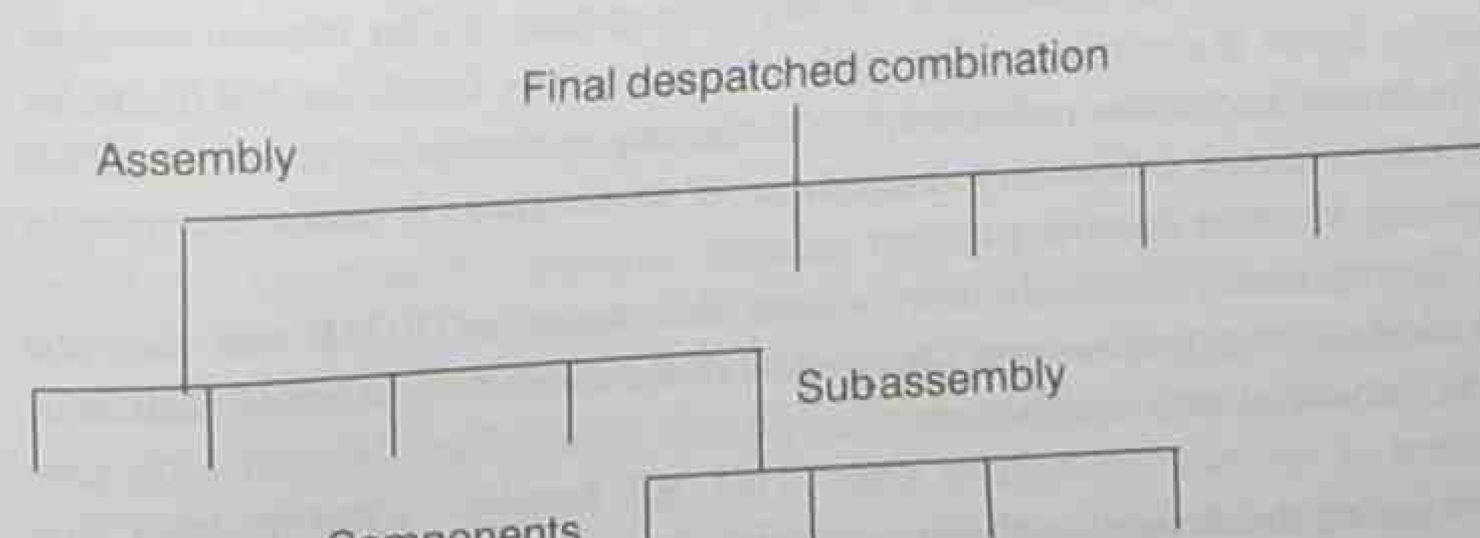


Figure 12.5 Structured demand – bill of materials

is as shown in Figure 12.4. This is a flat BOM. If the assembly and packing are separate, then the BOM is as shown in Figure 12.5. In a supply chain this would correspond to different suppliers of items which are to be sold together in the end.

Time delays between processes require recording and therefore add levels on to the BOM.

Planning the structure of supply

> **Example**
>
> *In many situations there is a balance between the usage of different items. For example, DIY stores know that someone fitting a door hinge will probably need screws. They will either have the screws added to the pack, or remind the purchaser that they are needed.*

Many activities result in the use of a variety of items, and if this usage is a fixed ratio, or even if the ratio of usages is consistent on average, then the ratio can be expressed in a loose 'bill of materials' .This could be used by a supplier providing a range of items to major customers. If the causes of use are understood, a planning (statistical) BOM can be created even though there is no manufacture involved. The use of this tool cuts down the amount of forecasting, since one forecast will provide the information for all the items included in the BOM.

> *Find similarity in demand, then use it.*

The inventory control BOM usually consists of two levels, the same as a JIT manufacturing bill. The top level is the demand profile for the collection of stock items. The level below is the dependent demand for each of the item codes. Since the relative usage rates are not entirely fixed, the bill has to be monitored and corrected according to the usage changes of the ratios. This situation is very similar to a manufacturer where there is major variation in process yield or there is product scrap.

The demand for each item within the statistical BOM will vary for individual customer orders. This means that actual safety stock of each item will consist of two components:

- safety stock against the error in the overall forecast
- individual line safety stock against the variances in the demand BOM.

In practice, the statistical BOM is an excellent and underutilized method for distributors to give enhanced service where customers buy a range of items.

Master planning

Master scheduling

Improvements over classic stock control can also be made by considering exactly when the demand will occur.

> **Example**
>
> *If there is a known demand for fifty this week, a hundred next week and sixty the week after, a knowledgeable traditional controller could work out the average demand as seventy per week with an MAD of twenty and then decide on the appropriate safety stock. A different approach would be to identify the demand in each week as fifty, a hundred and sixty and arrange supply accordingly.*

To do this requires a new type of record. Instead of the normal record of what has been issued each week, the requirement is for a list of what will be issued each week, showing individual weeks (or days) into the future. This is a master schedule. The result of using a master schedule instead of a review level is that the inventory is kept lower and large fluctuations in demand can be accommodated. There is still the option to keep some safety stock, calculated from the inaccuracy of the schedule each week.

The master scheduling process for receiving orders converts the existing demand forecast into firm customer demand by placing the order on the master schedule. This is a management function, not a clerical function. The master scheduler is responsible for arranging customer demands into an achievable plan. Normally, particularly outside the demand lead times, the forecast quantity will be greater than the total quantity ordered for each time period (a day or a week). However, if orders are received which exceed the forecast, decisions are required to deal with this.

Example

If the forecast is a hundred per week and the first order is for seventy in the first week, then it is no problem. If a second order is received for thirty for the first week it can also be supplied. However, if a third order is received for fifty for that week, this will have to be met from the forecast in the second week unless it is convenient to change the schedule.

The logic of allocation is, therefore, that the forecast is filled up by real customer orders to the forecast level each week. However, if the customers require more, then the master scheduler has to decide whether to:

- provide later delivery or part shipment to the customer
- reallocate goods already promised to another customer
- overload the forecast and cause panic buying and inefficiency.

The last option is commonly used but should be avoided since it upsets supplier relations. There is sometimes the option of increasing capacity by overtime, but the material supply chain may not be able to react to short-term schedule changes.

The master plan in its simplest form is a list of how many of each item the company is intending to supply to the customers each week (or each day for JIT business) – see Table 12.3.

Master planning takes the forecast and plans to supply parts to meet this demand at the right time. In the more detailed master plan (Table 12.4) the scheduled receipts are the delivery date of current firm outstanding supply orders, and the stock on hand is the inventory which results if no action is taken. The first three columns show the current situation. The remaining

Table 12.3 Master schedule

Item	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Product 1	1000	500	2300	600	1300	1300
Product 2	300	150	450	300	350	600
Product 3	3000	2000	3000	3000	3000	3000
Product 4	1500	~	~	~	~	~
Product 5	~	~	~	~	~	~
Product 6	~	~	~	~	~	~

Table 12.4 Planned supply – planned order release

Week	Forecast	Scheduled receipts (a)	Stock on hand	Planned order receipts	Projected available	Planned order release (b)
Current			500		500	0
1	3000	5000	2500		2500	0
2	4000	5000	3500		3500	5000
3	2000		1500		1500	0
4	2000	5000	4500		4500	5000
5	4000		500		500	5000
6	3000		–2500	5000	2500	0
7	1000		–3500		1500	0
8	3000		–6500	5000	3500	0
9	4000		–10 500	5000	4500	0
10	2000		–12 500		2500	

Note: Assumes a fixed batch size of 5000 and a supply lead time of four weeks.

three show what has to be done to meet the forecast demand (i.e. plan in new supply).

Set up supply using projected available.

The planned order receipts are calculated to minimize stock and, at the same time, give the information necessary to supply forecast. The resultant stock level is shown in the next column headed 'Projected available', i.e. the stock levels resulting from these actions. With a four-week supply lead time, the order release (last column) has to precede the order receipt by four weeks. If a planned order release occurs in the current week, then an order has to be placed. Alternatively the planned order release column is used as the time phased gross demand for the items on the next level down on the BOM, or as a supply schedule for purchasing.

Placing the order results in the order quantity being deducted from the planned order release and the 'planned order receipt' being transferred to a scheduled receipt. If a safety stock is appropriate instead of letting stock go down to zero, then the stock on hand can trigger a planned order receipt at the safety stock level instead of negative stock. A safety stock of 1000 would mean that stock would be too low by the end of week 5 (Table 12.4). The planned order receipt and release would therefore be required a week earlier.

Apply it to customers using available to promise.

Delivery promising

> The master plan shows what is available for customers and when.

When the customer orders are received, the forecast quantities should be available. The calculation of projected available and planned order release enables the supply to be planned. This calculation accepts the forecast and works out what supply is required to meet it. The amount of the item remaining to fulfil outstanding demand can be calculated in a similar way

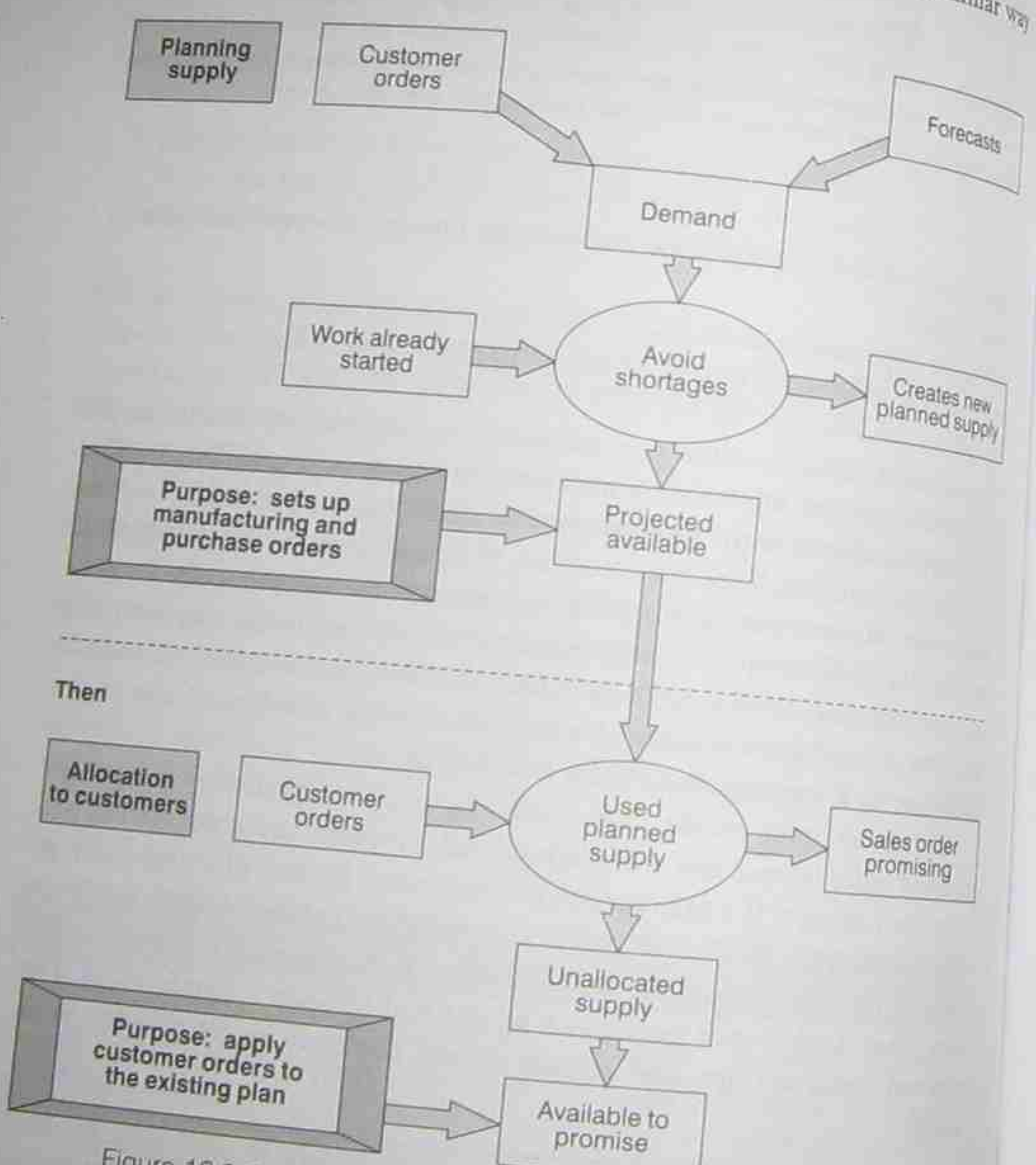


Figure 12.6 Control of supply and demand

Table 12.5 Available to promise

Delivery promising	Customer orders	Scheduled receipts	Stock on hand	Planned order receipts	Available to promise	Planned order release
Current stock			500			
1	2600	5000	2900		2900	
2	6600	5000	1300		1300	
3	1200		100		100	
4	3200	5000	1900		1900	5000
5	1200		700		700	5000
6	600		100	5000	5100	
7	4000		–3900		1100	
8	100		–4000	5000	6000	
9			–4000	5000	11 000	
10			–4000		11 000	

by looking at the actual orders instead of the forecast. The allocation of the projected available stock to real orders enables customer delivery times to be estimated using the MRP logic. This process is shown in Figure 12.6.

Creating an order book according to Table 12.4 gives the availability shown in Table 12.5. This shows the actual customer demand at the present moment but orders are constantly being received, so the situation may well change within a few minutes. The scheduled receipts and the planned order receipt are the values calculated from the projected available stock shown in Figure 12.6. The stock on hand and the available to promise columns are calculated from the real customer demand column.

> **Example**
>
> *In week 1 there are 5000 received and 2600 issued to customers. The stock which was originally 500 is therefore 2900 by the end of that week, and these could be used to fulfil customer demand. However in week 3 the available to promise is down to 100, and the supply lead time is four weeks. If more than 100 out of the 2900 are issued in week 1, then there will be a shortage in week 3. The delivery promise for quantities up to 100 can therefore be in the current week; the delivery promise for up to 1900 can be week 4, and greater quantities can be promised either up to a maximum of the forecast or in excess of the cumulative forecast, as long as the forecast is changed to match the demand.*

This change has to take account of the fact that the large demand is probably part of the forecast and therefore the forecast for other weeks will need to be reduced.

The two sectors – projected available and available to promise – ensure that customer demand is managed in line with supply. It avoids difficulties which result from accepting orders on a fixed or short lead time which then cause emergency sourcing or failing to provide delivery on time to other customers.

Batch sizes

The master schedule is an excellent tool for managing and monitoring stocks, especially where there is variable demand which can be estimated. It enables lower stock to be held than can be achieved by statistical stock-centred methods. It also links in very well with supplier scheduling. Getting the best out of master planning does require more effort because of the regular reviews and reforecasting. However, in a pure stock environment, it is only necessary to use MRP on the A class items and items where the demand is understandable and variable. The triggering of MRP by this schedule enables companies to forecast for whole ranges of stock with a single evaluation and give a balance of stock for customers who require co-ordinated supply.

(In classic MRP, a number of batch rules have been developed as many of these are based on 'economic batch sizes', which produce batch quantities which are unbalanced and too large for customer order requirements.)

The same logic can be used for MRP order quantities as for normal stock control. The three main alternatives are:

- lot for lot
- fixed batch size
- period cover.

Lot for lot is the first choice because it leads to a minimum of inventory.

> **Example**
>
> *The rule is to bring in the balance required in each time period. In Table 12.4 the scheduled receipts for weeks 1, 2 and 3 would have been 2500, (3000 forecast demand less the 500 stock) 4000 and 2000, and the projected on hand would always be zero.*

This gives a variable supply quantity which is inconvenient is some circumstances.

Fixed batch sizes are used widely to illustrate MRP. They are overused in practice since changes in demand level should be reflected in changes in order quantities. Fixed batch sizes are convenient to use but are only essential where there is a restriction in vessel size (distribution or manufacturing). Batch sizes based on price breaks are the result of too simplistic an approach to costing.

> *Choose batch sizes to minimize work and inventory.*

Period cover is a way of taking the demand for the next few periods and batching it up. It is the same as lot for lot but with several time periods added together. Again the order quantity is variable, and the stock level is significantly higher. This technique reduces the number of orders and increases their size at the expense of higher inventory. The application is therefore to C class items where this is the most beneficial option.

> **Key points**
>
> - Look for structure in demand patterns.
> - Working with dependent demand keeps inventory low.
> - MRP is a simple way to trigger future supply orders – especially where order levels vary.
> - Using 'available to promise' gives more accurate delivery dates.
> - Batch sizes have to be defined using rules which minimize total costs and workload.

13

The future – inventory and logistics

The basis of the lean supply chain

Extending inventory management

The science of inventory management has gradually developed, not because the theories have leapt forward, but because requirements change. No longer is inventory management a single problem but it is integrated into the whole of logistics of supply. The manager is now co-ordinating

- warehouse operations
- inventory control
- distribution
- suppliers
- customer operations

and attempting to optimize the profit for the supply chain rather than just keep delivery on time up and inventory down.

However, the targets are still the same but in a wider context:

- better customer service
- lower inventory
- decreased operating costs,

but now the focus is on the whole supply being lean, not on the individual echelon.

The development in information technology (IT) communications enables information to be gathered and exchanged immediately and accurately, and makes the information available to everyone. However, the challenge is now:

- collecting data accurately
- being willing to share it along the supply chain
- understanding how to use it most beneficially
- working for the good of the supply chain and not for political gain.

The techniques discussed in previous chapters are the basic methods to be used in each step of this chain. They then have to be put together to make the chain effective and competitive.

Lean supply

The concept of lean supply concentrates on the very essence of good inventory management:

- customer focus – exceeding the customers' expectations
- aim for perfection – not just better than the competition
- added value – only carry out activities which directly benefit the customer
- flow – so that inventory does not stop, it continues from one stage to the next
- unit movement – small batches (e.g. kanbans) moved frequently
- simplicity – in operation so that things are obvious and do not go wrong
- individual responsibility – at operator level for developing more effective operation
- integrated design approach – with additional features for the customer and ease of manufacture at low cost
- quality – always create the correct quality first time.

> Use lean supply ideas to see how to develop inventory control and logistics.

Although there are many principles here, they are all interlinked and working on one is supported by working on another. Improvement is a circular process of tightening up one area then moving on to the next and eventually back again to the original as it has become the weakest area again. When companies change to a lean attitude there are a lot of minor changes which add up to major developments. These improvements are made continuously, so the operation gets better and better.

Lean supply involves the whole business from purchasing through to dispatch. The achievement of good flow of supply requires this full participation. Purchasing has to ensure that goods are delivered on time, they are fit for purpose and labelled and packaged correctly and consistently. A design department has to ensure that a small number of products satisfy a wide range of customers and are composed of as small a number of components as possible. They should also ensure that components and modules satisfy the widest range of products. What is more the lean approach causes a continuous reassessment of the design, taking into account the change in customers' values and the needs to rationalize production processes and components.

The quality department does not inspect; it sets up the processes which either avoid mistakes occurring or enable the work to be checked as it is carried out. Purchasing is not concerned with lowest unit price of purchases, but is involved with minimizing the total cost to the customer and maximizing the value to the customer. Manufacturing need to focus on supporting customer requirement and not on throughput.

The break with traditional concepts is that improvement is part of normal work for everyone. The gradual development of the operations is led by the people carrying out the processes – the management play a supportive and guidance role rather than being the promoters of change.

The foundations of lean supply lie in the development of JIT processes described earlier in this book. Lean supply takes this concept wider so that greater benefits are gradually accrued.

Information accuracy

> More accurate records will reduce inventory holding.

All the work on inventory management will come to nothing if the base information is not correct. There are many companies where the quality of the information on the system is very poor. If the same quality standards were applied to products, then these organizations would quickly be out of business. Part of the development of better inventory control is therefore to ensure that the information on the formal system is

- real time
- accurate.

Computer systems are accurate, so it is a matter of making failsafe the information systems and making people believe that it can be correct. Inaccurate records have a high cost arising from continuously checking to see if items are there and holding extra inventory to avoid stockouts, not to mention the obvious cost when customers cannot have the items promised to them. It is another challenge to get stock records right. And this goes hand in hand with the development of inventory control.

Control methods

The achievement of tighter targets has largely been the result of better communication and simple controls. Customer pressure has improved delivery performance, and financial pressure has decreased operating costs.

The key tools have been the exception reports of 'these stocks are too low' and 'these stocks are too high', which keep the stock broadly correct so long as orders are placed at approximately the right time. The use of ABC analysis, dynamic review levels, forecasting techniques and JIT supply enables inventories to be improved significantly. Effective implementation of these techniques has been slow because conditions did not encourage it. In the past, supply lead time was unreliable so that review level calculations could not be relied upon. Now that suppliers are expected to guarantee reliable delivery the calculations give the correct results and there is no need to add a bit extra (lead time, inventory, etc). The techniques described in previous chapters have proved invaluable in meeting the demands of the new market environment. Once a company is operating these successfully it is ready to contribute to the logistic effectiveness of the supply chain.

There are other techniques, built into some advanced software, which are waiting to be used more widely, and more advanced inventory control techniques will show significant advantages for the user. (It has been known for decades that Newton's Law of gravity is not quite right, and Einstein's theory of relativity is nearer the truth, but Newton's Law is good enough for those of us who are not space travellers or who move below the speed of light. Similarly in inventory control, the review level calculations are not exactly correct and delivery frequency does affect the safety stock.)

> Aim to get it nearly correct now rather than perfect sometime in the future.

Better formulae are readily available, but from observation, stocks are so far out from the ideal that a small amendment to the safety stock level is not going to alter the real inventory. Similarly Gaussian statistics, simple availability calculations and independent demands are all concepts which are questionable theoretically, but which work in practice.

More advanced forecasting methods provide better answers when the future can be relied upon to mirror history (see Chapter 11). Crude tools like ABC analysis could be replaced by tools which take into account the potential different stock policies for high-value slow movers and low-value fast movers. Or ABC could be developed into a differential value analysis which gives an individual class for each item code and optimizes every inventory line. This evaluation will take place when companies have exhausted the current methods and realize that they can benefit.

Consignment stock (consignment inventory)

A popular technique in the supply chain is the creation of consignment inventory. This can be defined as 'Inventory owned by a supplier at the customer's site'. (There are a variety of similar arrangements to this, but the essential feature is that the inventory is near the customer and owned by the supplier.) On the face of it there seems to be great benefit. Sales people like it because:

- there is effectively a captured market
- the customer should not have shortages
- supply can be slow and regular to stock instead of short lead time to customer order.

Customers like consignment stock because they can have:

- immediate availability
- supply without problems
- ability to change requirements at will
- less administration when ordering.

However, consignment stocks are a bad thing. By reviewing the situation from a distance, it can be seen that a consignment stock will:

- require the supplier to have an extra stores stock
- have inventory in the stores which is relatively slow moving
- take the pressure off efficient supply
- transfer the responsibility for the inventory from customer to supplier.

Looking at these points in turn, the first arises because the consignment is stock in a remote location and, therefore, not available to other customers. There may be duplication of stock between the supplier's own stores feeding the consignment and the consignment inventory itself.

The second point is that a smaller stores has throughput because the MAD for small throughput is relatively large (decreases by the square root of the volume). Therefore the consignment will have a surfeit of stock, even if it is controlled very well. (The chances are that the inventory level is either set very high by the supplier or agreed with the customer at an inflated level.)

If the supply chain with the consignment is in competition with one without, then the one without will have lower cost, and therefore a competitive advantage. The consignment stock is a palliative to poor planning: it may however be a short term (few months) solution to cover a specific problem.

There are a few simple rules that should be used when applying consignments:

- Do not set stock levels; set availability targets instead.
- Review variety of inventory and levels regularly (every few weeks).
- Use consignments only for a few fast-moving item types.
- Arrange who is charged with inventory left over through change of design, demand pattern and any other non-moving inventory.
- Allow consignment to be available for other users if necessary.
- Have a formal agreement for a limited period of time.

> Consignment is only a temporary fix.

Logistics

Objectives of logistics

Logistics management presents new challenges to the inventory manager, who has to maintain and balance distributed stocks, additional inventories, manufacturers' work in progress, supply partnerships, and other new dimensions, including distribution costs. The need for this integration results from the changes in market structure given in Figure 13.1.

The development of supply chain strategy means that stock controllers from suppliers and customers should be sharing information and inventory in order to optimize the operation of the supply chain.

- >Stockholding has become strategic warehousing
- >Transport has become reliable
- >Distribution is faster
- >Operations are international
- >Outsourcing is the norm
- >Customers require specific solutions
- >Pressure on costs and inventory is greater
- >Warehousing, distribution and customizing costs are being optimized

Figure 13.1 Changes in the distribution environment

Management of logistics is a simple extension of inventory management, characterized by the same basic underlying themes. The most successful supply chains are:

- integrated
- consistent
- fast moving (short lead time)
- service oriented
- customer specific.

> Optimize the complete inventory structure, not each stage.

Supply chain structure

The management of a supply chain requires inventory to be held in different locations, even in different operating companies. The inventory controller can then optimize the stock across many locations. The structure of the supply chain can be altered to meet the needs of the market, and the overall balance between availability, stock investment and operating cost can be optimized by selecting the number of stages (echelons) in the supply chain and the number of stock sites in each echelon.

This can be established most simply by modelling the cost balance: generally it is more economical to hold inventory as near to the start of the supply chain as can be allowed by the demand lead time. Once the distribution structure has been decided, the proportion of stock held at each inventory level has to be calculated, as shown in Table 13.1.

Table 13.1 Supply chain inventory options

All stocks in a single central warehouse	Stock all held by end user or local stores
Minimum investment in stock	Items are available more rapidly
Accurate forecasting	Sensitive to local conditions
Higher demand rate	Duplication of stockholding
Reduced safety stocks	Poorer control
Low warehousing costs	Higher inventory investment
Better formal physical control	Greater risk of obsolescence.

The compensating factor in deciding where to hold stock is the distribution cost. In general, it is considered that:

- fast delivery costs more
- the greater the distance, the greater the cost
- the smaller the load, the higher, relatively, is the distribution cost.

However, there are other considerations to make, including the following:

- Regular shipments cost less than one-offs.
- Distribution costs are largely loading and unloading, mainly independent of distance.
- Consolidation saves cost and is easier for small, regular deliveries.
- Small loads need smaller transport and facilities which are cheaper.

The complexity arising from these latter considerations makes transport costing difficult. When considering JIT supply (Chapter 4) one of the principles discussed was continuous improvement (kaizen). This approach has to be used when reducing distribution costs and, therefore, inventory. However, to have a significant effect on the cost balance, bold leaps have to be made in distribution structure. The size and frequency of shipment have to be altered dramatically. This usually means a change in means of transport to provide smaller units faster. The savings in inventory and warehousing costs should be greater than any marginal increase in distribution costs.

The correct balance of stockholding and the location of the stock are different for each stock line, depending on their usage rates and values. Low-value, high-usage items are better nearer the point of use. High-value items with occasional usage can be held centrally.

The whole inventory system can be considered as a pipeline for supplying the fluid 'stock' to the point of sale (or use). The input to the pipeline is a continuous (scheduled) flow. For the high-usage items the pipeline can continue to the point of use. For low-usage items the movement includes a reservoir, a central stores from which the items are dispensed as required.

This analogy highlights some interesting aspects of the supply chain, and poses the questions:

- Should the system required be a stock control system or should it be a supply flow system?
- Which parts of the supply chain are best served by scheduled supply and which by batch ordering?
- If stocks are held at one echelon, do they need to be held at others as well?

For fast-moving goods, the central warehouse can act as a reservoir replenishing the local storage tanks. It is better if the central store is used purely as a transhipment point for these items. The pipeline is then connected to the distributed stocks. The central function becomes stock management co-ordinating ordering and supply, and ensuring consistent delivery.

Slow-moving items can all be held at the central stores, as long as there is the means to withdraw them and transfer them to the final customer within an acceptable time. This reduces the overall stockholding of these items. The resultant stock is structured with each stock-keeping unit (SKU) having a specified stock level based on its usage and value.

Multi-location supply chains are used commonly in some industries. Using an integrated stock information system, duplication and proliferation of stocks can be avoided. Distribution warehouses in different localities should be treated as a single level where all stock is considered as holdings of a single store apart from the safety stocks. The aim is to keep the appropriate amount at each location and if supply from the originator or manufacturer is rapid, then a central stock location can be avoided, with a corresponding reduction in stock. Careful supervision of delivery time is necessary to avoid duplication of stockholding at several locations.

The degree to which a company can take advantage of these techniques is limited by their organization and systems. Often it is necessary to make improvements as a two-stage process:

1 Introduce the 'stock to stock' system to get the control.
2 When control is good, reduce the stock at each stage in the system by improving the inventory control techniques.

Organizational control

For control of inventory in the distribution chain, the costs have to be minimized and service maximized. Distribution systems are neither 'pull' nor 'push' methods of stock control. Such methods imply that either distribution (pull) or manufacturing (push) are entirely responsible for the organization of distributed inventory. Modern techniques centre around an

- accessible database of information
- realistic schedules
- replenishment of distribution sites based on requirements, lead times and item availability.

> *Start from minimizing the total cost, then work out the nearest practical solution.*

If the supply chain is to be operated most efficiently, then there has to be some co-ordination and organization for the whole chain. This will generally require:

- good information fed back from all the distributed stores (it helps to have integrated systems)
- central decision-making of the allocation of stock.

This is a statistical reality. Assessment of demand is more accurate for the total demand than for individual warehouses, since the larger volume has a smaller MAD. (The MAD only increases as the square root of the volume.) This means that historical forecasting is most accurate when carried out by a central planner. However, if a demand pattern changes (new large customer, loss of contract, etc.) then the local knowledge of these events needs to be included in the forecast (see Chapter 9).

The techniques commonly used for managing the distribution channel are basically:

- fair shares (for allocating available stock)
- distribution requirements planning (DRP) (for planning supply).

Both the fair shares and DRP techniques are concerned with the quantity of stock to be shipped through the distribution network.

Distributed inventory

For an inventory planner controlling stock in several regions, the issues to resolve are:

1 Where the stock is to be held.
2 What level of stock to hold in each region.
3 How to maintain supply at the lowest cost.

The first question is answered by working out the general distribution and inventory cost balance from the factors just discussed in the first section in this chapter. The second issue is easily resolved using the classic inventory management approach described in Chapter 7. The third issue is to ensure that the transfers are carried out most economically within the framework arranged by (1) and (2). This is an optimization process. The cost of providing items at a remote warehouse depends largely on two factors:

- the item cost
- the distribution cost.

There is a unique cost for each item at each site, and SKUs provide the information required to evaluate distribution needs. The stock in some distributed warehouses is likely to be larger than necessary, particularly if they are the delivery points for outside supply, while others require supplying with stock to meet demand. Stock therefore has to be redistributed at the least cost from one location to another at the same echelon of the distribution network. If the costs of distribution between each of the stores is known, then the total costs can be optimized. This is particularly important where transport is expensive, or where there are significant differences in transport costs, as found in a multinational supply operation.

Fair shares

Fair shares analysis approaches the regional stock issue as an inventory control problem, assuming that distribution costs are not the prime concern.

The purpose of fair shares analysis is to distribute stock through the supply chain to alternative distributed storage. Stock is not held at a central warehouse but distributed to regional stores nearer to the customer demand. An implicit assumption is that the supply to the network can be

accomplished by short lead time orders. If the lead time is longer, the planning aspects of DRP have to be used.

When stock is delivered into a distributed stores network, it has to be reallocated to the various stores in the most equitable manner. The sensible approach is to cover priorities first and then share out the rest to meet future demand. Fair shares analysis is a method for achieving this using a standard procedure. The logical process is to identify the safety stocks in the supplying warehouse and in the other warehouses, and then to allocate the total available stock to meet the requirements in proportion to the demand rate.

The fair shares allocation of stock normally distributes stock in the following sequence:

1 Customer back orders unsatisfied.
2 Expected demand within supply lead time.
3 Safety stock at distributed stores.
4 Safety stock if required at supplying warehouse.
5 Allocation of remaining stock in proportion to demand rate.
6 Redistribution where a stores stock exceeds its share.

If stock is insufficient to meet the full requirement at any stage, it is allocated logic. The actual stock transfers used to achieve this distribution can be optimized through the transport algorithm. A simple example of the use of the fair shares logic is shown in Tables 13.2, 13.3 and 13.4.

A manufacturer sells items in batches of 400 to a central supply warehouse. This warehouse holds some safety stock, but distributes the items out to three regional stores. Estimates for the usage at the stores are given in Table 13.2 together with the safety stocks. If a delivery of 400 arrives from the manufacturer just at the start of week 1, the total stock available is 400 plus 200 central stock plus 280 at the regional stores (= 880). Table 13.3 shows the lead times, and develops the supply chain

Table 13.2 Stock and forecast demand

	Safety stock	Current stock	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
North stores	20	70	50	40	70	20	50	50
South stores	46	200	100	140	60	110	150	150
West stores	12	10	20	30	0	50	20	20
Total	78	280	170	130	130	180	220	220
Supply warehouse	37	200						

Table 13.3 Fair shares priority allocation

Location	Average demand	Lead time (wks)	Demand in lead time
North stores	46	1.2	56
South stores	118	1.0	119
West stores	23	0.4	10
Total	187	4.0	185

	Stock	Balance
Current stock	280	
Supply warehouse stock	200	
Current delivery	400	
Total	880	
Shortages	0	880
Demand in lead time	185	695
Stores safety stock	78	617
Warehouse safety stock	37	580

Table 13.4 Fair shares apportionment of stock[1]

Stores	3.1 weeks' demand	Demand in lead time	Safety stock	Current stock	Delivery required	Total stock
North stores	143	56	20	70	149	219
South stores	366	119	46	200	331	531
West stores	71	10	12	10	83	93
Total	580	185	78	280	563	843
Supply warehouse			37	600	−563	

Note: [1]Number of weeks' cover = 580/187 = 3.10 weeks.

calculation. There are no back orders, so the first allocation of stock is to cover regional demand in the lead time (185), then regional stores safety stocks (seventy-eight), then supply warehouse safety stocks (thirty-seven). Deducting these from the total available leaves 580 items to be split up across the regional stores. This can be done either by netting off the individual forecasts until the stock is exhausted (where these are accurate and contain real orders). Alternatively the average week's cover can be calculated. The total demand is 187 per week, so 580 will last 3.10 weeks on average. Multiplying the demand by 3.10 gives the fair share for each store. This is shown in Table 13.4.

Each store is then given a fair share consisting of:

Any customer back orders + Demand in the lead time
+ Safety stock + 3.10 weeks' demand – Current stock.

This is also shown in Table 13.4. All the stock and delivery to the supply warehouse has been allocated, apart from the safety stock (thirty-seven). If one regional store already had had sufficient, then either the excess can be transferred, or that regional store (and its stock) can be excluded from the analysis and a fair share worked out among the others.

Distribution requirements planning

Distribution requirements planning provides planners with a complete forward schedule for all dispatches, arranged by date order by warehouse. Each store at the demand end of the supply chain generates a forecast for demand which is used to identify the orders they are likely to place on their supply warehouse (supply). The forecast is for demand well in advance of the normal replenishment lead times. Since the supply is planned to meet the requirement well ahead, the demand rate may have changed so the resultant replenishment delivery may be too early or late by the time demand is met.

The delivery requirements calculated by the netting logic of DRP show the latest shipping dates associated with each warehouse requirement (see Figure 13.1). Distribution requirements planning is an extension of the MRP process for distributed warehouses. The netting logic used for DRP is exactly the same as for MRP (see Chapter 12, section on MRP logic):

- The future requirement is split into time periods of a week or a day.
- The expected demand each time period is identified.
- Starting with the current period, the demand is netted off against the available stock.
- If the remaining stock is above the safety stock then this quantity is carried forward to the next time period and the calculation is repeated.
- If the available stock is less than the safety stock, then a receipt is required. The quantity requested depends on the batch rules being used.
- The request generates a supply order a lead time earlier. (The lead time corresponds to the picking, packing, transfer and receiving time.)

By using DRP, supply planners can receive a complete forward schedule for all dispatches, arranged by date and by warehouse. Transport must be made to fit the schedule. Distribution requirements planning builds on that master scheduling, and looks at demand far past the normal lead time. The purpose of DRP is to produce a time-phased shipping requirement for the suppliers.

Distribution requirements planning is a method for planning the replenishment for the distribution network. The logic closely follows that of MRP. Like MRP the supply chain is planned by backward scheduling from the customer. This requires the establishment of fixed supply lead times and batch quantities. If this is organized by the manufacturer, the batch size can be large and the intervals of distribution correspondingly long. This, of course, adds stock throughout the supply chain and so the use of a just in time approach with very short lead times is a great advantage.

An example is shown in Table 13.5. In this table the items are distributed from a supply warehouse to three regional stores. The stores have given their estimates of demand (shown as 'forecast') and from these the stock and demand pattern needed to keep the 'projected on hand' positive are given in the 'Stock' and 'Supply' rows. The item is distributed in batches of 100 in this case. (This is the same calculation as for 'projected on hand' and 'planned receipts' for MRP in Chapter 12.)

There is a lead time of one week for delivery to the north and south stores from the supply warehouse. This is caused by the order processing, picking and transport: delivery is very fast to the west stores. The dispatches from the supply warehouse required to meet this are given in the lower half of the table, taking into account the week offset in the two cases. If the supply

Table 13.5 Example of distribution requirements planning

		Current	Wk1	Wk2	Wk3	Wk4	Wk5	Wk6
North stores	Forecast		50	40	70	20	50	50
	Stock	70	20	80	10	90	40	90
	Supply			100		100		100
South stores	Forecast		100	140	60	110	150	150
	Stock	200	100	60	0	90	40	90
	Supply			100		200	100	200
West stores	Forecast		20	30	0	50	20	20
	Stock	10	90	60	60	10	90	70
	Supply		100				100	
Supply warehouse								
Demand upon supply warehouse	North stores		100	0	100	0	100	?
	South stores		100	0	200	100	200	?
	West stores		100	0	0	0	100	0
	Total		300	0	300	100	400	0
	Stock balance	200	300	300	0	300	300	300
	Deliveries		400			400	400	

warehouse receives the goods in batches of 400, the required delivery schedule can be worked out as in the 'Deliveries' line. The resultant stock at the central warehouse is also shown. Once demand at this supply echelon has been determined, the demand on manufacturing can be placed into the master schedule. Again a supply lead time has to be allowed for the distribution process.

Distribution requirements planning is far simpler than fair shares and is easier to manage. An important practical difference between DRP and fair shares is that DRP relies on lot-sized quantities with pre-scheduled dispatch times. In DRP there is normally a resultant stock caused by the batch sizes, as illustrated in Table 13.5.

Some pitfalls which have to be avoided when using DRP rather than fair shares are:

- DRP is less flexible.
- It focuses on each site rather than the total system.
- Its sole concern is the balance of forecast and available stock, with little reference to customer service.
- It fixes and depends on the supply performance.

The technique is relatively rigid, since the quantities must be sent on the days scheduled. In this sense it is limited in value for transportation planners (transport must be made to fit the schedule). With DRP, quantities shown are planned, with the implication that stock is not usually currently available. Consequently, changes to shipment schedules have to be long term rather than immediate. There is little possibility of responding to shortages since the quantities shown are planned to be available, not currently available. Distribution requirements planning is best suited to situations of relatively steady demand: it is most successful where fixed delivery quantities are physically necessary (such as container loads).

Fair shares can be made to work in conjunction with DRP by making the best allocation of the shipments arriving into the network.

Wherever possible, the manufacturing and distribution systems should be synchronized, thus reducing the amount of inventory carried throughout the system. This could be achieved in the example shown in Table 13.5.

The interface between the computed lot-size requirements and the master schedule is the gross requirement from the supply warehouse, taking into account any finished goods stock held by the manufacturer. In practice, the DRP gross requirement can be exceeded by the master schedule and products can be planned in advance of the DRP schedule dates if it is beneficial for manufacturing planning and efficiency.

The distribution quantities calculated by the forecasting techniques and apportioned by fair shares often have to be adjusted because of other influences. The total shipment may have to be phased because of space limitation in a particular stores. The distribution quantities could well be rounded to the nearest pack size, or altered to take advantage of a distributor's cost structure.

Traditionally forecast demand would be made at the product level. Distribution requirements planning allows the forecast to be conducted at the SKU level. A consequence of this is the need to aggregate the local forecasts to generate forecast demand for each product and, ultimately, for product families.

It will have been observed from the example that DRP results in considerable unevenness of the gross requirements. This is caused by the fixed lot sizes. If the lot size were smaller, this effect would be minimized, but the number of deliveries would be increased.

Review

The development of inventory management is a dynamic activity, with new approaches being made, techniques being refined and new challenges being met. Successful inventory management is a balance between the use of basic techniques which are easy to use and give reasonable results, and sophisticated techniques which can give better control if they are used properly. Modern inventory managers have the opportunity to use better and more effective controls.

It is imperative that the challenges are met, since continuing competitiveness is based on ever-improving customer service and ever-reducing inventory costs. The approaches discussed in this book give companies major improvements in their efficiency across all types of business. Understanding the full implications and applications is a gradual process, but worth the effort because of the tremendous benefits.

The science of inventory management as developed in this book has to be carried out by an organized, aware and dedicated artist. The manager has to focus on the simple basic requirements of *creating the most profit for the organization through the best customer service*. This means:

- excellent forecasting
- low inventory value
- small operating costs.

The direction is towards perfect service, no costs and no inventory. To manage this it is essential to have operating controls and targets – monitoring

of current performance and tighter targets (which are achievable and agreed with those required to reach them).

There is also the balance between long-term and short-term requirements. Paying extra delivery costs to help the customer may be good for relations. As a one-off it is fine, but as a frequent activity it can damage the viability of the business. The inventory controller has always to look simultaneously at the long-term and short-term detailed picture and overall plan. By steering a consistent course, based on a set of operating principles, work is reduced and effectiveness increased.

As better control is introduced and the performance improves in one area, so it can be integrated with other activities and new challenges adopted.

Take a simple example of a manufacturer that used to hold raw materials inventory and then supply the finished items into a warehouse from where it was distributed to regional stores. Each stage of the process was struggling to gain good control in its own separate way. (Some people are still at base 1 and wrestling with these challenges.) There is a natural process of evolution from this setup:

1. Gain control over manufacturing (using MRP for example).
2. Add the finished warehouse stock to the manufacturing control process (form materials management activity).
3. Include distribution costs in the cost balance.
4. Form partnerships with suppliers to avoid duplicating the raw material stockholding.
5. Introduce flow through the business and work on the supply pipeline to eliminate work in progress and stores inventory.
6. Integrate the distributed regional stores into the inventory management system.
7. Optimize inventory along the supply chain.
8. Appoint supply chain management to optimize supply chain costs.
9. Extend the integrated management throughout all stages of the supply pipeline.

> *If the company is using the same methods as five years ago, it has a problem.*

This may be a large step for companies where the supply chain members are not all fixed and where there are many companies involved. However, it is necessary if businesses are to remain competitive. Other companies are developing and improving, so maintaining the present situation is not a viable option.

The key to success is simplicity of operation. The concepts can be complicated, just as reality is, and the inventory management system has to reflect this. It does not mean that the actual day-to-day operation should be difficult. In fact, the more sophisticated the system, the more reliable it should be and the less effort to operate. However, a good system requires setting up to operate properly and this requires good inventory management expertise, using the professional techniques described in this book. Using the forecasting described here, along with ABC analysis and stock level calculations, the author has saved millions of pounds worth of inventory in many companies. Most businesses want to generate more profits: many have not recognized the key role of inventory management in increasing turnover and reducing costs.

Inventory management is a collection of techniques developed by practical people and proven in the real world. It is no use using general good management practice in this area where there are specific inventory control methods required. These have taken many years to develop and refine.

There are refinements beyond those described in this book, but in practice the methods discussed give such good benefits that other areas become the focus of improvements before these are required. In these pages there is an attempt to illustrate the application of the methods as well as the techniques themselves. The reader should be able to gain significant improvements in the stockholding simply from implementing the concepts described – and that is the next challenge. Understanding the full implications and benefits is a gradual process, as is the implementation of the techniques. Taking up the challenge is definitely worth it because of the tremendous impact on the business.

Key points

- When the inventory in a stores has been improved, consider the supply chain as a whole.
- Minimize the logistic costs whilst exceeding customer expectations.
- Decide upon the best places to situate the stock and the amount of stock to hold in each, using inventory management principles.
- Employ the techniques described in this book for the best results, then add expertise to enhance those items which are important, new or exceptional.
- Set up your systems to take the work out of inventory control.

Questions and answers

Questions

1 What are the three key objectives of inventory control?

2 Which departments put pressure on inventory managers, and what are their objectives?

3 'More stock gives better availability.' Challenge this conventional view.

4 Two companies distribute fast moving consumer goods. Their results for last year were:

Annual accounts data	Ace Distributors (£ million)	Deuce Associates (£ million)
Sales turnover	8	26
Fixed assets	2	5
Cost of sales	6.2	21
Stock value	4	8

You have been offered the job of inventory manager by both companies. Which one would you take and, based on the data provided, explain why.

5 The aim for Ace Distributors (in question 4) is to reduce the stock to £1.5 million. What will be the return on assets then?

6 What change in the return on capital (as a percentage) will there be when a company has reduced its stock value by £500 000, given the following data:

$$\text{Return on capital} = \frac{\text{Annual profit}}{\text{Total capital employed}}$$

where

capital employed = stock + capital plant

sales turnover = £3 million
profit = £400 000
capital plant = £1.1 million
stock before reduction = £1.5 million

7 Discuss the ways in which the JIT approach differs from conventional inventory management.

8 How does a company give good customer service? A company is distributing a range of fast-moving items to stockists. The following is a list of dispatches for one day. What is their customer service level, assuming this to be a normal day?

Order No 1A

	Quantity requested by customer	Quantity dispatched from stores
Item no. 1	20	15
Item no. 2	100	100
Item no. 3	50	20

Order No. 2B

	Quantity requested by customer	Quantity dispatched from stores
Item no. 1	60	40
Item no. 2	150	100
Item no. 3	20	20

Order No. 3C

	Quantity requested by customer	Quantity dispatched from stores
Item no. 1	10	10
Item no. 2	70	70
Item no. 3	50	20
Item no. 4	25	25

9 Which items in the following list are in the A, B and C classes? How do you determine this?

Item code	Weekly demand	Unit price (£)
A901	2	3.00
B662	2	8.00
C355	20	4.00
D523	40	0.50
E191	10	65.00
F807	1	9.00
G010	3	1.00
H244	1	170.00
J459	10	3.50
K488	22	0.50

10 If the current stock of the items in question 9 were as shown below:

Item	Stock quantity
A901	6
B662	10
C355	50
D523	60
E191	50
F807	50
G010	12
H244	5
J459	20
K488	5

(a) What is the stock cover for each of the items?
(b) What would be the first task resulting from this analysis?
(c) What is the stock cover value for the total stock?

11 A stores is offering ex stock delivery to a wide variety of customers. Item 35721 is a non-seasonal item which has the following usage characteristics:

Weekly usage: 200.
MAD: 50.
Supply lead time: four weeks.

(a) If deliveries are brought in every week, what is the target stock level if 85 per cent availability ex stock is required?
(b) If the supplier is persuaded to deliver daily, what would be the new target stock level?
(c) How would the change affect the value of stockholding if the items cost £10 each?

12 The statistics on the stock of a range of items is as follows:

Item no.	Unit cost (£)	Annual usage
2A32	25.00	12
2B44	0.50	360
3D10	0.10	120
3E82	40.00	1
3F66	2.00	40
4G19	5.00	520
4H95	4.00	7
5J53	20.00	5
7N78	0.30	200
9P21	12.00	50

The stores controller is under pressure from the management to reduce the stock level to below half a month's usage. The stores personnel are adamant that they are not going to do any extra work. Show what the stock policy should be in these conditions and how the target is achieved. (Assume that safety stock is not necessary in this case.)

13 Stock levels are as follows:

A items: £28 000, 130 items
B items: £6500, 350 items
C items: £3500, 600 items
Obsolete items: £2000, 200 items.

Within the budget for next year, the estimated usage rate for each is:

A items: £133 000 per month
B items: £16 000 per month
C items: £2000 per month
Obsolete items: nil.

What is the best way to reduce the stock by £2000?

14 Do you believe that JIT simply passes the burden of holding stock on to the supplier?

15 Your managing director has asked for a short report on ways of reducing work in progress and throughput time. At present, batch sizes are about fifteen, setups average three hours, processing time per unit is two minutes, and WIP stands at thirteen weeks. Create a short report outlining in simple language the measures you would take.

16 (a) What are the financial benefits to a company through setting up a linear flow kanban system?
(b) A company is intending to set up a kanban system for production. Comment on the changes which will be required in:
responsibility
layout
operator training
quality management.

17 What is the role of safety stock, and why is it often not required in a JIT environment?

18 'JIT is a good theoretical idea but cannot work in our company' said the production director. 'We have customers who change their minds, a variety of processes, many products, and real people who cannot always be relied upon.' Set out the arguments you would make to persuade this director that JIT is a good idea for the company.

19 The data below is the stock record for item 345.

Date	Receipts	Unit cost	Issues	Balance
14/3	10	11		10
12/4			3	7
22/5			2	5
10/6	5	15		10
30/6			3	7
17/7			2	5

(a) What is the value of the last issue (17/7) of item 345, based on:
FIFO costing
average costing
LIFO costing.

(b) If you needed to reduce the stock value in your stores by £500 000, what items would you look at first and what techniques would you use to reduce the stock value?

20 Why is safety stock held?

21 What are the three factors which determine safety stock?

22 The usage for two items over the past five weeks has been:

Item	Week 1	Week 2	Week 3	Week 4	Week 5
D523	60	30	20	40	50
K488	30	15	0	65	0

(a) What are the values of average demand and variability for each?
(b) The items cost 50p each and the lead time from the supplier is two weeks. The inventory controller is deciding whether to try to provide a service level of 90 per cent or 95 per cent. What is the additional cost of the better service?

23 ABC Distributors have the UK dealership for a well known manufacturer of pre-packaged confectionery. Their main activity is the supply of goods to the local retailers ex stock. At the recent stocktake, ABC Distributors had 1280 different stock lines with a total value of £40 000. They order all their items during the week they fall below the reorder level.

The system works well except for the fact that the stores manager and his assistant are overloaded with work – in fact they are currently three weeks behind with the ordering. As they are ordering one month's requirement at a time, there are often times when the stock runs out. The suppliers take two months to deliver, and the shortfalls cannot therefore be rectified immediately.

What are the changes in operating practice which you would suggest to assist them? What controls should they use which will bring the changes into effect?

24 'Re-ordering systems are not designed to stop you running out of stock.' Give two reasons why this statement can be justified.

25 What factors do you consider when setting stock levels for an independent demand item?

26 The demand history of item XX1 is as follows:

Week	Demand
1	60
2	50
3	70
4	60
5	80
6	70

The items are supplied weekly in boxes of 100. Work out:

(a) the review level (reorder point)
(b) the average stockholding.

(You can assume a safety factor of 2 for 94 per cent customer service.)

27 Explain why, for some A items, we can expect the review level to always be higher than the physical stock level?

28 A company is putting expensive fountain pens into boxes at a rate of about 250 per week, together with ink cartridges. The stock controller has decided that the pens they buy are class A, while the cartridges and boxes are class C. Suggest order and delivery patterns and order quantities for the three types of item, and discuss why you have chosen them.

29 (a) It has been remarked that 'stock level for an item does not depend on its lead time'. Defend this remark using examples, and discuss why the statement is not precisely true.

(b) Calculate the reorder level for the supply of umbrellas given the following data:
Lead time = 2 months.
Customer service level expected is 90 per cent.
Demand over the last six months has been 11, 20, 8, 15, 10 and 14.

(c) Discuss the assumptions which you have made in using this calculation.

30 A company is proposing to schedule in supplies of twenty key items each week instead of ordering them monthly.

(a) What preparatory work do they need to do with the suppliers?
(b) If the purchase value of these items is £1 million per year, by how much will the average inventory be reduced?
(c) What are the benefits and extra jobs caused by scheduling?

31 What factors determine the delivery quantity?

32 Discuss ways in which stock can be reduced by suppliers and customers managing a supply chain together. What are the prerequisites for successful supply partnerships?

33 What is the difference between a target stock level system and a review level system?

34 Weekly sales of a product are:

Week	Demand
1	12
2	14
3	8
4	16
5	13
6	18

(a) Discuss what forecasting techniques can be used to determine the demand for the next week. Indicate which you would prefer to use and the reasons for your choice.

(b) If the initial forecast for week 1 was ten, work out the forecast demand for each week using exponential smoothing, and give your prediction for week 7. (You can use an α factor of 0.2 for this calculation).

35 What items of data (inputs) are required for the MRP calculation? For each, state the level and type of accuracy necessary to ensure reasonable results.

36 A watch manufacturer has a schedule for assembly covering ten weeks. The schedule is:

Watch X	Manufacturing orders for 1000 watches due in weeks 2, 4, 6, 8 and 10. Assembly lead time per batch is 1 week.
Watch Y	Manufacturing orders for 500 watches due in weeks 3, 6 and 9. Assembly lead time per batch is 2 weeks due to a special processing on the watch cover.
Watch Z	Manufacturing orders for 100 watches due every week. Assembly lead time is half a week. The parts for the batch of week 1 have already been issued.

Watches X and Y use the same module A and batteries (two per watch). Watch Z uses a different module B and the same batteries but only one per watch. Batteries and modules come in boxes of 500. There are seven boxes of batteries and three boxes of modules in stock. The buyer has outstanding purchasing orders for a total of 20 000 batteries and 6000 of each module, half quantities to arrive in weeks 4 and 10.

(a) What would you expect your MRP package to recommend for purchasing? (Assume that the package is not capable of recommending rescheduling.)
(b) What would you recommend should be done?

37 A product is manufactured in two stages. The material for the process consists mainly of one basic, bought out ingredient. Expected sales of the product are as follows:

Week	1	2	3	4	5	6	7	8	9	10
Sales	210	300	140	65	200	260	150	100	190	220

The stock situation is:

raw material	500
semi-finished	220
finished goods	650

The manufacturing and supply situation is:

Operation	Lead time
Purchasing	Three weeks
Bulk production	One week
Finishing	Two weeks

(a) What should the weekly purchase plan be on a lot-for-lot system?
(b) What should it be using with a 500 batch size?

38 (a) What different underlying assumptions are there between DRP and fair shares allocation of goods?
(b) What is the fair shares priority logic for distributing stocks to regional warehouses?

39 Prepare a manufacturing schedule, using the DRP technique, for the following forecast demands and lot size replenishment quantities.

	Lot size	Period demand							Current stock
		1	2	3	4	5	6	7	
London	600	150	320	370	190	150	320	370	600
Manchester	250	70	70	70	70	70	70	70	230
Newcastle	400	90	150	150	90	90	150	150	310
Internal sales	Monthly call-off	30	30	30	30	30	30	30	

Lot size for manufacturing store is 1000.
Current stock for manufacturing store is 1650.

Replenished lead time for all warehouses is two periods, internal sales are replenished immediately, manufacturing lead time is three weeks.

40 What are the advantages and disadvantages of keeping stock in retail stores rather than in supply depots?

41 The following ten items are the range of stock held in a small stores.

Item no.	Unit cost (£)	Annual usage
1	150	2
2	125	100
3	85	10
4	50	70
5	40	1
6	26	20
7	20	3
8	10	1
9	7	300
10	6	20

(a) Identify which are the A, B and C class items.
(b) Discuss how ABC is used in purchase ordering.
(c) Review another use of this Pareto classification.

42 The following data is available for the demand for an item:

Week	1	2	3	4	5
Demand	10	20	5	10	15

What would you estimate to be the average quantity held in stock to meet the demand for this item, treating it as:

(a) Independent demand (assuming that a 99.2 per cent customer service level is required).
(b) Dependent MRP demand (level 2) (assuming a one week lead time).
(c) JIT supply (assuming a one-day delivery interval).

Specify any assumptions you make.

Answers

1
Availability (delivery on time).
Minimum inventory cost.
Minimum operating cost.
See text, Chapter 1.

2
Purchasing, if they try to buy large batches to reduce unit cost.
Sales if they want large stocks to service possible demand.
Accounts who require lowest investment in inventory to improve cash flow.
In manufacturing, production traditionally want to make large batches and to have buffer stock in case of problems.
Also pressure comes from elsewhere – designers who specify items too late and management who keep changing the targets.
See text, Chapter 1.

3 High stocks result from poor control, and poor service results from poor management of supply and forecasts. The better the inventory management, the better the availability and the less stock cover is required.

4

Annual accounts data	Ace Distributors (£ million)	Deuce Associates (£ million)
Annual profit[1]	1.8	5
Return on sales	$\frac{1.8}{8}$	$\frac{5}{26}$
=	22.5%	19.2%
Assets are	2 + 4	2 + 4
Return on assets	$\frac{1.8}{6}$	$\frac{5}{13}$
=	30.0%	38.5%

Note: [1]Sales turnover – cost of sales.

Ace has better margin and more potential for improvement in stock control.

5 51.4 per cent.

Annual accounts data	Ace Distributors (£ million)
Annual profit	1.8
Return on sales	1.8/8 = 22.5%
New value of assets are	2+1.5 = 3.5
Return on assets	1.8/3.5 = 51.4%

6

Capital employed before reduction = Stock + Capital plant
= £1.5 million + £1.1 million
= £ 2.6 million

Capital employed after reduction = £ 2.6 million – £500 000
= £ 2.1 million

Return on capital before reduction = £400 000/£2 600 000
= 15 per cent

Return on capital after reduction = £400 000/£2 100 000
= 19 per cent

7 There are two key ways in which JIT is a different approach:

- The level of inventory is not a simple calculation – it is a negotiated quantity.
- JIT does not set a permanent inventory situation – it empowers the people to improve and reduce the levels required.

This means that those involved with managing the supply chain always have to see ways of improving reliability of performance, reducing timescales and costs and, therefore, removing the need for inventory.

The timescales of JIT operations are measured in hours, under the umbrella of a long-term relationship and smooth supply. Conventional inventory management works with days or weeks of lead time and inventory levels and any demand pattern.

8 Solving customer requirements by providing quality goods service, information and ongoing support, and by delivery on time (see text).

Demand satisfied on time:	Dispatched	Requested	%
By order	0	3	0
By order line	5	10	50
By item code:			
Item 1	65	90	72
Item 2	270	320	84
Item 3	60	120	50
Item 4	25	25	50
Total	420	555	
Average line availability			77
Average item availability			76

Different answers are used in different situations (see text).

9

Item code	Weekly demand	Unit price (£)	Turnover (£)	Rank	Class
A901	2	3.00	6	9	C
B662	2	8.00	16	6	C
C355	20	4.00	80	3	B
D523	40	0.50	20	5	C
E191	10	65.00	650	1	A
F807	1	9.00	9	8	C
G010	3	1.00	3	10	C
H244	1	170.00	170	2	B
J459	10	3.50	35	4	C
K488	22	0.50	11	7	C

See also text, Chapter 3.

10 (a)

Item	Stock	Usage	Cover
A901	6	2	3.0
B662	10	2	5.0
C355	50	20	2.5
D523	60	40	1.5
E191	50	10	5.0
F807	50	1	50.0
G010	12	3	4.0
H244	5	1	5.0
J459	20	10	2.0
K488	5	22	0.23

(b) Ensure that supplies of item K488 are arriving within one day.
(c) This is calculated on a financial basis.

	Stock	Usage	Unit value (£)	Stock value (£)	Turnover (£)
A901	6	2	3.00	18	6
B662	10	2	8.00	80	16
C355	50	20	4.00	200	80
D523	60	40	0.50	30	20
E191	50	10	65.00	3250	650
F807	50	1	9.00	450	9
G010	12	3	1.00	12	3
H244	5	1	170.00	850	170
J459	20	10	3.50	70	35
K488	5	22	0.50	3	11
				4963	1000

Stock cover for all stock = £4963/£1000
= 4.96 weeks

11

(a) TSL = Demand in lead time and review period plus safety stock in LT and RP
= 200*(4 + 1) + 1.3 * 50 * (√(4 + 1))
= 1000 +145.3
= 1145.3

(b) Assuming that the lead time is still four weeks and orders are placed daily:

TSL = 200*(4 + 0.2) + 1.3 * 50 * (√ (4 + 0.2))
= 840 + 133.2
= 973.2

(c) This would be a reduction of 172.1 in TSL. The average stock is TSL – half usage in (LT + RP). So reduction would be from 645.3 to 553.2. Reduction 92.1 units at £10 each = £921.

12 There are several ways of getting to this answer. The easiest is to perform Pareto Analysis and then apply batch sizes. Say A = every week, B = every fortnight, C = every four months. If every item is ordered each month, then there are 120 purchase events

Item no.	Unit cost (£)	Annual usage	Usage value per year (£)	Class	No of orders per year	Average stock value (£)	Percentage of value
2A32	25.00	12	300	B	24	6.25	7.5
2B44	0.50	360	180	C	3	30.00	4.5
3D10	0.10	120	12	C	3	2.00	0.3
3E82	40.00	1	40	C	1	20.00	1.0
3F66	2.00	40	80	C	3	13.33	2.0
4G19	5.00	520	2600	A	52	25.00	65.0
4H95	4.00	7	28	C	3	4.67	0.7
5J53	20.00	5	100	C	3	16.67	2.5
7N78	0.30	200	60	C	3	10.00	1.5
9P21	12.00	50	600	B	24	12.50	15.0
Total			4000		119	140.42	

Stock cover = 42.

From the table, the number of order events is now 119 (so no more work) and the stock is 0.42 of a month instead of 0.5. (See Chapter 3.)

13

Class	Stock (£)	Budget (£)	Stock cover (weeks)	% reduction required for £2000
A	28 000	133 000	0.91	7
B	6 500	6 500	1.76	31
C	3 500	2 000	7.58	57
	2 000	0	Infinite	100

Although the stock cover for A class is already under a week, it should be easier to gain £2000 by reducing this because it represents the majority of stock value. This can be achieved by scheduling or kanban. In practice some inventory reduction may also come from the other classes. (See Chapter 3.)

14 It does not pass on inventory, it eliminates it. However suppliers have to be able to respond rapidly to supplying to the smooth demand pattern required. (See Chapter 4.)

15 From the figures, the average batch only takes thirty minutes to produce but takes three hours to change over. This is inefficient. There needs to be a concerted effort on reducing the setup time. At the same time product design could be reviewed to see whether common parts can be created, so that set-ups can be avoided. There is also a problem with high WIP. Looking at the wasted time in between operations will show where this wasted time can be eliminated. The use of a kanban pull system would cause the WIP to decrease dramatically. (See Chapter 4.)

16 (a) The costs of manufacture and stockholding are reduced. Kanban create an environment where the WIP of the A class items is less than a day's worth between each process. The processes are linked together so that there is a flow of work. This enables efficient manufacture, and a reduction in waste, which contribute to lower operating costs.

The improvement in speed through the process means that any reject work is caught earlier and therefore can be rectified before too much material and capacity has been used up. There is also the direct reduction in inventory which removes working capital costs.

(b) *Responsibility.* Shop floor responsible for triggering supply of next item/batch without fail. Supplying operation responsible for replenishing in time.

Layout. Processes arranged in sequence with minimum distance between the processes.

Operator training. Operator responsible for their area and processes within it. Teamworking skills required as well as multiskilling. Operators need to be able to transfer from one process to the next in case there is a bottleneck.

Quality management. Excellent quality is a precursor to introducing kanban. Quality should be built into the process so that it will not go wrong. Quality is part of the manufacturing process, not a separate activity.

17 Safety stock decouples one process from the next. In JIT, the aim is to couple all the processes together to minimize timescales and inventory. Where JIT systems operate, the demand is smoothed to give similar requirements every day; consequently there is not the need for significant safety stock. Safety stock will be required where demand exceeds the flexibility of the manufacturing process.

18 Just in time throughput time is small, therefore there is not the need for detailed planning far ahead. This gives flexibility. By not having to forecast far ahead, then the forecasting process is more accurate. Manufacturing smaller batches means continuity of work and simpler planning, therefore less problems and less cost.

Many companies have been using JIT successfully for many years. They create the correct working environment, and applying a lean approach. They have great advantages over their competition because the costs of storage, interprocess stocks and non-value-added work are reduced.

19

Date	Receipts	Unit cost	Issues	Balance	No. left 14/3	From 10/6	Average cost
14/3	10	11		10	10		11
12/4			3	7	7		11
22/5			2	5	5		11
10/6	5	15		10	5	5	13
30/6			3	7	2	5	13
17/7			2	5	0	5	13

So, value of issue of the two items was:
(a) FIFO 22; average 26; LIFO 30.
(b) Highest stock value lines independent of usage rate. Reduce supply. See text, Chapter 5.

20 To absorb the different rates of supply and usage in two processes. The calculation of safety stocks normally only takes into account the variations in demand. The variation in supply is avoidable and should be controlled properly so that it is insignificant. The variation in demand is caused by the

'customer' and is therefore to be coped with and negotiated. The better the forecast of demand, the smaller the safety stock will have to be. (See Chapter 6.)

21 The calculation of safety stock depends on:

The level of service being offered (acceptable risk of running out).
The predictability of the demand (MAD as a measure of forecast accuracy).
The replenishment lead time (but only as the square root of lead time). (See Chapter 6.)

22 (a)

Item	Wk 1	Wk 2	Wk 3	Wk 4	Wk 5	Average	
D523	60	30	20	40	50	40.0	
K488	30	15	0	65	0	22.0	
						MAD	MADP
D523	20	10	20	0	10	12.0	30%
D488	8	7	22	43	22	20.4	93%

(b) Lead time two weeks:

CSF 90 per cent 1.6
CSF 95 per cent 2.06

	MAD	SS 90%	SS 95%	Increase	Extra cost	% of turnover
D523	12.0	27.15	34.96	7.81	3.90	20
K488	20.4	46.16	59.43	13.27	6.64	60
					10.54	

Answer is £10.54.

23 Apply Pareto's Law, then organize the supply to save work (see also question 12). Change the order pattern for the 128 A class parts to weekly and schedule in to avoid paperwork and time. Order the B class every four weeks as they fall below the review level, and order the 896 C class items every eight weeks to reduce the workload.

Then work out the better safety stocks, and flexibility of supply for the A class. It is important to measure customer service level and inventory levels while the change is happening, and also to make sure that the number of orders are reducing significantly. (See Chapter 7.)

24 The reordering system creates replenishment orders, which open the supply. However demand may not be satisfied because:

- *From the theory.* The review level is calculated assuming a customer service factor which gives a small risk of stockouts (zero risk is impossible).
- *In practice.* Supply may be later than expected and the stock is insufficient to cope with the longer lead time.

25 To set the safety stock requires demand. To set the average stock level or to order requires in addition:

delivery quantity variability of demand (MAD or SD)
level of availability required (customer service factor)
supply lead time.

To set the review level requires in addition, forecast.

Underlying the forecast and MAD calculations are therefore an assessment of demand pattern using historical information backed up by understanding of how future demand may vary from this.

26

Average 65: safety stock = 16.67.
MAD 8.33: review level = 81.67.
average stock = 131.67.

27 The reorder level is used to trigger new replenishment orders when the stock + supply orders outstanding become low. If the delivery frequency is shorter than the lead time then the stock is low, delivery quantities are small and there are always supply orders outstanding. Hence, for long lead time A class items, ordering is frequent and stock is always less than the review level.

28
A class – order in daily, if possible on Kanban.
C class – order in weekly.

See Chapter 7.

29 (a) Stock level depends mainly on supply delivery quantities, customer service and demand pattern. Lead time variation is only as the square root. (However, delivery performance may be lead time dependent also.)

Average stock = half delivery quantity + customer service factor × MAD × $\sqrt{\text{lead time}}$.

So changing the lead time has the least effect on stock levels.

(b) Average = 13.
MAD = 3.33.
Customer service factor is 1.6.
Safety stock = $1.6 \times 3.33 \times \sqrt{(2)} = 7.53$
Reorder level = $13 \times 2 + 7.53 = 33.53$

(c) Continuous demand.
Continuation of the same demand pattern.
Normal distribution and non-seasonal demand.
Supplier delivers on time.

30 (a) Work out with the supplier:

Benefits to supplier as well as customer (continuous demand, earlier payment, more consistent demand).
Delivery arrangements (transport, packaging).
Dispatch and receiving reorganization.
Information processing (schedules and invoicing).

(b) Reduction in inventory:

Average stock = half delivery quantity + safety stock.
Assume the safety stock remains the same.
Delivery quantity goes from four weeks' worth to one week's worth, so stock saving is two weeks to half a week.
At £1 million per year 1.5 weeks is £28 846.

(c) Benefits:

Lower inventory investment – better cash flow.
Less stores space required.
Easier movement within stores.
Improved information system (paperwork?) efficiency.
Extra employment in the long term.
More frequent (but smaller deliveries).
More receiving document processing.
More monitoring of inventory and co-ordination with suppliers.

31 The delivery quantity primarily depends on how frequently the supplier is prepared to deliver without significant increase in cost. It will depend upon:

The number and variety of other items bought from the same supplier.
The type of items (e.g. handlable, slow-moving).
The quantity in which they are to be consumed.
The turnover value of the item.

32

Better forecasting.
Treating demand as dependent and avoiding safety stock effects.
Measuring end user usage, structuring supply source.
Integrating systems
Square root of warehouses effect.

33 Target stock levels: variable order quantities fixed time interval between deliveries. Ideal where supply of many B and C class items are sourced from a single supplier. Organize purchasing to look at suppliers in rotation.
Review level system: interval between ordering is variable, order quantities are fixed by an ordering rule. Normal way to trigger orders. Good for A class items and slow-moving and one-product suppliers.
(See Chapter 8.)

34 (a) Extrinsic forecasts to give causes of demand.
Market analysis and surveys.
Historical techniques.

Moving average: OK as demand is random and varying much.
Exponential smoothing: better since there could be a slight trend.
Double exponential: will be overreactive until a trend is seen.

Base series: not appropriate as seasonality not established.
Regression: no advantage over other methods.
More sophisticated techniques: insufficient data for these.

(For details, see Chapter 10.)

(b) Exponential smoothing with $\alpha = 0.2$.

Week	Demand	Forecast
1	12	10.00
2	14	10.40
3	8	11.12
4	16	10.50
5	13	11.60
6	18	11.88
7	?	13.10

35 MRP requires:
master schedule: 95 per cent accurate.
bills of materials: 98 per cent accurate
inventory records: 95 per cent accurate

and

lead times (for suppliers and any manufacturing processes).
manufacturing routings (where appropriate).

36 *Components*

Watch	Modules	Battery quantity
X	A	2
Y	A	2
Z	B	1

Manufacturing orders

Watch type	Week	1	2	3	4	5	6	7	8	9	10
X			1000		1000		1000		1000		1000
Y				500			500			500	
Z		100	100	100	100	100	100	100	100	100	100

Component demand
Module A

Gross requirement		1500 0	1000	500	1000	0	1500	0	1000	0
Stock on hand	1500	0 0	−1000	−1500	−2500	−2500	−4000	−4000	−5000	−5000
Purchases outstanding				3000						3000
Projected on hand		0 0	−1000	2500	1500	1500	0	0	−1000	2000

New orders for module A of 1000 for week 3 and 1000 for week 9.
Module B: excessive stock of module B is building up.
Battery

Gross requirement		3100	100	2100	1100	2100	100	3100	100	2100	100
Stock on hand	3500	400	300	−1800	−2900	−5000	−5100	−8200	−8300	−10 400	−10 500
Purchases outstanding					6000						6000
Projected on hand		400	300	−1800	3100	1000	900	−2200	−2300	−4400	1500
Planned receipts				2000				2500			
Projected available		400	300	200	5100	3000	2900	2300	2200	100	6000

New orders for the battery of 2000 for completion in week 3 and 2500 for week 7.

(b) Recommend that schedule is smoothed. There are many ways of doing this. One option is:

Module A

Module A	Week	1 2	3	4	5	6	7	8	9	10
Gross requirement		1500 0	1000	500	1000	0	1500	0	1000	0
Stock on hand	1500	0 0	−1000	−1500	−2500	−2500	−4000	−4000	−5000	−5000
Reschedule receipts			**1500**	**500**	**1000**	**500**	**1000**	**500**	**1000**	**500**
Projected on hand		0 0	500	500	500	1000	500	1000	1000	1500

Try to reschedule the module supply as the schedule above.
Battery

Battery	Week	1	2	3	4	5	6	7	8	9	10
Gross requirement		3100	100	2100	1100	2100	100	3100	100	2100	100
Stock on hand	3500	400	300	−1800	−2900	−5000	−5100	−8200	−8300	−10 400	−10 500
Reschedule receipts			**1400**	**1400**	**1400**	**1400**	**1400**	**1400**	**1400**	**1400**	**1400**
Projected on hand		400	1700	1000	1300	600	1900	200	1500	800	2100

Try to reschedule the battery supply to 1400 per week

37 (a)

Week	Stock	1	2	3	4	5	6	7	8	9	10
Sales		210	300	140	65	200	260	150	100	190	220
FG	650	440	140	0	435	235	475	325	225	35	315
Order receipt					500		500				500
Order release		0	500	0	500	0	0	0	500		
Semi-finished	220	220	220	220	220	220	220	220	220	220	220
Order receipt			500		500				500		
Order release		0	500	0	500	0	0	0	500		
Raw material	500	0	0	0	0	0	0	0	0	0	0
Purchase receipt				500				500			
Purchase ordering		500				500					

(b)

Week	Stock	1	2	3	4	5	6	7	8	9	10
Sales		210	300	140	65	200	260	150	100	190	220
FG	650	440	140	0	0	0	0	0	0	0	0
Order receipt					65	200	260	150	100	190	220
Order release		0	65	200	260	150	100	190	220		
Semi-finished	220	220	155	0	0	0	0	0	0		
Order receipt				45	260	150	100	190	220		
Order release		0	45	260	150	100	190	220	0		
Raw material	500	500	455	195	45	0	0	0	0		
Purchase receipt						55	190	220			
Purchase ordering		0	0	55	190	220					

38 (a) Distribution requirements planning takes the forecast from warehouse and aggregates it into the master schedule. It therefore depends upon the warehouses to forecast in the first place.

Fair shares can manufacture against a forecast and then allocate the inventory against the warehouse requirement.

(b) The logic is:

1 Any unfulfilled demand from the warehouses.
2 Forecast usage which will be unfulfilled within the supply lead time.
3 Top up warehouse safety stock.
4 Allocate to central stock safety level if appropriate.
5 Distribute remainder in proportion to the forecast warehouse usage rates.
6 Transfer excesses to other warehouses if appropriate.

39

Week	Stock	1	2	3	4	5	6	7
London sales		150	320	370	190	150	320	379
Current stock	600	450	130	360	170	20	300	521
Order receipt				600			600	600
Order release		600	0	0	600	600	0	0
Manchester sales		70	70	70	70	70	70	70
Current stock	230	160	90	270	200	130	60	240
Order receipt				250				250
Order release		250	0	0	0	250	0	0
Newcastle sales		90	150	150	90	90	150	150
Current stock	310	220	70	10	10	10	40	70
Order receipt				90	90	90	180	180
Order release		90	90	90	90	180	180	
Internal sales		30	30	30	30	30	30	30
Total		970	120	120	720	1060	210	30
Current stock	1650	680	560	440	720	660	450	420
Order receipt					1000	1000		
Order release		1000	1000	0	0			

Manufacturing orders required in weeks 1 and 2.

40 Advantages:

Retail stores satisfy customer demand and, therefore, stock is available immediately to meet demand.

Any distribution problems do not affect availability.

No central stock is necessary.

Slow distribution method can be used to the retail stores.

Disadvantages:

More stock is required because there are more stock points.

Any change of products requires emptying the whole supply chain.

Redistribution of excess or supply to meet an individual shortage is more difficult.

41

Item no.	Unit cost	Annual usage	Turnover	% turnover	Cumulative % turnover
2	125	100	12 500	62.5	62.5
4	50	70	3500	17.5	80.0
9	7	300	2100	10.5	90.5
3	85	10	850	4.3	94.8
6	26	20	520	2.6	97.4
1	150	2	300	1.5	98.9
10	6	20	120	0.6	99.5
7	20	3	60	0.3	99.8
5	40	1	40	0.2	100.0
8	10	1	10	0.1	100.0
Total		527	20 000		

(a) A class: item 2.
B class: items 4, 9.
C class: items 3, 6, 1, 10, 7, 5, 8.

(b) ABC determines delivery quantity and reduces purchase workload (see Chapter 3).

(c) Discussion of optimizing profits, lead time reduction or cycle counting.

42 (a) Independent demand

Assume that a 99.2 per cent customer service level is required.

Average demand = 12.
MAD = 4.4.
Customer service factor = 3.
Safety stock = $3 \times 4.4 \times \sqrt{(1)}$
= 13.4
Delivery quantity = 12.
Average stock = Half delivery quantity + safety stock
= 6 + 13.4
= 19.4

(b) Dependent MRP demand (level 2), assuming no safety stock is required, stock would be:

Week	0	1	2	3	4	5
Demand	10	20	5	10	15	

Average stock = 12.

(c) JIT supply

Assume that demand is the same every day.

(Alternative assumptions are that there should be a safety stock calculated as (a) or as one-day lead time.)

Week	0	1	2	3	4	5
Demand	10	20	5	10	15	
Daily demand	2	4	1	2	3	

Assuming that demand is dispatched every day:
Average stock = half day's demand:

1, 2, 0.5, 1, 1.5

So average stock during the period is 1.2 (assuming a one-day delivery interval).

Index

MANAGEMENT

Best Practice in Inventory Management

Second Edition

Tony Wild

Good management of inventory enables companies to improve their customer service, cash flow and profitability. **Best Practice in Inventory Management** outlines the basic techniques, how and where to apply them, and provides advice to ensure they work to produce the desired effect in practice.

The book shows how inventory management techniques can be used in a wide variety of situations, particularly in stores where the inventory can be anything from fast moving products to slow moving spares. The discussion extends across distribution warehousing and manufacturers' operations.

The text is based on best theory and practice, which has been gradually developed by the inventory management profession over the years. It covers the inventory control aspects included in the courses for the DPIM, COM, ILTC, ILTD, CPIM and other professional and academic qualifications.

Readers develop their understanding o[illegible]ck control by seeing the techniques explained logically and [illegible]w inventory structuring, individual item control, forecasting and [illegible]tion provide the base for logistics management. This new edition [illegible] up-dated throughout and the final chapter, The Future – Inventory a[illegible]istics, has been re-written to reflect the developing applications of tech[illegible]ogy and changes in focus.

Tony Wild has spent a great deal of time solving real inventory issues for businesses in a variety of industries from aerospace to retail consumables, automotive to process chemicals. The result of this work has led Tony to produce and refine this book so that others can avoid some of the pitfalls. From an academic start (BSc, MSc, PhD) and a manufacturing and warehousing career in production control and materials management, Tony has specialized in this challenging field. He is Managing Director of Midas Consultants, who are leading experts in inventory management.

BUTTERWORTH HEINEMANN

An imprint of Elsevier Science
www.bh.com

ISBN 0-7506-5458

9 780750 65458