Modern Production Management

Managing the Operations Function 5th Edition

ELWOOD S. BUFFA

Modern Production Management

The Wiley/Hamilton Series in Management and Administration
Elwood S. Buffa, Advisory Editor University of California, Los Angeles

PRINCIPLES OF MANAGEMENT: A Modern Approach: Fourth Edition Henry H. Albers

OPERATIONS MANAGEMENT: Problems and Models, Third Edition Elwood S. Buffa

PROBABILITY FOR MANAGEMENT DECISIONS William R. King

MODERN PRODUCTION MANAGEMENT: Managing the Operations Function, Fifth Edition Elwood S. Buffa

CASES IN OPERATIONS MANAGEMENT: A Systems Approach James L. McKenney and Richard S. Rosenbloom

ORGANIZATIONS: Structure and Behavior, Volume I, Second Edition Joseph A. Litterer

ORGANIZATIONS: Systems, Control and Adaptation, Volume II Joseph A. Litterer

MANAGEMENT AND ORGANIZATIONAL BEHAVIOR: A Multidimensional Approach Billy J. Hodge and Herbert J. Johnson

MATHEMATICAL PROGRAMMING: Claude McMillan

DECISION MAKING THROUGH OPERATIONS RESEARCH, Second Edition Robert J. Thierauf and Robert C. Klekamp

QUALITY CONTROL FOR MANAGERS & ENGINEERS Elwood G. Kilpatrick

PRODUCTION SYSTEMS: Planning, Analysis and Control, Second Edition James L. Riggs

SIMULATION MODELING: A Guide to Using Simscript Forrest P. Wyman

BASIC STATISTICS FOR BUSINESS AND ECONOMICS, Second Edition Paul G. Hoel and Raymond J. Jessen

COMPUTER SIMULATION OF HUMAN BEHAVIOR John M. Dutton and William H. Starbuck

INTRODUCTION TO GAMING: Management Decision Simulations John G. H. Carlson and Michael J. Misshauk

PRINCIPLES OF MANAGEMENT AND ORGANIZATIONAL BEHAVIOR Burt K. Scanlan

COMMUNICATION IN MODERN ORGANIZATIONS George T. and Patricia B. Vardaman

THE ANALYSIS OF ORGANIZATIONS, Second Edition Joseph A. Litterer

COMPLEX MANAGERIAL DECISIONS INVOLVING MULTIPLE OBJECTIVES Allan Easton

MANAGEMENT SYSTEMS, Second Edition Peter P. Schoderbek

ADMINISTRATIVE POLICY: Text and Cases in the Policy Sciences Richard M. Hodgetts and Max S. Wortman, Jr.

THE ECONOMICS OF INTERNATIONAL BUSINESS R. Hal Mason, Robert R. Miller and Dale R. Weigel

BASIC PRODUCTION MANAGEMENT Elwood S. Buffa

FUNDAMENTALS OF MANAGEMENT COORDINATION: Supervisor, Middlemanagers and Executives Thomas A. Petit

QUANTITATIVE BUSINESS ANALYSIS David E. Smith

OPERATIONS MANAGEMENT: The Management of Production Systems Elwood S. Buffa

PERSONNEL ADMINISTRATION AND HUMAN RESOURCES MANAGEMENT Andrew F. Sikula

MANAGEMENT SCIENCE/OPERATIONS RESEARCH: Model Formulation and Solution Methods Elwood S. Buffa and James S. Dyer

MANAGEMENT PRINCIPLES AND PRACTICES Robert J. Thierauf, Robert Klekamp and Daniel Geeding

Modern Production Management

FIFTH EDITION

Managing the Operations Function

Elwood S. Buffa

University of California, Los Angeles

A Wiley/Hamilton Publication

John Wiley & Sons

Santa Barbara New York London Sydney Toronto

This book was set in 10 point Medallion by Dharma Press and printed and bound by Kingsport Press. Chuck Pendergast and Jean Varven supervised production. Cover design by Tri-Arts, illustrations by Graphics Two.

Library of Congress Cataloging in Publication Data

Buffa, Elwood Spencer, 1923–
Modern production management.

(The Wiley/Hamilton series in management and administration)
Includes bibliographies.
1. Production management. I. Title.
TS155.B723 1977 658.5 76-44547
ISBN 0-471-11829-x

ISBN 0-471-11829-x

Printed in the United States of America.
10 9 8 7 6 5 4 3 2 1

To Carl

About the Author

Elwood S. Buffa is professor of management science and operations management at the Graduate School of Management of the University of California, Los Angeles. He received his B.S. and M.B.A. degrees from the University of Wisconsin, and his Ph.D. from the University of California, Los Angeles. He worked as an operations analyst at the Eastman Kodak Company before entering the teaching profession, and has engaged in consulting activities in a wide variety of settings during the past twenty years. He has served as assistant dean and associate dean at the Graduate School of Management, Chairman of UCLA's Academic Senate, and has held visiting appointments at IPSOA in Turin, Italy, and at the Harvard Business School. Professor Buffa has published many research papers in management science and operations management. He is the author of other books published by Wiley and Wiley/Hamilton, including *OPERATIONS MANAGEMENT: The Management of Productive Systems*; *OPERATIONS MANAGEMENT: Problems and Models*; *BASIC PRODUCTION MANAGEMENT*; and, coauthored with James S. Dyer, *MANAGEMENT SCIENCE/OPERATIONS RESEARCH: Model Formulation and Solution Methods*. In addition, Professor Buffa is advisory editor for the Wiley/Hamilton series in management and administration. Currently, he serves on the board of directors of On-Line Decisions, Incorporated.

Preface

The fifth edition of *MODERN PRODUCTION MANAGEMENT: Managing the Operations Function* represents a major change from the previous editions, and this is partially indicated by the addition of a subtitle. Previous editions included material representing service and nonmanufacturing operations. However, the fifth edition represents a total reorganization, integrating into the conceptual framework materials dealing with service and nonmanufacturing operations throughout the book. There are two chapters that deal exclusively with service and nonmanufacturing operations. Chapter 14 discusses the operations, planning, and control of such systems, and Chapter 15, "Work Shift Scheduling," deals with the important problems of manpower scheduling in service operations. However, in most other chapters, service operations are considered along with manufacturing. For example, Chapter 5 develops the design of the service offered and its relationship with the productive system design; Chapter 6 considers the decisions involving the location of service operations; Chapter 8 integrates the layout of physical facilities for service operations with similar problems in manufacturing; and Chapter 16 considers the problems of defining the meaning of quality for services, and controlling quality, along with manufacturing operations.

There are other significant changes in the fifth edition that have considerable importance. Chapter 3 develops the nature of systems concepts and their importance for production/operations management somewhat beyond the material found in the fourth edition. Chapter 4, "Analytical Methods in Production/Operations Management," presents a new framework for relating the various analytical models. Chapter 5, "Design of Products and Services," has new important material on technological forecasting and its relationship to the design of productive systems. Chapter 6, "Distribution and Facility Location," develops the problem of product distribution and presents new material on warehouse location, as well as the new materials on the location of service operations that we mentioned before. Chapter 7, "Processes and Job Design," integrates these two related topics within a new framework and in addition includes important materials on sociotechnical systems. Chapter 9 on forecasting is now a separate chapter and includes materials on causal methods as well as exponential and adaptive forecasting methods. Chapter 11, "Inventories in the System," is now a separate chapter, and includes new materials on practical methods for determining buffer stocks. Chapter 12, which deals with industrial operations planning and control, combines materials that formerly were presented in two chapters, and introduces substantial material on "Material Requirements Planning." Finally, Chapter 16, "Maintaining System Reliability," integrates the materials dealing with quality control and maintenance into a unified framework.

Even the appendixes dealing with analytical models have been revised thoroughly. Appendix A, "Cost Models," combines materials that formerly occupied two chapters. Appendix B, "Linear Programming," presents substantially modified material that emphasizes formulation, and the managerial uses of the results of linear programming. In Appendix C, "Linear Programming—Distribution," Vogel's approximation method for establishing an initial solution has been added. In Appendix D, "Waiting Lines," the material dealing with infinite queues has been reorganized around a simple Poisson input model that accepts any service time distribution. The materials dealing with the multiple-channel case have been reorganized for problem solution through the use of a simple table, and materials dealing with finite queues have been added. Finally, the chapter end, Review Questions and Problems, have been revised thoroughly throughout, in keeping with the book's broader scope.

ACKNOWLEDGMENTS

A book of this length is necessarily based on a wide variety of sources. While I have made original contributions in some specific areas of analysis and application and in the conceptual framework, the bulk of the material on which *Modern Production Management* is based comes from original work by scores of colleagues throughout the country. The sources of these materials are cited where the materials are discussed. I hope I have made no omissions.

I have benefited greatly from reviews and comments on previous editions by well-known professors, such as Robert Albanese of Texas A&M University; William H. Bolen of Georgia Southern College; Robert W. Boling of the University of Tennessee; John D. Burns of DePaul University; Y. S. Chang of Boston University; Norbert L. Enrick of Kent State University; George J. Gore of the University of Cincinnati; Gene K. Groff of Georgia State University; S. T. Hardy of Ohio State University; Warren Hausman of the University of Rochester; Roy Housewright of Western Illinois University; James L. McKenney of Harvard University; William T. Newell of the University of Washington; J. A. Sargeant of the University of Toronto; Richard J. Tersine of Old Dominion University; and Thomas E. Vollmann of Indiana University.

I would like to thank Professor Gene K. Groff of Georgia State University and Professor Alan Krigline of the University of Akron for their considerable wisdom, good judgment, and suggestions in reviewing the final manuscript. Professors Ronald J. Ebert of the University of Missouri-Columbia, James A. Fitzsimmons of the University of Texas at Austin, Thomas E. Hendrick of the University of Colorado, Jarrett Hudnall, Jr., of Louisiana Tech University, Terry Nels Lee of Brigham Young University, D. Roman of George Washington University, and John E. Van Tassel, Jr., of Boston College helped immeasurably by reviewing parts of the manuscript and making suggestions.

The basic organization of the book retains the major division of materials into the strategic managerial decisions, and the shorter-run managerial decisions of operations planning and control. By placing the analytical materials in appendixes, we have been able to cover the basic concepts of production/operations management in sixteen substantive chapters. By placing the analytical chapters in appendixes, we have enabled instructors to assign these chapters in whatever sequence desired, or to not assign them if the instructors feel that students have had adequate prior preparation.

Elwood S. Buffa
Pacific Palisades, California

Contents

Part I

Introduction

Chapter 1

The Role of Production/ Operations Management

The creation of goods and services in our society is accomplished through productive systems. As late as 1965, the field concerned with the management of these productive systems was thought of in terms of manufacturing management. As we will note, however, the importance of service and nonmanufacturing systems has grown rapidly, and today "modern" production management must consider both kinds of systems and integrate them into a common conceptual framework.

We use the term "productive system" in the general sense to refer to any system that produces useful products and services. We also use the term "production/operations management," or sometimes simply "operations management," to refer to the management of any productive system, regardless of the type.

PRODUCTIVE SYSTEMS— A CRUCIAL ROLE IN MODERN SOCIETY

It would be difficult to overemphasize the crucial role that effective productive systems play in modern society and in our life style. Indeed, the term *developed* economy carries with it the image of large, highly organized, specialized, mechanized, efficient productive systems. On the other hand, the term *underdeveloped economy* includes the image of small, handicraft, inefficient productive systems powered mainly by the muscles of humans and beasts. Of course, this contrast of images is too simplistic to explain the difference between developed and underdeveloped economies. Transforming the nature of productive systems in an underdeveloped economy would not by itself create a developed one. Nevertheless, a modern society, with all its attributes of material wealth, is not possible without its characteristic productive systems.

As a matter of fact, it is unfortunate that we associate large output per man hour with a thrifty, industrious, hard working people, and that we assign the reverse personal characteristics to low output per man hour situations. Individuals in an economy typified by low output per man hour may actually work harder, in a physical sense, because there is no mechanization to accomplish the heaviest tasks.

Production Economics—A Problem of Cost Balance

It is also true that the high output per man hour economy is thought of as being efficient, while its opposite is thought of as being inefficient. But efficiency is a relative term meaning, essentially, "how effectively we employ the appropriate available resources (input) for a given unit of output." Thus, in developed economies, costs of machinery and equipment are relatively low, and labor costs are relatively high, reflecting the generally high output per work-hour of workers in that economy. In underdeveloped economies the reverse relative costs are generally true. (The cost of raw materials may be high or low in either situation, depending on a complex of factors.) Therefore, an efficient productive system in the developed economy is one that uses much mechanization and relatively little labor to yield a minimum combined cost of capital, labor, and raw material per unit of output.

For the identical product or service produced in an underdeveloped economy, an efficient productive system would reverse the relative amounts of capital and labor inputs in order to yield a minimum combined cost of capital, labor, and raw materials. Both systems could be efficient if they minimized the resource inputs per unit of output. This is an important point to understand about the effective design and operation of productive systems. Our goal is not always to employ the most sophisticated mechanized or automated technique known, but rather to employ the best balance of resources in each situation. Thus, even in the United States economy, a system designed for a relatively small output ordinarily will emphasize the use of labor as an input compared to capital. This general viewpoint, retained throughout the book, is most evident in Chapter 5, where we deal with production design of products and services; Chapter 6 on distribution and location; Chapter 7 on processes and job design; and Chapter 8 on layout of physical facilities.

Operating a Productive System—A Problem of Information and Decision Analysis

In a given productive system, successful management depends on plans, an information system concerning what is actually happening, and how we react to changes (make decisions). These changes might be in demand, inventory position, schedules, quality level, and product and equipment innovation. In forming plans for the management of a productive system, we are attempting to allocate the available resources in the most effective way for a given forecast of demand. The resources are units of productive capacity, such as number of man hours available at regular time and overtime, inventories available, and subcontracting. One can also conceive of negative capacity, that occurs when shortages or back orders take place. In constructing production plans, each of these capacities is provided at a cost, and the best plan is one that minimizes the sum of all costs over some future time span.

In attempting to meet the objectives of a plan, certain realities interfere, such as equipment failure, human error, discrepancies in the timing of order flow, quality variation, and so on. Therefore, systems for scheduling maintenance, quality control, and cost control are invented to help retain order where otherwise the system naturally would tend toward chaos.

Productive Systems—Conversion to Useful Products or Conversion to Pollution?

Productive systems commonly have been thought of as mechanisms for converting some sort of raw material into something useful. In the process, wastes normally are created, but in the past very little attention was paid to them. The emphasis was on the useful product or service, and wastes were disposed of in the cheapest possible way—dumped into rivers and streams or into the atmosphere. Only recently have we begun to realize that we may be fouling our own nests.

Today, socially conscious managers realize that the production function must include the processing of wastes so that they become benign or in themselves useful, rather than hazardous or even lethal. Waste conversion is a part of the productive process and must be included in our conceptual framework.

Service Systems Come of Age

Until fairly recently, the only productive systems thought to be significant were manufacturing systems. During the Industrial Revolution, we learned the first fundamentals of how to organize resources in order to produce something effectively. In the manufacturing setting, we focused our attenion on the production of goods to satisfy the basic human need and desire for material things, and focused our resources to develop manufacturing systems. Such systems were thus the original significant productive systems. Within the relatively short period of two to three hundred years, the industrial world has developed from handicraft production systems in the home to the fairly efficient industrial machine of today. Much of what we know about the *management* of productive systems was developed in this arena.

While past conditions focused resources and attention on manufacturing systems, current conditions have focused attention and resources on other sectors of society. They have broadened considerably the range of productive systems that are significant enough to warrant our study. It is fair to say that the operations phases of activities such as health care, education, transportation, and retailing were carried on at an almost handicraft level in the past. In some instances, these quaint systems had the advantages of the close personal relationships and warmth that often are associated with the family doctor or the corner grocery store. Suddenly, however, we were jolted into realizing that dramatic changes had developed. Health care and education, for example, grew into huge systems and attracted attention when their costs began to increase rapidly. As the productivity of our economy increased, a reallocation of personal expenditures took place, because the cost of services increased.

Currently, over 45 percent of personal consumption expenditures are for consumer nondurable goods. Consumer durable goods (dominated by autos and household equipment) have remained at an approximately level percentage since 1950. Thus, although the absolute levels of all expenditures have gone up, expenditures for services have increased at a somewhat faster rate than expenditures for goods.

A similar picture emerges for the U.S. economy as a whole. Services, as a percentage of Gross National Product, have increased from about 30 percent in 1948 to over 40 percent currently. The increase in the percentage for services relative to goods in GNP is due both to the reallocation of consumption expenditures and to the increase in the cost of services. Of course, the costs have increased for a variety of reasons; however, it is generally acknowledged that productivity in the service sector has not increased, as it has, traditionally, in manufacturing. For example, from 1956 to the present, aggregate productivity in U.S. industry increased by nearly 40 percent. During the same period, however, productivity in the U.S. Post Office increased by only approximately 3 percent.

In the last ten years, the aggregate costs of health care in the United States have more than doubled; they now exceed $70 billion per year. Expenditures have increased at an average rate of almost 13 percent per year, and the rate of increase itself is increasing. Productivity increases in health care, as in the post office, have been minimal. However, costs have risen because of the large increases in overhead costs required by advancing medical technology using expensive diagnostic and treatment equipment, and because of the fantastic increase in insurance premiums to finance the awards from malpractice suits.

Other kinds of nonmanufacturing systems that once were regarded as simple and insignificant have become worthy of careful study today. The corner grocery store has been replaced by the supermarket, where problems of forecasting, supply, inventory, layout, material handling, and so on can make the difference between profit or loss. Mass food preparation operations now parallel those of the supermarket. Other franchised operations, such as motels, face operations problems that are more significant today than they were when every manager was also an individual owner-entrepreneur. Banks have broadened their range of services and developed branch banking. In addition, banks have become one of the most important users of computers, and face large-scale office operations, that involve important problems of work flow and information processing. Other financial institutions, such as insurance companies, also face mass information processing problems.

The net result is that service and nonmanufacturing systems have taken their place with manufacturing as significant productive systems. They are important in society today, and therefore deserve the attention of the management scientist, the operations researcher, and the industrial engineer. Indeed, society demands that the service sector become efficient at converting its input resources to needed services, while maintaining certain quality levels. Productive service and nonmanufacturing systems are alike in the general nature of their problems and in the analysis and synthesis concepts and techniques that can be applied to them.

A HISTORICAL BRIEF

The great Scottish economist Adam Smith was the first to pay attention to production economics at the time the factory system was emerging. In 1776 he wrote *The Wealth of Nations*, in which he observed three basic economic advantages that resulted from the division of labor. These were: the development of a skill or dexterity when a single task was performed repetitively; a saving of the time normally lost in changing from one activity to the next; and the invention of machines or tools that seemed normally to follow when workers performed specialized tasks of restricted scope. Smith did not deduce these ideas in a theoretical way. Instead, he observed the cooperative approach of workers in the factory system.

Within the factory system, when a relatively large group of workers was brought together to produce in large quantity, division of labor logically developed as a common sense method of production. Smith noted these three advantages and wrote about them in his book. The book was a milestone in the development of production economics, not only because Smith's observations probably accelerated the division of labor, but also because a great scholar had recognized that a rationale for production existed. The actual development of the rationale took a long time. We are now, at last, in the true rapid developmental stage, in which operations management as a discipline is emerging from a purely descriptive phase and becoming an applied science.

Charles Babbage, an Englishman, augmented Smith's observations and raised a number of provocative questions about production organization and economics. Basically a mathematician who became interested in manufacturing, Babbage's inquiring mind and scientific orientation led him to question many existing practices. His thoughts were summarized in the book, *On the Economy of Machinery and Manufactures* (1832). Babbage agreed with Smith on the economic advantages resulting from the division of labor, but he observed that Smith had overlooked a most important advantage. For an example, Babbage used a study of pin manufacturing (the common straight pin) as it existed at that time. The level of specialization resulted in seven basic operations for making pins.

1. *Drawing wire.* This operation consisted of drawing wire through a die to reduce it to the desired diameter.
2. *Straightening the wire.*
3. *Pointing.*
4. *Twisting and cutting heads.*
5. *Heading.*

6. *Tinning or whitening.* This operation was comparable to a modern plating process to prevent rusting of the steel wire.
7. *Papering.* This consisted of placing the completed pins in papers or cards by piercing the paper; that is, the packaging of the pins.

Babbage noted the pay scale for these different specialties in shillings and pence (s.d.) (see Table 1). He then pointed out that if the shop were reorganized so that each person performed the entire sequence of operations, the wage paid to these workers would be dictated by the most difficult or rare skill required by the entire sequence. Thus, the enterprise would pay for the tinning skill, even when the worker was straightening, heading, or papering. With division of labor, however, just the amount of skill needed could be purchased. Therefore, in addition to the productivity advantages cited by Smith, Babbage recognized the principle of limiting skills as a basis for pay.

In the years after Smith's and Babbage's observations, the division of labor continued. It then accelerated during the first half of the twentieth century. Our production lines represent the principle of division of labor carried to its greatest extreme. In fact, it has come so far that some people are questioning the present level of application. Cost reductions based on *broadening* the scope

TABLE 1
ANALYSIS OF PROCESSES AND MANUFACTURING COSTS IN PIN MAKING*

Name of the Process	Workmen	Time of Making 1 lb of Pins, hours	Cost of Making 1 lb of Pins, pence	Workman Earns per Day s.	d.	Price of Making Each Part of a Single Pin, in Millionths of a Penny
1. Drawing wire (§ 170.)	Man	.3636	1.2500	3	3	225
2. Straightening the wire	Woman	.3000	.2840	1	0	51
(§ 171.)	Girl	.3000	.1420	0	6	26
3. Pointing (§ 172.)	Man	.3000	1.7750	5	3	319
4. Twisting and cutting	Boy	.0400	.0147	0	4½	3
the heads (§ 173.)	Man	.0400	.2103	5	4½	38
5. Heading (§ 174.)	Woman	4.0000	5.0000	1	3	901
6. Tinning, or whitening	Man	.1071	.6666	6	0	121
(§ 175.)	Woman	.1071	.3333	3	0	60
7. Papering (§ 176.)	Woman	2.1314	3.1973	1	6	576
		7.6892	12.8732			2320

*English manufacture. (178.) Pins, "Elevens," 5,546 weigh one pound; "one dozen," 6,932 pins weigh twenty ounces, and require six ounces of paper.
Number of persons employed: men, 4; women, 4; children, 2; total, 10.
Reproduced from *The Economy of Machinery and Manufactures*, 1832.

of jobs are being reported in the literature. A name has even been coined for this new trend: "job enlargement." Perhaps the optimal level of dividing tasks has been passed in some industries.

Frederick W. Taylor was undoubtedly the outstanding historical figure in the development of the production management field. Smith and Babbage were observers and writers, but Taylor was both a thinker and a doer. He was also an authoritarian with an indomitable will, a characteristic that drew much criticism but, at the same time, may have been the source of his great contributions. The practice of the day was to allow the workers themselves to decide the means by which production would be achieved. They determined how to produce a part, according to their skills and past experience, and the time and cost of production were guided by traditional methods. "Boondoggling" and spreading of the work were common.

Taylor was familiar with these practices because he entered the industrial system as a worker. In this capacity, he refused to go along with the other workers and instead produced as much as he could. He advanced rapidly and was later in a position to experiment with some of his ideas. To comprehend the extent of Taylor's accomplishments, we must understand that he was an innovator in a generally apathetic and strongly traditional managerial environment. The workers had free rein to determine manufacturing methods and the right to hold their knowledge as trade secrets. In this static environment, Taylor created a tidal wave of change in managerial philosophy that shook many organizations from top to bottom.

Essentially, Taylor's new philosophy stated that the scientific method could and should be applied to all managerial problems. He felt that the methods by which work was accomplished should be determined by management through scientific investigation. He listed four new duties of management that may be summarized as follows [Taylor, 1919]:*

1. The development of a science for each element of a person's work to replace old rule-of-thumb methods.
2. The scientific selection, training, and development of workers, instead of the old practice of allowing workers to choose their own tasks and to train themselves as best they could.
3. The development of a spirit of cooperation between the worker and management to ensure that the work would be carried out in accordance with the scientifically devised procedures.
4. The division of work between the workers and the management in almost

*Names with publication dates in brackets indicate corresponding references at the end of each chapter. When the name is used in a sentence, it will be followed by the publication date in brackets.

equal shares. Each group was to take over the work for which it was best fitted, instead of the former condition in which most work and responsibility fell on the workers.

Tay 51 per wgt

These four ideas, which led to much new thinking about managerial organization, are so much a part of present day organizational practice that it is hard to believe that the situation ~~was ever any~~ different. Taylor's work described under 1, above—the application of the scientific method—developed into the field of methods engineering and work measurement. In more recent years, this area has expanded greatly, with the help of experimental psychological and physiological researchers. Now the field known as "human engineering" has general application in production management. From 2 and 3 has developed the field of personnel, with its techniques of personnel selection and placement, together with the organizational function of industrial relations. Concept 4, the division of work between the worker and management, has had far reaching implications. The basic managerial functions of planning and control now cover much of the work formerly done by workers, leaving the first-line supervisor and the workers free to concentrate on the execution of carefully laid plans.

Taylor is also known for pioneering experiments in various areas. These include basic production organization; wage payment theory; and the development of fundamental procedures for tasks such as metal machining, pig-iron handling, and shoveling, which were common in the steel industry at that time. In his metal-cutting experiments, Taylor used thousands of pounds of metal over a ten-year period. The experiments resulted in specifications for the feeds and speeds that could be used for different metals and tool materials. In connection with these experiments, Taylor, in collaboration with Maunsel White, discovered high-speed steel. This discovery made him wealthy and allowed him to spend most of his later life furthering his philosophy. He believed that his general philosophy and approach to managerial problems were more important than any specific discoveries. He saw the latter as merely applications of "scientific management" to individual situations.

His followers were numerous. Carl Barth, Henry L. Gantt, Harrington Emerson, Frank and Lillian Gilbreth, and others worked within Taylor's general framework and philosophy. There were yet others who donned his cloak but lacked his knowledge and competence; and for a fast dollar they passed themselves off as consultants who could install the "Taylor system." Although Taylor disclaimed these "jackals," they did much harm. The bad name they gave "scientific management" probably slowed the development of good practice in the field. Little change occurred in Taylor's basic ideas, and the literature was filled with gimmicks and variants of his thinking, such as the wage payment plans, time study methods, and charts and mechanical control boards. But the science of production management as Taylor had envisioned it, was very slow to evolve.

There were many reasons for this slow development. Appropriate knowledge and tools were not yet available, and the misapplications of the post Taylor period had to be lived down. Measurements in production systems commonly display great variation. For example, how much output can we expect from an operation? Output depends on the person on the job and the job conditions. But even for a given person on the job, we can expect great variation from hour to hour, day to day, week to week, etc. To describe such a system we need probability concepts and a knowledge of statistical methods. For years, people tried to approach such problems by representing the output of a worker or a worker-machine system by a single number, as was common in most engineering problems. However, this method simply did not fit the situation. In fields such as mechanical engineering, electrical engineering, and chemical engineering, variability of measurements was small and the deterministic models yielded fairly good results. In production problems, however, variability was characteristic. Today, with the general knowledge of statistics and probability concepts, our models of productive systems are closer to reality than ever before.

Another serious difficulty that beset investigators in the period after Taylor was the complexity of the large-scale problems that appeared. All variables of any problem seemed completely interdependent. It was obvious that mathematical techniques were needed, but none was available to give the kinds of solutions required. Even if they had been available, the time required to develop solutions manually would have been measured in worker lifetimes. Modern high-speed digital computers were needed, but these were not to become available, even for the biggest and strongest companies, until the 1950s. An attempt at mathematical analysis was made in 1915 by F. W. Harris, who developed the first economic lot size model for a simple situation. This was further developed by F. E. Raymond and others, but applications of the idea in industry were not general.

The Current Era

The current rapid development of concepts, theories, and techniques in production/operations management began shortly after World War II. War operations research sponsored by the armed forces produced new mathematical and computational techniques, and created an awareness of how to apply old techniques to war operations problems. These problems seemed to parallel problems that occurred in productive systems; new approaches to operations problems began to be carried over into industry, thus introducing the modern era of management science and operations research.

As in the scientific management era, the original proposals for the application of operations research to industry revolved around a broad systems approach to managerial problems, and the techniques that were developed were conceived as modes of implementation. However, for a considerable period of time, emphasis was put on the new and powerful analytical techniques, and much of the broader systems view of problems was overlooked. Research and application became centered on such things as inventory control, mathematical programming, PERT/CPM, scheduling techniques, simulation, and waiting line theory. The broad philosophical framework of systems analysis seemed to be something to talk about rather than something to do, except in the field of computers and information processing.

As with scientific management, there are probably good explanations for the current era's emphasis of technique. It is much easier to work on a smaller problem of restricted scope and make progress with it than it is to grapple with large-scale problems. The systems concept required models that consider higher-order interactions between organization units. Not only is this difficult to accomplish conceptually, but it may also prove impossible to implement organizationally.

Currently, however, there is renewed interest in the broad conceptual framework of systems as well as in the enlarged view of what constitutes a productive system. At some point in the future, we can expect to achieve a truly integrated systems view of production/operations management problems. Figure 1 places these milestone events in perspective with time.

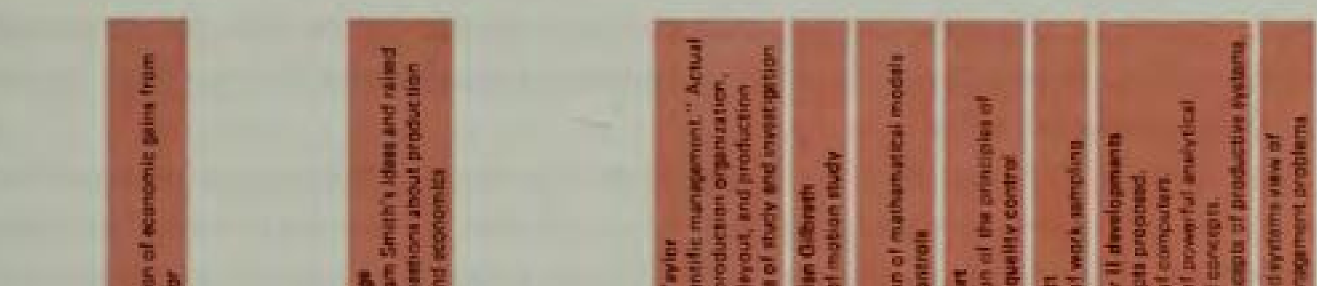


Figure 1. Developments in production/operations management

WHERE WE STAND TODAY

Almost two centuries have passed since Adam Smith's day. What have we learned about production/operations management in that time? Where does our knowledge stand today? In assessing the past, we can say that the results speak for themselves. Productivity and total productive capacity have expanded tremendously. Life for the average person in our Western civilization has been transformed from mere survival to a scale of living undreamed of by Smith. During this period, production/operations management has developed largely as an *empirical* applied science. In these two hundred years, we have responded to the expanding market and to the growth of increasingly large business units by division of labor and by progressive mechanization in order to take advantage of the economics of large-scale production.

Through the years we have learned to design better work places, better material handling equipment, and better buildings to house productive activities. We have created production lines and even automatic machines. We have learned basic principles of production economics and thereby have learned to employ labor, materials, and machines in a delicate balance to match the changing, relative values of these basic components of production. We have learned to control the production systems that we have designed, so that products or services meet quality standards and are available when needed, at a cost that may be fairly well predicted. Most of this development has been evolutionary; we improved existing systems through a process of trial and error.

Only in the last twenty years have we begun to evolve principles that enable us to design facilities and control systems with some degree of predictability as to their performance. Here is a true measure of our knowledge. Today, we are beginning to develop answers to problems of limited scope, knowing that the result is the best possible (that is, optimal, not simply better than the previous solution). This is real progress, and it indicates that the applied science Taylor envisioned is developing.

Where we stand today, then, in terms of production/operations management as an applied science, is at the threshold of the rapid developmental phase. Increased knowledge about some particular field often is related to a growth and saturation curve, where initial developments are slow and difficult. As bits of knowledge fall into place, the rate of growth accelerates to the rapid developmental phase and finally levels off as it approaches a saturation level. We have diagrammed this growth curve in Figure 2. In the years to come, the scope of the problems for which we can find provable optimal solutions will increase. The theory of productive systems will be pervasive and will embrace integrated systems as a whole, not just segments of a system; our ability to design facilities and control systems with predictable characteristics will increase. The use of computers to simulate systems will become common, as will numerically controlled (computer-controlled) production processes of various types.

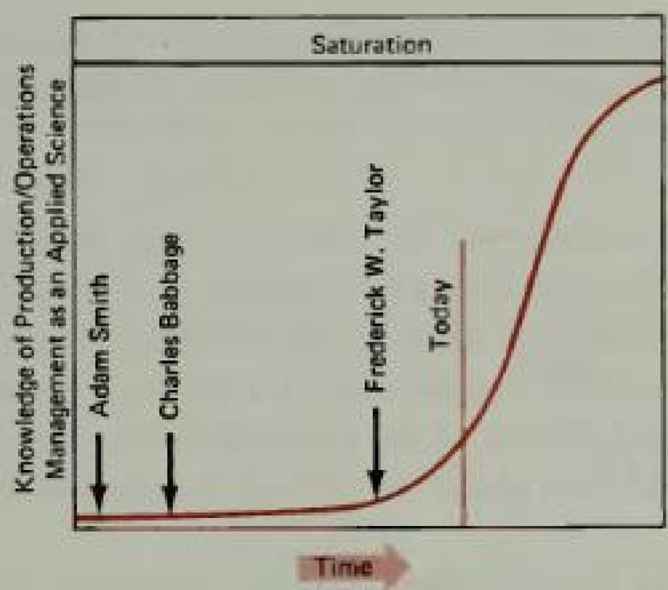


Figure 2. Growth curve of knowledge of production/operations management as an applied science

THE BIG ORGANIZATION VERSUS THE SMALL ONE

The following comment often is made: "All this talk about the new look is fine for the big organization—what about the small organization?" The basic principles of production economics and facility design and control are as applicable to the small organization as to the large. Implementation of these principles, however, must change with the size and financial strength of the organization. For example, a large organization may use a high-speed computer to help generate a schedule that meets requirements with minimum inventory. The small organization may be compelled to approximate a solution that is similar to its schedule problems, using hand methods and graphical aids. Both organizations may be attempting to apply the same principles of scheduling. The techniques are different, but the principles are general.

Review Questions

1. Select some country that is generally thought of as underdeveloped. Is it characterized by: industrialization; high or low productivity per work-hour; labor intensive activities; high or low real wages? What makes a country underdeveloped?

2. Underdeveloped countries could make more rapid progress toward development by investing in the most advanced labor saving technology. Discuss.

3. Which of the following can be thought of as part of the system for producing a good or service: raw-material exploitation and reduction; transportation of raw materials to the productive system; direct labor; indirect labor; productive equipment; scrap; by-products and waste; distribution of output? Which of the foregoing items are likely to be under managerial control for a physical product; for a service?

4. Why have services become significant as productive systems in the past ten years or so? Can you think of types of systems that currently are largely ignored as productive systems of significance—i.e., that are not the subject of inquiry and analysis? Why do we pay little attention to them?

5. What were the contributions of Adam Smith to the development of production economics? What were those of Charles Babbage?

6. Frederick W. Taylor developed four new duties of management as part of his philosophy of scientific management. Relate each of these four duties to functions that are routinely accepted today.

7. What pioneering work marked the beginning of mathematical model-building as a mode of analysis and decision in productive systems? What events led to the development of management science and operations research?

8. Taylor felt that his broad principles and philosophy were the important parts of scientific management; however, the movement came to be characterized by techniques such as wage payment plans, time and motion

study techniques, and charts and mechanical control boards. Are there parallels in the development and application of management science and operations research?

9. Appraise our current position in the march of knowledge about productive systems. In which general areas is important progress likely to be made in the future?

10. The principles and methods of production/operations management ~~are~~ mainly applicable to large organizations where improvements in operations can be significant to the organization, primarily because of the economics of scale. Discuss.

References

Babbage, C. *On the Economy of Machinery and Manufactures*. London: Knight, 1832.

Buffa, E. S. *Operations Management: Problems and Models*. 3d ed. New York: John Wiley & Sons, 1972.

———, and J. S. Dyer. *Management Science/Operations Research:* Model Formulation and Solution Methods. Santa Barbara, Ca.: Wiley/Hamilton, 1977.

Litterer, J. A. *Organizations*. 2d ed. New York: John Wiley & Sons, 1969.

Petit, T. A. *Fundamentals of Management Coordination*. New York: John Wiley & Sons, 1975.

Pickle, A., and R. Abrahamson. *Small Business Management*. Santa Barbara, Ca.: Wiley/Hamilton, 1976.

Taylor, F. W. *Principles of Scientific Management*. New York: Harper & Bros., 1919. Reprinted in *Scientific Management*. New York: Harper & Bros., 1947.

Thierauf, R. J., and R. C. Klekamp. *Decision Making Through Operations Research*. 2d ed. New York: John Wiley & Sons, 1975.

Urwick, L., and E. S. L. Brech. *The Making of Scientific Management*. Vol. 1. *Thirteen Pioneers*. London: Management Publications Trust, 1945.

Chapter 2

Decision Making in Production / Operations Management

Management's primary function is to make decisions that determine an organization's future course of action over both the short and the long term. While a wide variety of other managerial activities certainly exist, their function is to support the objective of improving decisions. Managers make plans in order to help make the best decisions. They set up controls to ensure that the plans and decisions that have been made are carried out or to alert themselves to the entry of new factors that might require a change of plans and decisions.

Decision making is complex because the systems with which we deal are complex and normally involve multiple goals and criteria. We will use criteria to determine which solutions are best. For example, a solution that minimizes cost or maximizes profit is guided by the criterion of cost or profit. However, we must always be aware of complexity in decisions, and consider all factors that may impinge on a decision problem. For this reason, we will introduce systems concepts and methods at length in Chapter 3 and will continuously attempt to maintain a systems point of view throughout this book.

NATURE OF DECISION MAKING

The obvious implication of a decision is that alternatives exist. In the process of decision making, we select from these alternatives the course of action to be carried out. The simplest way to make decisions is, of course, to let some chance system determine the choice, such as the flip of a coin. Such a decision making system, however, obviously would not be appropriate unless all outcomes of decisions were equally desirable. The problem of judging desirability immediately points up the need for (1) a purpose, and (2) criteria for measuring or comparing the desirability of the alternatives in relation to the purpose.

The decision process becomes more difficult when we realize that we are invariably talking about future values and, unfortunately, these future values often conflict. Each alternative may have both desirable and undesirable aspects, and these conflicting values must be reconciled. Since the results expected exist in the future, how can we be sure of obtaining those exact results? What is the probability of obtaining the results for each alternative? Or, conversely, what is the risk of not obtaining the forecasted result?

Figure 1 is a diagram of the structure of our decision making situation with alternatives, data from the real world, and criteria and values representing inputs to the decision maker. The predicted relative values of the alternatives, based on the criteria and values, are computed inside the decision making system. Finally, the relative desirability of the alternatives may be determined by weighing the values according to the probability of attainment. The course of action that maximizes this final desirability is the most rational decision.

The extent of sophistication of decision making for a given area depends on the level of knowledge within the area and the complexity of the decisions to be made. Sometimes we find that criteria and values are clear and straightforward, data are readily obtainable, future values are quite predictable, and risks are

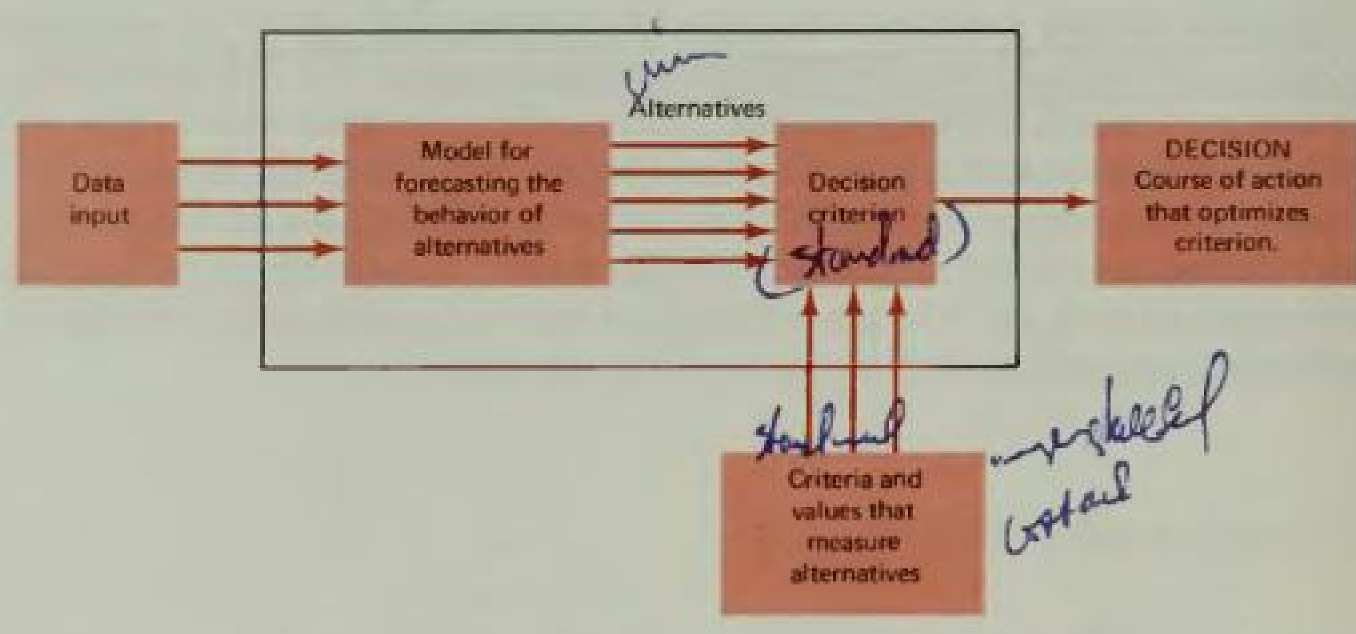


Figure 1. Use of models in decision making

fairly clear. In these instances, decision making seems scientific, mathematical, and almost automatic. In many other cases, criteria and values are vague and take several forms, the comparability of which is difficult to establish. Predicting risk and future performance may be even more difficult.

Judgment is the device by which we balance conflicting values, assess risks, and finally select a course of action. At all times, however, decision making is the attempt to choose those courses of action that have the greatest net desirability.

Complex Decisions Require Multiple Criteria

There are reasons for emphasizing the complexity of decisions. Both objective and subjective factors are important, and we need to view decision problems within a systems context. First, with the growth and development of quantitative methods of analysis, people have been tempted to make decisions based on seemingly objective factors. A mathematical or simulation model is attractive and persuasive. However, while model builders have been extremely clever in including a wide variety of variables in their models, it is not yet possible to include most subjective and behavioral variables, and perhaps it will never be.

Decision makers should welcome the objective analysis, but should not assume that all variables have been included. Indeed, perhaps the main reason that most decision makers are people rather than machines is that judgment must be used in making trade-offs.

The second reason for emphasizing the complexity of decision making is that the tools for dealing with this complexity are being developed rapidly.

We now will present a framework for decision making. It would be easy to interpret Figure 1 and our comments about decision making in terms of mathematical models and objective criteria only. We wish to emphasize that, in the following discussion of decision making (and throughout the balance of the book), we will assume that important objective and subjective factors both exist, and that the decision maker should consider trade-offs even if the model builder does not do so explicitly.

Therefore, when, in later chapters, we develop the objective factors of a model whose solution points to a given decision, we must recognize that we have temporarily narrowed our horizon so that we can look at local behavior of certain variables. We may emerge with such concepts as economic order quantities, line balancing, or a hospital admissions rule. These concepts lack the context of a specific organization situation. They are, in fact, inputs to the broader decision process, and are subject to trade-off—and possibly even dominance—by factors that are not in the model. Most important decisions have multiple objectives or goals, and the trade-off process can be objective in a broader concept of the decision process (see Buffa and Dyer [1977], and Easton [1973]).

Models

Figure 1 shows the relationship of the major elements of a decision making structure. The function of the model is to describe behavior, to forecast behavior of the alternatives that might be considered, or to scan the effects of alternatives and pick out the most desirable. There are several reasons why we might construct a model to accomplish these objectives.

In analyzing the design and operation of productive systems, we may be interested in the functional interaction among the various components of the system, and the sensitivity of any particular conception to variations in the form or content of such interactions. In addition, we may seek a basic understanding of the structure and functioning of the system being studied. Or, finally, we may be interested in the prediction of performance of the system under similar or different inputs. While we might choose to experiment with the real world system, this may not be possible or economical. Therefore, a model can be helpful in any one of the following five ways [Elmaghaby, 1968]:

1. As an aid to thought.
2. As an aid to communication.
3. As a tool for prediction.
4. For control purposes.
5. For training and instruction.

The model is always some sort of abstraction from the process for which it was modeled. The model may be physical, graphic, pictorial, schematic, analytical, simulation, or some other type. In operations management, all these kinds of models are useful for particular kinds of problems.

A good model will forecast performance that correlates well with data from the real situation; and, of course, models need to be validated. Using the model, we can generate alternative solutions to be appraised by the decision criterion. Here is where we frequently have trouble, because the model builder often builds a criterion function into the model, using a cost or profit function. For example, the linear programming algorithm uses a criterion function to generate the *optimal solution*, as do many other optimizing methods. There is nothing wrong with this. In fact, it epitomizes the model builder's art, for it presents the best solution for the model, given the *single* valued criterion.

The difficulty occurs after that point in the process. A so-called optimal solution to a problem seems to say, "This is *the* answer," and often intimidates the decision maker into either accepting it as it is or rejecting it. Too often, decision makers reject the solution because they do not trust the model, and/or because they legitimately feel that the model does not take into account other important factors that enter the problem. A much more appropriate presentation of a model's results would be an indication of a range of solutions, including the optimal solution, and an indication of the costs associated with each. This puts the decision maker in a position to see both how trade-offs can be made and the cost of these trade-offs.

Figure 1 indicates that the criterion has multiple values and should be thought of as separate from the model even when the model contains its own criterion function, as is the case with many optimizing methods.

THE PRODUCTION FUNCTION

Production is the process by which goods and services are created. We find productive processes in factories, offices, hospitals, and supermarkets. Production/operations management deals with decision making as it relates to productive processes, so that the resulting good or service is produced according to the specifications, in the amounts and by the schedule demanded, and at minimum cost. In accomplishing these objectives, production/operations management is associated with two broad areas of activity: the strategic or longer-run decisions focused in the design of the system, and the day-to-day decisions of operations.

efficiency

A GENERALIZED DESCRIPTIVE MODEL OF PRODUCTION

In light of our broad definition of productive systems, let us construct a model of what we mean by production. According to our definition, the factory, the office, the supermarket, the hospital, etc., all represent special cases that have special characteristics. Our productive system has inputs that represent the material, parts, paper work forms, and customers (or patients, as the case may be). The inputs are processed in some way by a series of operations whose sequence and number are specified for each input. The operations may vary from only one to any number, and may take on any desired characteristics; they may be mechanical, chemical, assembly, inspection and control, dispatching to the next operation, receiving, shipping, personal contact such as an interview, and paper work operations.

The outputs of our system are completed parts, products, chemicals, service to customers or patients, completed paper work, etc. Our system provides for storage after the receipt of the input and between each operation in the system. The time in storage may vary from zero to any finite value. Inputs are transported among all operations in the system. Any means of transportation may be supplied, including self-transportation in the case of clients and customers. Our model also has an information system and a decision maker. The information system interconnects all physical activities and provides a basis for management decision. These functions provide the equivalent of a "nervous system" for the model. Figure 2 represents our productive system.

Systems may occur in series. For example, when completed products are shipped from a factory to a warehouse, they leave the factory system only to arrive at a second productive system, called a warehouse. In this way, the two

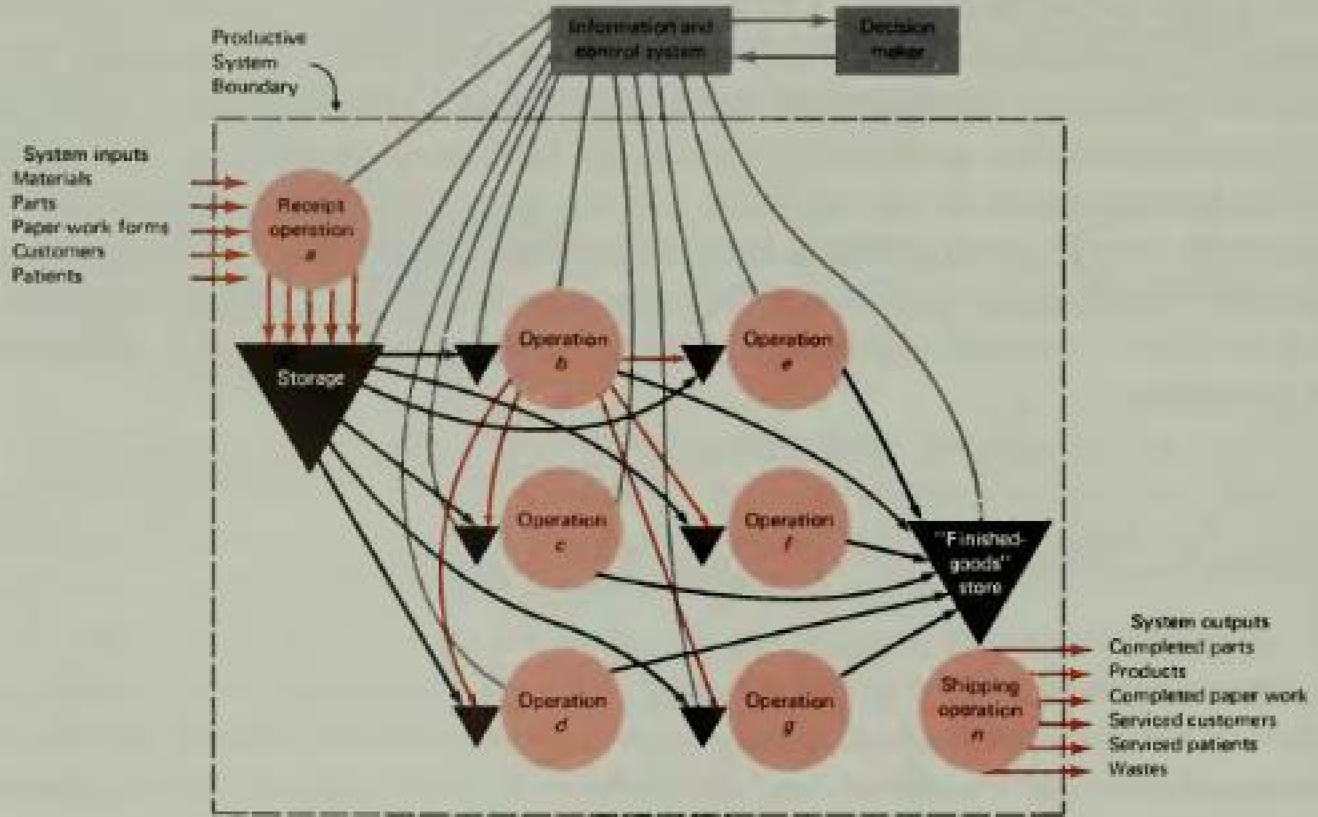


Figure 2. Diagram of a generalized productive system. Inputs may be processed in any specified sequence of operations and are transported between operations. The number of operations may vary from one to any finite number. Storage occurs between all operations, and the time in storage may vary from essentially zero to any finite amount. *Note:* There are interconnections among all combinations of operations *b* through *f*, although only those originating at *b* are shown. The information system interconnects all activities and provides the basis for management decision.

systems may be considered as part of one large system. Systems also may occur in parallel, such as when a number of factories produce similar products to supply several market areas. For solving some types of problems, we may consider these factories as one large productive system. Similar combinations may occur with service oriented systems.

Now let us examine what happens to an input to the system. After being received, the input goes into storage to take its turn in the processing. By some set of priority rules, it is drawn from storage to begin processing. These rules might be: first in–first out, time or date required for completion or delivery, urgency, or some other system of priority rules. It then is processed according to a predetermined sequence. Let us assume the sequence *b*, *d*, *c*, *n*. From initial storage it goes to operation *b* and is placed in temporary storage to await processing there. We assume that operation *b* has already some assigned work or load and, therefore, our input takes its place in line (in storage) to be processed at *b* according to the priority decision rules established. After being processed at *b*, it is transported to *d*, placed in storage, drawn out according to priority rules, processed, etc. One operation may be an inspection. The operation just preceding shipment may be a packaging operation, preparing the item for shipment.

If we are speaking of a high-volume, standardized fabricated item, the operations may be placed in sequence and interconnected by conveying equipment. The storages would take place on the conveyors themselves, and the decision rules are first in–first out. If we are speaking of the hospital, the storages may take place in waiting rooms or in hospital beds. The priority rules may be first in–first out, with urgency exceptions. In the instance of the hospital, many tasks are mobile, such as when nurses give shots or medication.

In a supermarket, products are received and stored on display shelves. Customer receiving is practically nonexistent as an operation. The customer picks the desired items from shelves, transports them, and takes his/her place in line for the single operation of check-out.

In soap manufacturing by the continuous process, materials are received and stored. They are then taken from storage in large quantity, dissolved, and pumped through a series of chemical operations, so that transportation and storage between operations occurs in the pipes. The material is processed *chemically* while it moves, and it emerges as soap. It is then packaged and shipped.

The job shop manufacturing situation, where custom products are fabricated and assembled, is undoubtedly the most complex type of productive system. For example, to build a missile, thousands of individual parts must be fabricated and assembled into subassemblies and then final assemblies. This activity must be dovetailed to fit into a complex schedule, so that operation time is available when it is needed to provide parts for subassemblies and final assemblies. The pattern or flow for the multitude of parts from operation to operation is so complex that it can only be visualized by some abstract means of representation. Many parts require operation time on the same machines, but the operations occur at different times in the overall manufacturing cycle. The problem of loading the operations in such a way that the machines can be utilized effectively is obviously a difficult one.

CONTINUOUS VERSUS INTERMITTENT MODELS

Continuous flow situations are those where inputs are standardized, so that the facilities can be standardized as to routings and flow. Therefore, a standard set of processes and sequence of processes can be adopted. Continuous models are represented in practice by production and assembly lines, large-scale office-operations processing forms by a standard procedure, continuous flow chemical operations, etc.

Intermittent production situations are those where the facilities must be flexible enough to handle a wide variety of products and sizes, or where the basic nature of the activity imposes changes in important characteristics of the input (e.g., change in product design). In instances such as these, no single sequence pattern of operations is appropriate; thus, the relative location of the operations must be the best compromise for all inputs considered together. Transportation facilities between operations also must be flexible enough to accommodate a wide variety of input characteristics, as well as the wide variety of routes that the inputs may require. Because the flow is intermittent, these conditions commonly define an intermittent production situation. It is intermittent because the flow is intermittent. Considerable storage between operations is required in order that individual operations can be carried on somewhat independently. This results in ease of scheduling and fuller utilization of workers and machines. In practice, intermittent systems are represented by custom- or job order type machine shops, hospitals, general offices, batch-type chemical operations, etc.

As we have shown, our generalized descriptive model in Figure 2 can be made to fit both the intermittent and continuous flow situations by the specification of some detailed characteristics. We have assumed intermittent flow in our model, and defined it in general enough terms that continuous flow conditions would be determined by the specification of a fixed operation sequence, the specification of continuous flow type transportation facilities, the assumption of low storage times between operations, and a first in-first out set of priority decision rules would all determine continuous flow conditions. The continuous flow situation is common enough today, however, for us to think of it as a separate case from intermittent flow models.

PROBLEMS OF PRODUCTION/OPERATIONS MANAGEMENT

Using our generalized model as a background, let us outline the nature of problems generated in a production system. These problems require two major types of decisions: one that relates to the design of the system, and one that relates to the operation and control of the system (that is, both long-run and short-run decisions). The relative balance of the emphasis on such factors as cost, service, and reliability of both functional and time performance depends on the basic purposes of the total enterprise and on the general nature of the good or service being produced. Thus, in general, economic enterprises probably will emphasize cost, consistent with quality and delivery commitments. Hospitals may emphasize reliability and service, consistent with cost objectives etc.

Long-run strategic decisions related to the design of the productive system are:

1. *Selection and design of products.* There are strong interactions between product selection and design with productive capability, and vice versa.
2. *Selection of equipment and processes.* Usually, alternate equipment and processes are available for a given need. Management must make decisions that commit capital of the enterprise and establish its basic approach to production in the facility design.
3. *Production design of items processed.* Cost interacts strongly with the design of parts, products, paper work forms, etc., or the design of the service offered. Design decisions often set the limiting characteristics of cost and processing for the system.
4. *Job design.* Job design is an integral part of the total system design, involving the basic organization of work, as well as the integration of human engineering data, to produce optimally designed jobs.
5. *Location of the system.* Location decisions can, in some cases, be important if the balance of cost factors determined by nearness to markets and material supplies is critical. In service systems, location in relation to users may be of significance.
6. *Facility layout.* Decisions related to design capacity, basic modes of production, shifts, use of overtime, and subcontracting must be made. In addition, operations and equipment must be located in relation to each other in a pattern that minimizes overall interaction cost or meets the

requirements of some more complex criterion. The latter requirement is most difficult for the complex intermittent model where routes vary. Many detailed problems are associated with one another in the process of specifying the layout of a productive system adequately. These include heating, lighting, and other utility requirements; the allocation of storage space, aisle space, etc.; and the design of the building to house the layout.

Short-run decisions related to day-to-day operation and control:

1. *Inventory and production control.* Decisions must be made concerning the allocation of productive capacity, consistent with demand and inventory policy. Feasible schedules must be worked out, and the load on workers and machines and the flow of production must be controlled.
2. *Maintenance and reliability of the system.* Decisions must be made regarding maintenance effort, the random nature of equipment breakdown, and the possibility that machine down time may, itself, be associated with important costs or loss of sales.
3. *Quality control.* Decisions must be made to set the permissible levels of risk; sometimes bad parts are produced and shipped, errors are made, or good parts are scrapped. Inspection costs must be balanced against probable losses resulting from passing defective material or services.
4. *Labor control.* Labor is still the major cost element in most products and services. Production planning requires an appraisal of the labor component; thus, much effort has gone into developing work measurement and wage payment systems.
5. *Cost control and improvement.* Supervisors must make day-to-day decisions that involve the balance of labor, material, and some overhead costs. In controlling the activities, costs may be controlled.

The relative importance of these problems of production/operations management varies considerably, depending on the nature of individual systems. Nevertheless, every system has these problems to some degree. For example, equipment policy may occupy a dominant position in productive systems where the capital investment per worker is very large, as in the petroleum industry. On the other hand, equipment policy may occupy a minor role in a productive system that is represented by a large labor component or a large material component. Part of the art of production/operations management involves sensing the relative importance of these various problems within a given situation.

The chapters that follow are organized into three broad sections. Part 2 deals with general analytical methods that are of particular value in production/operations management. These methods serve as general models for decision making and may be altered to fit specific situations. Parts 3 and 4 deal with the general problems that are faced in the management of productive systems.

Review Questions

1. A manager of a small precision manufacturing concern in the field of electro mechanical devices is in the process of deciding on the employment level and aggregate production rate for the next month, July. He knows by past experience that July represents a seasonal low and, in addition, that finished goods inventories are higher than expected because sales in May and the first half of June were lower than the forecast. The forecast for July, however, is relatively high, based on industry indicators and the current flow of orders.

   Much of the fabrication work requires highly skilled machinists, and even the assembly operations require careful adjustments. The machinists' union is well organized, and layoff policies are controlled by a union agreement. In addition, the manager was once a machinist himself, and he recognizes the value of his skilled employees. The company has developed excellent cost and performance standards, and the manager feels that, in making his decision, he wants to be able to appraise the cost and profit impact of alternatives, and to request estimates of the cost performance of various combinations of numbers of employees, use of overtime, and activity rates. Some of the alternatives involve different numbers of employees laid off, and some involve the further accumulation of finished goods inventories.

   Place the manager's problem within the decision making structure of Figure 1, and identify: the input data; the system for forecasting performance; criteria and values that measure alternatives; the alternatives themselves; and the decision criterion. On what basis can the manager determine which alternative will maximize overall desirability?

2. What are the appropriate criteria for deciding on the employment and activity levels for Question 1? Are any of the criteria in conflict with one another? If so, how can conflicting criteria be reconciled? In your judgment, can all the criteria and values be reduced to the equivalent monetary units?

3. The following are decision problems in production/operations management that occur in various types of systems. List for each the kinds of criteria you think should be important:

a. Location of fire stations.
b. The amount of raw material to order at one time for a manufacturer.
c. The number of tellers' windows to maintain open for service in a bank.
d. The scope of jobs (degree of specialization) in automobile assembly, or in the claims department of a large insurance company.
e. The number of flight attendants assigned to the coach section of a transcontinental flight.

228

4. What is the function of a model in production/operations management problems? In what ways can a model be useful?

5. Under what conditions does the optimal solution to a model also represent the best solution to the real world problem the model purports to represent? Under what conditions might optimal model and real world situations disagree?

6. Figure 2 is a diagram of a generalized productive system. What is the nature of flow for each of the following types of system outputs?
   a. A high-volume manufactured product, such as beer cans.
   b. A low-volume manufactured product, such as a large electrical distribution transformer.
   c. Customers in a supermarket.
   d. Patients in a hospital.
   e. Waste products in a steel mill.

7. What are the characteristics of intermittent and continuous productive systems? How would you classify each of the following, and why?
   a. Registration at a university.
   b. Services available at a city hall.
   c. Aircraft assembly.
   d. A computer center, from the viewpoint of the user.
   e. A cafeteria.
   f. A full service hotel.

8. Define the following terms:
   a. Production design of products.
   b. Job design.
   c. Facility layout.
   d. Production control.

9. Classify each of the following kinds of decisions as having long-run or

short-run significance for a productive system:

a. Quality control.
b. Inventory management and control.
e. Equipment selection.
d. Labor cost control.
e. Design of products.

References

Alcalay, J. A., and E. S. Buffa. "A Proposal for a General Model of a Production System." *International Journal of Production Research*, March 1963.

Bross, I. D. F. *Design for Decision*. New York: Macmillan, 1953.

Buffa, E. S., and J. S. Dyer. *Management Science/Operations Research: Model Formulation and Solution Methods*. Santa Barbara, Ca.: Wiley/Hamilton, 1977.

Easton, A. *Complex Managerial Decisions Involving Multiple Objectives*. New York: John Wiley & Sons, 1973.

Elmagraby, S. E. "A Role of Modeling in IE Design." *Journal of Industrial Engineering* 19(June 1968): 292–304.

Groff, G. K., and J. F. Muth. *Operations Management: Analysis for Decisions*. Homewood, Ill.: Richard D. Irwin, 1972.

Miller, D. W., and M. K. Starr. *Executive Decisions and Operations Research*. 2d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Reisman, A., and E. S. Buffa. "A General Model for Production and Operations Systems." *Management Science* 11(September 1964).

Schoderbek, P. P. *Management Systems*. 2d ed. New York: John Wiley & Sons, 1971.

Vollmann, T. E. *Operations Management: A Systems Model-Building Approach*. Reading, Mass.: Addison-Wesley, 1973.

Part II

Analytical Methods

Chapter 3

Systems Concepts

The concepts and methodology of "systems" help unify and relate the complexities of managerial problems. Although these concepts have been applied more to the analysis of productive systems than to other fields, the value of the concepts in management is clear.

The word *system* is used so commonly in the general literature of management that it has almost lost its value. Yet it describes so well the general interaction of the myriad elements entering managerial problems that we can no longer talk of complex problems without using this word. And in fact we use the term liberally throughout this book. It is important to distinguish the general use of the term from its specific use as a mode of structuring and analyzing problems.

One of the great values of systems concepts is that they help us order and structure a very complex situation. Systems concepts reduce complexity in managerial problems to a block diagram that shows the realtionship and interactions of the various elements affecting the problem at hand. At their present state of development and application, systems concepts are most useful in helping us understand problems. A second, and very powerful, contribution of systems analysis is that it provides a basis for generating solutions to problems, evaluating their effects, and designing alternative systems.

In 1947, the idea of a general systems theory was proposed by L. V. Bertalanffy. General systems theory asserts that there are properties of systems that do not derive directly from the components themselves but from the unique combination of components that make up the whole. Furthermore, these properties make the whole (the system) add up to more than the simple sum of its parts. A human being is more than an assembly of cells, tissues, organs, etc;* an economy is more than a group of industries; a productive system is more than processes, people, and materials. The behavior and performance of each is an expression of its totality. In every instance, the combination of components, whose interrelationships are unique, produces a whole that is unique.

To understand the whole, begin with the whole, not with the components! This is not to deny the importance of understanding the components, but it reorients the focus of our examination of the components; in a systems orientation, we view a component, in terms of its reason for being, as a part of the system. Thus, systems concepts take advantage of the results of both viewing the system as a whole, and analyzing the proper role of components within the system.

Because of system goals and objectives, and complex relationships and interactions among components, we may need what might seem like poor performance from a certain component. For example, physical machine maintenance is one of the necessary components in operating a computer center. If we examined the maintenance operation, we might be shocked to realize that highly paid skilled technicians are idle a good deal of the time. Thus, maintenance seems to be a very inefficient operation. If we examined ways of improving the efficiency of the maintenance operation, we might reduce the crew size, or have maintenance work performed only on the night shift, using one-third the total labor required. The result would be a maintenance labor force that was utilized more effectively. However, if we examined the values of the broader system (operating the computer center as a whole), we would see that, clearly, the role of the maintenance component is to keep the system operating. The job of designing the maintenance subsystem thus takes on a different character; the maintenance subsystem takes its goals from those of the system as a whole.

EXAMPLES OF SYSTEMS

Two examples will illustrate the complex nature of productive systems. The first is a manufacturing system, which is focused on the planning, scheduling,

*Regarding the behavior of whole systems, Buckminster Fuller made the following quip: "It helps explain the fact that an intelligent woman can love $4.98 worth of chemicals arranged into a form called man."

and control problems of manufacturing. The second is a community health care system, which focuses on the problems of operating a maternal and infant care clinic. Both are significant productive systems.

A Manufacturing System

Suppose that we are examining the production of integrated circuits (see Reitman [1971, Chap. 9]). The process involves a sequence of operations and inspections on *batches* of silicon slices. The silicon slices are approximately one-and-a-half inches in diameter and may yield up to one thousand separate electronic circuits. The processes involve the use of various types of equipment, such as mechanical polishers, chemical etchers, diffusion furnaces, photoengraving apparatus, and testing equipment.

Because the technology is new, the production process is subject to change. Circuits rarely are made the same way for two years in a row, and most processes change within one year. All silicon slices have the same processing sequence until a branching point is reached; at this point, the slice becomes a unique circuit. From here on, the details of processing may depend on the nature of the circuits and the slice. The yields of good product are highly variable and rather low.

The operations manager has many problems, but the variable nature of the finished silicon slice, the differences in order sizes, and the low and variable yields place particular emphasis on production controls. Thus, if we follow an order through the system, we can develop a block diagram, similar to Figure 1, which shows the major decision points and activities required. Decisions or branches in the sequencing are shown as diamonds, and activities are shown in boxes or blocks.

Note that Figure 1 does not represent the physical production process. Rather, it shows what happens to a typical order in the processing sequence. In this way, we determine whether or not the order can be filled from stock, whether or not identical items are already in process, etc. If none of the foregoing is an alternative, a new batch, is started whose size is based on yield expectation. This batch is processed through the sequence of operations that finally produces a completed order.

Figure 1 is a simplified diagram of what happens as an order is processed. The level of detail shows the complex interrelationships of the various decision and branch points, and the main inputs and outputs of each block. A complete diagram would follow the same general structure, but each activity might be represented by a series of steps, and perhaps by branches and decision points within activities.

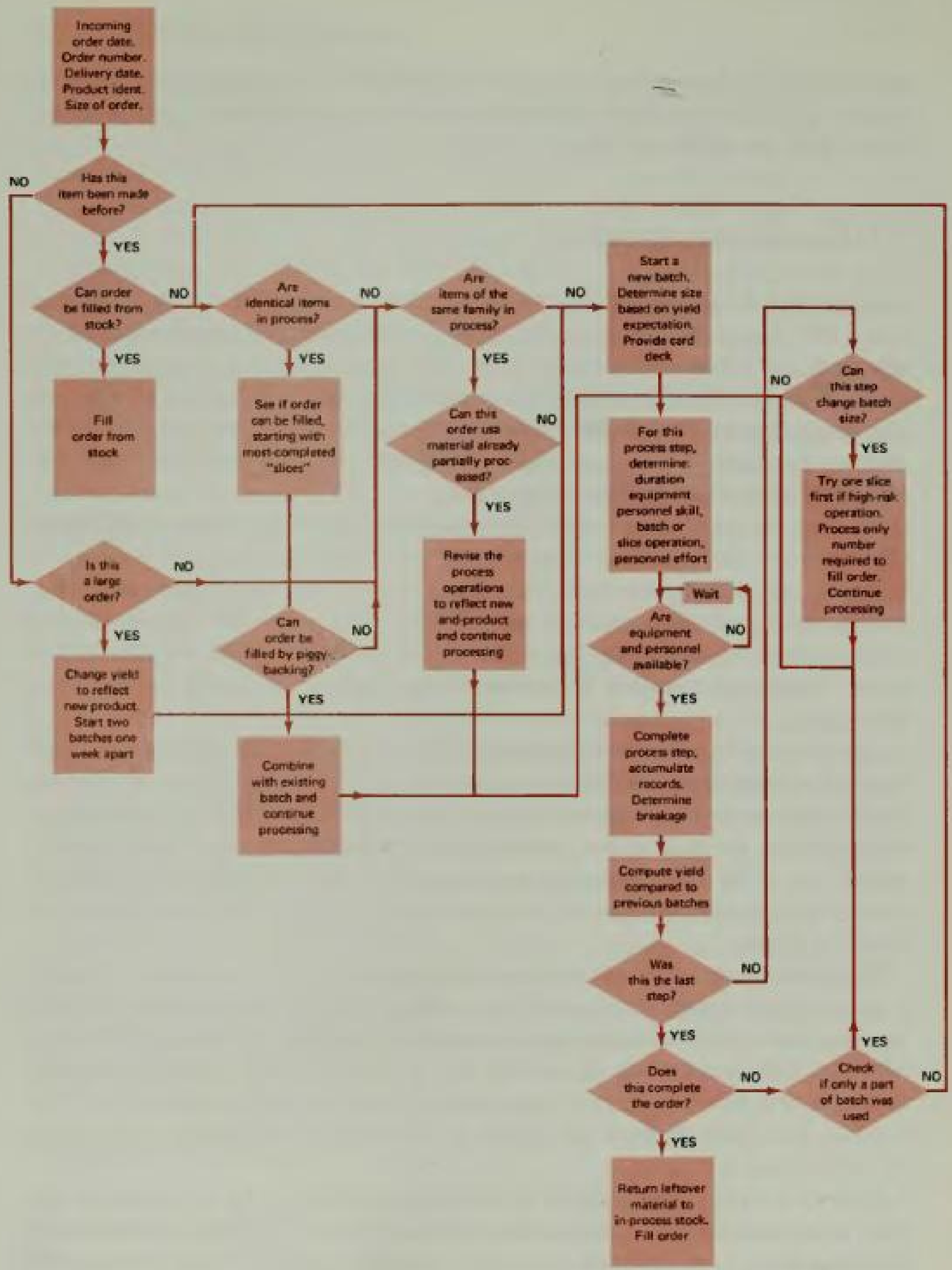


Figure 1. Simplified logic diagram of manufacturing system

SOURCE: J. Reitman, *Computer Simulation Applications* (New York: John Wiley & Sons, 1971).

Now, if Figure 1 describes the system, with its complexities and interactions among components, can we use it as a basis for predicting system behavior, given stated inputs? Yes, by developing a system simulation computer program.

Simulation. A simulation program was developed for the manufacturing system of Figure 1 [Reitman, 1971]. The great advantage of using system simulation is that we can experiment on the representation of the system rather than on the actual system. Or, if we are designing a system, we can predict system performance before expensive installation, possibly changing some elements of the design.

Results. In our case, the simulated manufacturing system was driven with typical order input data to see how it would behave. We will show only samples

QUEUES ON MONDAY 8.00 AM

FACILITIES QUEUES (1 EACH EXCEPT FOR 6 PHOTO)

#/ orders +/B processed at each facility

FACILITY	USED BY STEP(S)	QUEUE LENGTH
SLUG DIFFUSION	1	.
OXIDATION	2,5B,11	45.
PHOTO	3,12,14,16,18,20	.
ETCH	4,22	22.
PYROLYTIC	5B	83.
XTAL GROWTH	6	.
DEBURR	7	.
BACK LAP - FRONT LAP	8,9	.
POLISH	10	.
BASE DIFFUSION - STEP A	13	.
BASE DIFFUSION - STEP B	13	11.
EMITTER - STEP A	15	.
EMITTER - STEP B	15	6.
GAIN ADJUSTMENT	17	3.
ALUM EVAP	19	.
INTERCONN ALLOY	21	.
GOLD EVAP	22	.
GOLD ALLOY	24	.

QUEUES FOR PEOPLE

PERSONNEL SKILL	NUMBER TRAINED	QUEUE LENGTH
DIFFUSION	7	.
PHOTO	6	43.
POLYCRYSTALLINE	1	21.
LAPPING & POLISHING	1	76.
BACKING	1	32.

Figure 2. Report of system status showing orders waiting to be processed (queues) by facility and by type of skill

SOURCE: J. Reitman, *Computer Simulation Applications*, (New York: John Wiley & Sons, 1971)

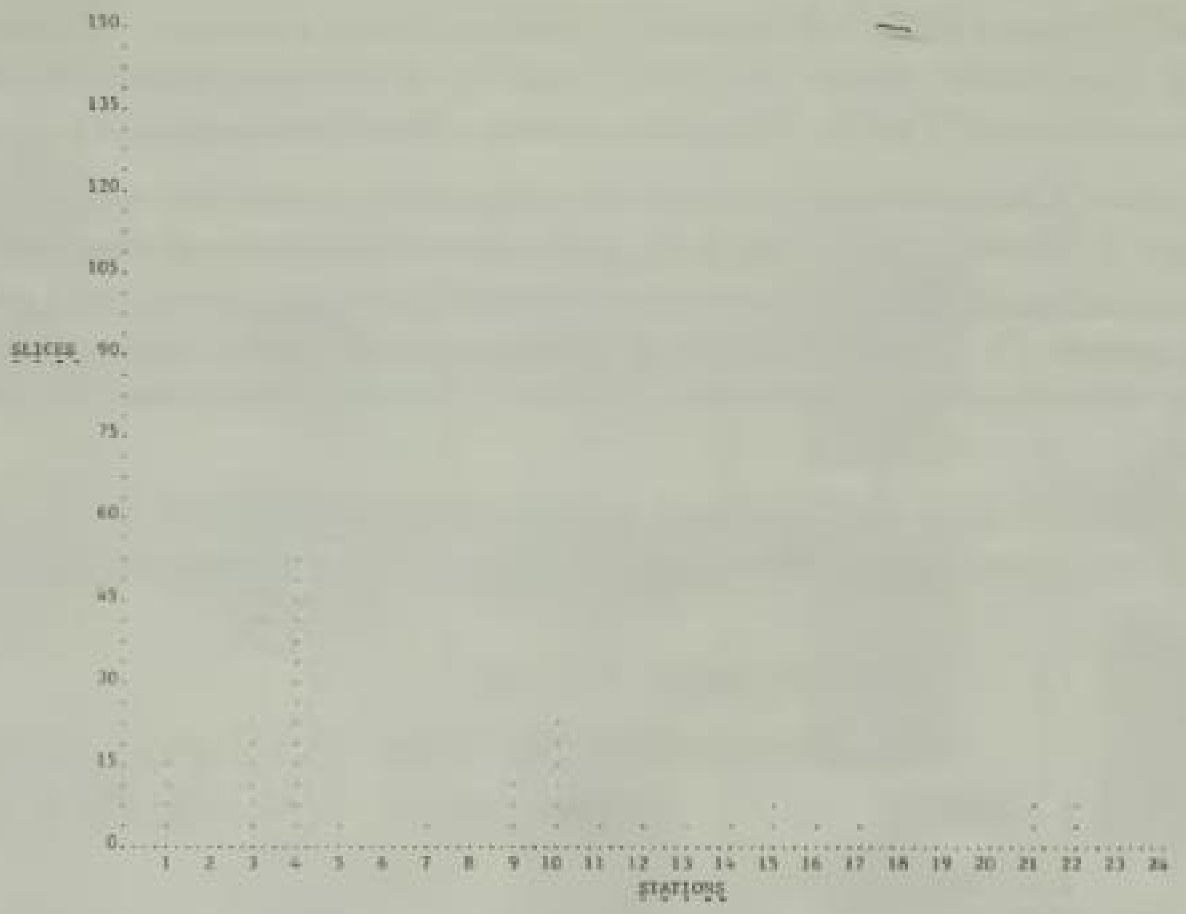


Figure 3. Output report showing status of orders which have not moved since a stated date

SOURCE: J. Reitman, *Computer Simulation Applications* (New York: John Wiley & Sons, 1971)

of the kinds of information that resulted. For example, Figure 2 shows the status of the system at 8:00 A.M. Monday. For each facility in the system, Figure 2 shows the steps for which it is used and the queue length (number of orders waiting to be processed at each facility). Shown at the bottom of Figure 2 is a listing of how orders waiting (queues) are related to types of skills.

Figure 3 shows graphic computer output concerning the orders that have been held up and therefore have not progressed in the system since a particular date. Another report shows how long each order has been held. Figure 4*a* shows summary graphic results concerning the number of days that orders spent in the system. Figure 4*b* shows what happens under overload conditions: the order-completion time becomes progressively longer.

Other kinds of output could also be abstracted: the in-process and finished goods inventory levels; the labor, material, and total production costs; the yields; and so on. All these outputs, and perhaps others, are measures of system performance. They are products of the complex interaction among system components, and cannot be derived in any simple way.

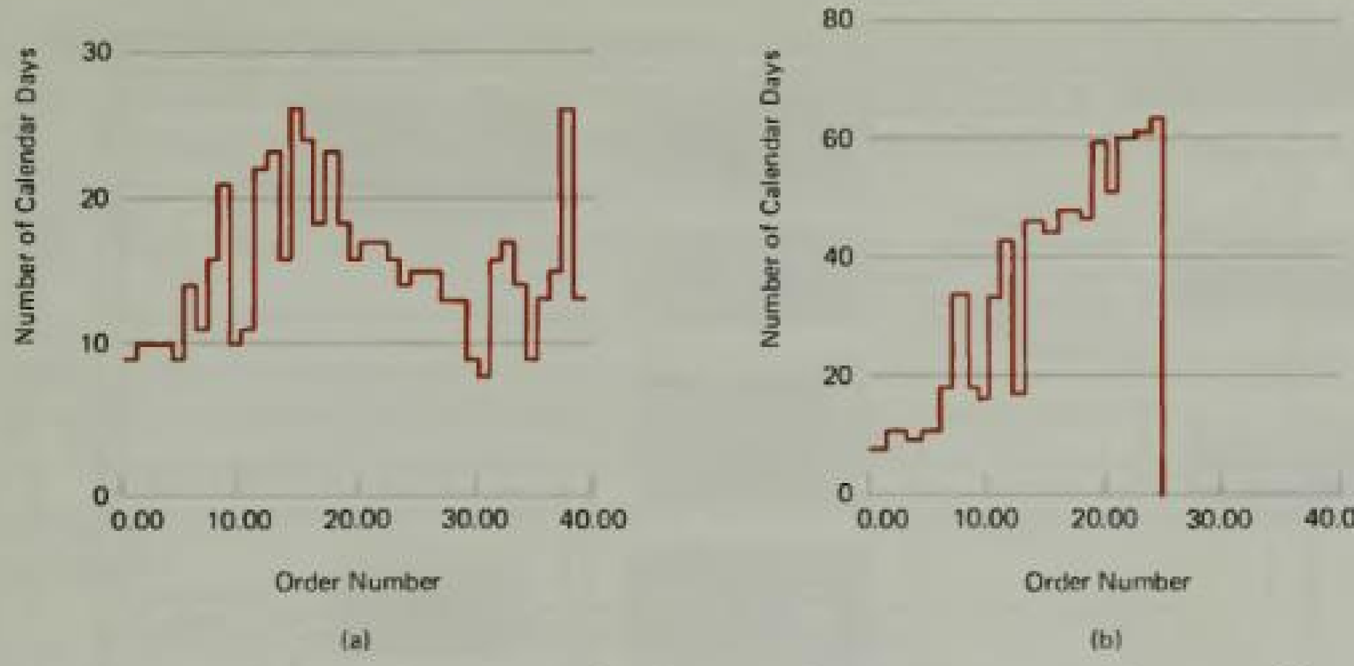


Figure 4. (a) Order completion times for the manufacturing system under normal load, and (b) under heavy load.

SOURCE: J. Reitman, *Computer Simulation Applications* (New York: John Wiley & Sons, 1971)

A Health Services System

The operations phases of health delivery systems have been catapulted into prominence because of the great cost increases, as we noted in Chapter 1. Medical care costs are increasing at a faster rate than any other consumer price indexes. Because health care systems are complex, simple prescriptions for improving their performance have seldom been effective.

Maternal and infant care have been examined as a system [Kennedy, 1969]. Figure 5 shows a block diagram developed for the North Carolina State Project for Comprehensive Maternal and Infant Health Care. Demands on the system are created by pregnancies and ill children. They are fed to the health care facilities—Infant Care Clinic, Pediatrician Office, Hospital Outpatient Clinic, and the Maternal Care Clinic. The functions of diagnosis, treatment, prevention, and promotion are associated with each facility (promotion refers to programs for disseminating information about community health resources, health standards, etc.).

Flows from the facilities are (1) rehabilitation, and (2) delivery and care of mother and child, both of whom ultimately are recycled to become potential health needs again.

The magnitude of the various flows in the system was established on the basis of regression analyses of data collected. For example, regression analysis

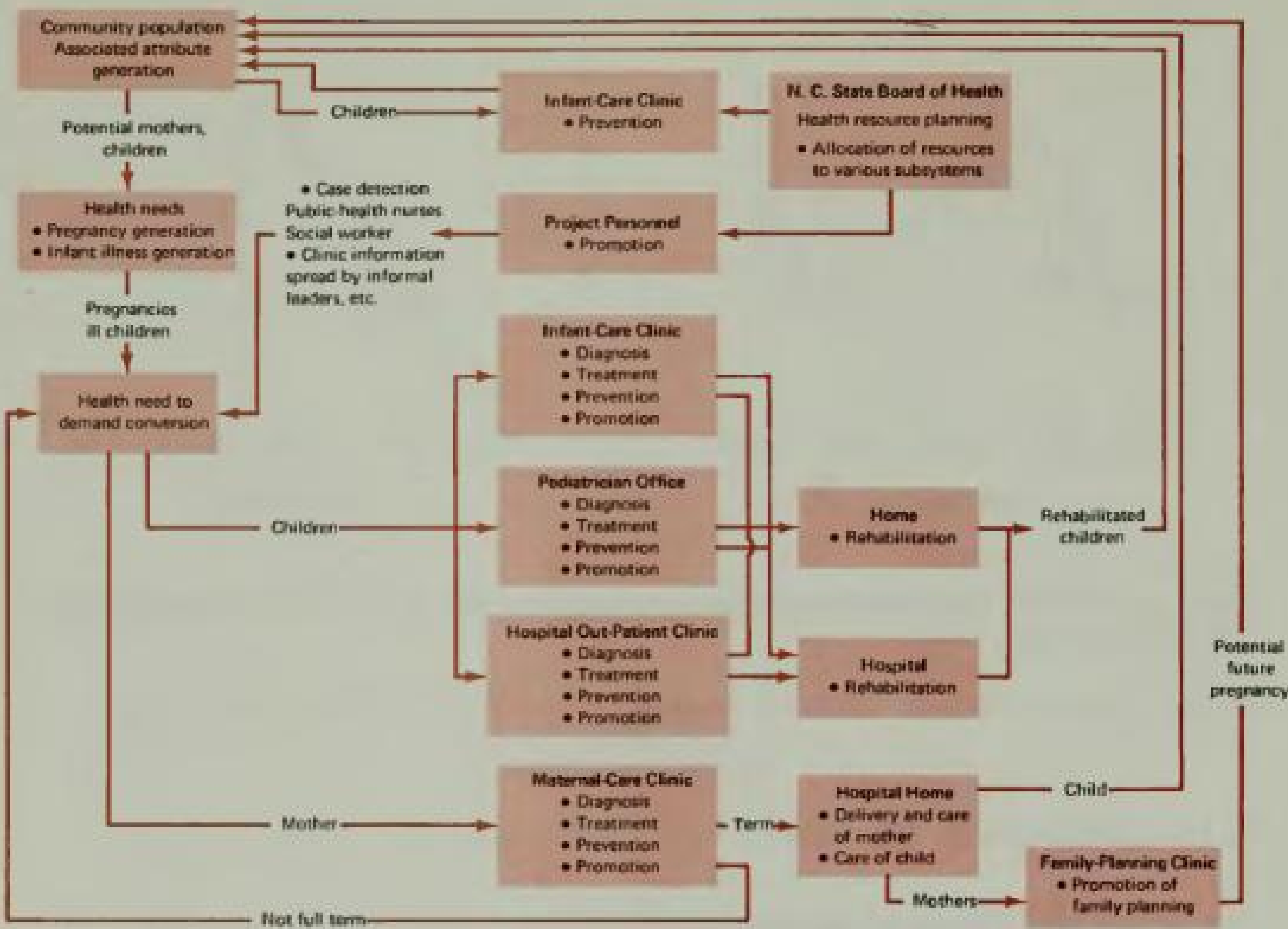


Figure 5. Block diagram for a generalized maternal- and infant care health system

SOURCE: F. D. Kennedy, "Development of a Community Health Service System Simulation Model," *IEEE Transactions on Systems Science and Cybernetics* SSC-5, no. 3 (July 1969): 199–207.

yielded the following equation for predicting the number of visits by mothers to the Maternal Care Clinic.

$$\begin{aligned}\text{Visits} = &-(0.224 \times \text{number of weeks pregnant at first visit})\\ &+(0.3547 \times \text{length of gestation at birth, weeks})\\ &+(0.0002 \times \text{annual income})\\ &-2.427\end{aligned}$$

Other data, analyses, and assumptions about needs and demands—the number of home and extra visits, birth weight, number of infant illnesses, service time, etc.—establish the quantitative relationships within the system.

Evaluation of Alternative Programs. Alternative programs were evaluated on the basis of various measures of effectiveness, as follows:

1. The ratio of scheduled to actual visits.
2. Personnel utilization ratio (hours used divided by hours allocated). This is a composite measure, reflecting relative cost as well as service level.
3. Abnormality ratio compared to national rates.
4. Infant illness ratio in relation to national rates.

Simulation and Results. Given the identification of the system in Figure 5, together with the quantitative relationships established, a computer simulation program was developed as a basis for evaluating proposed plans. The demands on the system were converted into demands for personnel and facilities. The simulation model could be used to assist in several dimensions of planning. For example:

1. Sensitivity of clinic caseload to pregnancy rate.
2. Sensitivity of clinic caseload migration rates.
3. Sensitivity of program effectiveness to personnel staff levels.
4. Sensitivity of clinic costs to social parameters of acceptance.

Figure 6 shows the results of two sets of simulated experiments indicating the effect on caseload of variations in the pregnancy rate and the migration rate. Figure 6*a* shows the effect of a $\pm$ 15 percent variation in pregnancy rate, indicating that caseload is quite sensitive to this parameter. Figure 6*b* shows the effect of a $\pm$ 40 percent variation in migration rate, indicating that caseload is fairly insensitive to migration rate. Note also the apparent seasonality in caseload.

Implications of the Manufacturing and Health Care Examples

Now let us consider why the methodology of the two studies was so important. First, in both examples, the entire system was considered. In the manufacturing example, this consideration included a series of decisions concerning ways of handling the order efficiently by examining finished goods inventories, as well as the availability of facilities and personnel, yields, etc. In the health care example, the process for generating demand, the facilities, the impact of preventive measures, family planning, and the recycling of patients were all related in the system. In both examples, changes in each component of the overall system can affect the other components, as some of the sample results indicated. Any attempt to evaluate a proposal solely in terms of its components must fail, because we cannot predict the complex interacting effects if we examine only the components.

Second, the criteria for evaluating proposals are of a complex, multiple char-

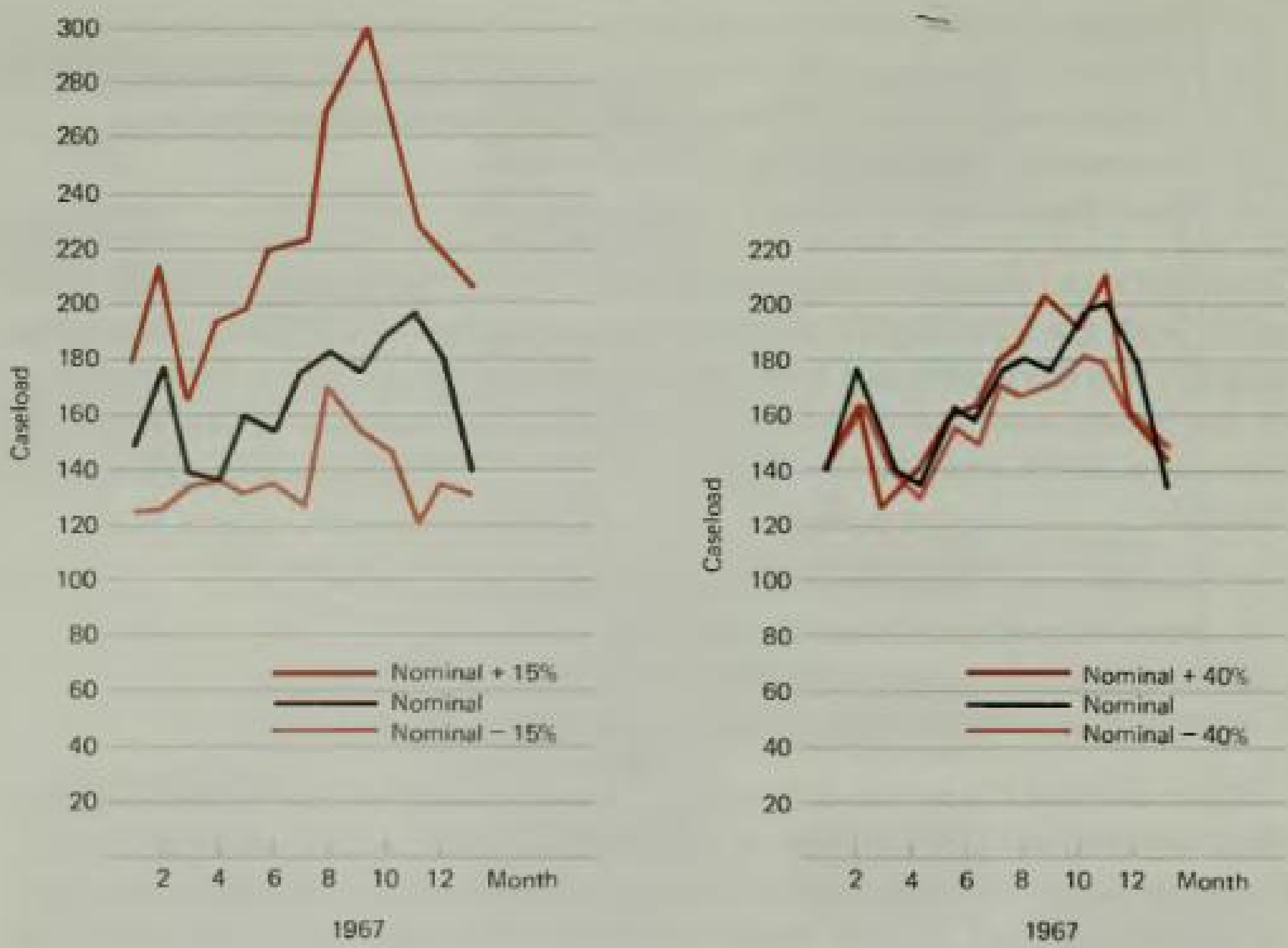


Figure 6. Simulated sensitivity of caseload to (a) estimates of pregnancy rate, and (b) migration rate

SOURCE: F. D. Kennedy, "Development of a Community Health Service System Simulation Model," *IEEE Transactions on Systems Science and Cybernetics* SSC-5, no. 3 (July 1969): 199–207.

acter. Thus, the cheapest medical system may not be the best. Even in the manufacturing system, it would be an oversimplification to conclude that the system was operating best when unit costs were lowest, since Figure 4*b* indicates rapid growth of order-delivery time as load increases. In addition, a complex interaction occurs between yield and method of processing.

Third, there is more than one way of representing and describing alternative systems. Two of the most useful ways, in our examples, are the block diagram and the computer simulation program. The block diagram representation is valuable for its graphic, common sense appeal: it helps us to visualize the complex interactions, inputs, and outputs. Although the simulation program is useful largely as a computational device, it is more important than simply this: it represents, in great detail, all the complex interactions that make the manufacturing concern or the health care unit function as a system.

HALLMARKS OF SYSTEMS

Let us abstract from the two examples the characteristics, or *hallmarks* of systems. Some of the characteristics are prominent and even obvious in the manufacturing or health care examples; however, all the characteristics are important in systems (see Litterer [1969]).

1. *Interrelatedness of components, attributes, and events.* The characteristics of any of the components, attributes, or events depend on the other existing components, attributes or events. In addition, a change in any one signifies some adjustment or change in the others. Both our examples are filled with these kinds of interrelationships. In the manufacturing example, the in-system time for an order depends on existing inventories, existing orders, yield, load, and so on. In the health care example, caseboad depends on pregnancy rates, the number of mothers' visits, infant illnesses, migration, season of the year, etc. For both systems, the effects are complex and interrelated.
2. *Wholism.* The system is an undivided whole and its performance must be viewed as an integrated system rather than as a collection of subsystems. In its functioning and behavior, the whole system is greater than the sum of its parts. One subsystem affects the others; therefore, system performance will not be represented adequately by simply summing up the performance of each subsystem.
3. *Goal seeking.* Large complex systems appear to have considerable stability. For example, our manufacturing system may produce regularly at about the same man-hour level, and aggregate output may be quite stable. Such stability, however, does not imply that other elements of the system are equally stable. They may vary greatly. As a matter of fact, because yield is highly variable, a special step in the process is designed for high-risk operations. In the health care example, quality of care is quite stable, since it is designed to meet certain standards; however, rescheduling of personnel and patients may be required to absorb caseload fluctuation. The important point is that stability is a characteristic of the whole, not of the components.
4. *Inputs and outputs.* All systems involve inputs and generate outputs. All systems produce outputs that are needed by other systems. In closed systems, the inputs are determined and new inputs do not enter the system at later points. In open systems, new inputs can be taken from the environment to influence the end result. In operations management, all systems

representative of reality are open systems. In the manufacturing example, there are several kinds of inputs, including: energy; information; various kinds of resources; and, of course, customers requesting integrated circuits. The outputs are completed products and waste products.

5. *Transformation*. All systems transform inputs into outputs, and every productive system is centered in this process. Transformation processes may be of any conceivable type that produces goods and services. In the manufacturing example, the transformation is the conversion of physical form and electrical and chemical characteristics. In the medical care example, the transformation is from the state of being ill to the state of being healthy, and from the state of being pregnant to the state of having responsibility for a child.
6. *Entropy*. Every transformation process involves the degradation or use of energy and resources. In order to keep going, a system must be open to the infusion of new energy and resources. Processes tend to run down from a state of order to disorder. With the infusion of energy and resources, however, the process can be maintained. Management has many counterparts to this conceptual framework. For example, the purpose of organization is to produce order.
7. *Regulation*. The fact that systems are goal seeking implies that they are regulated, adaptive, or even self-regulating. Regulation implies stability and the existence of mechanisms to keep variables within defined limits. A thermostat is a device to maintain temperature stability. The production system operates in a stable fashion because of some control concepts and devices. The feedback of information regarding performance is used to adjust and control performance in order to continually provide the service level desired by the health care system and the required quality and quantity of integrated circuits produced by the manufacturing system.
8. *Hierarchy*. One system may contain within it several other systems. This nesting of systems-within-systems is referred to as a hierarchy. The fact that complex systems can be "decomposed" into smaller and usually less complex systems offers one of the main avenues of analysis. A hierarchy is involved in the manufacturing system. There are various subsystems, such as the scheduling subsystem and the inventory control subsystem. In the health care system, each of the facilities represents a subsystem. Specialization of the subsystems is important to the rationale of the manufacturing and health care systems as a whole. Without specialization, the integrated system might offer poorer service at higher cost.
9. *Equifinality*. In complex systems, an initial state can have several possible final states, and the same final state may be arrived at from several initial

starting points. One result can have different causes. In the manufacturing system, a completed order may involve different yields or different resource inputs; it may come from finished goods inventory or be piggybacked on an existing order; and so on.

These hallmarks of systems represent not only a philosophy or viewpoint (interdependence, wholism, goal seeking), but also a basis for modeling systems (inputs and outputs, transformation, regulation, hierarchy, differentiation, and equifinality). Both groups of characteristics are important in operations management. If we are to design and manage productive systems effectively, we must take care that we are designing and managing the whole with certain goals and objectives in mind. If we do this, the components and subsystems will fall into place, taking their individual objectives from the broader ones.

In terms of general systems theory classification, there are two basic types of systems: open and closed. We alluded earlier to this dichotomy, and we noted that closed systems are of little interest in operations management. In a closed system, there is no interchange with the environment once the process starts. No materials, energy, or information cross the system boundaries during the transformation process. Closed systems have applications in thermodynamics and perhaps elsewhere, but productive systems inevitably have various flows that cross the system boundaries, often continuously, during the transformation process. These flows include people, energy, materials, money, information, and other productive resources that are necessary to keep the process going. Therefore, we will focus only on open systems and, from this point on, any systems we discuss will be assumed to be open systems.

THE INPUT-TRANSFORMATION-OUTPUT MODULE

The hallmarks of systems allows us to derive a fundamental concept that is the basis for much of the methodological framework of systems analysis: the input-transformation-output module. The most common mode of system identification and description, the block diagram, is built from component modules. Each block signifies an elemental transformation process.

The most rigorous situation is illustrated by Figure 7*a*, where the input and output are related mathematically by the transfer function *K*. The transfer function is simply the ratio of output to input. The relationship between output and input could be simple or very complex. The block is analogous to an opaque "black box"; we may or may not entirely understand what goes on inside the box, but we can observe the input and the output and infer the transformation.

Figures 7*b* and 7*c* show that we need not be mathematically rigorous in order

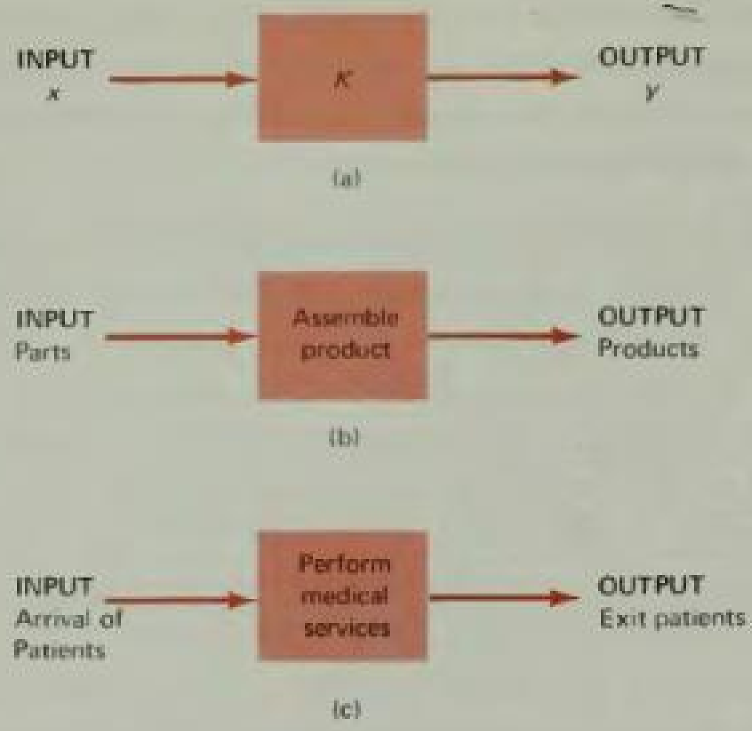


Figure 7. Input-transformation-output module. (a) For a continuous operation, the input x is converted to the output y by the transfer function k. (b) The process is the assembly of a product, grouping all the details of assembly, with parts as inputs and assembled products as outputs. (c) The process is medical services, with arriving patients as inputs and treated patients as outputs.

to describe the transformation—we can simply describe the process with its inputs and outputs. Therefore, the transformation could be simply a description; for example, assemble product, write purchase order, perform medical tests, or cook hamburger.

The nature of the transformation in the module can involve any level of detail we choose. The module represents some grouping of activities, depending on our needs or purpose. Thus, if we are studying the flow of work in a factory, each process might be represented by a block. In representing a system for scheduling production to departments within the plant, however, the individual processes might be subsumed within a department module representing all processes in the department. If we were dealing with a multiplant allocation of output to plants, the departments might be aggregated within a single block representing each plant. The block diagraming methods that use the module as an element are flexible and, by defining the level of aggregation, they accommodate the existing hierarchy of systems.

ADAPTIVE SYSTEMS

Some of the most interesting systems are adaptive: they react to what is happening in a way that contributes to the continued operation of the system. The general concepts used in systems theory derive from and are parallel to the role of *homeostasis in biological systems*. Homeostasis is the property of living systems that maintains essential variables within physiological limits [Ashby, 1969]. Cold produces shivering, which in turn helps raise body temperature. Heat causes the capillaries to expand, bringing blood flow to the surface of the skin where heat can be exchanged with the environment; and heat also produces sweating, evaporation, and cooling. There are hundreds of examples of homeostasis in natural systems involving physical and chemical responses that have the effect of *keeping essential variables within physiological limits*.

The counterpart of homeostasis in human organizations is illustrated in the entire field of managerial planning and control. Through the plans that are developed, planning and control systems set the limits of variations for variables, essential to the continued healthy operation of the social system involved. When variables go outside these limits, someone or some system takes action to correct the condition. As we know, managers often will try to make these control systems function automatically. This enables the managers to give their attention to other important problems. The computer and present day management information systems can help in this process. MBE IMS

Only recently have the general concepts of homeostasis and adaptive behavior been applied to the design of machines with any sophistication. Yet, as we know today, there are automatic and semiautomatic machines. Since the machines are vastly simpler than biological systems, the cause and effect mechanisms for achieving adaptive behavior are fairly comprehensible. We will use these mechanisms to help explain adaptive system behavior, stability, goal seeking, and feedback. Then we will use the resulting concepts to understand the behavior of productive systems.

Negative Feedback and Goal Seeking

The common examples of feedback control are ~~the~~ 1 Watts governor and the 2 thermostat. Each reduces the feedback and control processes to their bare essentials. The elements of a basic feedback control loop are shown in Figure 8, in which we assume that the productive system itself has inputs that it processes to produce outputs of goods or services. Conceptually, the feedback loop is

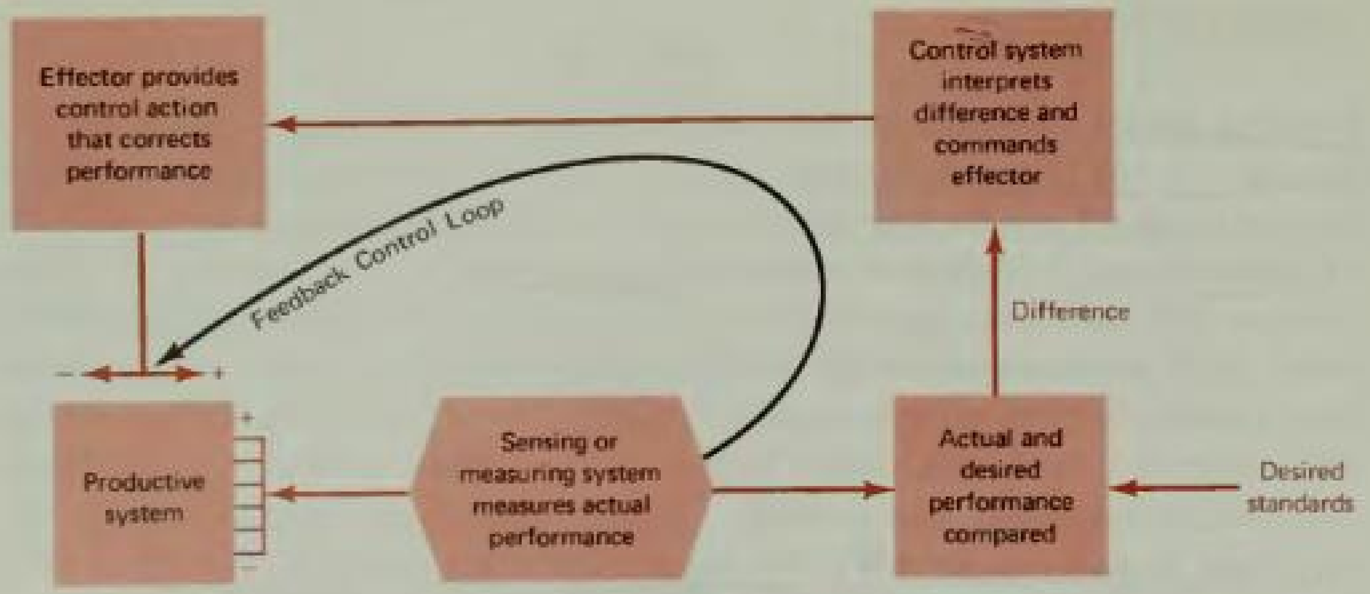


Figure 8. Elements of simple feedback-control loop

comprised of: a *sensing unit*, that measures the output of the variable being controlled; a *comparator*, that compares the actual output with the desired level; a *decision maker*, that interprets the error information and gives commands; and an *effector*, that carries out commands to make a correction in the proper magnitude and direction so that output will meet standards. Every feedback control loop involves these four functions in some way. Although they may not be obvious in terms of separate units or mechanisms, the functions are being performed.

The word *negative* is used because the connections between sensing, comparison, decision making, and effecting are made in such a way that the effector causes a change in a negative direction compared to the output difference.

In order to control the performance of a system, as in Figure 8, measurements are made on aspects of the output, perhaps in both quantity and quality terms. These measurements are compared with the desired performance, and the control system interprets the difference, commanding the effector to correct performance. Thus, a control system always involves some kind of measurement of what is actually happening. This information is fed into some kind of data-processing system, where the measurements are compared with standards that have been derived. Directly or indirectly, these standards are based on managerial policies.

For example, a bank may want to offer customer service at the windows so that the average waiting time is only one minute. If the waiting time exceeds the one-minute standard by perhaps more than 10 percent, then additional tellers' windows are opened. Alternatively, the control system might consider a physical measurement, such as a quality control on the dimension of a part or the chemical composition; or the rate of output might be the measurement to be controlled. In some instances, feedback controls may require high technology and be virtually automatic; in other instances, the control may require managerial intervention. Whether they are automatic or not, however, the same basic principles of feedback control are being used.

Combinations of Simple Feedback Systems. In Figure 9*a*, we represent the simple subsystem module of an input-output transformation process, involving information feedback to provide control of the process. These simple modules

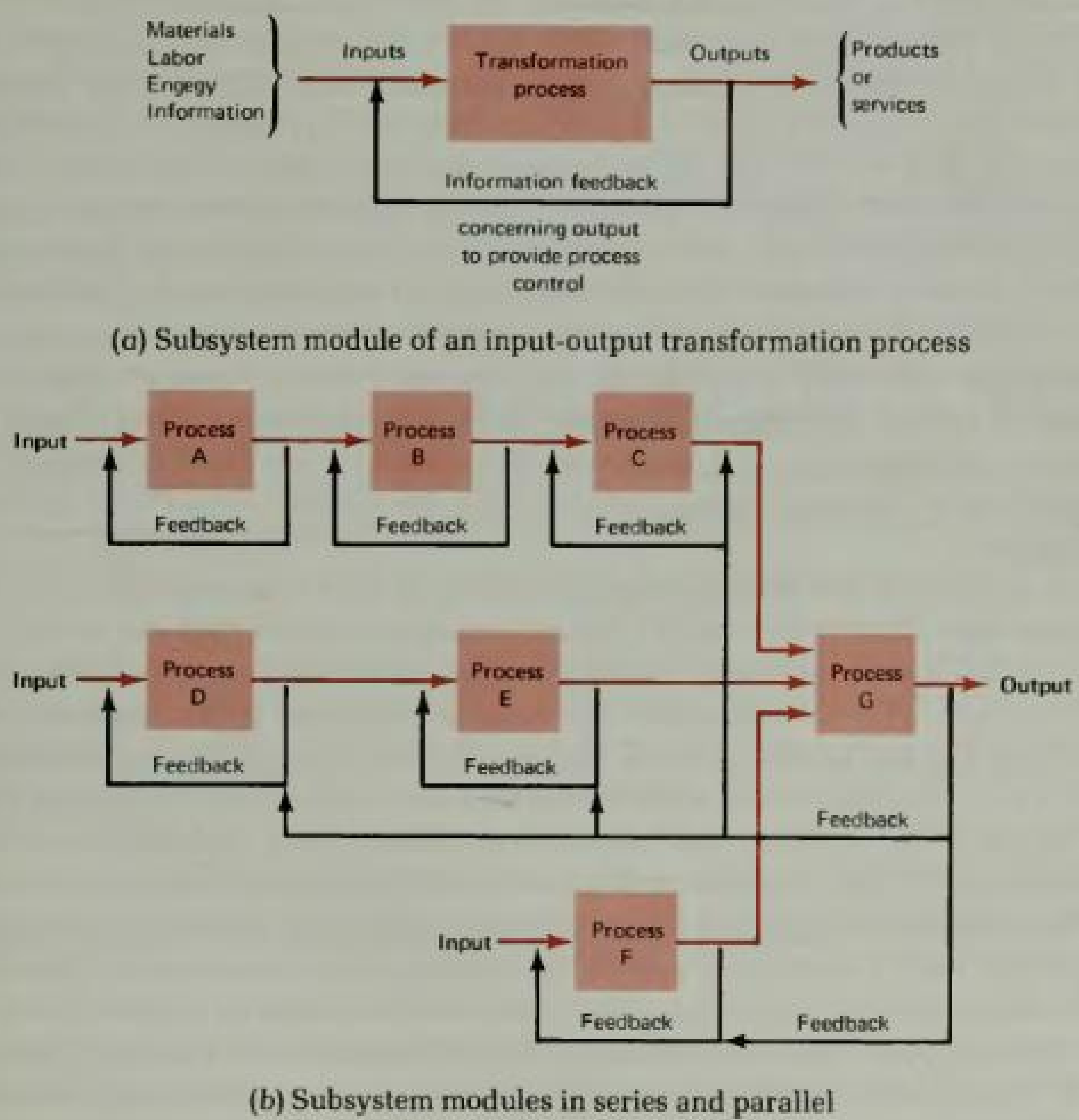


Figure 9. Input-output concept of systems and subsystems with information feedback to provide the basis for control

may be combined in series and in parallel, as shown in Figure 9b. Here, we see that each process has its own feedback loop for control. In addition, however, the final output of the system of processes also provides feedback for control purposes to the critical processes upstream. These elements of managerial control might be in terms of the quality of output, the quantity of output, or some other aspect of performance.

Feedback–Time Lag Effects. How a system under control by information feedback responds depends on a number of factors. Referring back to Figure 8, assume that: the difference between actual and desired performance is fed back with no time lag, the effector commands the system to change settings immediately by the observed difference, and the system responds immediately. Then the system would be completely responsive and presumably exhibit perfect control (since any deviation is corrected immediately). Even in completely automatic, high technology, process control systems, however, this ideal is not attained. How the system responds to control depends on: time lags in the feedback-response system and in the transformation process itself; how much control action is called forth; and the sensitivity of the system under control.

The time lags inherent in the feedback control system, and in the process itself, are an extremely important factor affecting how the system output responds to control. If it takes substantial time for the feedback control system to respond and command an adjustment, and if the process itself is sluggish in responding to changed control settings, then the system output may exhibit oscillation.

For example, if you were attempting to drive an auto and you could only see the rear view (through the mirror), the information feedback that you needed in order to make a turn would occur after you had passed the curve. To adjust, you might then turn the steering wheel. But suppose that it was a winding road; and that, once you had turned, you saw that the road subsequently turned the other way. You would need a very wide road to drive under these conditions and stay on the road. In any case, you would be oscillating back and forth over the smooth line that you would drive if you could see ahead through the windshield.

The sensitivity of the system refers to the response in the output to a change in the input. For example, in the steering mechanisms of most autos the ratio between the amount of rotation of the front wheels and a given degree of turn of the steering wheel is modest. Therefore, the driver does not oversteer (overcontrol). The opposite situation has been experienced by most of us in the "bumper car" autos in amusement parks. In these cars, the steering ratio has been deliberately set to cause the front wheels to turn considerably when the steering wheel is turned only slightly. The result is a loss of control (as well as some fun). Thus, a system may be relatively sensitive or insensitive to control. How much control action is called for is related to the sensitivity of the system.

Feedback and Managerial Systems. Managerial control systems suffer from all the unfortunate effects: time lags, sensitivity, and amount of control action. In general, long time lags and low sensitivity are the big culprits. However, there are situations where supersensitivity is difficult to deal with. For example, when we study service systems, that involve the theory of waiting lines, we will find that the number of people waiting in line at the check-out counter of a supermarket is very sensitive to utilization factors above 0.75 (ratio of arriving rate of customers to the rate at which they can be processed).

Consider the operations manager attempting to control the employment, production, and inventory levels of operations, using month-old information feedback on the progress of retail sales. If the manager reacts to an increase in retail sales recorded one month ago, he/she might increase employment and production levels at a time when sales in the field actually were decreasing. Also, retail sales might exhibit random fluctuations. If the manager were to respond to the retail sales increase (whether the time-lag were long or short), he/she might respond by making small increases in employment and production level or by taking a wait-and-see attitude (to learn whether the increase is permanent or only represents a random change in sales level).

Positive Feedback

If the Watts governor somehow were assembled backwards—if, as the engine speeded up, the effect was ~~to open~~ the throttle rather than to close it—this would be called *positive feedback*. Its effect would be to generate growth rather than to control it. In the case of an engine, increases in speed would cause even greater increases in throttle opening, and so on in a vicious cycle—vicious because ultimately the engine would disintegrate. Thus, the difference between positive and negative feedback is basically in the *polarity* of the connections. Although this is a conceptually simple difference, it creates fantastic differences in performance. In positive feedback situations, action builds a result that generates still greater action; and under certain circumstances, this may be desirable. The concept of compound interest is an example. Interest generated in the current period becomes a part of capital, which generates greater interest in subsequent periods. Also, if we were attempting to develop the sales of a product, we might include some positive feedback loops in our system design.

Stability of Systems

Negative feedback systems exhibit stability and, as we have just noted, positive feedback systems exhibit growth or instability. It is important to note that "a system's stability (or instability) is a property of the whole system and can be assigned to no part of it" [Ashby, 1969]. Some intriguing examples of this concept follow:

1. Two systems may be joined so that they act on and interact with one another to form a single system. To know that the two systems, when separate, were both stable is to know nothing about the stability of the system formed by their junction; it may be stable or unstable.
2. Two systems, both unstable, may join to form a whole that is stable.
3. Two systems may form a stable whole if joined in one way and may form an unstable whole if joined in another way.
4. In a stable system, fixing a variable may render the remainder unstable.

The four preceding examples may indicate some of the reasons why managers may be startled by the effects of individual actions taken. The presence of system stability implies some coordination of the actions among the component elements. Indeed, an important function of management in a productive system is to coordinate a complex of people, materials, and other resources to achieve some goal. Whether that system will operate in a stable fashion or not depends on the way in which the parts are coordinated.

REGULATION AND CONTROL OF COMPLEX SYSTEMS

The negative feedback systems that we have discussed exist at a basic level. They react to their own performance and, through an error correction process, they maintain the system variables within limits. They are "statemaintaining" systems [Ackoff, 1971]. The functioning of the simple feedback systems are important both in concept and in application to machines and some machinelike systems. While more complex systems can be constructed by assembling a network of components made up of processes with and without feedback, it is wrong to assume that the complex managerial systems with which we deal in operations management are merely an assembly of these simple building blocks. It would be even more of a mistake to assume that the regulation and control of the large system is conceptually just a complex error correction machine that operates like a simple device and is only an assembly of the simple devices whose interaction is complex.

There are different levels of feedback, and we must understand the distinctions between the levels in order to understand the regulation and control of complex systems. Feedback control systems that are typified by the thermostat are called first-order feedback systems. They feed back the actual performance of the system in question, by measuring the errors and connecting the linkages so that control action is opposite in sign to the error. These systems hone in on a fixed goal. The mechanisms for achieving control may be complicated in some sense, but the conceptual framework for control is simple and direct. Second- and third-order feedback systems are more sophisticated, and may have more lofty goals.

Second-Order Feedback Systems

The basic difference between first- and second-order feedback is that the latter has a memory. This capability enables the system to change response patterns automatically, depending on specific, preprogrammed conditions. Second-order feedback systems can anticipate and seek goals [Ackoff, 1971].

Perhaps the simplest and most comprehensible example of second-order feedback control is the clock-controlled thermostat, that has different daytime and nighttime settings for home heating. The settings are stored, usually mechanically, and selected at the appropriate times by the clock mechanisms. An even more sophisticated system might include an outside thermometer and wind-speed indicator. The additional information from these sensors enables

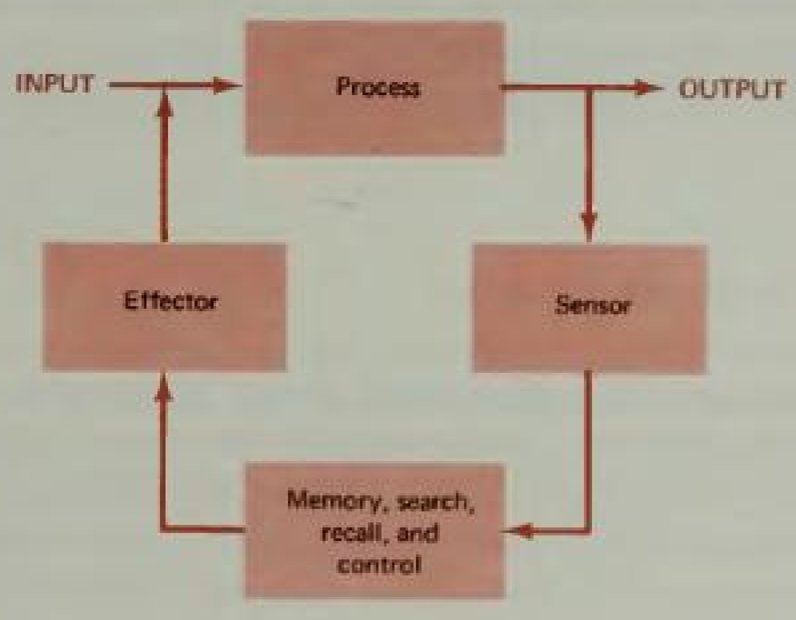


Figure 10. Second-order feedback control, including memory, to facilitate automatic goal-changing and anticipation

the heating system to improve its operation by anticipating future heating requirements [Hare, 1967, p. 113]. Figure 10 schematically shows the addition of the memory control device to the feedback circuit.

In scheduling production, for example, suppose that, from the records we keep, we determine that product demand follows a seasonal pattern, with variations from period to period around the basic seasonal level. Now, we can put the seasonal pattern in the memory of our system and recall the appropriate index when calculating the production to be scheduled in each period. In this way, we would be anticipating the seasonal variation. Or, carrying the idea a step further, we could store in the memory the cost relationships for: production at regular and overtime; inventory cost; and the cost of expanding or contracting the labor force. Now we would be asking our feedback decision process not only to react to the current conditions and the estimate of new demand adjusted seasonally but also to compute production to be scheduled in such a way as to minimize these relevant costs. We could continue expanding this particular system by including more and more relevant information in the memory.

Refining the process with second-order feedback further reduces the error between results and the objective (or goal) of the system. This reduction is caused by automatically changing subgoals, depending on the conditions. The second-order system handles a wider variety of conditions, but it is still at the preprogrammed decision rule level. Conceptually, this system says: if condition 2 is present, use plan B; but if condition 3 is present, use plan C. The second-order feedback system's *memory* is what enables it to select the contingency plan that matches certain input conditions. While it might seem, at this point, that we are talking about computing systems, this is not necessarily so. The system might involve people for implementation; the memory could be a human memory. Operations management is filled with examples of second-order feedback control systems in scheduling operations, inventory control, quality control, and so on.

Third-Order Feedback Systems

"If an organization can collect information, store it in a memory, and then reflect upon or examine the contents of the memory for the purpose of formulating new courses of action, it will have reached a new level of autonomy. The mechanism that considers various goals and courses of action can be called the *consciousness* of the organization. Such systems are purposeful [Ackoff, 1971]. Reflective decision making takes place in such third-order feedback systems. The action of the organization begins to approach what we would expect of an actual industrial or human organization.

Thus, while first- and second-order feedback systems deal with the error correction concepts related to the process itself, third-order systems deal with broader system control and longer-term optimization. In short, third-order feedback loops in productive systems commonly have sensors, comparators, decision makers, and effectors that are not machines but human beings—i.e., managers. While it is true that the well trained manager uses computers, management information systems, decision models, and every conceivable aid available to increase effectiveness, he/she is still the "reflective goal changing unit" in the system.

Thus, it is by means of third-order feedback controls that we cope with larger systems. Of course, in the process of managing, managers do everything they can to reduce areas of control to second- and even first-order automatic systems so that they can concentrate their attention on the broader problems that, to date, have not been reduced to automatic control. The methodology of this reduction is arrived at through the broad methods of management science. At this point, it is academic to speculate on the final "takeover" by computers of reflective goal changing activities that require consciousness now available only to individuals and organizations of people.

SYSTEMS METHODS

The methods of systems analysis are not simply those of management science and operations research, although these general methods have an important function in systems methods. The methods associated with systems analysis are those of problem definition in systems terms, systems description and representation, and prediction of systems performance.

Problem Definition

Earlier in this chapter, we stated that a problem initially must be defined in systematic terms if solutions that deal effectively with the problem are to be designed. We further said that the starting point should be the system as a whole. If the problem is defined narrowly from the outset, the system designed to solve the problem will reflect this narrowness and very likely be inadequate.

Chestnut [1967] puts the process of problem definition in system terms, using the block diagram, as shown in Figure 11. The process is an interactive one, in which feedback of information about the projected (or existing) system affects not only how the system is structured, but also how the problem itself is

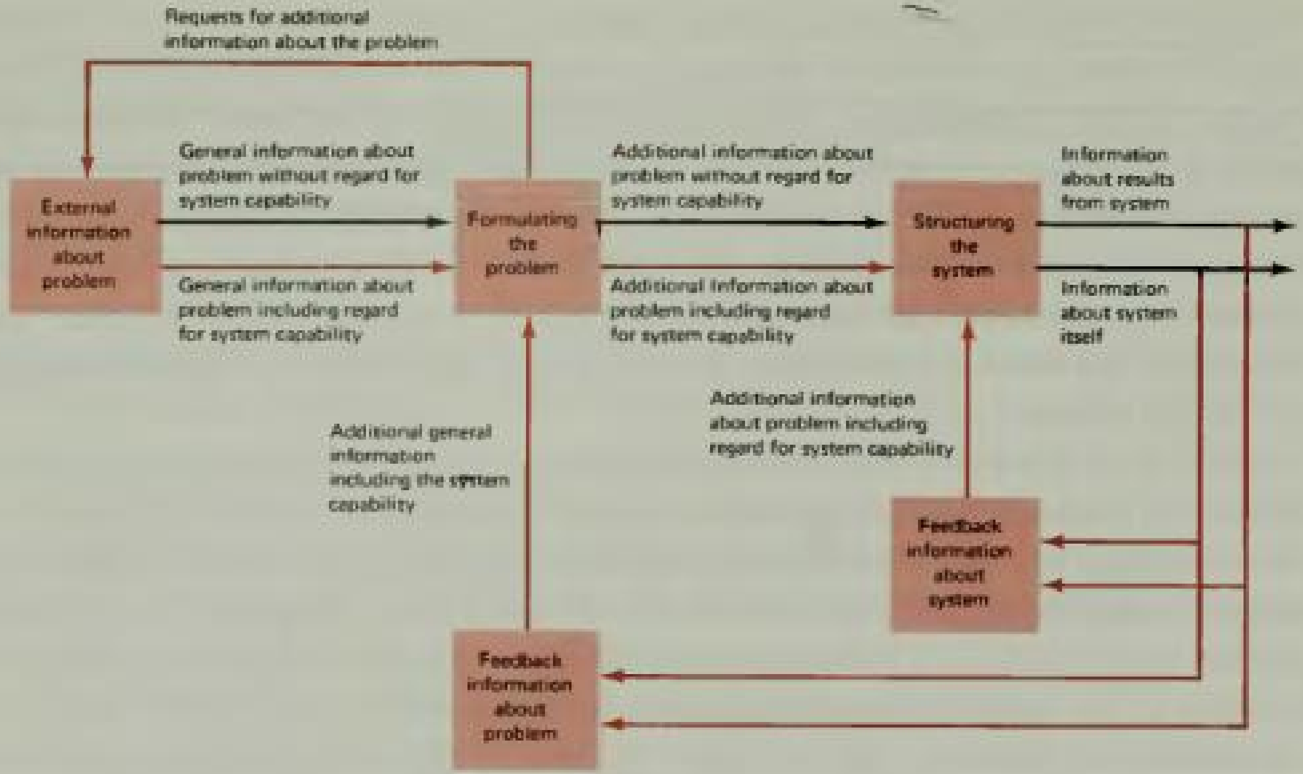


Figure 11. Information flow diagram showing relationships among formulating, structuring, and other aspects of the systems problem

SOURCE: H. Chestnut, *System Engineering Methods* (New York: John Wiley & Sons, 1967), Figure 1.4–1, p. 19

formulated. Furthermore, there is an information feedback loop between the external information block and problem formulation. Therefore, systems methods show us that problem definition is a dynamic process rather than a static one. The defining of the problem itself must remain open to new information that might require making changes. Thus, problem definition and formulation overlaps with systems development, as indicated by the information feedback process of Figure 11. The more complex the problem and the resulting system, the more likely that, as the process develops, problem definition will need to be modified.

Suboptimization. Maintaining the systems point of view helps to minimize the effects of suboptimization. Suboptimization occurs when we choose the solution that appears best when viewed from within the subsystem, but not when viewed from within the system as a whole. Examples of suboptimization within productive systems are legion. We could easily suboptimize by taking a short-range view of maximizing profits, or an organizational, narrow gauge view. A production manager with a short-range view might produce according to the rise and fall of the sales curve. But a manager with a broader horizon would look for trends and seasonal variations, and would develop a production program to meet the specific requirements. In stabilizing production levels, the latter view might create higher total inventory costs but substantially lower hiring and layoff costs. The short-range view would be a suboptimization, since it would focus exclusively on the payroll cost but ignore the longer time effects and costs of production fluctuation and inventories.

To take a service system example, the post office might be urged to install letter sorting machines for all sorting of first class mail, because they would operate faster and seemingly create a lower labor cost per thousand letters. However, if the problem definition is enlarged to include the entire system of sorting for all kinds of first class mail (which have varying arrival patterns), a combination of manual and machine sorting appears to offer a superior allocation of resources.

Organizational suboptimization is common where the production function and the distribution function are operated essentially as two separate enterprises. Using this approach, the factory will try to minimize its costs independently, as will the sales and distribution function. Thus, sales and distribution will be faced mainly with an inventory management, shipping, and customer service problem, and will try to minimize the associated costs. On the other hand, the factory will be faced with minimizing production cost.

Each suborganization optimizing separately probably will result in a combined cost that is somewhat larger than it would be if the attempt were made to optimize the combined system as a whole. The reasons are fairly obvious since, in minimizing the costs of inventories, the sales function transmits most of the effects of sales fluctuation directly to the factory, instead of absorbing these fluctuations through buffer inventories. Suboptimization is the result. By coordinating their plans, however, the various suborganizations may achieve some balance between inventory costs and the costs of production fluctuation.

System Boundaries. Systems involve transformation with inputs and outputs. The concept of the system boundary is an effective tool for problem definition. In the input-transformation-output modules represented in Figure 7, the blocks themselves represent boundaries around a process. The boundary drawn around the process facilitates a precise definition of what goes on inside the block. We can see what happens by noting the input, the output, and the composite transfer function. The boundary around the process means that the performance of the simple system is described completely by the input and the transfer function or process, or by the input and the output. If there are any other possible inputs to the system, they are assumed to have no effect. Thus, the system boundary highlights the inputs and outputs. Anything not explicitly

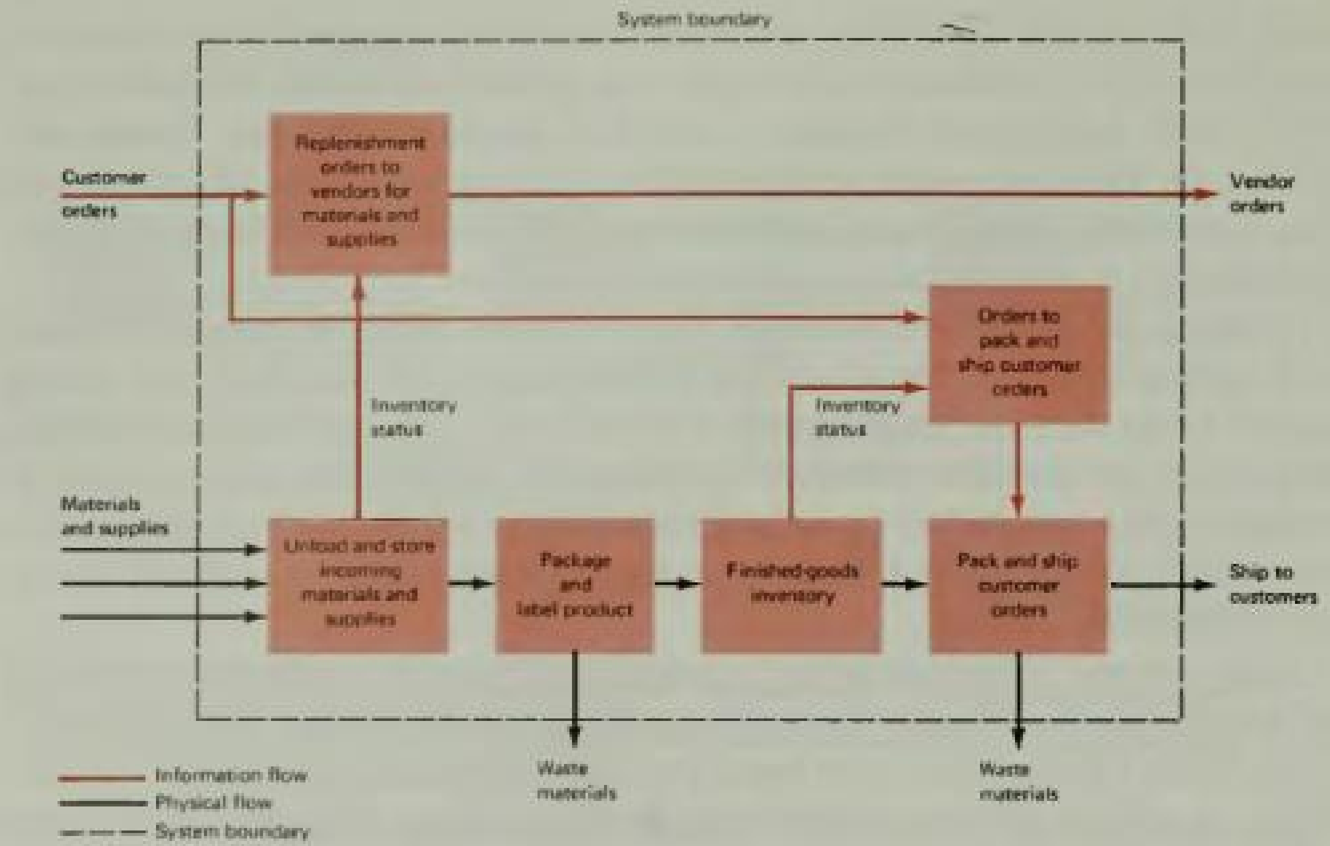


Figure 12. Information and physical flow in relation to system boundary for a one product, mail order operation

shown as an input or an output is not part of the problem represented. Any environmental effects either are part of the input or are assumed to have no effect.

Let us take a simple, one product, mail order operation as an example. The enterprise does no manufacturing. Rather, it takes a bulk material and packages, labels, and ships it to customers on mail order. For our purposes, the decisions involved directly with operations are simple, concerning only the ordering of materials and supplies to sustain the customer order levels. Figure 12 shows the simple system with physical inputs of materials and supplies, and outputs of shipments plus waste materials. There are information inputs of customer orders, and outputs of orders to vendors for new materials and supplies, based on incoming order rates and inventory status. The system boundary defines what is inside and outside the system. It also helps us account for what happens to material inputs, by means of the physical law of conservation of matter: what comes in either goes out or is stored in some form. In our simple system, incoming materials and supplies may be stored temporarily as materials and supplies or as finished goods, and may leave the system as customer orders or as waste materials.

The system boundary, however, does not necessarily bound the enterprise system, as Figure 12 might suggest. The boundary reflects the problem under consideration. For example, if we were concerned with overall delivery time to customers—from the time the firm received and ordered the product until the customer actually received it—the system boundary would have to include the transportation and delivery systems and possibly the vendor systems. Under such circumstances, these "external" systems become part of the system under study. Of course, other focal interests change the boundaries. For example, the "package and label" operation could be subjected to analysis of the processing methods used. Thus, we see that even in small, simple systems, there is a hierarchy of systems, each with its own inputs and outputs which ties it in with other systems. *The output of every system is an input to another system*. The boundary of the system under study is a key to defining that system.

As a practical matter, analysts do not necessarily always draw in the boundaries as we have in Figure 12. Nevertheless, they think in these terms when defining system problems. They may accomplish the equivalent by defining relevant variables. The specification of boundaries, however, is graphic and helps to indicate the intended demarcation between the system and its environment, showing clearly where in the system an interchange with the environment exists.

Systems Description and Representation

Some of the most important systems methodology occurs in the ways systems are described or represented. These techniques help us elaborate the problem itself, as well as the complex interactions among system components. At the broadest and, usually, most highly aggregated level of description, we have diagrams and charts that often convey general impressions. They have initial value in describing a system, but normally they do not transmit enough information to be effective in problem solving.

The tools of system methodology, however, are block diagrams and network diagrams. They are all models of the system; they all describe the system in some way. We use these mechanisms to methodically set down the step-by-step transformations that take place (or would take place) in designing a system. Each step has inputs and outputs, and by defining the steps together with the inputs and outputs, we define the interrelationships.

Block Diagrams. The block diagram is built up from component input–transformation-output modules. Each block has the significance of an elemental transformation process. The most rigorous situation is illustrated by Figure 7*a*, where

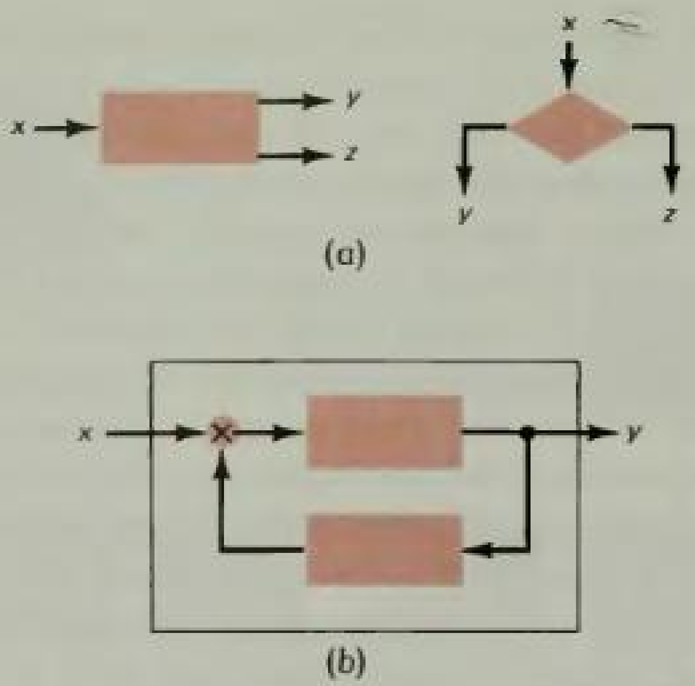


Figure 13. Building blocks used to construct complex block diagrams. (a) Two forms of the decision block, and (b) the feedback block. The third common element is the transformation block shown in Figure 7 which converts an input x to an output y, according to the rules specified by the transfer function

the input and output are related mathematically by the transfer function K. The transfer function is simply the ratio of output to input. The relationship between output and input could be simple or very complex.

There are three kinds of elementary blocks which, when used in combination can describe very complex systems. First, there is the transformation block referred to in Figure 7. Figure 13*a* is the decision block used to achieve branching, depending on decision rules. Either form of the decision block is satisfactory; it is easier to write descriptions in the rectangle, however, the diamond suggests alternatives more graphically. When we come out of decision blocks, we can loop back to previous blocks in the sequence, or branch further, in the form of a decision tree. The concept of the feedback block shown in Figure 13b is identical to that discussed earlier, and is used most commonly in describing the control functions of productive systems.

Network Diagrams. An alternative form of system description is the network diagram. The block diagram and the network diagram are equivalent in terms of components and relationships represented. In network analysis, the transfer functions are represented by arrows, and the system variables are the nodes (usually represented by small circles). Network diagrams are simpler to use when detailed algebraic analysis is required. Later in the book, we will discuss an important application of network analysis: the PERT model for planning, scheduling, and controlling large projects.

METHODS OF PREDICTING PERFORMANCE

One of the most important functions of models is to predict the behavior or performance of a system. In operations management, analytical and simulation models are the most useful.

When a mathematical model is used to predict performance of a complex system, we assume that we know the mathematical transfer functions for each of the blocks, and that we are able to combine them in a mathematically tractable way. An excellent example of predicting large-scale system performance by a mathematical model was developed in the analysis of the criminal justice system (see Blumstein and Larson [1969]).

The complexity of many systems, together with the need to connect very different methods, as well as empirical data, within the same system mathematically, often makes it either impossible or impractical to predict the performance through analytical models. As a result, simulation is the common system methodology for performance prediction. Simulation was used earlier in this chapter to predict performance for both the manufacturing- and health-care system examples.

In Chapter 4, we will survey those analytical and simulation methods that are most useful in operations management, and provide a mapping for relating types of problems and methods.

Structural Models of Complex Systems

The process of isolating feedback loops in a large system often is difficult because a loop may be completed through a complex sequence of variables acting on one another, rather than through the more simple, direct systems we have discussed so far. One technique for developing the structure is the signed digraph (directed graph).

As an example, let us develop a signed digraph for a set of variables involving employment stabilization for a small applicance manufacturer. The manager has been concerned about the seasonal hiring and layoff of employees that has resulted from seasonal demand. First, let us write down the main variables that seem important: (*D*)emand, (*P*)roduction level, (*L*)abor productivity, (*W*)ork-force size, (*I*)nventory level, and (*C*)ost of product per unit. There may be other important variables to add later, but for now let us see what broad structure or interaction results from this set of variables.

We will represent the variables as the nodes of a network (see Figure 14). Starting with demand, we ask the question, "What is the effect of an increase in demand (*D*) on production level (*P*), other factors remaining equal?" We are concerened only with whether the impact is plus or minus. In this case, we mark the *DP* arc with a plus sign, indicating that the effect of *D* on *P* is positive. Similarly, the *PW* arc is positive.

The other arcs shown in Figure 14 have been developed similarly, though it is not always clear that the arc should be labeled in this way. For example, we have marked the *LW* arc as minus, meaning that an increase in labor productivity will result in a decrease in work force size, and this may depend on whether or not work rules permit layoffs.

The markings of arcs may, in some instances, be controversial, and may depend on factors or assumptions that must be made explicit. Thus, alternate structures may result, depending on the particular conditions and assumptions.

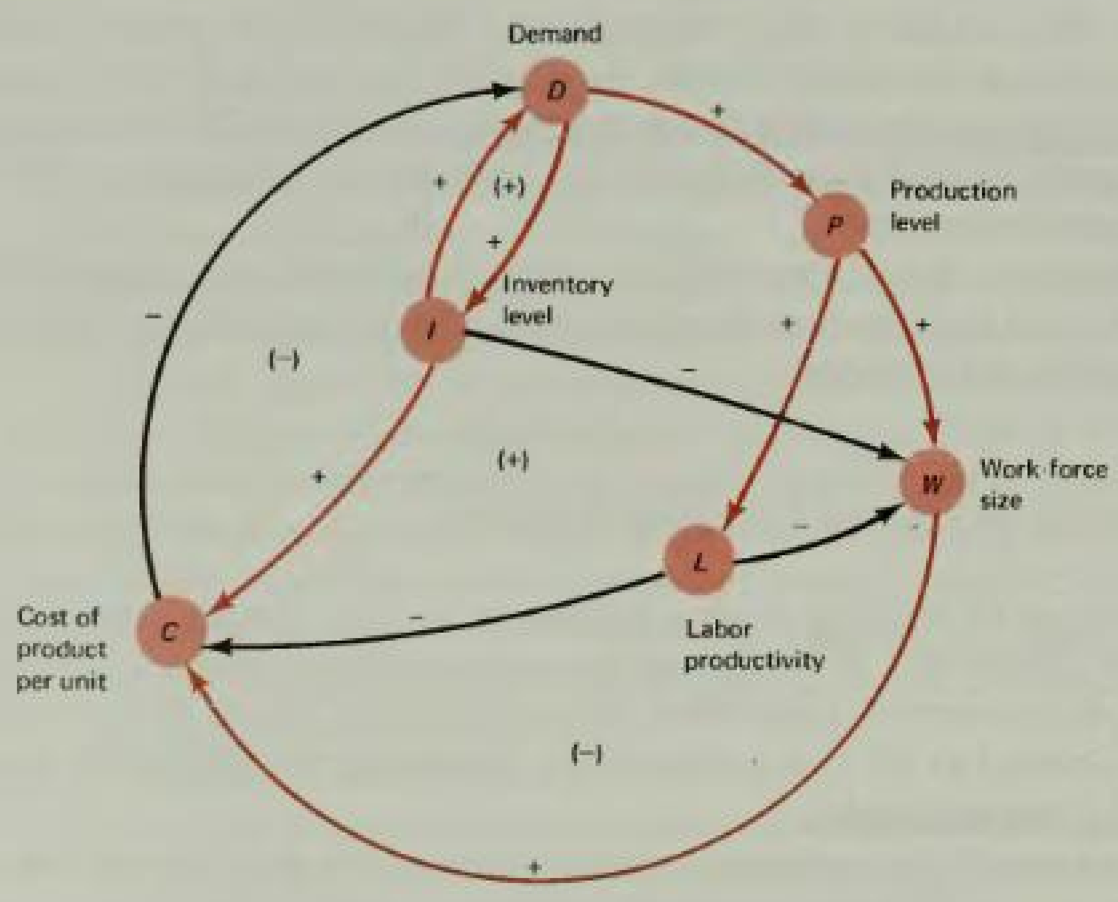


Figure 14. Signed digraph for the variable demand (*D*), production level (*P*), work force size (*W*), labor productivity (*L*), inventory level (*I*), and cost per unit (*C*)

Nevertheless, let us assume that, for our situation, the marking of arcs represents a concensus.

Identify Feedback Loops. Now, are there closed loops in Figure 14? Yes, we can identify four: *DPWC*, *DPLC*, *DIC*, and *DI*. Are these closed loops positive or negative in their net feedback effects? We can tell simply by observing whether or not there is an even or odd number of arcs marked with minus signs in the loop. An odd number of negative arcs identifies a negative feedback loop. Consequently, *DPWC* is negative, *DPLC* is positive, *DIC* is negative, and *DI* is positive. The signs of these feedback loops are marked with plus or minus signs in parentheses in Figure 14. The basic loop involving *W* is negative; that is, it is stabilizing. However, this does not mean that it is stable; we already know that it oscillates, but the limits of oscillation are regarded as too great.

Note that the *DPLC* loop is positive. This, of course, is one of the kinds of growth loops that business people have emphasized since Adam Smith's era—building market size or demand partially through lower costs that result from productivity increases.

Simulating Dynamic Effects. Given the signed digraph as a structural model, we may wish to determine the effects of changes in the variables over a period of time. Two simulation systems have been developed that produce dynamic output for structural models. The first was developed by Forrester [1961] and is commonly called systems dynamics. Systems dynamics methods develop functional equations that relate the variables, and the variables take on values computed at each point in time. The well known popularized study, *The Limits to Growth*, was based on a world model developed by Forrester [1972] within the framework of systems dynamics.

The second, and more recent, simulation system is called KSIM, and was developed by Kane, Vertinsky, and Thompson [1973]. It depends on arc ratings to develop a cross-impact matrix. The cross-impact effects are computed as bounded variables between zero and one, and dynamic effects then are printed. A summary of methods for dynamic structural models is contained in Buffa and Dyer [1977].

THE SYSTEMS APPROACH

Let us abstract the meaning of the "systems approach to problems" from the concepts we already have introduced. First, we presented the ideas involved in systems analysis in terms of the manufacturing and health care examples.

There, we emphasized that the problem was defined as the system as a whole, rather than simply as an assembly of the pertinent subsystems.

In representing the system as a whole, certain techniques were effective in showing the interactions among all the subsystems. These were block diagrams, network representations, and computer simulation programs. Indeed, representing these complex interactions is absolutely crucial if we are to evaluate the effectiveness of the system accurately. In determining effectiveness, it is impossible to evaluate each of the subsystems independently, because such a technique could not represent the interactions among subsystems.

In evaluating the manufacturing system, for example, there were many dimensions of performance: the number of orders waiting to be processed at each facility; in-process and finished-goods inventory levels; production costs; yields; and so on. In the health care system, the dimensions of performance were: personnel utilization; abnormality ratios compared to national rates; infant-illness ratio in relation to national rates; sensitivity of caseload to pregnancy and migration rates; and so on. Thus, the systems approach recognizes the need for careful compromises and trade-offs among competing criteria and values. Within this framework, the evaluation includes the concept of optimization. But the optimization is of the multiple-criterion function, and currently must be accomplished by methods other than mathematical. Finally, the systems approach commonly uses simulation and analytical models to predict performance before money is spent to actually create or alter the system. In summary, the systems approach to problems includes:

1. Originally defining the problem in systemic terms.
2. Taking the system as a whole as the starting point.
3. Representing the complex interactions among the components.
4. Recognizing the need for compromise and trade-off among competing criteria and values; attempting to optimize jointly.
5. Predicting performance of the system as a whole before creation or alteration of the system.

Review Questions

1. Explain the significance of the statement, "The whole adds up to something more than the simple sum of its parts." Give examples of situations described by the statement.

2. Based on the results of the simulation of the manufacture of integrated circuits, can you identify possible bottlenecks in the system?

3. Of what value might an output report (similar to Figure 3) for the integrated circuit manufacturer be to a manager?

4. Characterize the response of the manufacturing system described in the text to both normal and heavy loads. What other measures of performance would be useful to a manager, that might suggest how to maintain good performance under heavy loads?

5. Based on the forecasting equation for the Maternal Care Clinic, how many visits would be expected by a mother who was six weeks pregnant at her first visit, had a gestation period of 38 weeks, and an annual income of $4,000?

6. Using the health services system as an example:
   a. In what ways are some system components, attributes, and events dependent on other system components, attributes, and events?
   b. What are the most important inputs and outputs?
   c. What kinds of transformation take place in the system?
   d. Is there a hierarchy of systems within the system as a whole? Describe.
   e. Are there feedback loops within the system?

7. Suppose that we are dealing with an input-transformation-output module of the Figure 7*a* type, where the input is 24 units and the output is 18 units. What is the value of the transfer function?

8. Following are descriptions of inputs and outputs. Based on the descriptions, what transformation is implied? (Input is described first.)
   a. Machine breaks down. Machine is running again.
   b. Students form a line at the administration building. Students leave the administration minus $500 to $1,500 in cash reserves.
   c. Woman enters hospital. Woman leaves hospital with a crying bundle that makes water.
   d. Data concerning revenues and costs. Profit-and-loss statement.
   e. Enter Cleveland Airport. Leave JFK Airport.
   f. Coke and iron ore. Cast iron ingot.

9. Define the terms:
   a. Homeostasis.
   b. Adaptive system.
   c. Negative feedback.
   d. Positive feedback.
   e. Transformation.
   f. Equifinality.
   g. Second-order feedback.
   h. Third-order feedback.

10. Suppose that we had a simple rule for scheduling the number of units to produce next month. The rule is that we compute the number to schedule P_t, in these proportions: $0.9 \times P_{t-1}$, plus $0.1 \times D_t$, where P_{t-1} is the amount actually scheduled last month, and D_t is an estimate of current monthly demand. Thus, if last month's production was $P_{t-1} = 100$ units, then for several possible estimates of current demand, the computed value of P_t is as follows:

current demand — # to schedule

D_t	P_t	
120	102.0	
110	102.8	101
95	102.0	99.5
75	99.3	97.5
130	102.4	103

Would you characterize the response (calculated P_t) as being relatively stable or unstable to different inputs of D_t?

11. Using question 10 as a background, what is the input, the output, and the exact statement of the transfer function?

12. Now let us generalize on the production scheduling rule stated in Question 10, since we wish to experiment with several combinations of the proportions of D_t and P_{t-1} to use in computing the amount of production, P_t, to schedule. Since the rule used proportions of the two components D_t and P_{t-1}, let us designate a as the fraction of D_t, and $1-a$ as the fraction of P_{t-1}. The rule then becomes: $P_t = aD_t + (1-a) P_{t-1}$. Then, if $a = 0.1$, we have the rule used in question 10.
   a. If we use an extremely small value of a (perhaps $a = 0.01$), does the system respond in a more or less stable way?

b. Suppose that we increase a to 0.5. Does the system respond in a more or less stable way?

c. Is there feedback in this simple scheduling system?

13. Suppose that we make one simple change in the scheduling rule stated in Question 12: we reverse the sign of a in the P_{t-1} term, resulting in $P_t = a D_t + (1+a) P_{t-1}$. If $a = 0.1$ and $P_{t-1} = 100$ (as in Question 10), the computed values of P_t for the same input values of D_t are

D_t	P_t
120	122.0
110	145.2
95	169.2
75	193.6
130	226.0

Would you characterize this response as being relatively stable or unstable?

14. Using the rule stated in Question 13, and $a = 0.1$, suppose that we computed scheduled production each month for the next five months. With our estimate of current demand remaining constant at 100 units, and $P_0 = 100$ units, the results would be as follows:

t	D_t	P_t
0	—	100.0
1	100	120.0
2	100	142.0
3	100	166.2
4	100	192.8
5	100	222.1

How would you characterize the response to the constant input? Is feedback operating? Explain.

15. What is meant by "sensitivity" in feedback systems? by time lag effects? by amount of control action?

16. What is the practical significance of Ashby's statement that a system's stability or instability is a property of the system as a whole and can be assigned to no part of it?

17. Differentiate between first- , second- , and third-order feedback systems. What are the general kinds of applications of each?

18. What are the unique characteristics of problem formulation and definition when placed in the context of systems analysis?

19. Define the term *suboptimization*. What is its significance in relation to problem definition? What is organizational suboptimization?

20. What is the function of system boundaries?

21. Define the functions and uses of the following elements of block diagrams:
    a. Decision block.
    b. Transformation block.
    c. Feedback block.

22. What are the methods for predicting system performance prior to the installation of a new system?

23. What is a signed digraph? How can we identify the sign of a feedback loop in a signed digraph?

24. What is the *systems approach*?

Problems

1. The vice-president in charge of manufacturing of the Wash-N-Dry Appliance Company is pondering his problems of employment and production scheduling. He has been receiving static from the union because, over time, employment levels have fluctuated considerably. He replied that sales also fluctuate, so why should the union expect employment to be stable? The union president maintained that employment levels actually fluctuate more than sales, and charged that the reason is bad management.

   The company board of directors has just held its bimonthly meeting. Among the reports presented to the board were status reports concerning operations, a portion of which are shown in Table 1. The board was upset

TABLE 1
ABSTRACT FROM OPERATIONS REPORT TO
BOARD OF DIRECTORS, WASH-N-DRY COMPANY

	Current, 80th week	8 weeks ago	52 weeks ago
Order backlog, aggregate units	5,200	5,300	5,400
Finished goods inventory, units	2,500	1,750	2,250
Finished goods inventory, dollar value	$37,500	$26,250	$33,750
Number of employees	750	1,010	625

over the rapid growth in finished goods inventories. Finished goods inventories had increased by almost 42 percent in the last eight weeks, and were even 27 percent higher than a year ago. The comparison with the figures for the previous year at this time was particularly unsettling, because current sales levels actually were lower than a year ago. The board chairman instructed the president to take action immediately to reduce finished goods inventories. The president defended himself by pointing to the fact that employees were already being laid off at a rapid rate, and that the employment level was currently almost 26 percent below the level reached only eight weeks ago.

Immediately after the board meeting, the president met with the VP-Manufacturing and "raised hell" (using other colorful language not normally included in textbooks). The president demanded that finished goods inventory be slashed drastically, even if it meant a plant shutdown. "After all, that was how the auto industry reduced its inventories in the 1974–75 depression." The union president walked in at that point, having heard the president's order and justification. The discussions that followed (also censorable) involved strike and boycott threats.

Working late that evening, the VP-Manufacturing resolved to "get a handle on this problem" (or else he might lose his position). He assumed that he could muddle through the current crisis, but that he would probably not survive if it happened again. The next morning, he put his staff to work gathering statistics that he hoped might help shed light on the basic problem. He plotted some of these data, as shown in Figure 15.

He knew that he had been operating by moving from one crisis to the

Figure 15. Indexes of sales, inventory, backlog, and employment for the Wash-N-Dry Company

next, but even so he was startled by Figure 15. The union president was correct; employment levels seemed to fluctuate more than sales. But the inventory and backlog curves went off the chart.

a. What is the basic problem?
b. Should the VP-Manufacturing take the minimization of finished goods inventory as his goal? Why or why not?
c. Should the VP-Manufacturing take the stabilization of employment as his goal? Why or why not?

2. In terms of the impact of one variable on another, what are the relationships of the following: fertility; industrial output per capita; births per year; and population? Relate these variables in a signed digraph, labeling each arc with a plus or minus sign to indicate the nature of the impact of one variable on the other. If there is a feedback loop, is it positive or negative?

3. An appliance manufacturer is attempting to deal with the range of inventory problems she confronts. She produces a line of small household appliances that sell in the \$20–\$75 range, and the demand patterns are seasonal in nature. The raw materials are approximately 10 percent of the manufacturing cost, and labor is approximately 60 percent.

   Labor is semiskilled; workers usually can be trained to be normally productive within three months. While not maternalistic toward her employees, the owner feels a strong social responsibility for providing good working conditions and stable employment. Labor turnover, however, has been high in spite of the fact that labor was paid higher than average wages for similar work.

The production cycle is relatively short—approximately two weeks for the average item—and the owner has taken advantage of this fact by keeping finished goods inventories to no more than one month's supply for any time during the seasonal demand pattern. She accomplishes this close control by keeping abreast of market trends and seasonals through a well conceived forecasting system.

The owner has commissioned a study by an outside consultant and has focused attention on minimizing her inventories throughout the process (that is, raw material, in-process, and finished-goods inventories). However, the consultant has manipulated the owner into agreeing that the first phase of the study should be a report that defines the scope and objectives of the study. Draft a statement of the portion of the consultant's phase-I report that defines the statement of the problem and what the scope and objectives of the study should be.

References

Ackoff, R. L. "Towards a System of System Concepts." *Management Science* 17(July 1971): 661–671

Ashby, W. R. *An Introduction to Cybernetics*. New York: John Wiley & Sons, 1956.

——— "Stability and Adaptation." In *Organizations: Systems, Control and Adaptation*. Vol. 2. 2d ed. Edited by J. A. Litterer. New York: John Wiley & Sons, 1969.

Bertalanffy, L. V. "General System Theory—A Critical Review." In *Organizations: Systems, Control and Adaptation*. Vol. 2. 2d ed. Edited by J. A. Litterer. New York: John Wiley & Sons, 1969.

Blumstein, A., and Larson, R. "Models of a Total Criminal Justice System." *Operations Research* 17(March–April 1969): 199–232.

Boulding, K. E. "General Systems Theory—The Skeleton of Science." In *Management Systems*. 2d ed. Edited by P. P. Schoderbek. New York: John Wiley & Sons, 1971.

Buffa, E. S., and J. S. Dyer. *Management Science/Operations Research: Model Formulation and Solution Methods*. Santa Barbara, Ca.: Wiley/Hamilton, 1977.

Chestnut, H. *Systems Engineering Methods*. New York: John Wiley & Sons, 1967.

Churchman, C. W. *The Systems Approach*. New York: Delacorte Press, 1968.

______; R. L. Ackoff; and E. L. Arnoff. "Analysis of the Organization." In *Organizations: Systems, Control and Adaptation*. Vol. 2. 2d ed. Edited by J. A. Litterer. New York: John Wiley & Sons, 1969.

Drake, A. W.; R. L. Keeney; and P. M. Morse. *Analysis of Public Systems*. Cambridge, Mass.: MIT Press, 1972.

Forrester, J. W. *Industrial Dynamics*. Cambridge, Mass.: MIT Press, 1961.

______. *Principles of Systems*. Cambridge, Mass.: Wright-Allen, 1970.

______. *World Dynamics*. 2nd ed. Cambridge, Mass.: Wright-Allen, 1972.

Hall, A. D., and R. E. Hagen. "Definition of a System." In *Organizations: Systems, Control and Adaptation*. Vol. 2. 2d ed. Edited by J. A. Litterer. New York: John Wiley & Sons, 1969.

Hamilton II, W. F., and D. K. Nance. "Systems Analysis of Urban Transportation." *Scientific American*, July 1969.

Hare, Jr., V. C. *Systems Analysis: A Diagnostic Approach*. New York: Harcourt, Brace & World, 1967.

Kane, J.; I. Vertinsky; and W. Thomson. "KSIM: A Methodology for Interactive Resource Policy Simulation." *Water Resources Research* 9 (1973).

Kennedy, F. D. "Development of a Community Health Service System Simulation Model." *IEEE Transactions on Systems Science and Cybernetics*. SSC-5. 3(July 1969): 199–207.

Litterer, J. A. *Organizations: Systems, Control and Adaptation*. Vol. 2. 2d ed. New York: John Wiley & Sons, Inc., 1969.

Machol, R. E., and R. F. Miles, Jr. "The Engineering of Large-Scale Systems." In *Systems Concepts*. Edited by R. F. Miles, Jr. New York: John Wiley & Sons, 1973.

Miles, Jr., R. F. *Systems Concepts*. New York: John Wiley & Sons, 1973.

Morse, P. M. *Operations Research for Public Systems*. Cambridge, Mass.: MIT Press, 1967.

Ramo, S. "The Systems Approach." In *Systems Concepts*. Edited by R. F. Miles, Jr. New York: John Wiley & Sons, 1973.

Rapoport, A. "Mathematical Aspects of General Systems Analysis." In *Organizations: Systems, Control and Adaptation*. Vol. 2. 2d ed. Edited by J. A. Litterer. New York: John Wiley & Sons, 1969.

Reitman, J. *Computer Simulation Applications*. New York: John Wiley & Sons, 1971.

Rudwick, B. H., *Systems Analysis for Effective Planning*. New York: John Wiley & Sons, 1969.

Schoderbek, P. P. *Management Systems*. 2d ed. New York: John Wiley & Sons, 1971.

———, A. G. Kefalas; and C. E. Schoderbek. *Management Systems: Conceptual Considerations*. Dallas, Tex.: Business Publications, Inc., 1975.

Simon, H. A. "The Architecture of Complexity." In *Organizations: Systems, Control and Adaptation*. Vol. 2. 2d ed. Edited by J. A. Litterer. New York: John Wiley & Sons, 1969.

Chapter 4

Analytical Methods in Production/ Operations Management

Perhaps the most notable thing about the current production/operations-management field is the degree to which analytical methods have been applied. These methods have been used both to establish a conceptual framework and to solve practical problems. Early methods of analysis represented graphic and schematic models of various aspects of a productive system. Since World War II, however, there has been an increased use of more sophisticated techniques, such as analytical, statistical, and simulation models, as well as an increased emphasis on systems and the systemic point of view that we discussed in Chapter 3.

Systems concepts and methods provide the general framework for analyzing productive systems. The input-transformation-output module is the building block that is used to describe any component process. By relating component processes through block or network diagrams, we can represent the complex interactions among components. Analytical methods may be used both to gain insight into the behavior of the components as well as the system as a whole, and to help predict system performance. Thus, the analytical methods useful in production/operations management are embraced by the general philosophical and methodological concepts of systems.

Analytical methodology in production/operations management essentially follows the framework of scientific method, and uses as a base various kinds of models that represent some system or some part of the system under study. The discussion of models and their place in decision making in Chapter 2 is pertinent in this regard.

KINDS OF ANALYTICAL MODELS

Analytical models for management fall into a system of their own (see Figure 1). Box 3 contains a model concerned with predicting system performance. Ideally, such a model will help managers to answer the question: "If I take a certain action, then what will be the result?" Some models that fall into this category are specifically called "What if . . ." or "If . . ., then . . ." models. As Figure 1 shows, this *predictive* model will be affected by the environment, and can be manipulated by the manager. A second type of predictive model (box 1 in Figure 1) may be used to forecast what the environment will be like in the future, and may provide some of the important data needed to drive the predictive model (box 3).

The evaluation model (box 4) represents a third category of models in Figure 1. As we noted in Chapter 2, a manager needs a way of evaluating the results of a predictive model. Such models may be implicit in simple cases involving decisions that produce only a single, certain or deterministic effect on the system. However, in cases involving risk, where different effects may occur with different probabilities, or in cases involving multiple effects that must be evaluated

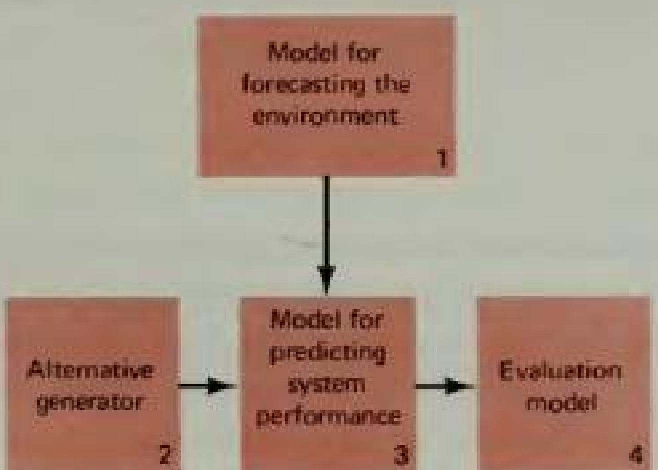


Figure 1. Kinds of analytical models

SOURCE: E. S. Buffa and J. S. Dyer, *Management Science/Operations Research: Model Formulation and Solution Methods* (Santa Barbara, Ca.: Wiley/Hamilton, 1977) according to multiple criteria, a manager may require an explicit evaluation model in order to select the alternative.

Finally, we have an alternative generator (box 2). In many instances, the manager (or the equivalent) is the alternative generator. In raising the question, "What happens if . . .," the manager is generating an alternative. In other instances, we may devise ways of scanning an entire class of alternatives, or even enumerating all possible alternatives, whose effects are reflected through the predictive model and evaluated by the evaluation model.

Finally, one important type of model (that Figure 1 does not illustrate) can be formulated for problems exhibiting special characteristics. This model type actually allows the rules of mathematical analysis to generate and investigate the sets of all alternative solutions, selecting according to a specific evaluation model, the one with the best predicted outcome. These optimizing models are termed *normative* (or *prescriptive*) models, since they result in a proposed solution to the problem.

We will discuss three main types of analytical methods: predictive models, evaluation models, and normative models.

PREDICTIVE MODELS

Some of the models that managers use most widely are those which predict the performance of a system. For a given alternative, the model predicts the outcomes that must be evaluated in terms of the criteria and values.

Forecasting Models

As already noted, the model used to forecast the behavior of the environment is also a predictive model. Forecasting models focus attention on the prediction of the exogenous (outside) factors that normally are not within our control but have a tremendously important influence on what happens. For most models of productive systems, these outside factors are likely to be the demand for the product or service with which we deal. Demand is a fundamental driving force, whether the system produces a product or a service, and whether the enterprise is profit making or nonprofit making. Forecasting data represent important inputs to models that predict system performance, and greatly affect the behavior of such systems.

Knowing about the demand function may suggest some of the most significant strategies for productive systems. For example, knowing that the demand function for the maternal and infant care health system is seasonal (see Figure 6 of Chapter 3) suggests important ways in which to meet the work load. Furthermore, knowing something about the arrival rate pattern of individuals who come to the clinic for care provides important data for setting short-term capacities. If we take an even longer-term viewpoint, which factors in our society control the need for maternal and infant care in the future? How do we forecast the longer-term needs for facilities, and for systems appropriate to future, possibly much larger, loads?

These questions apply to all kinds of systems. There are short-range, medium-range, and long-range variations in the demand for products and services. In order to predict system performance, we must forecast demand as an input. This book covers forecasting methodologies in two places. In Chapter 5, we discuss forecasting methodologies appropriate for designing a product or service, locating productive systems, and determining longer-term capacities. In Chapter 9, we discuss statistical forecasting methods useful in day-to-day operations.

Computer-Based Corporate Models

Managers now use computer-based predictive models commonly as vehicles for planning at various levels. The applications include steel making, forest products, various consumer products manufacturing, cigarette manufacturing, publishing, leasing companies, banking, aerospace companies, electric utilities, and governmental agencies.

We can view the problem of predicting system performance (including the performance of the preceding forecasting models) in the context of any productive process, as suggested by our input–transformation–output module, discussed in Chapter 3. The function of the predictive model is to describe in detail what goes on within the "black box" that represents a transformation process.

These kinds of predictive models have commonly been developed in computer-based interactive mode to facilitate a "manager-active" situation, in which results from one query may stimulate new questions. This creates a powerful combination of decision maker and predictive model in a loop. Very complex computations, including data input, computing, and output—reflecting assumptions about volume, price, costs, or the effects of a labor dispute—can be handled in a short turnaround time (as little as a few minutes).

Basically, the mathematical relationships for these kinds of models are as simple as the flow of accounting data.

Application at the Inland Steel Company. The Inland Steel Company uses computer-based predictive models that focus on the production process. The models that have been developed, and their relationships, are shown in Figure 2.

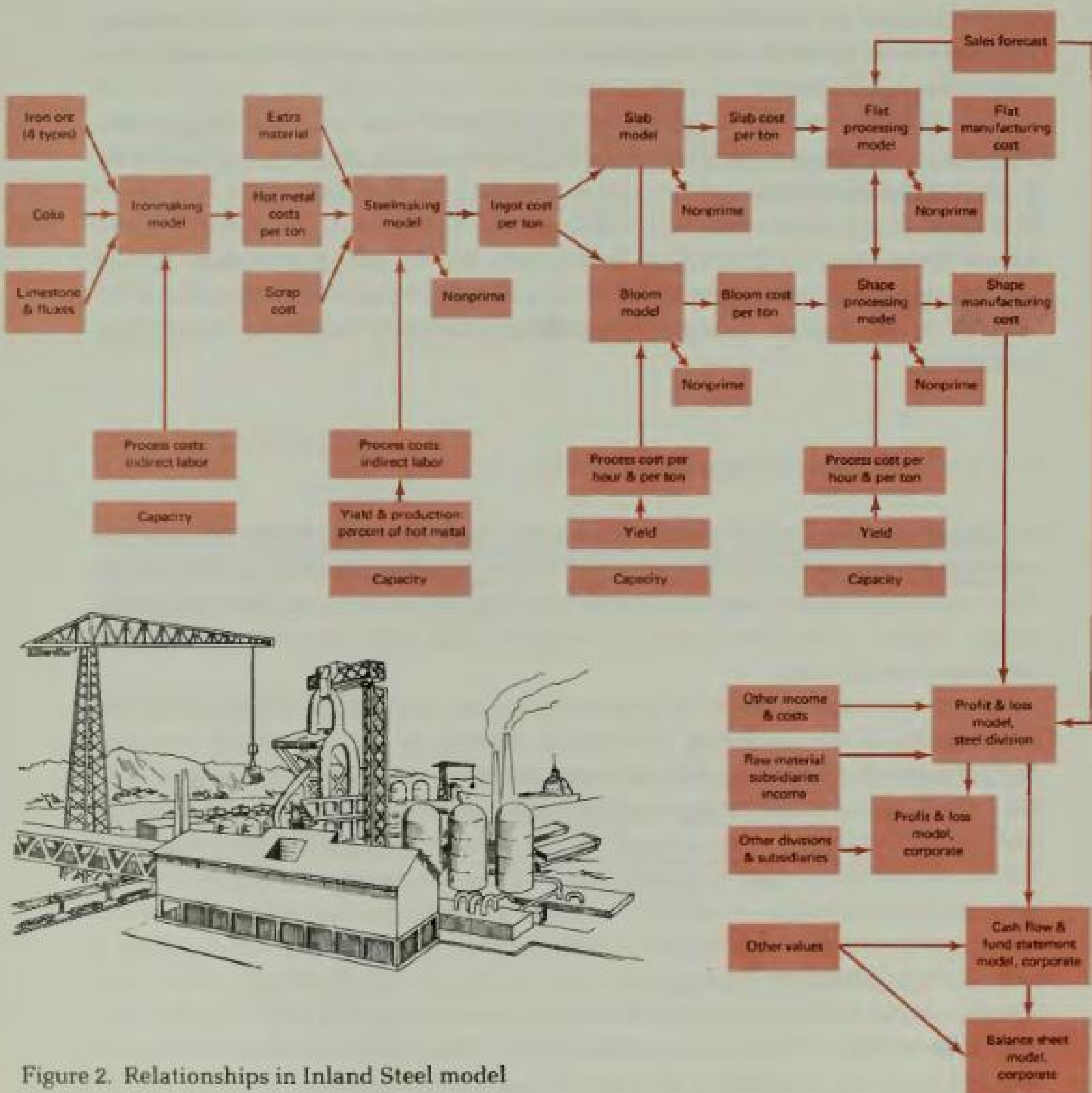


Figure 2. Relationships in Inland Steel model

SOURCE: J. B. Boulden and E. S. Buffa, "Corporate Models: On-Line, Real-Time Systems" *Harvard Business Review*, July-August 1970.

Using the models, corporate planners can simulate very quickly the effects of a wide variety of planning assumptions. Each model deals with a basic process in the sequence from raw material to finished product. The models simulate the various costs incurred in converting ores to molten iron, converting molten iron to steel ingots, processing ingots, and finishing the steel to various end-products.

The types of questions asked by Inland Steel management, using the models, are typified by the following: How much raw material is required to meet production forecasts? What are the cost effects of various hot metal to scrap ratios and the resulting yield under various assumptions of raw material costs? What are the capacity requirements for proposed levels of operation?

The models have been used widely by Inland Steel in the development of its annual profit plans and related five-year profit-and-cash projections. In approaching the profit-planning cycle, company management has used the models to simulate operations by means of different scenarios, particularly those involving labor disputes in 1971 and their possible effects on operations [Boulden and Buffa, 1970].

Systems Dynamics Predictive Models

In Chapter 3 we referred to the effects of positive and negative feedback, time-lags, system sensitivity to control, and the amount of control action. The interesting characteristics of such systems may be found in the dynamic effects on outcomes; that is, how the system responds over time.

Such models have been used to study the effects over time of inventory control policies on system inventories, different distribution structures, and so on [Forrester, 1961]. More recently, systems dynamics models have been used to predict the effects of urban decay [Forrester, 1969], as well as of the limits on growth imposed by finite resource pools and by exploding population [Meadows, et al., 1972]. The same techniques have been used to debunk the "Limits to Growth" scare [Cole, et al., 1973]. Since in dealing with various problems in operations management, our interest may lie in the dynamic effects, the systems dynamics methodology may be useful.

Predicting the Effects of Risk

In many systems and subsystems in operations management, the outcomes can be predicted only on a probabilistic basis. Two ~~of the most useful types of~~ predictive models for productive systems that consider risk are waiting line (or queuing) models, and Monte Carlo simulation models.

Waiting Line or Queuing Models

Many production problems are related in some way to the build-up of waiting lines. In these types of problems, we have people, parts, or machines that need some sort of service at random time intervals. The activity required to service them may take a variable amount of time. Under certain conditions of arrival rates and servicing rates, a waiting line builds up. Waiting line theory provides a means of forecasting the *probable length* of the waiting line and the *probable delay or waiting time*, as well as other important data. Knowledge of these facts makes it possible to develop predictive models and make more intelligent decisions about the following types of questions.

1. How many ambulances are needed to service a given geographic area, giving service within five minutes 90 percent of the time? Demand for service is highly variable. The demand distribution specifies the arrival rates and the variable times to respond to calls for service. Demand distribution, plus the time to render the service, define the total service time. Delay or waiting time depend on the probablistic nature of demand for service and service time. Waiting line models can predict the average waiting times, and provide the basis for determining the system's capacity to provide the needed service.
2. Trucks arrive at a truck dock to load or unload. Such activities occur commonly in industry, at wholesale grocers, and many other distribution types of activities. How big should the truck dock be? That is, how many trucks should the dock be able to accommodate at one time? Waiting line models can predict the average waiting time of trucks, given data on the distribution of arrival rates and the times to service the trucks at the dock.
3. What should be the size of the repair crew to maintain equipment used in a productive process? The equipment breakdown rate depends on some probabilistic process and represents the arrivals for service. The time to repair the machines represents the service time. Waiting line models can predict the average machine down time for different sizes and configurations for the repair service. The relative costs for providing the service versus the down time costs provide a basis for evaluating the most appropriate size and configuration of repair crew service.

An Example

Let us take as an example the situation of the tool crib where mechanics come to obtain special tools. The attendant checks out the tool to the mechanic in

TABLE 1
RANDOM ARRIVAL TIME OF MECHANICS,
SERVICE TIME CONSTANT AT 3 MINUTES

Arrival Time of Mechanics	Service Begins *at*	Service Ends *at*	*Idle Time of Attendant**	*Waiting Time of Mechanics†*	*Length of Line, Excluding Mechanic Being Serviced*
10:00	10:00	10:03	0	0	0
10:09	10:09	10:12	6	0	0
10:13	10:13	10:16	1	0	0
10:19	10:19	10:22	3	0	0
10:34	10:34	10:37	12	0	0
10:36	10:37	10:40	0	1	1
10:37	10:40	10:43	0	3	1
10:38	10:43	10:46	0	5	2
10:39	10:46	10:49	0	7	3
10:42	10:49	10:52	0	7	3
11:02	11:02	11:05	10	0	0
11:03	11:05	11:08	0	2	1
11:05	11:08	11:11	0	3	1
11:05	11:11	11:14	0	6	2
11:09	11:14	11:17	0	5	2

*Attendants' idle time = 32 minutes.
†Mechanics' waiting time = 39 minutes.

exchange for a "tool check," which has the mechanic's name or number on it. We assume that the mechanics arrive at random times at an average rate of 12 per hour or 1 each 5 minutes. The attendant can service the mechanics in 3 minutes, and initially we shall assume that this service time is constant for all mechanics. Table 1 shows a record of the servicing of 15 mechanics. The average interval between arrivals is slightly less than 5 minutes for this sample. We see that the attendant had 32 minutes of idle time and that the various mechanics waited a total of 39 minutes.

What happens if the service time is increased from 3 to 4 minutes, assuming the same schedule of arrivals? The outcome, shown in Table 2, is predictable. The attendant is busier now, as evidenced by the smaller idle time, but the waiting time of the mechanics is almost twice as long as before. We immediately see that the waiting time and the length of the waiting line can increase dramatically because of such a time change. When the average rate of arrival equals the rate at which the mechanics can be served, theoretically the length of the waiting line will increase and become infinitely long.

Now let us introduce random service times, using the same arrival schedule of the mechanics. This makes the problem a fairly realistic one, since the time required to service a mechanic should vary, depending on the number of tools

TABLE 2
RANDOM ARRIVAL TIME OF MECHANICS, SERVICE TIME CONSTANT AT 4 MINUTES

Arrival Time of Mechanics	*Service Begins at*	*Service Ends at*	*Idle Time of Attendant**	*Waiting Time of Mechanics†*	*Length of Line, Excluding Mechanic Being Serviced*
10:00	10:00	10:04	0	0	0
10:09	10:09	10:13	5	0	0
10:13	10:13	10:17	0	0	0
10:19	10:19	10:23	2	0	0
10:34	10:34	10:38	11	0	0
10:36	10:38	10:42	0	2	1
10:37	10:42	10:46	0	5	2
10:38	10:46	10:50	0	8	2
10:39	10:50	10:54	0	11	3
10:42	10:54	10:58	0	12	3
11:02	11:02	11:06	4	0	0
11:03	11:06	11:10	0	3	1
11:05	11:10	11:14	0	5	2
11:05	11:14	11:18	0	9	3
11:09	11:18	11:22	0	9	2

*Attendants' idle time = 22 minutes.
†Mechanics' waiting time = 64 minutes.

the mechanic needs and the size, weight, and location of the tools. Table 3 is a record of such a situation: the mechanics arrive at random times, and the service times vary randomly, averaging 3.07 minutes for the sample given. For a particular sample, the attendant's idle time and the mechanics' waiting time will depend on how well the arrivals of the mechanics match up with the longer and shorter service times. If several mechanics arrive almost simultaneously with requests that require long service times, their waiting time will be relatively long while the attendant is continuously busy. If the long service time requests are distributed in a particular sample, their effect on the mechanics' waiting time will be smaller. Since waiting line theory can forecast the *probable* waiting time and the *probable* length of the line, management can decide on the optimum allocation of personnel and equipment in order to minimize incremental costs. To do this, we must know the nature of the distributions of arrivals and service times.

Looking at Table 3, let us make some typical calculations, recognizing that the sample is probably too small for decision making purposes. If the attendant is paid \$4.00 per hour, then the cost of the attendant's idle time is:

$$\frac{26}{60} \times \$4.00 = \$1.73$$

TABLE 3
RANDOM ARRIVAL TIME OF MECHANICS
AND RANDOM SERVICE TIMES

Arrival Time of Mechanics	Service Begins at	Service Ends at	*Idle Time of Attendant**	*Waiting Time of Mechanics†*	*Length of Line, Excluding Mechanic Being Serviced*
10:00	10:00	10:01	0	0	0
10:09	10:09	10:14	8	0	0
10:13	10:14	10:15	0	1	1
10:19	10:19	10:22	4	0	0
10:34	10:34	10:40	12	0	0
10:36	10:40	10:44	0	4	1
10:37	10:44	10:51	0	7	2
10:38	10:51	10:53	0	13	3
10:39	10:53	10:57	0	14	4
10:42	10:57	11:02	0	15	4
11:02	11:02	11:03	0	0	0
11:03	11:03	11:04	0	0	0
11:05	11:05	11:07	1	0	0
11:05	11:07	11:08	0	2	1
11:09	11:09	11:12	1	0	0

*Attendants' idle time = 26 minutes.
†Mechanics' waiting time = 56 minutes.

for the sample of 72 minutes, or:

$$\frac{480}{72} \times \$1.73 = \$11.53$$

per day. If the mechanics' wages average $7.00 per hour, the cost of their idle time is:

$$\frac{56}{60} \times 7.00 = \$6.53$$

for the sample, or:

$$\frac{480}{72} \times 6.53 = \$43.53$$

per day. The total cost of idle and waiting time is $11.53 + $43.53 = $55.06 per day.

Now the question is: Is it economical to add a second attendant? If the second attendant is added, mechanics' waiting time will be reduced, but the total cost for attendants will be increased. If the net effect on the cost of the attendants plus waiting time is negative, the second attendant is justified.

Analytical Solutions for Waiting Line Problems. In Tables 1, 2, and 3, we have simulated simple waiting line problems. In cases where the distributions of arrivals and service times fit certain standard mathematical distributions, equations have been developed that give directly the items of data in which we are interested, such as the average length of the waiting line, average waiting time, and average number of units in the system. The important restrictions are the distributions. If the actual distributions of arrivals and service times fit the standard mathematical distributions fairly well, very little work is required to determine a solution. On the other hand, if they do not fit, a great amount of work could be needed to develop a mathematical solution. In such instances, simulation methods can be used to develop a solution, regardless of the nature of the actual distributions. These methods are discussed on the next page of this chapter.

Poisson Arrivals. The mathematical function known as the Poisson distribution is commonly used to describe arrival rates. Figure 3 shows a Poisson distribution for arrival rates averaging 12 per hour. The simulated examples which we have been using had arrival rates that were based on this distribution. If the average arrival rate were 12 per hour and the service rate were constant at 20 per hour (3 minutes per mechanic served), as in our first simulated example (Table 1), mathematical solutions would give directly:

Average length of waiting line	= 0.45 persons in line, excluding mechanic being served
Average waiting-time	= 0.0375 hour/person, or 2.25 minutes/person

(See Appendix D for formulas and a more complete discussion of waiting lines, and Chapter 14 for an analysis of service systems.)

Our limited sample gave an average line-length of 0.50 persons and an average waiting time of $\frac{39}{15} = 2.6$ minutes per person. Had our sample been larger in the simulated example, we would have approached more closely the answers given by the formulas.

We mentioned earlier that, theoretically, when the mean arrival and service rates are equal, the waiting line length and, therefore, the waiting time become infinitely long. However, this is true only theoretically, because in a practical situation several things may happen to prevent it. If people are the units involved in the waiting line, new arrivals are apt to be discouraged by an excessively long line and leave. Also, where people are involved, working hours limit the line length, since it falls to zero when the work day is over. Finally, the people performing the services tend to react to the build-up of a waiting line by

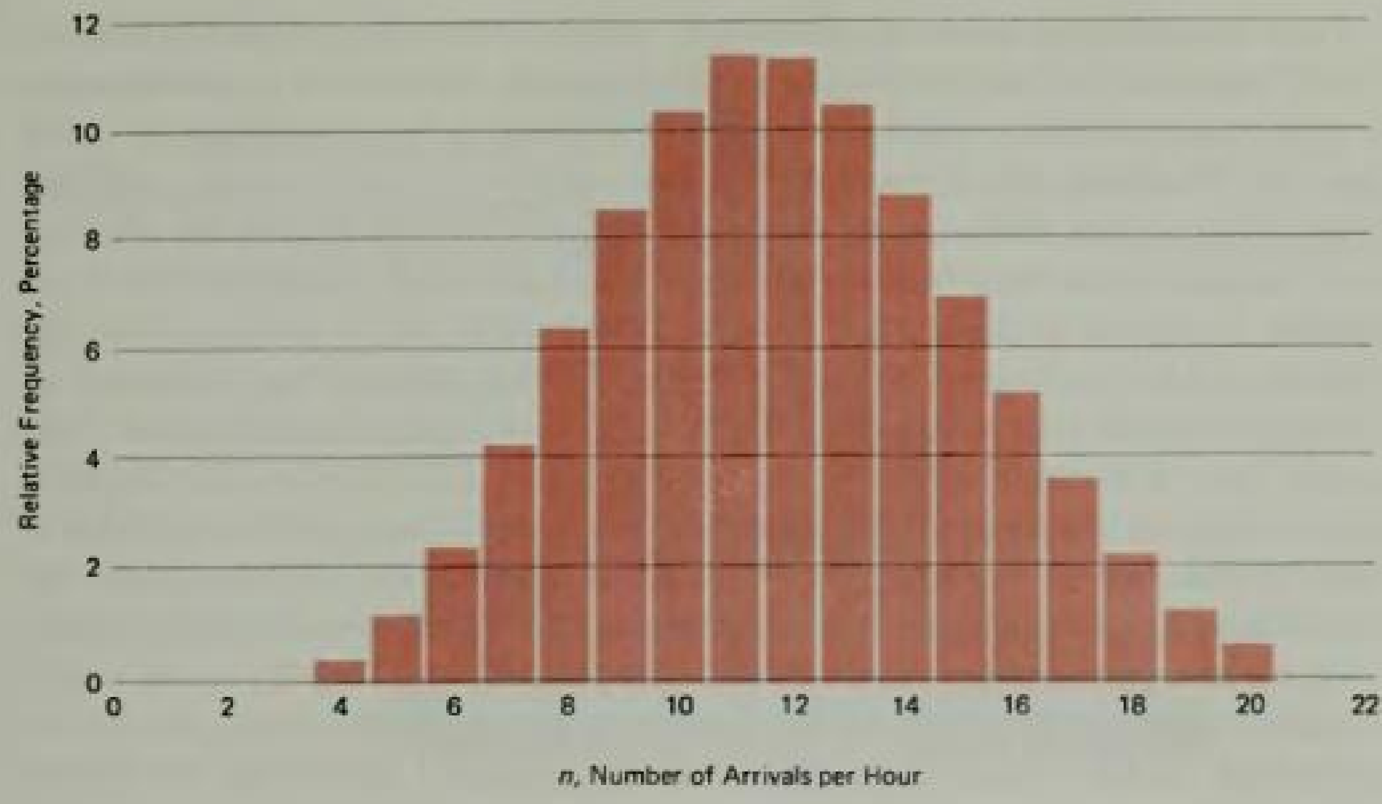


Figure 3. Poisson distribution of arrival rates averaging 12 per hour

speeding up the rate of service. Where the units requiring service are parts, or perhaps machines, some of these compensating effects do not take place.

Simulation Models

Simulation of operations management problems is a rapidly growing technique. Although the fundamental ideas behind simulation can stand alone, simulation's rapid growth actually has been paced by the high-speed electronic computer, because the arithmetic work required for practical problems ordinarily is too great for hand computation. This approach to problems sets up a simulated experiment and then carries it through completely on paper (or in a computer) to see the effect of the variables of interest on the criterion. Comparisons among alternatives can be made by means of the simulation model, but the analyst must set up the alternatives. It is a systematic trial and error method for solving complex problems.

Where is simulation useful? Examples might be: (1) waiting line problems where standard distributions do not approximate actual distributions for arrival and processing rates; or (2) complex waiting line problems in which each station has several channels or stations in tandem (such as on a production line). On a large scale, this technique may be used to simulate the entire operation.

Here lies the great future of simulation. Imagine the power of such a tool for the decision maker faced constantly with a legion of alternatives regarding work scheduling, equipment, layout, methods, inventory policies, changes of work load, etc. The basic objectives of the decision maker are clear: to make decisions that will maximize return on investment or minimize costs. If only the effect of the various alternatives faced could be seen clearly and quickly, without requiring a waiting period to see how they will work out in practice. For the decision maker, a situation that lets one experiment with an idea, without first risking or committing company funds, is ideal; a simulation model and a computer allow a decision maker to try out dozens of alternatives. The decision maker also can learn a great deal about the interdependence of the variables in the complex system. For example, the effect of a change in inventory policy on plant capacity and labor cost may be observed, thus teaching the decision maker to make many decisions jointly instead of separately and thereby avoid costly mistakes of suboptimization. With the aid of high-speed computers, simulation ultimately makes available to operations management personnel an experimental laboratory.

Simulated Sampling. The sampling method, known generally as Monte Carlo, is a simulation procedure of considerable value. In the last section, under waiting line theory, we spoke of distributions that did not fit the standard ones for which the theory had been worked out, and of complex waiting line situations in which the mathematical complexity eliminates analytical solutions. The required data can be built up through simulation of the conditions of the problem.

Let us assume that a product is being assembled by a two-station assembly line. There is one operator at each of the two stations. Operation *A* is the first of the two operations. The operator completes approximately the first half of the assembly and then sets the half-completed assembly on a section of conveyor, where it rolls down to operation B. It takes a constant time of 0.10 minute for the

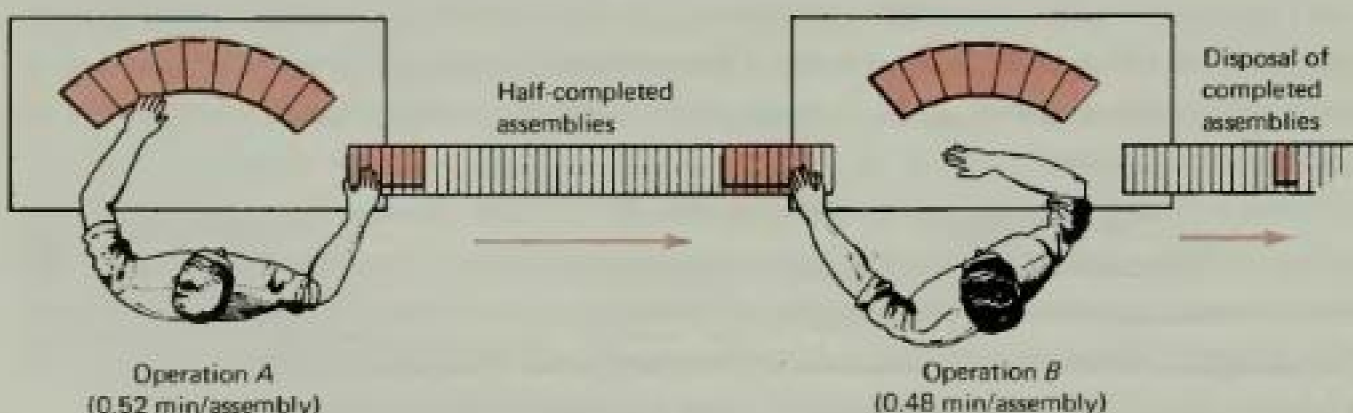


Figure 4. Layout of two-station assembly line

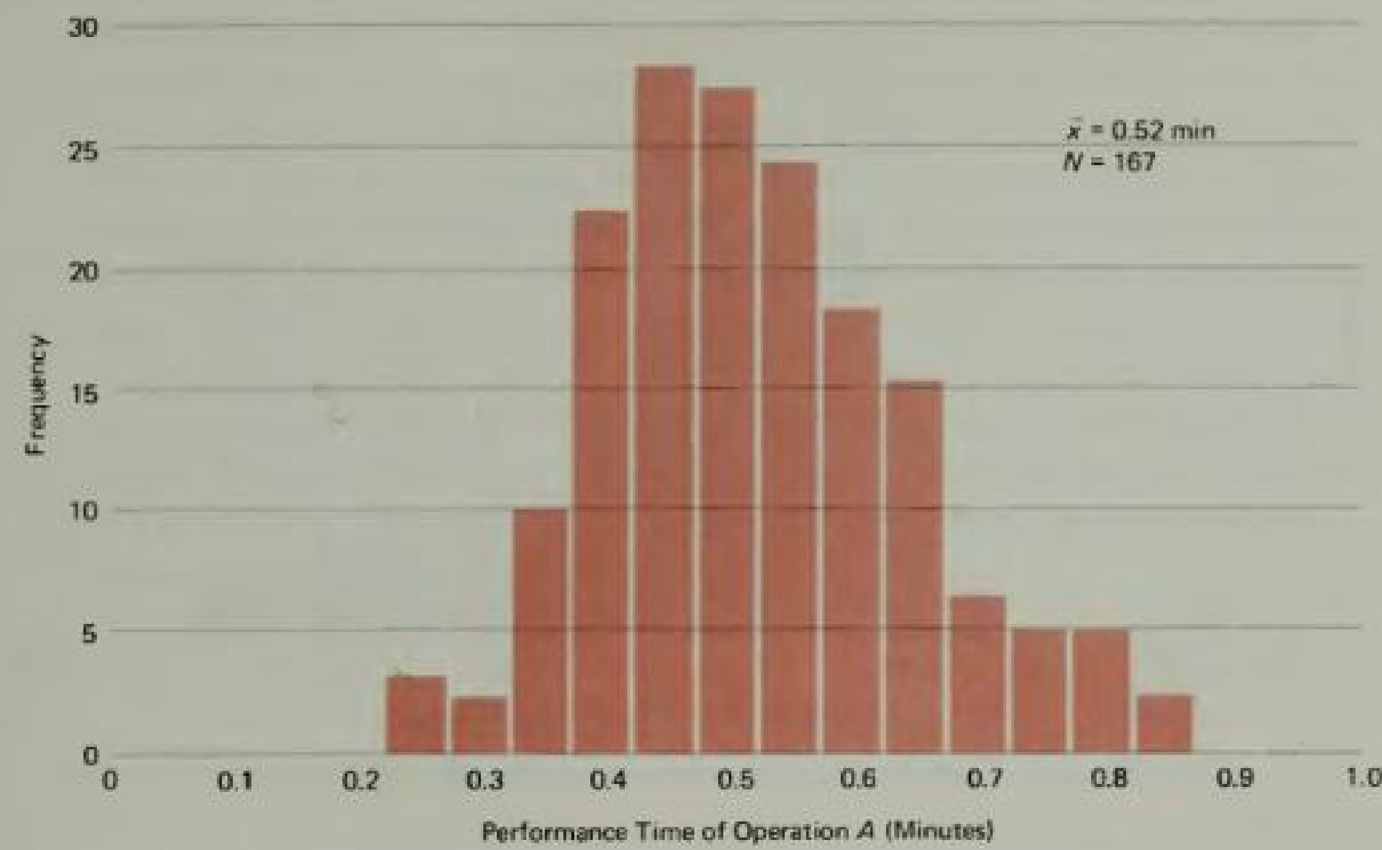


Figure 5. Distribution of performance times for operation *A*

part to roll down the conveyor section and be available to operator *B*. Operator B then completes the assembly. The average time for operation *A* is 0.52 minute per assembly, and the average time for operation *B* is 0.48 minute per assembly. The simple setup is shown in Figure 4. We wish to determine the average inventory of assemblies that we may expect (average length of the waiting line of assemblies) and the average output of the assembly line. This may be done by simulated sampling, as follows:

1. The distributions of assembly time for operations *A* and *B* must be known or procured. For example, these are represented by Figures 5 and 6. A study was taken for both operations, and the two frequency distributions resulted. For Figure 5, the value 0.25 minute occurred three times, 0.30 occurred twice, etc. The two distributions do not necessarily fit standard mathematical distributions, but this is not important.
2. Convert the frequency distributions to cumulative probability distributions. This is done in Figures 7 and 8 by (1) summing the frequencies that are less than or equal to each performance time, and (2) plotting them. The cumulative frequencies then are converted into percentages by assigning the number 100 to the maximum value. As an example, let us take Figure 5 and convert it to the cumulative distribution of Figure 7. Beginning at the lowest value for performance time, 0.25 minute, there were three observa-

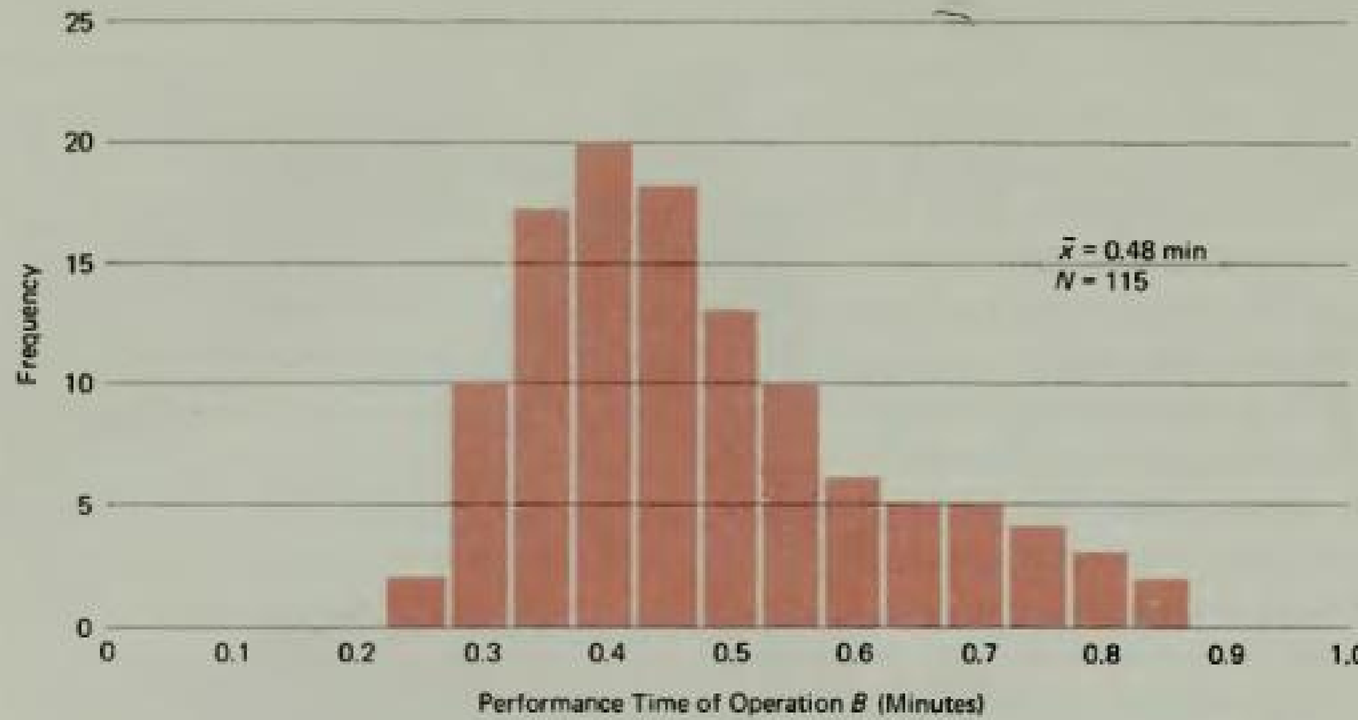


Figure 6. Distribution of performance times for operation B

tions. Three is plotted on the cumulative chart for the time 0.25 minute. For the performance time, 0.30 minute, there were two observations; however, there were five observations that measured 0.30 minute or less, so the value (5) is plotted for 0.30 minute. For the performance time, 0.35 minute, there were ten observations recorded: however, there were fifteen observations that measured 0.35 minute or less. Figure 7 was constructed from Figure 5 by proceeding in this way. When the cumulative frequency distribution was completed, a cumulative percentage scale was constructed on the right of Figure 7 by assigning the number 100 to the maximum value, 167, and dividing the resulting scale into equal parts. This results in a cumulative probability distribution. From Figure 7 we can say that 100 percent of the time values were 0.85 minute or less, 55.1 per cent were 0.50 minute or less, etc.

3. Sample at random from the cumulative distributions to determine specific performance times to use in simulating the operation of the assembly line. We do this by selecting numbers between 0 and 100 at random (representing probabilities or per cents). The random numbers could be selected by any random process, such as drawing numbered chips from a box. The easiest way is to use a table of random numbers such as the one in Appendix E, Table 5.

   The random numbers are used to enter the cumulative distributions in order to obtain time values. In the example shown in Figure 7, the random number 10 is used. A horizontal line is projected until it intersects the distribution curve; a vertical line projected to the horizontal axis gives the midpoint time value associated with the intersected point on the distribution curve, which happens to be 0.40 minute for the random number 10. Now we can see the purpose behind the conversion of the original distribution to a cumulative distribution. Only one time value now can be associated with a given random number. In the original distribution, because of the bell shape of the curve, two values would result.

Sampling from the cumulative distribution in this way gives time values in random order which will occur in proportion to the original distribution, just as if assemblies actually were being produced. Table 4 gives a sample of twenty time values that were determined from the two distributions in this way.

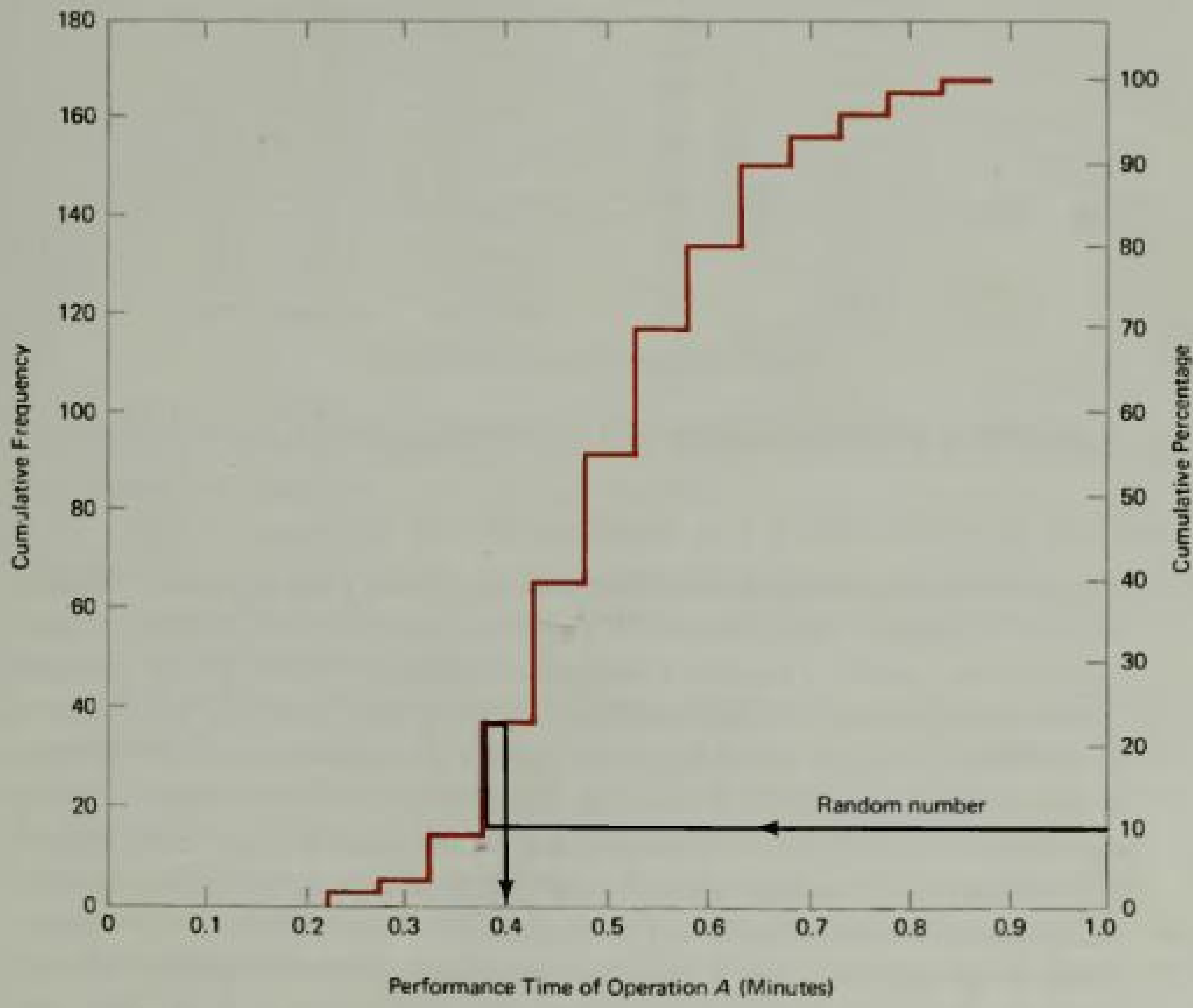


Figure 7. Cumulative distribution of performance times for operation A

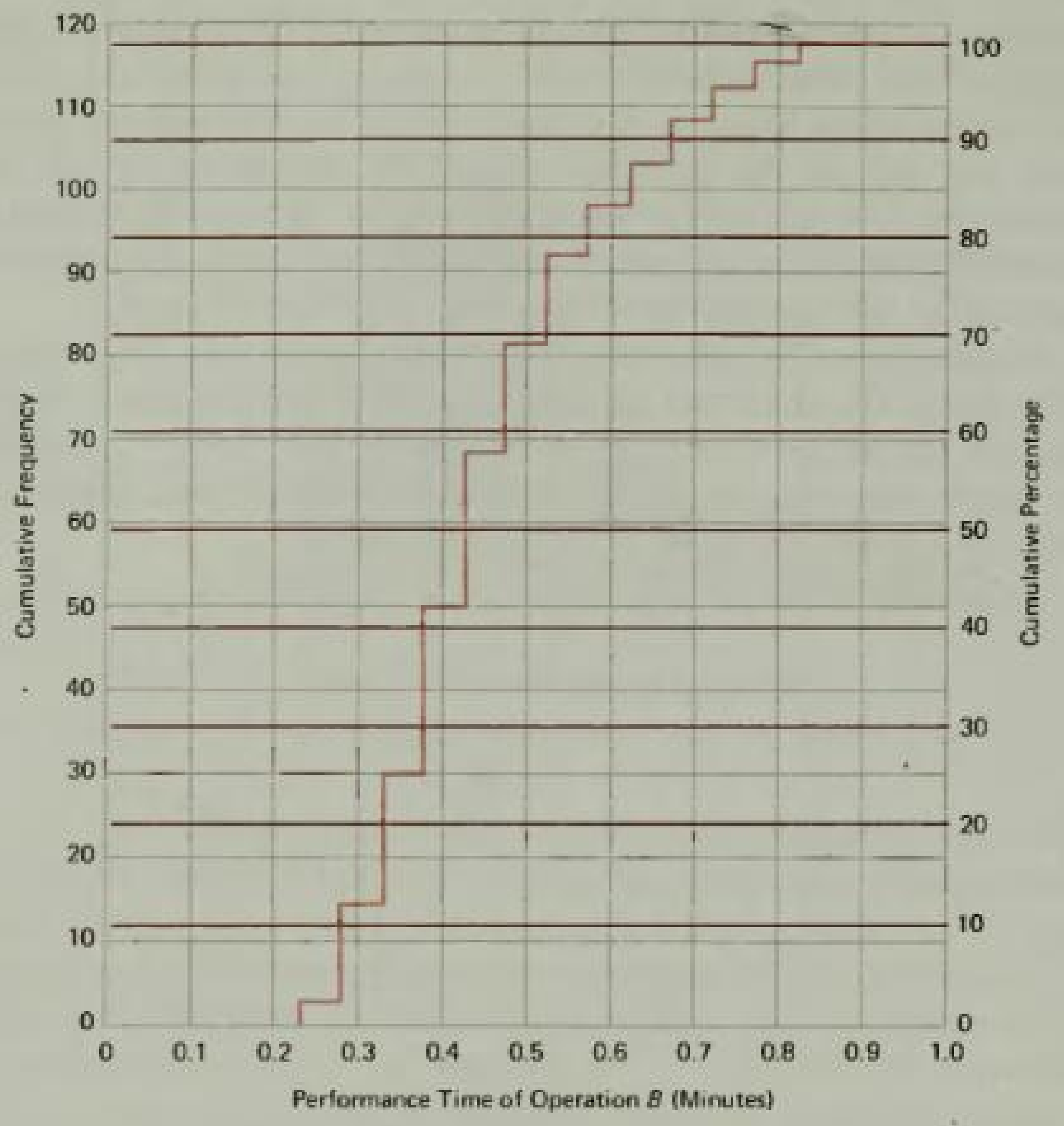


Figure 8. Cumulative distribution of performance times for operation B

4. Simulate the actual operation of the assembly line. This is done in Table 5, which, of course, is very similar to the examples of waiting line problems that we used earlier. The time values for operation *A* (table 4) first are used to determine when the half-completed assemblies would be available to operation *B*. The first assembly is completed by operator *A* in 0.40 minute. It takes 0.10 minute to roll down to operator *B*, so this point in time is selected as zero. The next assembly is available 0.40 minute later, and so on. For the first assembly, operation *B* begins at time zero. From the simulated sample, the first assembly requires 0.60 minute for *B*. At this point, there is no idle time for *B* and no inventory. At time 0.40, the second assembly becomes available; but since *B* is still working on the first, the assembly must wait 0.20 minute. Operator *B* begins work on it at 0.60. From

TABLE 4
SIMULATED SAMPLE OF 20 PERFORMANCE TIME VALUES FOR OPERATIONS *A* AND *B*

Operation A		*Operation B*	
Random Number	*Performance Time from Figure 7*	*Random Number*	*Performance Time from Figure 8*
10	0.40	79	0.60
22	0.40	69	0.50
24	0.45	33	0.40
42	0.50	52	0.45
37	0.45	13	0.35
77	0.60	16	0.35
99	0.85	19	0.35
96	0.75	4	0.30
89	0.65	14	0.35
85	0.65	6	0.30
28	0.45	30	0.40
63	0.55	25	0.35
9	0.40	38	0.40
10	0.40	0	0.25
7	0.35	92	0.70
51	0.50	82	0.60
2	0.30	20	0.35
1	0.25	40	0.40
52	0.50	44	0.45
7	0.35	25	0.35
Totals	9.75		8.20

Table 4, the second assembly requires 0.50 minute for *B*. We continue the simulated operation of the line in this way.

The sixth assembly becomes available to *B* at time 2.40, but *B* was ready for it at time 2.30. *B* therefore was forced to remain idle for 0.10 minute because of lack of work. The completed sample of twenty assemblies is worked out progressively.

The summary at the bottom of Table 5 shows the result in terms of the idle time in operation *B*, the waiting time of the parts, the average inventory between the two operations, and the resulting production rates. From the average times given by the original distributions, we would have guessed that *A* would limit the output of the line (since it was the slower of the two oprations). Actually, however, the line production rate is less than that dictated by *A* (116.5 pieces per hour compared to 123 pieces per hour for *A* as an individual operation). The reason is that the interplay of performance times for *A* and *B* do not always match up very well, and

TABLE 5

SIMULATED OPERATION OF THE TWO-STATION ASSEMBLY LINE WHEN OPERATION *A* PRECEDES OPERATION *B*

Assemblies Available for Operation B at	Operation B Begins at	Operation B Ends at	Idle Time in Operation B	Waiting Time of Assemblies	Number of Parts in Line, Excluding Assembly Being Processed in Operation B
0.00	0.00	0.60	0	0	0
0.40	0.60	1.10	0	0.20	1
0.85	1.10	1.50	0	0.25	1
1.35	1.50	1.95	0	0.15	1
1.80	1.95	2.30	0	0.15	1
2.40	2.40	2.75	0.10	0	0
3.25	3.25	3.60	0.50	0	0
4.00	4.00	4.30	0.40	0	0
4.65	4.65	5.00	0.35	0	0
5.30	5.30	5.60	0.30	0	0
5.75	5.75	6.15	0.15	0	0
6.30	6.30	6.65	0.15	0	0
6.70	6.70	7.10	0.05	0	0
7.10	7.10	7.35	0	0	0
7.45	7.45	8.15	0.10	0	0
7.95	8.15	8.75	0	0.20	1
8.25	8.75	9.10	0	0.50	1
8.50	9.10	9.50	0	0.60	2
9.00	9.50	9.95	0	0.50	2
9.35	9.95	10.30	0	0.60	2

Idle time in operation B = 2.10 minutes
Waiting time of parts = 3.15 minutes
Average inventory of assemblies between A and B = 3.15/9.35 = 0.34 assemblies

Average production rate of A $= \frac{20 \times 60}{9.75} = 123$ pieces/hour

Average production rate of B (while working) $= \frac{20 \times 60}{8.20} = 146$ pieces/hour

Average production rate of A and B together $= \frac{20 \times 60}{10.30} = 116.5$ pieces/hour

Note: In the above computations, 20 is the total number of completed assemblies; 9.75 is the total work time of operation A for 20 assemblies from Table 4; 8.20 is the total work time, exclusive of idle time, for operation B for 20 assemblies from Table 4.

sometimes B has to wait for work. B's enforced idle time plus total work time actually determine the maximum production rate of the line.

A little thought should convince us that, if possible, it would be better to redistribute the assembly work so that *A* is the faster of the two operations, thus reducing the probability that *B* will run out of work. This is demonstrated by Table 6, which assumes a simple reversal of the sequence of *A* and *B*. The same

TABLE 6
SIMULATED OPERATION OF THE TWO-STATION ASSEMBLY LINE WHEN OPERATION *B* PRECEDES OPERATION *A*

Assemblies Available for Operation A at	*Preparation A Begins at*	*Operation A Ends at*	*Idle Time in Operation A*	*Waiting Time of Assemblies*	*Number of Parts in Line, Excluding Assembly Being Processed in Operation A*
0.00	0.00	0.40	0	0	0
0.50	0.50	0.90	0.10	0	0
0.90	0.90	1.35	0	0	0
1.35	1.35	1.85	0	0	0
1.70	1.85	2.30	0	0.15	1
2.05	2.30	2.90	0	0.25	1
2.40	2.90	3.75	0	0.40	1
2.70	3.75	4.50	0	1.05	2
3.05	4.50	5.15	0	1.45	2
3.35	5.15	5.80	0	1.80	3
3.75	5.80	6.25	0	2.05	3
4.10	6.25	6.80	0	2.15	4
4.50	6.80	7.20	0	2.30	4
4.75	7.20	7.60	0	2.45	5
5.45	7.60	7.95	0	2.15	5
6.05	7.95	8.45	0	1.90	5
6.40	8.45	8.75	0	2.05	5
6.80	8.75	9.00	0	1.95	6
7.25	9.00	9.50	0	1.75	5
7.60	9.50	9.85	0	1.90	6

Idle time in operation *A* = 0.10 minute
Waiting time of parts = 25.75 minutes
Average inventory of asemblies between *A* and *B* = 25.75/7.60 = 3.4 assemblies

Average production rate of *A* (while working) $= \frac{20 \times 60}{9.75} = 123$ pieces/hour

Average production rate of *B* $= \frac{20 \times 60}{8.20} = 146$ pieces/hour

Average production rate of *A* and *B* together $= \frac{20 \times 60}{9.85} = 122$ pieces/hour

sample times have been used, and the simulated operation of the line has been developed as before. When the faster of the two operations is first in the sequence, the output rate of the line increases and approaches the rate of the limiting operation, and the average inventory between the two operations increases. Once there is a higher average inventory, the second operation in the sequence is almost never idle owing to lack of work. With regard to the balance of assembly lines, this conclusion is fairly general: the best labor balance will be achieved when each succeeding operation in the sequence is slightly slower than the one before it. This minimizes the idle time that is created when the operators run out of work because of the variable performance times of the various operations. In practical situations, it is common to find safety banks of assemblies between operations in order to absorb these fluctuations in performance.

Perhaps we wanted to build a more sophisticated model of the assembly line. Our simple model assumed that the performance times were independent of other events in the process. In the actual situation, the second operation in the sequence might tend to speed up when the inventory began to build up. We could have included this effect if we had known how inventory affected performance time.

Having followed this simulation example through carefully, we may feel strongly that it would work but be very tedious for problems of practical size. Even for our limited example, we probably would wish to have a larger run on which to base conclusions. In addition, there might be other alternatives to test. For example, several alternative ways to distribute the total assembly task between the two stations might exist; or more than two stations could be considered. Which of the several alternatives would yield the smallest incremental cost of labor, inventory costs, etc.? To cope with the problem of tedium and excessive work-hours to develop a solution, a computer may be used.

Statistical Analysis

A knowledge of statistics and probability provides a framework for the rigorous handling of data. Thus, we can not only draw conclusions based on the *predictive model* constructed, but also assess the risk that the forecasts or predictions may be in error. Therefore, a correlation study of two factors—such as paint viscosity and paint defects—may indicate that, although the two factors do move together, the probable error may be very large if we attempt to forecast the number of paint defects expected from the viscosity measured. Similarly, a statistical control chart of errors in a clerical operation or of labor turnover may indicate that we are experiencing an abnormal situation that is either above or below accepted limits of variation. However, we also know that the chance of being mistaken in assuming that the process is out of control is only 0.27 percent if the usual probability control limits have been assigned.

One great contribution of modern statistics to production analysis lies in the general field called statistical inference. Here, we find the development of a methodology by which to test hypotheses formally. This field is immensely valuable because it allows us to deal with problems that exhibit great variation in the measured values of the factors or variables that may define the system. Yet we can draw conclusions about the system, and these conclusions may be quite precise. Thus, it may be possible to test the hypothesis that a new method for a clerical operation is faster or produces fewer errors. Or we may test the somewhat more complex hypothesis that, of several factors that may contribute to poor quality, only one is significant.

Suppose, for example, that there are four machines used to perform the same operation in the production of a part, and that we are experiencing difficulty with the quality of output. The foreman maintains that his four machines are not equally reliable. The plant superintendent accuses the machine operators, on the theory that one simply cannot obtain qualified help these days. The workers say that the raw material is sometimes faulty, especially when it comes from a certain supplier. When the quality reports are examined, it is not at all obvious whether the machines, workers, or suppliers are at fault. However, by resorting to a statistical inference technique, *analysis of variance*, it is possible to perform a simple experiment that will tell us—with a preassigned level of confidence—the contributions toward poor quality from each of the three sources.

Statistical analysis in production/operations management has its own large field of application. For example, techniques such as *work sampling* and *statistical quality-control*, are based on statistical models. As a general tool of analysis, however, statistical concepts often help us to apply some of the other kinds of analytical techniques that we already have described. For example, in attempting to determine what factors or variables are significant in a simulation model of flow through a production system, we may test the hypothesis that the time required for flow through the system depends on the number of operations to be performed. Or, in studying the nature of an existing wage structure, we may wish to determine the regression line of a measure of job difficulty versus wages paid. In neither of these instances is statistical analysis the central model used: however, it contributes to the overall analysis.

Cost Models

Cost models are still the most common mode of predicting performance in managerial problems. They take many forms and are based on a knowledge of the behavior of cost factors. Predicting performance is not simply a matter of gathering cost information from accounting data: many of the accounting data are significant only within the accounting system. We are interested in the actual behavior of pertinent costs under the alternatives considered. *Break-even analysis* utilizes the differences in behavior of certain costs as the volume of operation varies. Such analysis helps us establish a proper volume of operation, diagnose problems, and evaluate the appropriateness of a productive process.

One of the most valuable of the simple techniques is *incremental cost analysis*, which deals only with those costs affected by alternate policies or actions considered. The problem of alternate production programs is an example of incremental cost analysis. We do not attempt to evaluate the alternative costs by calculating total operating costs with alternate plans. Such an approach would have been tedious and time consuming. Instead, we determine the levels of only those costs that are affected by the alternative programs: inventory costs, labor turnover costs, overtime premium, and extra subcontracting costs. Incremental cost concepts are useful in virtually all areas of analysis of productive systems. They commonly are found in some of the formal mathematical models, such as linear programming, waiting line analysis, and others.

Another kind of cost problem occurs when capital assets are considered in the alternatives generated. Cost behavior is somewhat different in nature under these circumstances, and requires special treatment. The question of when to replace a capital asset is a related problem. Since almost all productive systems depend heavily on the use of capital equipment, the handling of these kinds of cost items is important.

Cost-predictive models, including incremental costs, breakeven analysis, and capital costs, are covered separately in Appendix A.

EVALUATION MODELS

Given the predictive model that produces outcomes for various alternatives, the manager needs a way to evaluate the alternatives. Here, the criteria and values discussed in Chapter 2 become the focus.

The choice of the appropriate evaluation model depends on the circumstances under which the decision is being made. One of the important considerations is whether the outcome is known with certainty or only in terms of a probability distribution. When the outcomes of the predictive model are known with certainty—and when we have a single decision criterion, such as minimum costs or maximum profits—the selection of the appropriate decision is relatively simple. We select the alternative that produces the minimum cost or maximum profit. In decision theory terms, the evaluation model is termed *decision making under certainty*.

When the outcomes from the predictive model are known only in terms of a probability distribution, the outcomes must be weighted according to their respective probabilities of occurrence, and the decision criterion becomes the expected value of outcomes. In decision theory terms, such evaluation models are termed *decision making under risk.*

A second important consideration in defining the appropriate evaluation model is the number of different criteria that are relevant for a decision's desirability. As we noted in chapter 2, many, if not most, managerial problems involve multiple goals; therefore, multiple criteria must be considered. The appropriate evaluation model then must consider the trade-offs among the criteria. When both quantitative and nonquantitative criteria exist, a common treatment is to set up a formal procedure to establish the manager's utility function, relating the amount of increased cost that the manager is willing to accept in order to achieve other goals (see Chapters 2, 3, and 4 in Buffa and Dyer [1977]).

NORMATIVE (OPTIMIZING) MODELS

Normative models are among the most exciting and potentially valuable methods of management science: given a criterion, these models combine the elements of alternative generator, predictive, and evaluation models in such a way that the *best possible* solution can be determined. Best possible means that, given the criterion, no superior combination of the decision variables for the model esixts. These are management science's powerful optimizing methods, such as mathematical programming, network optimization, inventory models, and others.

Implicit in the normative or optimizing kind of model is the assumption that the submodel that predicts system performance validly represents how the system works, and that the evaluation model reflects the decision maker's utility function. These are important assumptions. It may be difficult to meet them completely when we are combining the functions of alternative generator, system performance prediction, evaluation, and optimization into one, large-scale system.

Let us recall the scheme of models in management science presented in Figure 1, and shown in slightly different relationships in Figure 9. In Figure 9, the alternative generator, predictive model, and the evaluation model are enclosed to form a system; and a line connects the evaluation model to the alternative generator. We state that the output of the combined system is an optimum solution. Normative models combine these three elements in such a way that we can determine the best combination of the decision variables. This is what

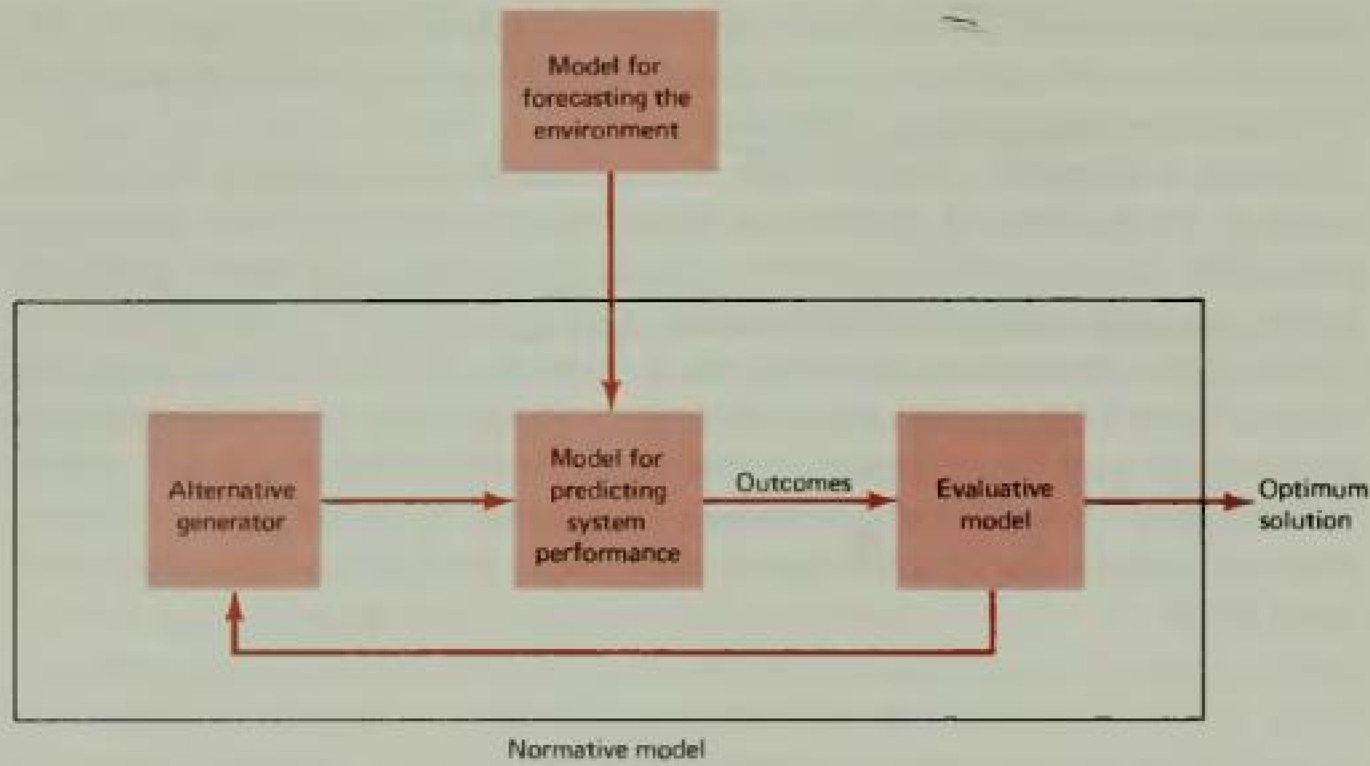


Figure 9. Elements of a normative or optimizing model

SOURCE: E. S. Buffa and J. S. Dyer, *Management Science/Operations Research: Model Formulation and Solution Methods* (Santa Barbara, Ca.: Wiley/Hamilton, 1977)

normative means—a way of establishing a norm or a standard. In the normative models that are useful in operations management, each of the different elements of Figure 9 will not have a separate and distinct identity; yet the functions will be performed.

Figure 10 shows a flow chart that contains all the elements of Figure 9. However, below the evaluation model, we have inserted a test for optimality by which to determine whether or not the solution just produced can be improved. If it can be improved, we need some mechanism to determine the direction and amount of change we should make in the decision variable for the next iteration. This information is sent back to the alternative generator to produce a new alternative. System performance is predicted and evaluated, and the cycle is repeated until the test for optimality indicates that no further improvement is possible. We have diagrammed the process as an iterative one. However, although some of the solution techniques actually do iterate in the manner we have described, others combine these steps and go directly to the optimum solution (as with certain optimum inventory models).

Now we will review some of the most important normative models. These methods will be described in more detail, in terms of their development and use, later on in the text.

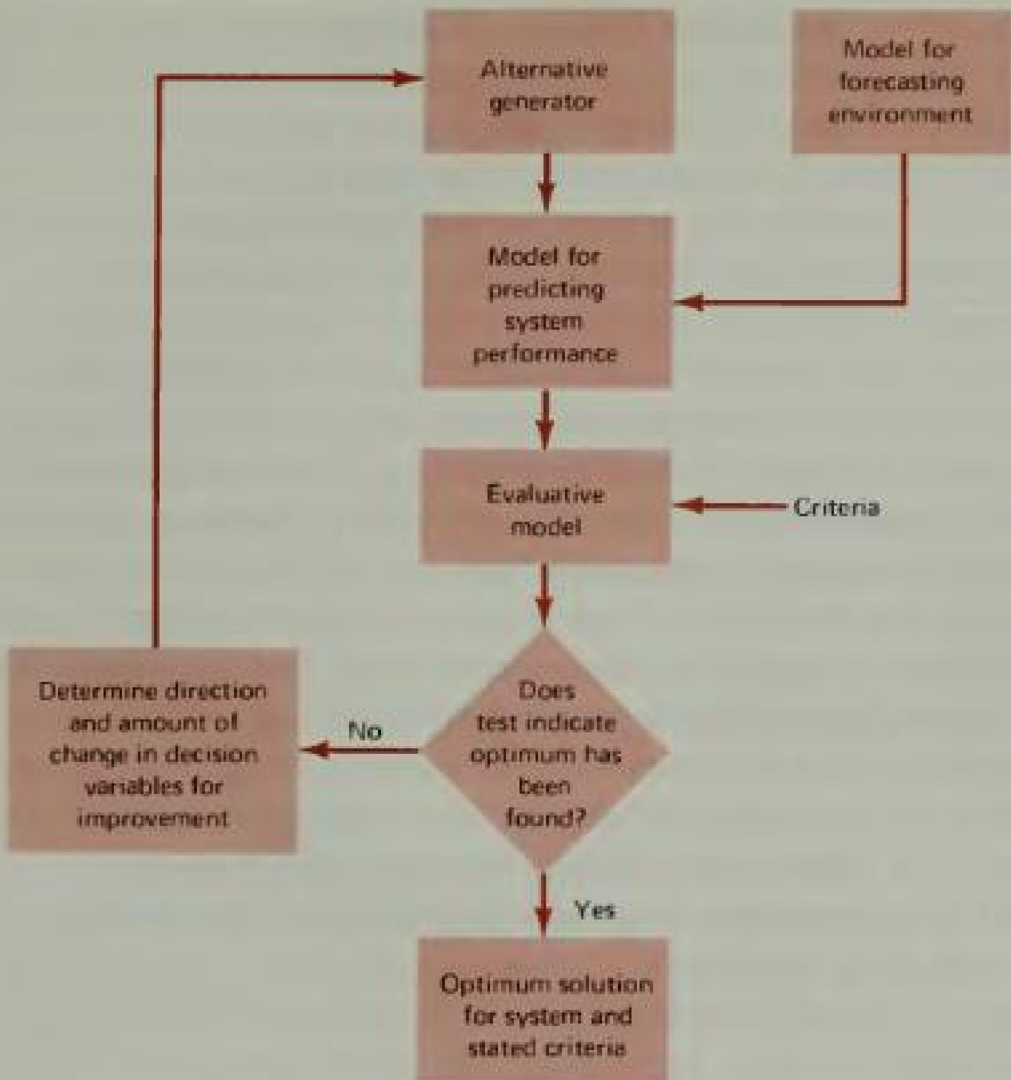


Figure 10. Flow chart showing test for optimality and coupling to alternative generator

SOURCE: E. S. Buffa and J. S. Dyer, *Management Science/Operations Research Model Formulation and Solution Methods* (Santa Barbara, Ca.: Wiley/Hamilton, 1977)

Linear Programming

Linear programming is representative of the new methodologies that come to us from mathematics and economics. The field of application of this relatively new and very important general model is in dealing with the problems of allocating an organization's scarce resources.

Simplex methods of linear programming are discussed in Appendix B, and distribution methods are covered in Appendix C. Here is a discussion of a few typical problems in which linear programming may be useful:

1. Allocation of limited raw materials used in a variety of products so that total profit is maximized, meeting market demands as far as possible.

2. Allocation of production facilities when alternate routings are available. Given the: unit machine time for the alternate machine routes; total hours available on the different machine classes; requirements for the number of each product; and unit revenue for each product, linear programming can give a solution that maximizes some profit function, minimizes incremental costs, or meets some other management objective.
3. Blending problems. For example, a paint manufacturer may need to prepare paint vehicles that are a blend of several constituents. The constituents, such as oil and thinner, are available in limited quantity and in commercial blends of fixed proportions. Costs per gallon of the various possible raw materials are known. The problem is to determine the amount of each raw material so that required amounts for the new blends are obtained at minimum cost. Another similar problem is the blending of animal feed to provide certain minimum nutrient values at minimum cost.
4. Maximizing material utilization. Many times, different stock sizes must be stamped or cut from standard raw material sizes. The problem is to determine the combination of cuts that will meet requirements for the amounts of different sizes with a minimum of trim loss.
5. Developing a program for production when demand is seasonal. Here we are attempting to allocate to various production periods available capacity for the products to be produced in such a way that requirements for all products are met, and combined incremental inventory and production costs are minimized. The incremental production costs may include overtime premium, turnover costs, and extra subcontracting costs.
6. Product mix problems. This is an interesting class of problems. If we have production facilities that can be used to produce several different items which may have different costs, revenues, and market demands, we wish to know how best to allocate the available capacity to various products within the limitations of market demand.
7. Long-range planning. The general capacity planning problem usually is met by ownership, leases, or short-term contracts. A model cast in a linear programming framework may help answer questions such as: (a) the effect of a given demand forecast on capacity plans; (b) the effect of changes in ownership cost; (c) the effect of changes in the costs of leased capacity; and (d) the sensitivity to forecast error of various decisions and costs.
8. Distribution of products from a set of origin points to a number of destinations so that demand at each destination and supplies available at the origins are satisfied, and that total transportation costs are minimized.

9. Distribution of products from factories to warehouses, similar to 8, but minimizing combined production and distribution costs. If the products have different revenues in the various marketing areas, we could maximize a function of revenue minus production-distribution costs.
10. Multiple plant location studies where common products are produced in a decentralized complex of plants. Here we wish to evaluate various alternate locations for the construction of a new plant. Each different location considered produces a different allocation matrix of product from the factories to the distribution points because of differing production-distribution costs. The best new location is the one that minimizes total production-distribution costs for the entire system. This is not necessarily the location that seems to have the lowest production costs.
11. Locational dynamics for multiple plants. The problem is somewhat similar to 10, but here the question is which plants to operate at what levels for a given total demand. Since additional capacity at each location normally can be obtained through overtime, and since certain overhead costs can be saved by shutting down a plant, there are conditions in which shutting down a plant and supplying the total demand from the other plants minimizes total costs, even though this incurs overtime costs. The plant to be shut down is not necessarily the high production cost plant. The relative importance of production and distribution costs determines which plant should be shut down.
12. Redistribution of empty freight cars from their existing locations to points where they are needed in a way that minimizes transportation costs.

Network Planning Models

The post-World War II importance of research and development, and other large-scale, one-time projects in our economy, have called forth special planning techniques. These techniques, commonly known as PERT, represent the work that is required to be done in terms of a network of activities which recognizes that the timing and phasing of operations are interdependent. From the basic network model, schedule statistics can be computed that make it possible to determine which activities must be done on time (are critical) and which can allow for schedule slack. The result is a concept known as *the critical path schedule*, which is based on principles of network optimization. The network of activities and the schedule statistics then become a basis for planning the deployment of resources to the project.

We shall discuss network planning and scheduling methods in Chapter 13, under the topics of planning, scheduling, and controling large-scale projects.

Computer Search Methods

Another relatively new approach for obtaining excellent solutions to some very complex problems is through the use of computer search methods. In the past, many operations management problems were so complex that the model builder had to either build only simple models that could be solved by analytic techniques, or resort to simulation or possibly heuristic methods. Today, however, the computer has enabled us to use new quasi-analytic and heuristic search techniques. These techniques have significantly increased the probability of finding the global optimum of complex models. One such optimum-seeking technique is known as a direct computer search procedure. "Direct search" methods sequentially examine a finite set of feasible trial solutions of a criterion function. A single trial evaluation is produced by specifying values for each independent variable, evaluating the criterion function, and recording the result. Each trial value is compared to the best previous value; if an improvement is observed, the trial value is accepted and the previous best value is rejected. The procedure continues in this way until no further improvements can be found (when a predetermined number of trial evaluations has been made or when the computer time limit is exceeded). At this point, the computer program prints out the best combination of independent variables that has been found.

The advantage of using direct search methods is in building the model of the criterion function. Constraints of mathematical form (for example, linearity) are less important. To date, problems have been solved with as many as 120 independent variables, using modest computer time, and it appears that the possible number of independent variables in such programs can be expanded even more. In operations management, direct search methods have been applied to the aggregate planning and scheduling problem which we shall discuss in Chapter 10, and to the network scheduling problem with limited resources, that we shall discuss in Chapter 13.

Summary

In this chapter we have attempted to lay out a general framework for analyzing productive systems, and to discuss the nature and application of some prominent modes of analysis. In most instances, specific methods of analysis have been discussed as an introduction to other sections or chapters in the book.

General systems concepts, and the scientific method approach to any problem, provide the general framework for analysis. The scientific method, as applied to operations management, is summarized as follows:

1. Define the system under study.
2. Define a measure of effectiveness.
3. Construct a model where E, the effect, can be expressed as a function of the variables that define the system.
4. Generate alternatives which are based on analysis.
5. Weigh and decide.

The analytical methods useful in production/operations management are classified by Figure 1. They fall into the general categories of predictive, evaluation, and normative models. Normative (or optimizing) models involve joining the alternative generator, predictive model, and evaluation model to produce optimum solutions for the model.

Review Questions

1. Outline the analytical process for production/operations management problems. Where in the process do we open ourselves to the dangers of suboptimization?

2. Compare the analytical process with the kind of model, such as alternative generator, forecasting model, model for predicting system perfor-

mance, evaluation model, and normative or optimizing model. In which ways are there parallels between the two?

3. Define the following terms:
   a. Alternative generator.
   b. Predictive model.
   c. Evaluation model.
   d. Normative or optimizing model.

4. Discuss the nature of forecasting models in terms of the kinds of applications.

5. Classify computer based corporate models in terms of the "kinds of models" scheme that we have established. Why have they been developed in interactive mode?

6. Contrast the nature and use of the following predictive models: systems dynamics, queuing models, computer based corporate models.

7. Describe each of the following as a queuing or waiting line problem:
   a. Doctor's office.
   b. Toll bridge.
   c. Teller's window at a bank.
   d. Assembly line.

8. As simulation models, what is the difference between the computer based corporate models described and Monte Carlo?

9. Outline the process for simulated sampling.

10. Discuss the concept of incremental cost analysis. How does the concept couple with formal models, such as waiting line, simulation, and linear programming models?

11. Define the terms:

a. Decision making under certainty.
b. Decision making under risk.

12. Describe how normative or optimizing models combine the alternative generator, predictive model, and evaluation model into an integrative system to produce the best possible answer for the model.

13. What is the general field of application for linear programming? Network planning models? Computer search methods?

Problems

1. The Caribbean Tea Company is a nationwide chain of supermarkets, with headquarters in Chicago, Illinois. The vice-president in charge of western operations is located in Los Angeles, and controls all areas west of the Rocky Mountains. The largest portion of the western operations is centered in the greater Los Angeles metropolitan area.

Although the Los Angeles population is increasing at a rapid rate, total company sales have been leveling off. The vice-president believes that the primary cause of this condition is increased competition. Recent association publications report that the number of supermarkets is growing at a faster rate than the population.

The vice-president believes that, to meet this impending crisis, the only answer lies in becoming more competitive. In pursuit of this objective, he has formed a task force to deal with the following problems areas:

a. Product line.
b. Pricing.
c. Labor saving equipment.
d. Advertising.
e. Improved customer service.

As part of this effort, the vice-president calls you in and says: "I want you to investigate customer service at the check-out stands and make recommendations. I don't want to influence your judgment, but it seems strange to me that when I go into a store I frequently see long lines and about one-half the check-out stands not being used. I think we need to do some hard thinking about how many checkers we should have. I want you to prepare a preliminary proposal of how you intend to investigate customer service at the check-out stands, and submit it at the next task force meeting."

a. Translate the vice-president's instructions into a meaningful statement of the problem as you see it.
b. Outline a proposal to solve the problem that you have stated in Question a. Prepare your proposal in sufficient detail so that a third party could implement it with no further instructions from you.

2. You are studying a two-station assembly line in which each operator does approximately half the work. The distributions of assembly times are as follows:

Cycle Time (Minutes)	Frequency of Cycle Times for Operation *A*	Frequency of Cycle Times for Operation *B*
0.20	4	2
0.30	9	5
0.40	28	17
0.50	22	10
0.60	15	8
0.70	10	2
0.80	8	3
0.90	4	3
	100	50

a. Construct a cumulative probability distribution that could be used in the simulation of the two-station assembly line.
b. Simulate the assembly of ten parts.
c. What is the average length of the waiting line of the half-completed assemblies ahead of station *B*?
d. What is the average output per hour of the assembly line?

3. A sample of 100 customer arrivals at a check-out station of a small store is represented in the following distribution:

Time between Arrivals (minutes)	Frequency
0.5	2
1.0	6
1.5	10
2.0	25
2.5	20
3.0	14
3.5	10
4.0	7
4.5	4
5.0	2
	100

A study of the time required to service the customers by adding up the bill, receiving payment, making change, placing packages in bags, etc., yields the following distribution:

Service Time (minutes)	Frequency
0.5	12
1.0	21
1.5	36
2.0	19
2.5	7
3.0	5

Set up the foregoing data in a form that is useful for simulation. Determine a simulation procedure that will use this revised data to furnish an estimate of the mean number of customers serviced per minute. Develop a flow chart that completely specifies the decision rules used for the simulation.

4. The WASH-N-DRY Manufacturing Company produces a line of washing machines and dryers. The major manufacturing departments are the Stamping Department, the Motor and Transmission Department, and final assembly lines for each product. The Stamping Department fabricates a large number of the metal parts both for the washer and the dryer. The Motor and Transmission Department produces the drive units for both product lines. Monthly department capacities are:

Department Name		
Stamping Department	500 washers or 500 dryers	
Motor and Transmission Department	400 washers or 600 dryers	
Washer Assembly Line	300 washers	—
Dryer Assembly Line	—	250 dryers

Plot the restrictions on a graph. Which department capacities are likely to limit the output of the plant as a whole? What is likely to be the best number of washers and dryers to produce, assuming that the market is not restricting? Why is this number likely?

References

Bierman, H.; C. P. Bonini; and W. H. Hausman. *Quantitative Analysis for Business Decisions*. 5th ed. Homewood, Ill.: Richard D. Irwin, 1977.

Boulden, J. B., and E. S. Buffa. "Corporate Models: On-Line, Real-Time Systems." *Harvard Business Review*, July-August 1970.

Buffa, E. S. *Operations Management: Problems and Models*. 3d ed. New York: John Wiley & Sons, 1972.

———, and J. S. Dyer. *Management Science/Operations Research: Model Formulation and Solution Methods*. Santa Barbara, Ca.: Wiley/Hamilton, 1977.

Cole, H. S. D.; C. Freeman; M. Jahoda; and K. L. R. Pavitt, eds. *Models of Doom*. New York: Universe Books, 1973.

Forrester, J. W. *Industrial Dynamics*. Cambridge, Mass.: MIT Press, 1961.

———. *Urban Dynamics*. Cambridge, Mass.: MIT Press, 1969.

Hoel, P. G., and R. J. Jessen. *Basic Statistics for Business and Economics*. 2d ed. Santa Barbara, Ca.: Wiley/Hamilton, 1977.

Levin, R. I., and C. A. Kirkpatrick. *Quantitative Approaches to Management*. 3d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Meadows, D. L.; D. H. Meadows; J. Randers; and W. W. Behrens, III. *The Limits to Growth*. New York: Universe Books, 1972.

Naylor, T. H., and H. Schauland. "A Survey of Users of Corporate Planning Models," *Management Science* 22, May 1976, 927–937.

Reisman, A. *Managerial and Engineering Economics*. Boston: Allyn and Bacon, 1971.

Thierauf, R. J., and R. C. Klekamp. *Decision Making Through Operations Research*. 2d ed. New York: John Wiley & Sons, 1975.

Part III

Design of Productive Systems

Chapter 5

Design of Products and Services

The existence of productive systems depends on the existence of some needed product or service. Therefore our discussion of the design of such systems must consider the effect of the product or service design on the system. In reality, the effect occurs in both directions: the productive system must be designed to suit the product or service being produced, and the design of the product or service must take account of process capabilities, costs, materials, organization, and so on. It is an endless cycle, and the most successful organizations take advantage of the interactions among these functions to design a product-process system.

TECHNOLOGICAL INNOVATION

The application of scientific discovery to products, services, and processes commonly is thought of as a chain of events: scientific discovery, invention, development, innovation, and finally, application. Long time lags frequently occur in this chain, as enabling conditions pace developments. For example, the general processes for solar energy conversion have been known for some time, but economic factors did not justify their costly development. Similarly, shale-oil extraction processes are well known, but cheaper sources of crude oil and other energy forms have been used instead. The developing shortage of conventional energy sources and the rapid increase in energy prices makes the application of shale-oil processes more likely.

The chain-event model does not help explain the driving forces that produce the innovations, that are converted to applications near the end of the chain. Abernathy and Townsend suggest that product innovation, process innovation, and changes within a productive segment of industry appear to feed on one another:

> No single external force, such as market factors or technological factors is dominant in stimulating technological innovation. Sources of stimulation that arise within a productive segment are more frequently the critical factor that sparks technological innovation. . . . Historical patterns of development in several productive segments suggest that the efforts of engineers and managers in improving production processes themselves, may be a key factor in stimulating technological innovation [Abernathy and Townsend, 1975].

An Industrial Example of the Structure of Technological Innovation

It commonly has been held that most innovations are market stimulated, and usually are applied to new products rather than to productive processes. The results of one study in the railroad industry are shown in Table 1 [Meyers and Marquis, 1969]. Note that the largest stimulation source of innovations is in the market, and the greatest application impact is on products.

Abernathy and Townsend, however, analyzed the same data, taking into account the vertical integration structure of the industry. Their results suggest a very different conclusion. Figure 1 shows the flow of innovations allocated to two levels of industrial integration for the railroad industry, based on the data from Table 1. The dotted lines with arrows leading into the circles show the source of innovation stimulation, and the solid lines emerging from the circles indicate the frequency and area of application impact. Note that "market factors" represent the primary source of stimulation for 53 percent of innovations produced by the equipment manufacturers. However, market factors in that industrial segment are the same as process equipment needs for the transportation companies and users of transportation service. The greatest impact of innovation is upstream. Abernathy and Townsend note that:

> Most innovations in the lower levels of the vertical integration chain are *product* innovations. These are at the same time *process* innovations for processes at higher levels of vertical integration and as such have direct productivity implications. In fact,

TABLE 1
SOURCE AND IMPACT OF SUCCESSFUL INNOVATIONS IN THE RAILROAD INDUSTRY

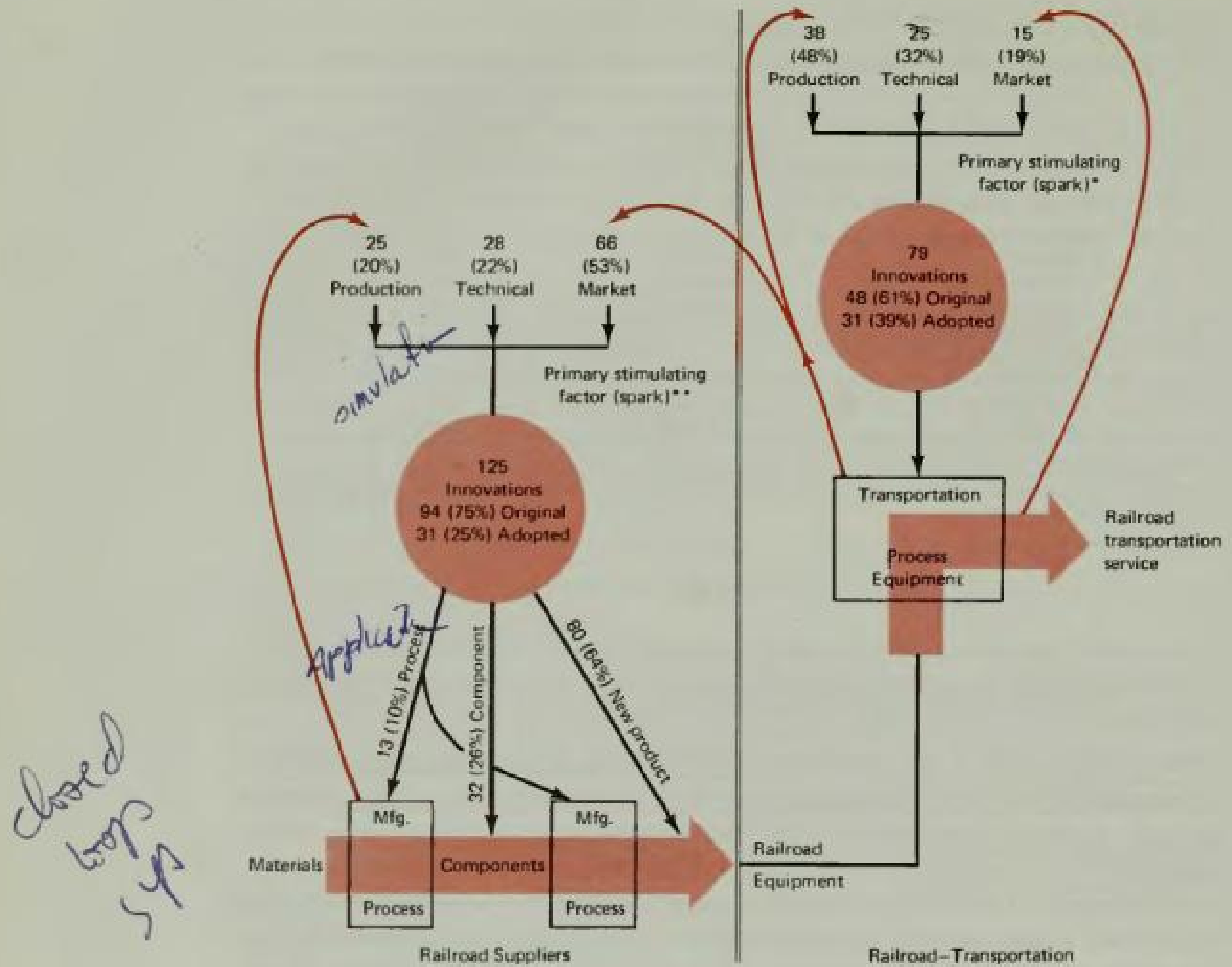
	Equipment Manufacturers	*Transportation Companies*
Number of innovations	125	79
Stimulation source		
Market	66 (53%)	15 (19%)
Production	25 (20%)	38 (48%)
Technical	28 (22%)	25 (32%)
Administrative	6 (5%)	1 (1%)
Application impact		
Product	80 (64%)	
Component	32 (26%)	
Process	13 (10%)	79 (100%)

SOURCE: S. Meyers and D. Marquis, *Successful Industrial Innovations* (Washington, D. C.: National Science Foundation, 1969), pp. 69–70.

from a strict perspective of the process at the highest level of vertical integration (users of transportation), *all* of the innovations considered here are process innovations and *all* have implications for process productivity [Abernathy and Townsend, 1975].

The culmination of the industrial structure is a service industry—providing transportation service to users—provides an example of the interlinking between a service industry and those manufacturing industries that provide equipment for the processes. Figure 1 in effect represents a closed-loop system for innovation in the industry. The meaning of market stimulation is innovation stimulation resulting from an opportunity to serve the process needs of an organization whose position in the integration structure is high. For railroads as a whole, 129 or 63 percent of all innovations were stimulated by opportunities for process change, either within the particular industrial segment or in another process-related segment (66 market-stimulated by process needs in the transportation companies, and 25 plus 38 stimulated by production within equipment manufacturer and transportation companies). Market factors, if external to the process, are the source of stimulation in 15 innovations (7 percent of total innovations), and technical factors account for 53 innovations (26 percent of the total).

Abernathy and Townsend [1975] in addition performed a comparable analysis of the computer industry (which also culminates in a service) with similar results.



* An additional 1% were stimulated by "administrative factors."
** An additional 5% were stimulated by "administrative factors."

Figure 1. Technological innovation in the railroad industry

SOURCE: W. J. Abernathy and P. L. Townsend, "Technology, Productivity and Process Change," *Technological Forecasting and Social Change* 7 (1975). Data from Meyers and Marquis [1969].

Interaction Between Product- and Process Innovation

We already have alluded to the iterative nature of the interaction between product design and productive system design, in which each affects the determination of the other's design. This iterative process also takes place on a macro level within industries (as shown by Figure 1), and the concept includes the design of services as well as manufactured products. In other words, the nature of services offered is affected by the productive process and vice versa, and so on back through the chain. We will develop this interaction process more fully later in the chapter. At this point, however, we will discuss the interaction as part of the overall innovation process.

A Model of Process- and Product Innovation

Utterback and Abernathy [1975] developed a dynamic model of process- and product innovation in firms, and tested it on empirical data developed by Meyers and Marquis [1969]. The model relates the product- and process innovations to three stages of development.

Stage 1. The first stage begins early in the life of products and services and of processes; initially, the innovations are stimulated by needs in the marketplace. Process innovations also are stimulated by the need to increase output rate (see Figure 2). In terms of innovation rate, product innovation is high and the initial emphasis is on performance maximization, perhaps in the anticipation that new capabilities will, in turn, expand requirements in the marketplace.

While we may think largely in terms of physical products, service innovations are quite comparable: for example, the initial introduction of innovative services such as social security, no-fault auto insurance, comprehensive health services (e.g., Kaiser Permanente), fast-food services, and so on.

Utterback and Abernathy call the first phase "performance maximization" for products and services, and "uncoordinated" for processes. High product innovation rates increase the likelihood that product diversity will be extensive. As a result, the productive process is composed largely of unstandardized and manual operations, or operations that rely on general purpose equipment. The productive system is likely to be of the intermittent type, but the characterization—uncoordinated—is probably justified in most instances because the relationships between the required operations are still not clear.

Stage 2. Price competition becomes more intense in the second stage, as the industry or product and service group begin to reach maturity. Productive system design emphasizes cost minimization, as competition in the marketplace begins to emphasize price. The productive process becomes more captial intensive and more tightly integrated through production planning and control. At this stage, the production process often is segmented in nature, partly because

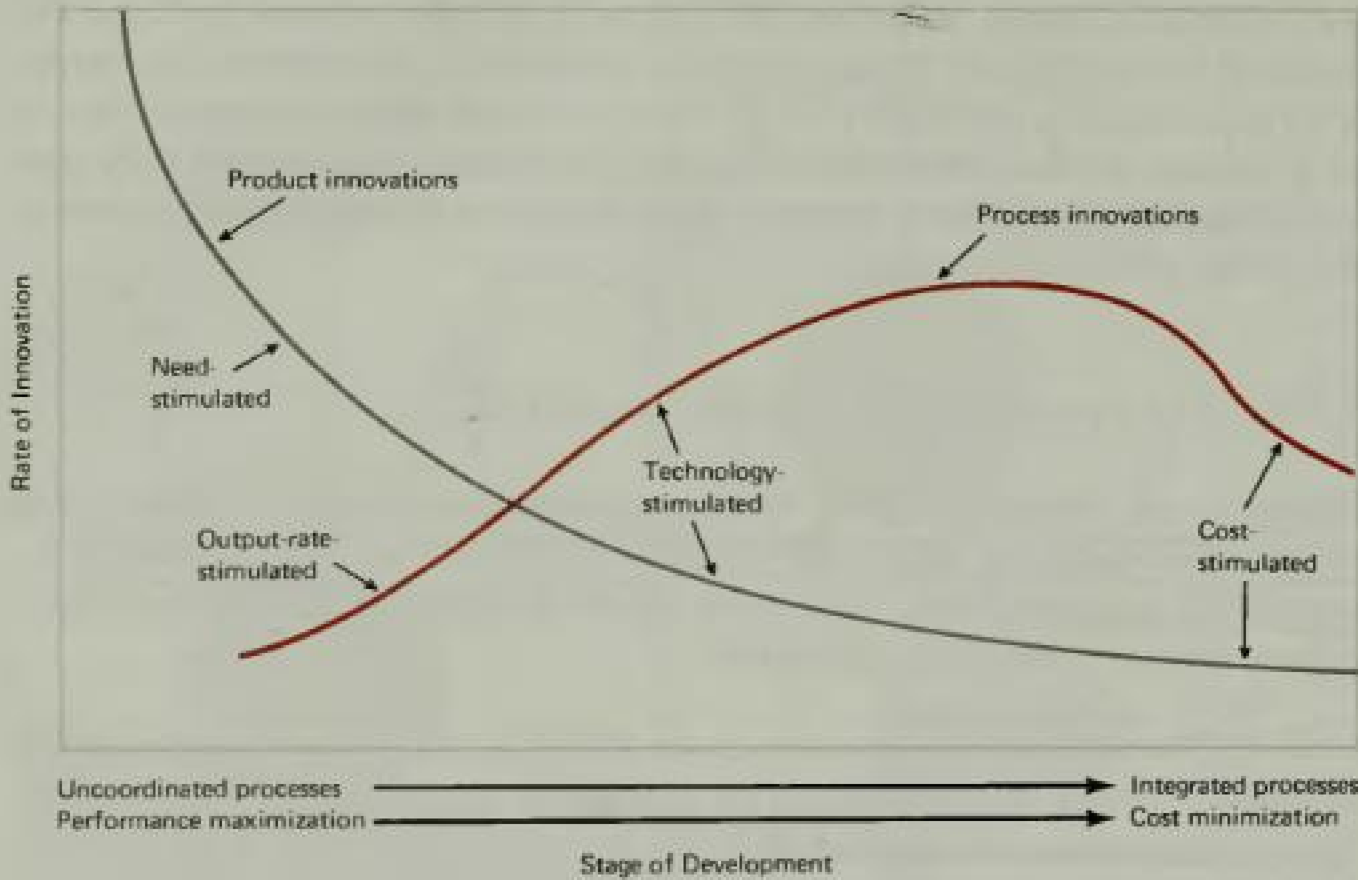


Figure 2. Relationships of product- and process-innovations in a dynamic model

SOURCE: J. M. Utterback and W. J. Abernathy, "A Dynamic Model of Process and Product Innovation by Firms," *Omega*, 1975.

integration is taking place at a broader level through managerial control systems and partly because the dominant system type is the intermittent system. As shown in Figure 2, process innovations dominate; however, both process- and product innovations are stimulated by technology.

Stage 3. Finally, as the entire system reaches maturity and saturation, innovations tend to be largely cost stimulated, as indicated in Figure 2. Further price competition puts increasing emphasis on cost minimizing strategies, and the production process becomes even more capital intensive. The productive process becomes more highly structured and integrated, as illustrated by automotive assembly lines, continuous chemical processes, and such highly automated, large-scale, service systems such as social security. The productive process becomes so highly integrated that it is difficult to make changes because any change at all creates significant interactions with other operations in the process.

The Utterback–Abernathy model of innovation indicates the close relationship between the design and development of products and services, and the productive system design. In fact, during the third, or cost minimizing, stage, the effects of innovation on product cost follow a surprisingly clear pattern. This is indicated in the Ford study which follows.

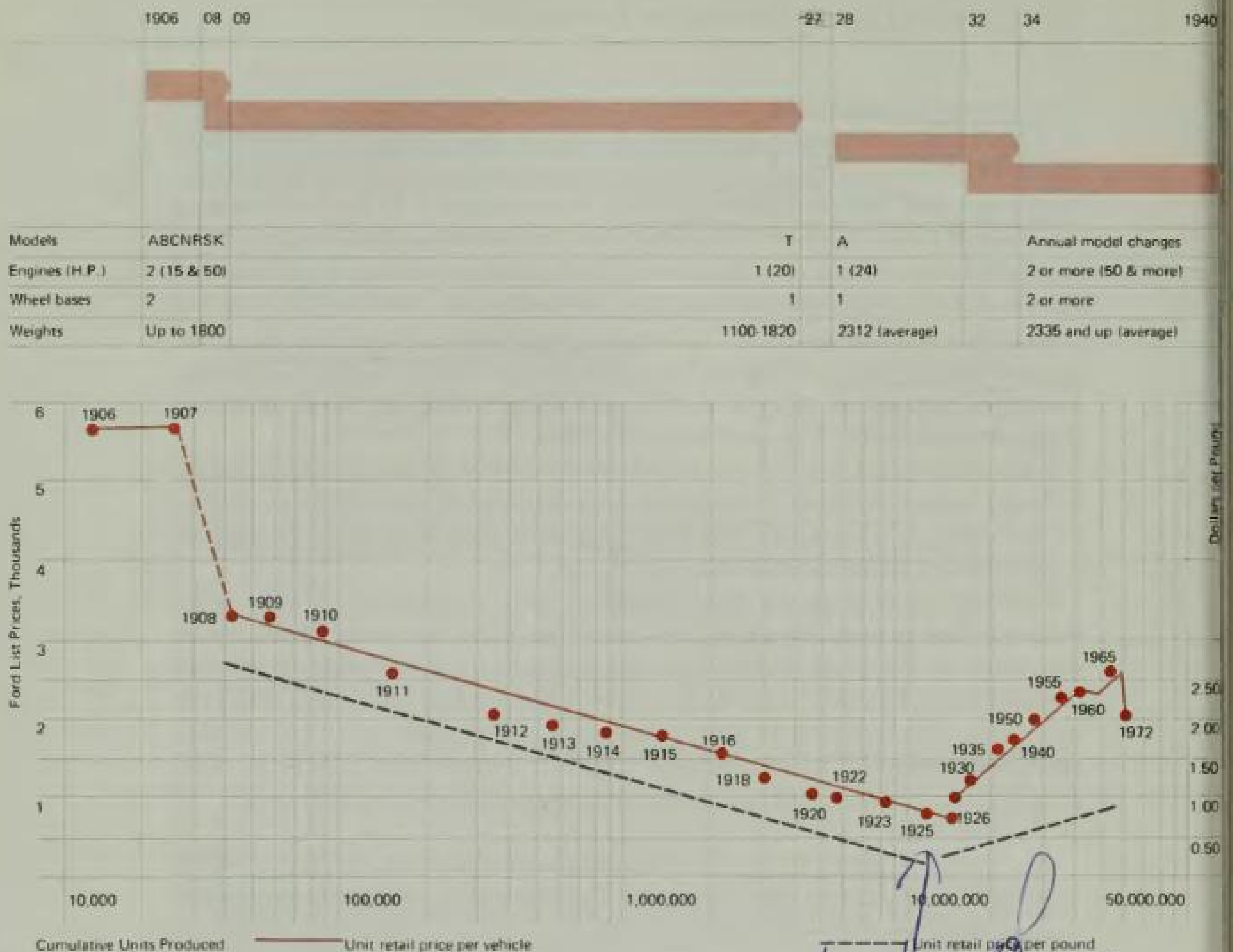
The Learning Curve—Innovation Interaction

One conceptual framework that many manufacturing organizations have used as a basis for a production-marketing strategy is the learning (or experience) curve. The concept is that product costs decline systematically by some fixed percentage with the doubling of volume. For example, a 90 percent learning curve would be one in which the product cost for a doubling of volume would be 90 percent of the former cost. Given such a relationship, managers can increase market share through price competition, depending on costs to decrease according to the learning curve. The learning curve model formalizes the "economics of scale" concept. Abernathy and Wayne [1974] developed a fascinating history of innovation, process change, and organization for the Ford Motor Company, part of which is summarized by Figure 3.

Figure 3 shows the price decline (in 1958 dollars) of the famous Model T during its long product life cycle. The price decline culminated in a costly conversion to the Model A, and finally, the price increases associated with the annual model changes began in about 1932.

Beginning in 1908, Henry Ford embarked on a conscious policy of price reduction that reduced the price from more than $5,000 to nearly $3,000. From that point on, the price decline was characterized by an 85 percent learning curve during the Model-T era. Market share increased from 10.7 percent in 1910 to a peak of 55.4 percent in 1921 [Abernathy and Wayne, 1974]. During this spectacular period of stable product design, innovations were largely process- or production oriented. Abernathy and Wayne chronicled them as follows:

> The company accomplished savings by building modern plants, extracting higher volume from the existing plant, obtaining economies in purchased parts, and gaining efficiency through greater division of labor. By 1913 these efforts had reduced production throughput times from 21 days to 14. Later, production was speeded further through major process innovations like the moving assembly line in motors and radiators and branch assembly lines. At times however, labor turnover reportedly ran as high as 40 percent per month.
>
> Up to this point, Ford had achieved economies without greatly increasing the rate of capital intensity. To sustain the cost cuts, the company embarked on a policy of backward and further forward integration in order to reduce transportation and raw materials costs, improve reliability of supply sources, and control dealer performance. The rate of capital investment showed substantial increases after 1913, rising from 11 cents per sales dollar that year to 22 cents by 1921. The new facilities that were built or acquired included blast furnaces, logging operations and sawmills, a railroad, weaving mills, coke ovens, a paper mill, a glass plant, and a cement plant.
>
> Throughput time was slashed to four days and the inventory level cut in half, despite the addition of large raw materials inventories. The labor hours required of unsalaried employees per 1000 pounds of vehicle delivered fell correspondingly some 60 percent



Δ here began Chevrolet entrance

Figure 3. The Ford experience curve (in 1958 constant dollars)

SOURCE: W. J. Abernathy and K. Wayne, "Limits of the Learning Curve," *Harvard Business Review*, September–October 1974, pp. 109–119.

during this period, in spite of the additions to the labor force resulting from the backward integration thrust and in spite of substantial use of Ford employees in factory construction.

Constant improvements in the production process made it more integrated, more mechanized, and increasingly paced by conveyors. Consequently, the company felt

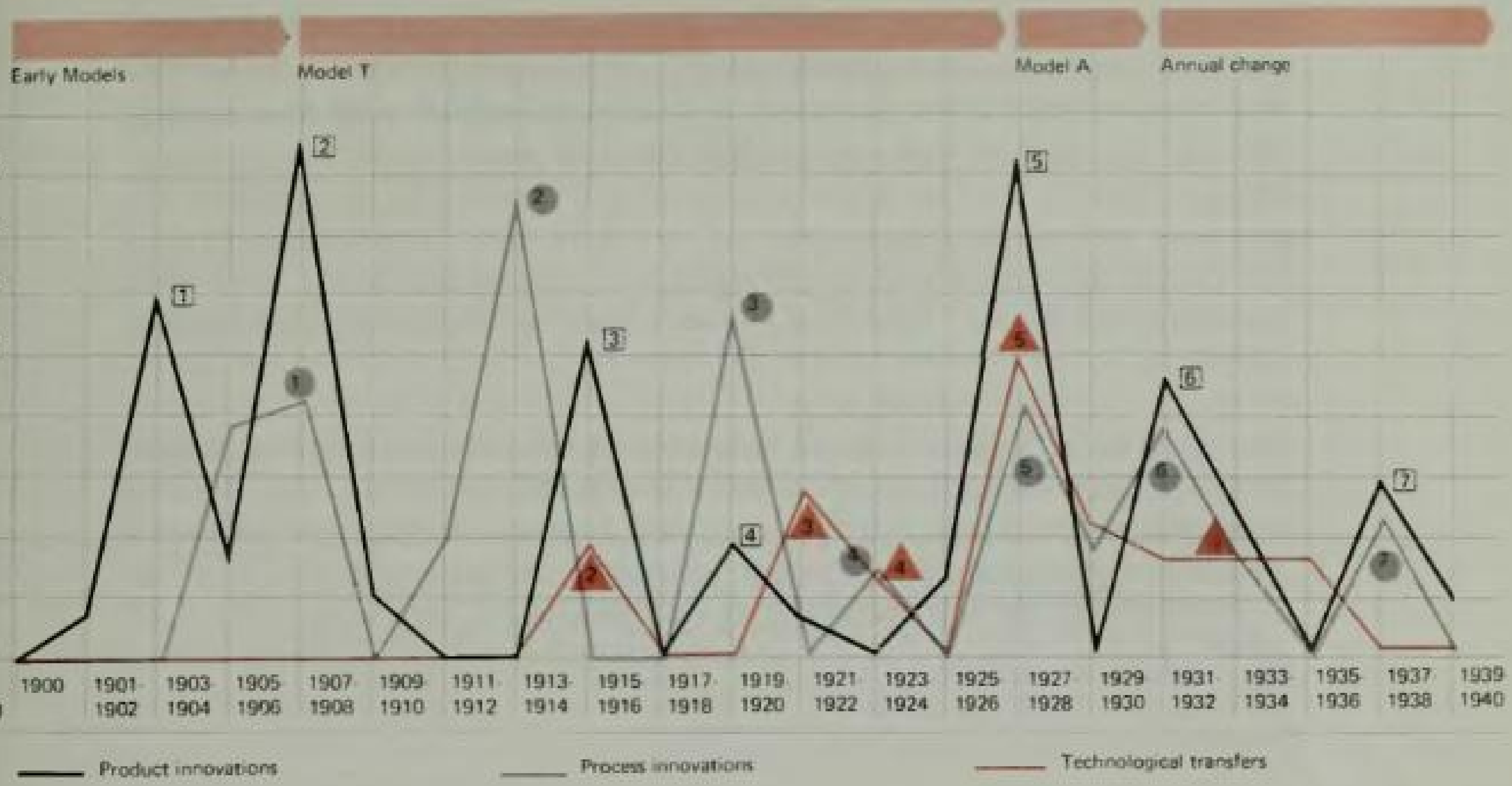


Figure 4. Innovation and process change at Ford

SOURCE: W. J. Abernathy and K. Wayne, "Limits of the Learning Curve," *Harvard Business Review*, September–October 1974, pp. 109–119.

> less need for management in planning and control activities. The percentage of salaried workers was cut from nearly 5 percent of total employment for 1913 to less than 2 percent by 1921; these reductions in Ford personnel enabled the company to hold in line the burgeoning fixed cost and overhead burden [Abernathy and Wayne, 1974].

Beginning in the middle 1920s, however, General Motors successfully focused the competitive arena in product innovation. The Ford Company was so completely organized to produce a standardized product that the effects of the change in consumer demand nearly sunk the enterprise.

Plotting Innovation Cycles. Even in a period of stable product designs, there appears to be an innovation cycle. Figure 4 shows a plot of product- and process innovations, and technological transfers over a thirty-eight year period at the Ford Motor Company. Ford-initiated innovations were rated on a scale of 1 to 5 by four independent industry experts. The innovations ranged from the introduction of the plastic steering wheel (average rating 1) in 1921 to the power driven final assembly line (average rating 5) in 1914. Figure 4 shows that new product applications occurred in clusters associated with new models, followed by a decline as the new designs became standardized. Process innovations peaked after the product innovations, presumably to integrate the processes with existing operations and to reduce costs. Technological transfers plotted in Figure 4 refer to the transfer of process technology to or from associated industries. These transfers increased as Ford undertook vertical integration.

Managing Technological Change. One of the results of the Abernathy-Wayne study was that the need was recognized for balance between a cost-reducing strategy and new product innovations, which were at odds. "The ability to switch to a different strategy seems to depend on the extent to which the organization has become specialized in following one strategy and on the magnitude of change it must face. An extreme in either factor can spell trouble."

answers to overspecialization

This balance can be achieved by periodically inaugurating major product innovations, stressing cost reduction along the learning curve between model changes. An alternative mode of achieving balance is to decentralize within the corporate structure; to have separate organizations follow different strategies within the same general product line. One organization might follow the Ford Model-T strategy of cost reduction and volume expansion. Another might develop innovative products or processes, which, when developed, might follow the cost reduction strategy, perhaps finally displacing the former products.

PREDICTING MARKETS FOR PRODUCTS AND SERVICES

Rational plans for products and services (including their productive systems) cannot be made without estimating the size of the market. This is true whether we are dealing with profit or non-profit enterprises. If we want to determine whether or not a new product or service should be launched, then the data on potential market size is crucial. If the basic decision to produce the product or service already has been made, then the information is just as necessary to finalize designs both for the product-service and the productive system. When we are dealing with an on-going situation, in which products and services are in the growth- or even saturation phase of their life cycles, refinements of the product design and probably additions to physical capacity, and/or relocating or rebalancing physical capacity, usually are involved. All these kinds of plans are of a longer-term nature, and often require the commitment of large investments. Therefore, insight into the future is required.

Prediction and Forecasting

We will use two different terms to describe methodologies for estimating future demand: prediction and forecasting. *Prediction* suggests a crystal ball. When we predict, we are integrating a great deal of subjective and objective information to form our best estimate of the future. We use prediction methods when we have little experience on which to base estimates of the future. *Forecasting*, on the other hand, connotes a statistical technique for *casting* the historical record *forward*. Forecasting depends on sufficient historical data to enable us to describe the record in statistical terms, and on reasonably stable market-generating factors that make it logical to forecast the future based on the past. Both prediction and forecasting have their place.

PREDICTION AND FORECASTING METHODS

We will group the available methods into predictive, causal, and time-series forecasting models. In general, *prediction* applies to the more qualitative methods, and *forecasting* to the causal and time series models. Table 2 summarizes some of the most prominent and useful methods under each of the three main headings. Note that, in terms of general application, the predictive methods apply to new product introductions and longer-range predictions; the causal methods to short- and medium range forecasting; and the time series models to shorter-range forecasts for operations. (See Chambers, Mullick, and Smith [1971] for an excellent survey of methods and applications.)

Predictive Methods

In this age of management science and computers, why must we resort to qualitative methods to make some of the most important predictions of future demand for products and services—predictions on which hinge the greatest risks involving large investments in facilities, and in market development? The answer is that, where we have no historical record, statistical methods have no validity. Perhaps the best that we can do is to find out what people think, sample

TABLE 2
METHODS OF PREDICTION AND FORECASTING

Method	General Description	Applications	Relative Cost	References*
Predictive methods:				
1. Delphi	Expert panel answers a series of questionnaires; the answers of each questionnaire are summarized and made available to the panel to aid them in answering the next questionnaire.	Long-range predictions, new products and product development, market strategies, pricing and facility planning.	Medium-high	Claycamp, 1969; and Liddy; North and Pyke, 1969.
2. Market surveys	Testing markets through questionnaires, panels, surveys, tests of trial products, analysis of time series.	Same as above.	High	Ahl, 1970; Bass, 1969; Claycamp and Liddy, 1969.
3. Historical analogy and life cycle analysis	Prediction based on analysis of and comparison with growth and development of similar products. Forecasting new product growth based on the S-curve of introduction, growth, and market saturation.	Same as above.	Medium	Bass, 1969; Chase and Aquilano, 1973; Chambers, Mullick, and Smith, 1971, 1974; Gerstein, 1971.
Causal forecasting methods:				
Regression analysis	Forecasts of demand related to economic and competive factors which control or couse demand, through least squares regression equation.	Short- and medium-range forecasting of existing products and services. Marketing facility planning.	Medium	Chambers, Mullick, and Smith 1971, 1974; Claycamp and Liddy, 1919; Evans; Parker and Segura, 1971; Spencer, Clark, and Hoguet, 1961.
Econometric models	Based on a system of interdependent regression equations.	Same as above.	High	Evans, 1969; Chambers, Mullick and Smith, 1971, 1974.
Time series forecasting models:				
Moving averages	Forecast based on projection from time series data smoothed by a moving average, taking account of trends and seasonal variations. Requires at least two years of historical data.	Short-range forecasts for operations such as inventory, scheduling, control, pricing, timing, special promotions.	Low	
Exponential moving averages	Similar to moving averages, but averages weighted exponentially to give more recent data heavier weight. Well adapted to computer application and large numbers of items to be forecast. Requires at least two years of historical data.	Same as above.	Low	

*References refer to items in reference list at the end of chapter.

how they react to markets tests, know about consumer behavior, and make analogies to similar situations. Given this situation, the most scientific approach involves bringing as much order to these kinds of judgments as possible. We cannot create hard demand data that do not exist. Qualitative methods are therefore very important, since they provide a basis for some important decisions.

Delphi Methods. *Technological forecasting* ~~is a term used to~~ describe the longest term predictions. The *Delphi technique* is often used as a vehicle, the objective being to probe into the future in the hope of anticipating new products and processes in today's rapidly changing culture and economy. As we shall see, in the shortest range of such predictions, this technique also can be used to estimate market sizes and timing. (See Gerstenfeld [1971]; North and Pyke [1969].)

The Delphi technique draws on a panel of experts in order to eliminate the possible dominance of the most prestigious, the most verbal, the best salespeople, etc. The intention is to use expert opinion in the form of a consensus, rather than a compromise. The result is pooled judgment; the range of expert opinion and the reasons for differences of opinion are shown. The Delphi technique first was developed by the RAND Corporation to achieve these kinds of results. In comparison with conferences and panels, where the individuals are in direct communication, this technique eliminates the undesirable effects of group interaction.

The panel of experts can be contructed in various ways. Often, it includes individuals both inside and outside the organization. Each panel member may be an expert on some aspect of the problem, but often no one is an expert on the entire problem. In general, the procedure involves the following:

1. Each expert in the group makes independent predictions in the form of brief statements.
2. A coordinator edits and clarifies these statements.
3. The coordinator provides a series of written questions to the experts, which combine the feedback supplied by the other experts.

One of the most extensive probes into the technological future was reported by TRW, Inc. [North and Pyke, 1969]. The project involved the coordination of fifteen different panels, corresponding to fifteen categories of technologies and systems that were felt to affect the company's future. Anonymity of panel members was maintained to stimulate unconventional thinking. A Delphi method then was used to question and requestion the experts, as follows:

Round one. The experts were asked to list probable technical events in their categories that could have a significant impact on the company. Each event was weighted on the basis of desirability, feasibility, and timing. After duplications were edited and eliminated, a total of 1,186 predictions resulted.

Round two. Each panel member received a composite list of the edited predictions contributed by the panel in his/her category, plus those from other panels that were related to that member's category. Each panelist was asked to evaluate all events on the basis of the same three factors of desirability, feasibility, and timing.

Round three. Wide differences of opinion concerning events and dates were eliminated by discussing predictions individually with the panelists involved. The result was a composite rating of each event on the basis of its desirability, its feasibility, the probability that the event would occur, and the probability estimates of the timing of occurrence. The extensive results then were formed into logic networks.

Market Surveys. Market surveys and the analysis of consumer behavior have become quite sophisticated, and the resulting data have become an extremely valuable input for predicting market demand. In general, the methods involve the use of questionnaires, consumer panels, and tests of new products and services in various kinds of surveys. The entire field is a specialty in itself, and one that is beyond our scope. A considerable amount of literature deals with the estimation of new product performance based on consumer panels [Ahl, 1970], using analytical approaches [Bass, 1969; Claycamp and Liddy, 1969] as well as simulation and other techniques [Bass, King, and Pessemeier, 1968]. Proposed products and services may be compared with the known plans of competitors, and new market segments may be exploited with variations of product designs and/or quality levels. In such instances, comparisons can be made with data on existing products. These kinds of data often are the best available for refining the designs of products and facilities for new ventures.

Historical Analogy and Life Cycle Analysis. Market research studies sometimes can be supplemented by referring to the performance of an ancestor of the product or service under consideration, and applying an analysis of the well-known S-curve. A typical S-curve is shown in Figure 5, where demand in the initial phases of market development accelerate to the middle growth period, culminating in market saturation. Of course, following saturation, there may be an actual decline. For example, it was assumed that color TV would follow the general sales pattern experienced with black and white TV, but that it would take twice as long to reach a steady state [Chambers, Mullick, and Smith, 1974]. Such comparisons provide guidelines during the initial planning phases, and may be supplemented by other kinds of analyses and studies, as initial actual demand becomes known. In studying the problems of production management, Chase and Aquilano [1973] focused their attention on the life cycle of products.

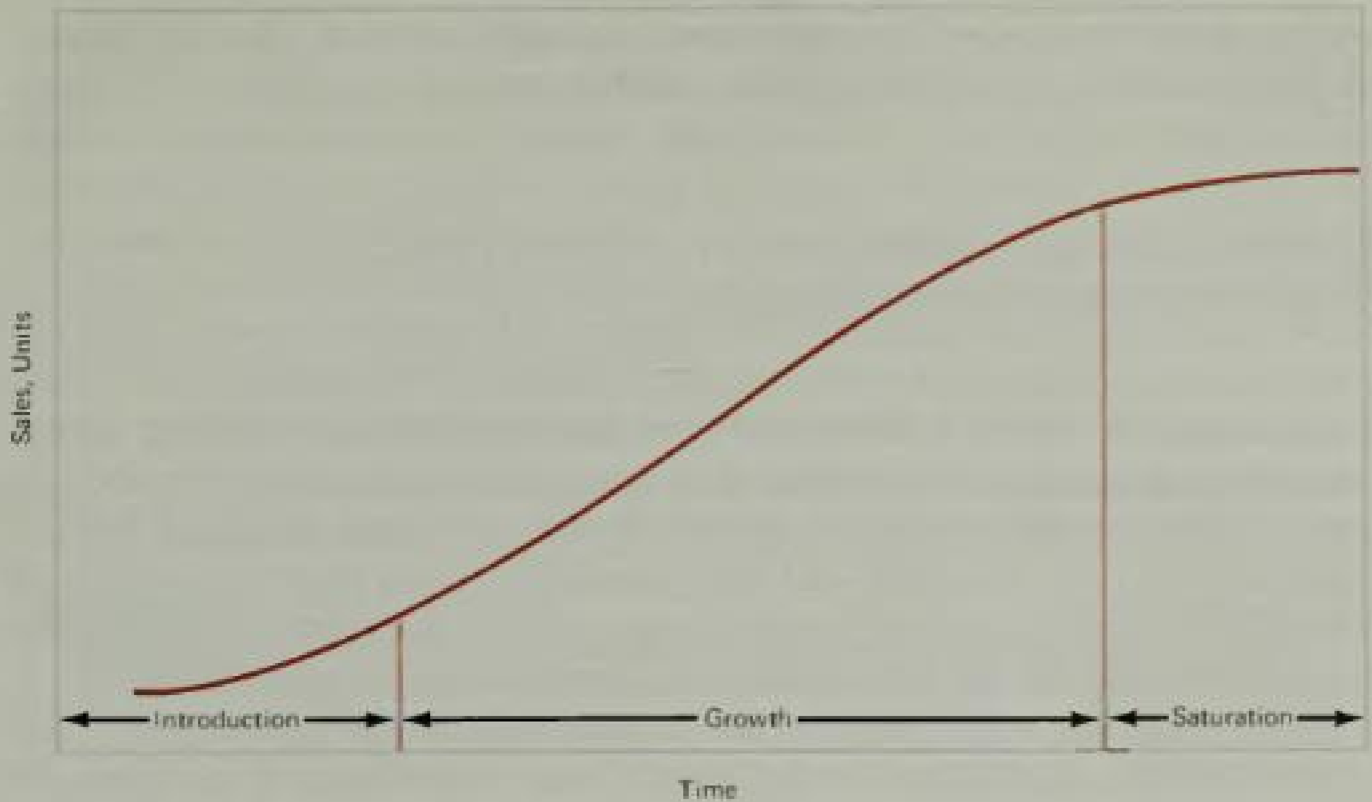


Figure 5. Typical S-curve of the introduction, growth, and market saturation in the life cycle of a product or service

Causal And Time Series Forecasting Methods

With enough historical data and experience, it may be possible to relate forecasts to those factors in the economy that *cause* the trends, seasonals, and fluctuations. Thus, if we can measure the causal factors, and can determine their relationships to the product or service of interest, then we can compute forecasts of considerable accuracy.

Time series forecasting models are based on projections or extrapolations from time series data that have been processed by statistical methods. Statistical processing is designed to discount random fluctuations to some degree, and to take into account trend and seasonal variations. Time series methods require at least two to three years of historical data to be reasonably accurate and useful.

As shown in Table 2, causal and time series forecasts mainly are applied to short- and medium range forecasts. Therefore, these kinds of forecasts are most valuable for operations; however, causal forecasting methods may also be valuable in connection with decisions that are related to product and service design, system location, and productive system design. Since the quantitative methods of forecasting are best considered together, we will delay our coverage of causal and time series forecasting methods until Chapter 9, when we discuss such methods in connection with operations planning and control.

INTERACTION BETWEEN PRODUCT-SERVICE DESIGN AND PRODUCTIVE SYSTEM DESIGN

In examining the nature and flow of innovation, we saw that, on the macro level, innovation in products at lower levels of the vertical integration chain fed into process innovations (productive system design) at higher levels of vertical integration within an industry. At this broad, industry wide level, there is an interaction between product-service design and the design of the productive system. Both the design of services and products are involved because, in both the computer and transportation industries examined by Abernathy and Townsend, the industrial process culminates in services rendered rather than in a physical product.

A similar process takes place within an enterprise in which the design of products and services partially depends on the productive system design, and the productive system design strongly depends on the nature of the products to be produced and services to be rendered. The concept is so well recognized in the mechanical industries that a name has been coined for the process of designing products from the point of view of producibility—*production design*. We shall discuss and illustrate production and service design and redesign.

Production Design

The producibility and minimum possible production cost of a product originally are established by the product designer. The cleverest production engineer cannot change this situation, but can only work within the limitations of the product design. Therefore, the obvious time to start thinking about basic modes of production for products is while the products are still in the design stage. This conscious effort for producibility and low manufacturing cost is called production design, as distinguished from functional design. To be sure, the product designer's first responsibility is to create something that functionally meets requirements. However, once this has been accomplished, design alternatives ordinarily can be found. Which of these alternatives will minimize production costs? A well conceived design already has narrowed the available alternatives, and has specified, for example, a sand casting (if that is appropriate in view of both function and cost considerations).

Once the design has been chosen, process planning for manufacture must be carried out to specify, in careful detail, the processes required and their sequence. Production design first sets the minimum possible cost that can be achieved through the specifications of materials, tolerances, basic configurations, methods of joining parts, and so on. Final process planning then attempts to achieve that minimum by specifying the processes and their sequence to meet the exacting requirements of the design. Here, the process planner may work under the limitations of available equipment. But if the volume is large, or the design stable, or both, the planner may be able to consider special-purpose equipment (including semiautomatic and automatic processes) and special-purpose layout. In performing these functions, the process planner is setting the basic design of the productive system.

The thesis of a production design philosophy is that design alternatives nearly always exist that still meet functional requirements. For the projected volume of the product, then, what differences in cost would result? Here, we must broaden our thinking, because there are likely to be more areas of cost that can be affected by design than we would imagine. There are the obvious cost components of direct labor and materials; but perhaps not so obvious are the effects on equipment costs, tooling costs, indirect labor costs, and the nonmanufacturing costs of engineering.

Indirect costs tend to be hidden, but suppose that one design required thirty different parts, while another required only fifteen (for example, the reciprocating automobile engine versus the rotary engine). There are differences in indirect costs due to greater paper work and the cost of ordering, storing, and controlling thirty parts instead of fifteen for each completed item. The indirect cost for each design is composed of those items that are necessary to process parts through the paper work system, and includes such items as: planning; tool ordering; material purchasing; shop and assembly order writing; storing material; dispatching material, tools, and parts; order control; accounting, transportation, inspection; etc. Studies in industry have attempted to relate the relative amount of various cost categories that are under the direct control of the product design engineer [Hahir, 1953; Papen, 1954].

Design and Redesign

The design process is an iterative one, and, in a sense, it is never done. New information feeds in from users, and we find ways to improve designs from the viewpoint of the productive system—usually ways that reduce production costs, though the quality criterion often is an objective as well.

As a general example of production design and redesign, Bright [1958] describes the development of electric light bulb manufacturing during the period from 1908 to 1955. Initially, a batch process was used, involving manual operations and simple equipment. The conversion from batch to continuous operation was achieved by adopting systematic layout, standardizing operations, and effecting operation sequence changes. Then, however, the light bulb itself was redesigned a number of times to facilitate process changes and permit mechanical handling. Finally, an evolution took place in which individual standardized manual operations were replaced by mechanical operations; and these operations in turn were integrated to produce an automated, fully integrated process.

The following practical examples of contrasting designs and costs illustrate the kinds of interactions that may take place and the kinds of cost factors affected by differences in basic product design. They also suggest some ways in which productive system and technology requirements affect product design and redesign. The examples reflect redesign of products from the viewpoints of: processes and materials used; methods of joining parts; tolerances; design simplification; and techniques for reducing the amount of processing.

Processes and Materials. Figure 6 shows an improved design for a bracket for a Westinghouse electric motor. Originally, it was machined from a casting. Analysis of functional and economic considerations revealed that it could be made in three pieces (as shown), and could be stronger, lighter, and more economical to manufacture. The body and hub were designed to be drawn to shape from sheet steel, and the flange was designed to be formed in a press. The volume required of such brackets justified the expenditure of forming dies, and resulted in a superior functional design as well.

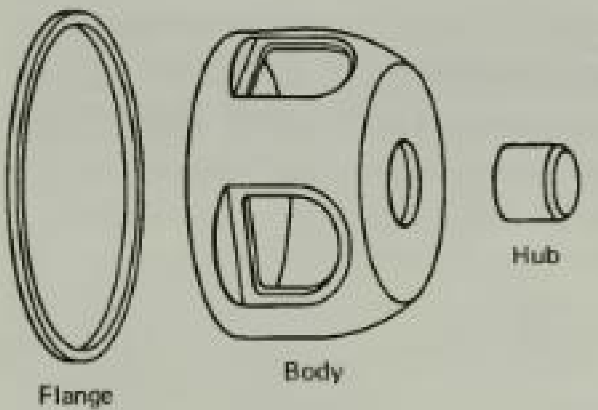


Figure 6. Motor bracket designed to be formed in presses instead of from a casting resulted in a stronger, lighter design which was more economical to manufacture (Courtesy Westinghouse Electric Corporation)

Figure 7 shows two designs of a cam switch assembly. The design on the right involved a brass casting with the other parts assembled to it. The design on the left is from an extrusion and was functionally equivalent to the assembled design. The extruded design eliminated most of the machining and, in addition,

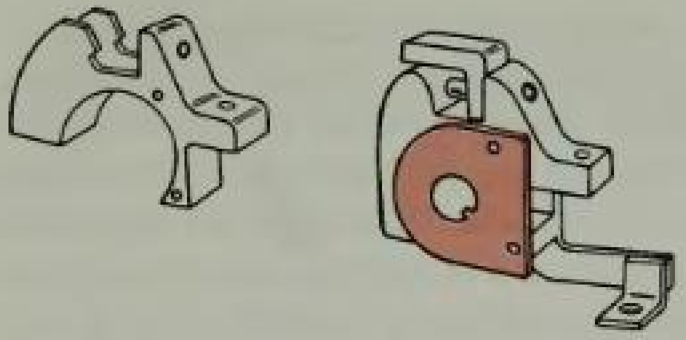
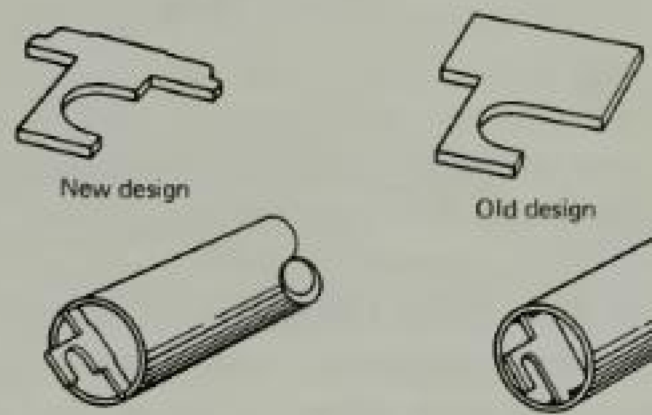


Figure 7. The extruded design on the left cost only 55 percent of the cast and assembled design. (Courtesy Westinghouse Electric Corporation)

cost less in raw material. The net difference was dramatic. The extruded design cost only 55 percent of the assembled design to manufacture.

Joining of Parts. The methods by which parts are joined together can result in important cost differences. Figure 8 shows contrasting designs of a resistor tube. The design on the left makes it possible to spot weld the bracket at a substantially lower cost. Figure 9 makes a similar point, showing contrasting designs for a tank liner. The riveted design required that holes be made in the liners to admit the rivets and washers. The stapled design eliminated the need for the holes, and the stapling operation itself was much faster and less costly.

Figure 8. The design of this resistor was functionally acceptable as either an arc-welded joint (right) or a spot-welded joint (left). However, spot-welding was substantially more economical. (Courtesy Westinghouse Electric Corporation)



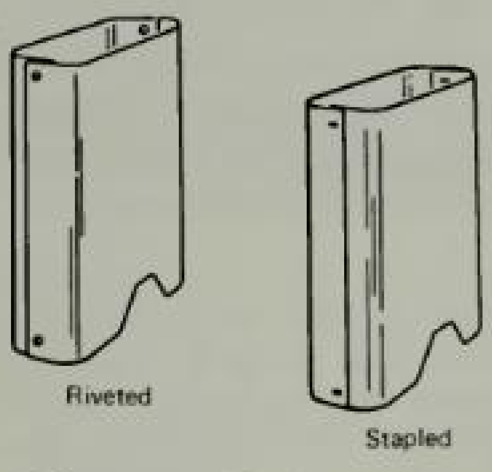


Figure 9. Both riveted and stapled tank liners meet functional requirements, but stapling eliminates the need for holes to be made, rivets and washers to be positioned, and rivets to be upset. On balance, the stapled design is much more economical. (Courtesy Westinghouse Electric Corporation)

Tolerances. Design engineers specify the acceptable limits of certain dimensions. These limits are called *tolerances*. In general, achieving a tolerance of ± 0.0001 inch is likely to be more expensive than achieving a tolerance of only ± 0.0010 inch, because: greater skill is required; better equipment and increased scrap are needed; or such operations as grinding may have to be added in order to achieve the closer tolerance. However, sometimes a closer tolerance will save adjusting and fitting time during the assembly of parts. Figure 10 illustrates such a case. The holes at the ends of the three contact arms are required to be in line within ±0.005 inch. These arms are mounted on a mircarta bar with a steel center, and are held in place by a part called a staple. The square bar could not be held to a tolerance closer than 0.012 inch and, as a result, the arms often fit too loosely or too tightly, and require adjusting operations. A closer tolerance specification of ±0.002 inch for the bar eliminates the adjusting. The bar is molded oversized and broached to a closer tolerance. A special assembly fixture permits the broaching to be accomplished as a part of the assembly operation.

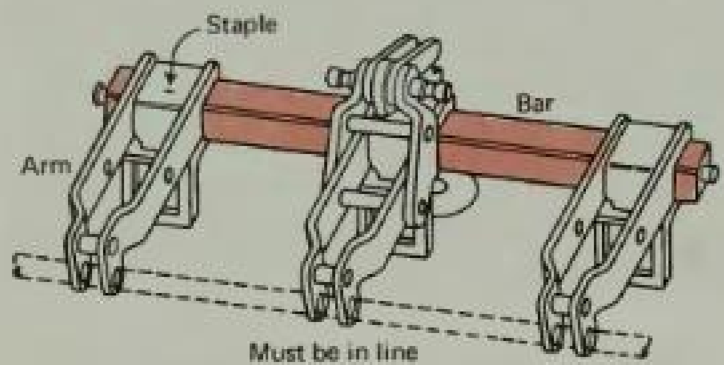


Figure 10. Specification of closer tolerances reduced overall production costs on this assembly by eliminating costly adjusting operations. (Courtesy Westinghouse Electric Corporation)

Simplified Designs. When two or more parts finally are assembled together rigidly, the question of whether of not the assembled unit could be designed as one piece is raised. One-piece design often is feasible when a single material meets service requirements for all surfaces and cross-sections of the part. Figure 11 shows an example for a contact arm of a circuit breaker. The design on the left is made of two parts, a casting and a punching. The parts are machined separately and then joined together with two screws. The one-piece design shown on the right involves a precision casting that eliminates all but the very close machining. The one-piece design requires 74 percent as much machining time, 50 percent as much assembly time, and a substantial decrease in material cost. Because of very high volume, the one-piece design is much less expensive to produce.

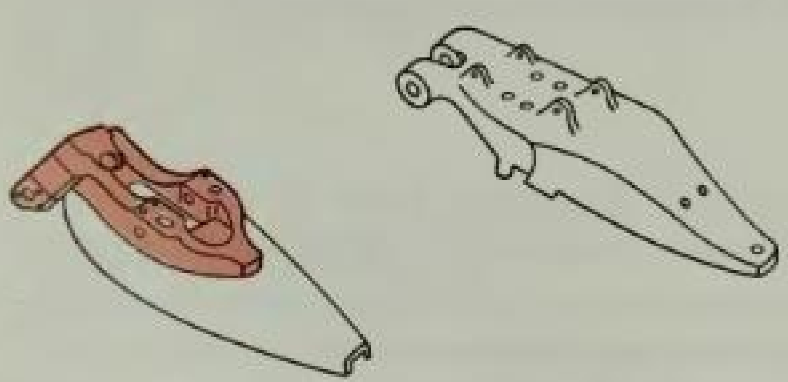


Figure 11. The simplified design on the right involves a precision casting which requires only 74 percent as much machining time, 50 percent as much assembly time, and a substantial decrease in material cost, compared to the design on the left. (Courtesy Westinghouse Electric Corporation)

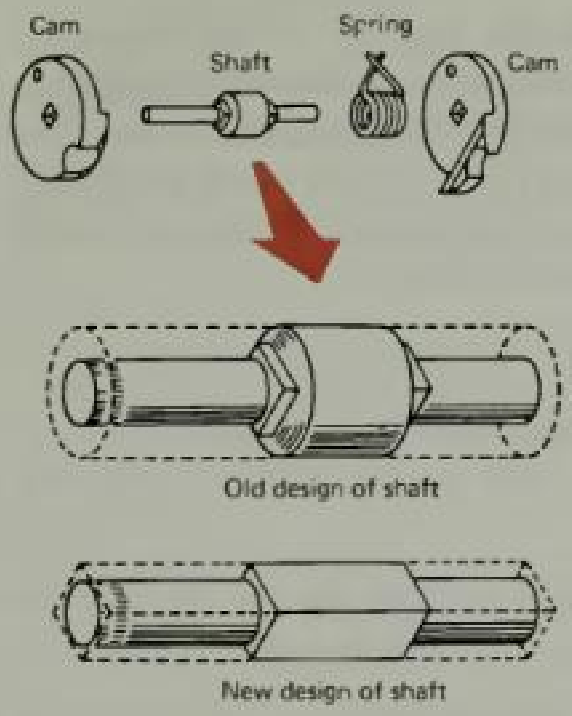


Figure 12. New and old designs for a cam subassembly of a limit switch (Courtesy Westinghouse Electric Corporation)

Reduced Amount of Processing. Figure 12 shows a contrasting design for a cam subassembly of a limit switch. The difference is based on a substitution of raw materials. In the original design, the shaft was machined from a solid bar. The large diameter middle section of the shaft centered the spring and separated the two cams. A square was milled on each side to key with the square holes of the cams. The new design used square raw stock, and the size of the raw stock matched that of the square that was formerly milled. A sleeve was provided as a spacer for the cams, which slipped over the square of the shaft. The milling operation was eliminated, and the amount of material to be removed was reduced greatly. The cost reductions more than offset the cost of manufacturing the new sleeve.

The several preceding examples of production design applications show how ingenuity and understanding of the productive system design's details can have considerable effect on productivity and product design.

Design of Services Offered

While no term has been coined yet to describe it, a process similar to production design exists and represents the interaction between the design of services to be offered and the productive system design. The motivation for altering the service that currently is offered may be a desire to accommodate cost factors and service quality, as measured by time performance and other dimensions. Some of the kinds of accommodation in the nature of services offered are as follows:

1. Transfer some of the activity involving the service to the client or customer. This, of course, has been one of the most common techniques for lowering costs. In hospitals, supermarkets, mass food-services, etc., the client or customer performs some of the activities that formerly were part of the service. The result is usually a lower labor cost and the elimination of some activities performed by the productive system.
2. Eliminate some aspects of service entirely.
3. Change the mix of services offered.
4. Change the reaction time for service; that is, give poorer service.
5. Change other aspects of service quality; that is, the range of services offered.

Table 3 indicates some of the interactions between the design of services offered and the productive system design for several well known types of systems. The kinds of accommodations listed for both services and productive systems are not intended to be exhaustive.

Unfortunately, there seems to be a consensus that the quality of all kinds of services has deteriorated throughout the country and perhaps elsewhere as well. The reasons are probably that most service operations are labor intensive, and that labor saving devices are not common in service operations. The result is that rising wage costs are not compensated for by increases in productivity. A simulation study of computing service provides as example that combines important estimates of potential market demand, together with a consideration of the interaction of services to be offered, with equipment requirements [Yourdon, 1970].

DECISION PROCESSES AND SYSTEM DESIGN

Let us now step back from the details of technological innovation, prediction and forecasting, and designing products and services to see how they all fit into system concepts of design, and where crucial decision processes are called for. While the preceding materials have not been organized according to the systems approach to design problems, nevertheless those concepts have been implicit.

The Systems Approach to Design

The product–service design problem is not independent, but is intertwined with: market needs and demands; relationships with other products and services

TABLE 3
INTERACTIONS BETWEEN THE DESIGN OF SERVICES OFFERED AND THE PRODUCTIVE SYSTEM DESIGN FOR SEVERAL TYPES OF SERVICES

Type of Service	*Types of Alteration of Service to Accomodate Needs of Productive System*	*Types of Alteration of Productive System to Accomodate Needs of Service*
Nursing care in Hospitals	Specialization by levels to reduce costs, e.g., nurses aides.	Changes of layout and activity scheduling.
Food markets	Self-service to reduce cost.	Change in layout and flow to facilitate. Balance of numbers of check-out stands to maintain waiting time maximums.
Postal service	Reduction of services to reduce costs (e.g., number of deliveries per day).	Introduction of some semiautomatic equipment for sorting. System improvement, etc., to improve overall delivery time.
Food service	Elimination of waiters and waitresses to reduce cost, as in cafeterias, and to reduce waiting time and cost in mass food outlets.	Change of layout and flow.
Computing service, e.g. in service bureaus	Change mix of types of services offered, e.g., time share, work batch processing, etc.	Change mix of equipment for basic machine and peripheral equipment.
Fire protection	Increase or decrease time to react and provide service.	Relocate stations and/or add or delete stations to maintain a reaction time standard.
Police protection	Increase or decrease time to react and provide service.	Increase or decrease size of police staff, reallocate staff based on changing crime patterns.
Emergency medical service	Increase or decrease time to react and provide service. Change mix of services available on an emergency basis.	Relocate ambulance stations and/or add or delete stations to maintain a reaction time standard. Change equipment and/or level of training of paraprofessional medical personnel.
Airline cabin service	Change average ratio of passengers to hostesses. Eliminate services and reduce quality of services offered to reduce cost.	Increase public relations efforts through the media to try to convince passengers that services are the same or better.

offered by the enterprise, as well as by competitors; and the design of the sociotechnical system that produces the item. Thus, we originally defined the design problem in systemic terms, taking the system as a whole as our starting point. We attempted to represent the complex interactions between forecasts and longer-range plans, the productive system design, and the design of the product or service itself. We included the feedback of information from users and the on-going productive process, which results in redesign of both the

summary of chapter thus far

products and services, and the productive system, in a dynamic process. We recognized that the criteria and values that enter the decision processes are multidimensional, and that the decisions require trade-offs in an attempt to jointly optimize the product and productive systems.

Now let us consider the prime decision areas for management, and what decision processes support them.

Decisions and Decision Processes

The decision points are focused on the approval of (1) specific designs and product mixes, and (2) required alterations to the productive system. The drive to gain advantages through modularity of design and standardization often is involved. Thus, managers are again forced to look at any product-service design decision within a systems context. How will a new or redesigned product or service be received? What impact will it have on the existing product line? What is the best new product mix? What is the impact on the productive system, its design and capacities? How will the new design affect operating schedules? And, of course, what are the effects on revenue, costs, and profits? Some of these effects can be reflected in forecasts and estimates of incremental revenues and costs, but the manager must make judgmental trade-offs. In this regard, the decision framework discussed in Chapter 2 is appropriate.

The Product Mix Problem. One way of expressing the system of problems surrounding product design is to ask. What happens to the optimum mix of products when a new product or product line is introduced? The question confronts the differential revenue issue and the aggregate impact on capacities to produce.

The product mix problem is general throughout industry, and solutions should reflect the most economical allocation of capacity to demand. For example, in oil refining there are interdependencies in the quantities of different products that are to be produced. If more of one product is to be produced, e.g., aviation gasoline, then less of some other products will be produced. The profitability of various products may be different, and there are limits to the markets for each. The result is a complex programming problem for determining the best product mix to produce. In oil refining the interdependencies always exist, because the basic raw material, crude oil, can be processed into many different products. Therefore, an increase or decrease in one product always means a change in the quantities of some others.

In the mechanical industries, the product mix problem is not so obvious.

Nevertheless, the mix problem can be acute if we are operating near capacity with time-shared facilities. At this point, the interdependencies become critical, and increases or decreases in the amount of one product produced can mean changes in the amounts of other products produced. These problems have been approached through linear programming and variations.

Summary

The design of products and services is of great interest in operations management because the nature of the product greatly affects the design of the system that will produce it, and vice versa.

Innovation is the generator of new products and services. Studies of innovation within industries indicate that, while a majority of innovations may be classified as product innovations, when we look at the vertical integration structure of an industry, we find that product innovations of lower levels in the hierarchy are in fact process innovations for the higher levels in the structure. Thus, there is an interaction between product design and productive system design, even in these macro terms.

The details of product design, system location, and productive system design all flow from predictions and forecasts of markets. The methods for making market demand estimates are classified as predictive, causal forecasting, and time series forecasting. While predictive methods lack mathematical rigor, they are perhaps most important for new products that lack an historical record. They also are valuable where long-term predictions are needed as a basis for plant locations and investments. Causal methods are quite accurate for short- and medium-range forecasts, but are relatively expensive. Time series methods are quite accurate for short-range forecasts, and are more appropriate for operations rather than for decisions relating to productive system design.

The term *production design* is used in industry to connote the important interaction between product and system design. Examples of production design are shown where the accommodation was to processes and materials, methods of joining parts, tolerances, design simplifications, and reduced amount of processing.

The accommodations by the design of services offered most often are to cost considerations, and often result in having the client or customer do part of the activity, receive less service or a different mix of services, or wait longer to get service.

Review Questions

1. Contrast the chain event model of technological innovation with the Abernathy–Townsend model.

2. What is the significance of the statement that "most innovations in the lower levels of the vertical integration chain are *product* innovations. These are, at the same time, *process* innovations for processes at higher levels of vertical integration"?

3. Describe the succession of innovation events in Figure 1 that complete a feedback loop within the vertical integration structure of the railroad industry.

4. What is a learning curve in manufacturing? What is the basic principle of economics on which the concept of a learning curve is based?

5. Explain how Ford accomplished savings in manufacturing during the Model-T era.

6. How is it possible that Ford's inventory level was actually cut in half while, at the same time, system volume was expanding rapidly?

7. Explain the reasons for the reversal of the Ford price curve during the period 1930–1965, shown in Figure 3.

8. What results did Ford achieve through backward integration during the Model-T era? What were the advantages and disadvantages of the backward integration policy?

9. Describe a typical innovation cycle from Figure 4. What is the seeming effect on innovation cycles of vertical integration?

10. Define the following two terms as used in this book:
    a. Prediction.
    b. Forecasting.

11. Describe the general field of application of:

a. Regression analysis.
b. Delphi methods.
c. Exponential moving average.
d. Market surveys.
e. Historical analogy and life cycle analysis.

12. Describe the Delphi process. Why does it result in a consensus, rather than simply an averaging of viewpoints, or compromise?

13. For the following products or services, where would you pinpoint the current placement of the item in the product life cycle? Justify your answer.
    a. Automobiles.
    b. Computers.
    c. Banking.
    d. Mass food service.
    e. Television.
    f. Airline service.

14. Define the term *production design*.

15. Describe the impact of each of the following process or productive system innovations on the design of the product or service offered:
    a. The supermarket check-out stand.
    b. The automotive moving assembly line.
    c. Computer airline reservation systems.
    d. The cafeteria.
    e. Computer check clearing.

16. Describe the impact of each of the following changes in the design of the product or service offered on the design of the productive system required to produce the product or service:
    a. Solid-state TV and high fidelity sound systems from electronic tube designs.
    b. The "pocket" electronic calculator, compared with the earlier mechanical design.
    c. Automatic bank tellers.
    d. Limited menu restaurants.
    e. Fluorescent light bulbs, compared with incandescent bulbs.

Problem

Nels Jensen started his grocery business thirty-five years ago in the Lake Tahoe resort community. The combination of good management and a market among well-to-do patrons produced an independent supermarket of unsurpassed quality. One of the hallmarks of the Jensen success model is his emphasis on and definition of the service aspects of his store. Those service aspects are: ten hours of operation every day, and minimum customer hassle and waiting time to obtain the desired purchases and get checked out.

Personal ↓ impersonal

Tracing some of the major changes that occurred over the years revealed that both the nature and quality of service and the system design successively affected each other. When the store was small, employees were stationed in areas to help customers select items, check out, bag, and transfer purchases to the parking lot. Later, with a much larger volume of customers, the system transferred virtually all the selection process to the customer, and the system design focused on the check-out stand.

By the number of check stands in operation, Jensen set a service standard that he tried to maintain. The standard was set in terms of the time the customer had to wait before he/she was served. Jensen felt that, for his clientele, a waiting time of 1.5 minutes was about as long as would be tolerated without complaint. In fact, however, he tried to control waiting time by keeping an eye on the size of the waiting lines. Any time the lines had two or more people waiting, he would open up another checkstand, even if he had to operate it himself. He then tried to schedule checker shifts to accommodate the peak shopping hours. He also trained some of the checkers to do other work, such as pricing, storing stock on display shelves, etc., so that they could be flexible in accommodating unexpected peak loads.

Jensen tried several variations with the check-out system. Originally, employees helped unload carts to the checkers; but, as labor costs increased, this activity was transferred to the customer. Customer complaints resulted. Then, having changed the design of the checkstand, Jensen tried a system in which the checker worked directly from the cart. However, the checker complained of backaches from constantly having to lean over to obtain items from the carts.

Jensen then modified the system with a new cart and checkstand design system. This still enabled the checker to work directly from the cart, but did not require him to lean over to obtain items. The new cart-checkstand system raised the working level. The cart had a hinged end, which the checker opened, thus placing the bottom of the cart at the check-stand working level. In this way, all aspects of service could remain the same in terms of waiting time, what the customer had to do to obtain service, and the check-out time.

The next cycle of system design involved what is called *front-end automation*. Measurements made in a survey by Jensen indicated that the average time for a customer in the check-out process was 5.5 minutes, including 1.5 minutes of waiting time. The 4.0 minutes for actual check-out included 2.5 minutes for ringing up the sale and placing the purchases in bags. The balance was payment, which very often involved check cashing, as well as some chit-chat and other miscellaneous activities. With the advent of item scanning systems, Jensen saw the possibility of improving overall service time by reducing the check-out time itself and possibly simultaneously reducing labor costs.

The scanning systems required that a universal code be placed on products (the bar-pattern codes now commonly used on product packages), that would be read by the scanners. An entire customer's order could then be moved over the scanner rapidly. The system required the customer to load his/her purchases on a belt, that fed the purchases to the checker, who repositioned them to move past the scanning eye. The scanner read and transmitted the information to a computer, which translated the information into prices and a total bill, including sales tax. The checkers' activities thus were confined to scanning, bagging, collecting, and making change.

While the scanning system had other operating advantages, Jensen was most interested in a posisble service improvement, assuming that a productivity increase might justify the scanner's installation on a reasonable basis. The system was installed, partially as a test. Measured results indicated that check-out time, exclusive of waiting time prior to check-out, was reduced to an average of 2.4 minutes. In addition, checker productivity increased from an average of $252 of sales per hour to $360, and check-out errors were reduced by 65 percent. Checkers were paid $4.50 per hour.

Jensen could install the scanning system to cover the present twelve check-out stands for a lease cost of $7,000 per month. Even though only an average of six check-out stands were in use (current system), if he were to decide to install the scanners, he would want the entire system to be automated. His concern was that service would be improved, in the sense that the customers' time in the system would be reduced. On the other hand, the scanner system represented a step backward, in that the customers would have to unload their carts and load the conveyor belt. He was not sure how customers of his type of clientele would react.

Should Jensen install the scanner system? Justify your answer.

if less customers then not worth it

compare before : after to $7000

Nonmonetary operating advantages

References

Abernathy, W. J. "Production Process Structure and Technological Change." *Decision Sciences* 7(October 1976).

______, and P. L. Townsend, "Technology, Productivity and Process Change." *Technological Forecasting and Social Change* 7 (1975.)

______, and K. P. Wayne. "The Limits of the Learning Curve." *Harvard Business Review*, September–October 1974, pp. 109–119.

Ahl, D. H. "New Product Forecasting Using Consumer Panels." *Journal of Marketing Research* 7(May 1970): 159–167.

Bass, F. M. "A New Product Growth Model for Consumer Durables." *Management Science* 16(January 1969).

______; C. W. King; and E. A. Pessemeier. *Applications of the Sciences in Marketing Management*. New York: John Wiley & Sons, 1968.

Box, G. E. P., and G. M. Jenkins. *Time Series Analysis, Forecasting, and Control*. San Francisco: Holden-Day, Inc., 1970.

Bright, J. R. *Automation and Management*. Boston: Harvard University, Graduate School of Business Administration, 1958.

Buffa, E. S., and W. H. Taubert. *Production-Inventory Systems: Planning and Control*. Rev. ed. Homewood, Ill: Richard D. Irwin, Inc., 1972.

Chase, R. B., and N. J. Aquilano, *Production and Operations Management*. Homewood, Ill.: Richard D. Irwin, Inc., 1973.

Chambers, J. C.; S. K. Mullick; and D. D. Smith. *An Executive's Guide to Forecasting*. New York: Wiley-Interscience, 1974.

______; ______; ______. "How to Choose the Right Forecasting Technique." *Harvard Business Review*, July–August 1971, pp. 45–74.

Claycamp, H. J., and L. E. Liddy. "Prediction of New Product Performance: An Analytical Approach." *Journal of Marketing Research* 6(November 1969): 414–421.

Evans, M. K. *Macro-economic Activity: Theory, Forecasting and Control.* New York: Harper & Row, 1969.

Gerstenfeld, A. "Technological Forecasting." *Journal of Business* 44 (January 1971): 10–18.

Hahir, J. P. "A Case Study on the Relationship Between Design Engineering and Production Engineering." *Proceedings, Fifth Annual Industrial Engineering Institute.* Berkeley–Los Angeles: University of California, 1953.

Linstone, H. A., and D. Sahal. *Technological Substitution: Forecasting Techniques and Applications.* New York: Elsevier/North Holland, 1976.

Martino, J. P. *Technological Forecasting for Decision Making.* New York: Elsevier/North Holland, 1972.

Meyers, S., and D. Marquis. *Successful Industrial Innovations.* Washington, D.C.: National Science Foundation, 69–17, 1969.

North, H. Q., and D. L. Pyke. "Probes of the Technological Future." *Harvard Business Review,* May–June 1969.

Papen, G. W. "Minimizing Manufacturing Costs through Effective Design." *Proceedings, Sixth Annual Industrial Engineering Institute.* Berkeley–Los Angeles: University of California, 1954.

Parker, G. G. C., and E. L. Segura. "How to Get a Better Forecast." *Harvard Business Review,* March–April 1971, pp. 99–109.

Starr, M. K. *Product Design and Decision Theory.* Englewood Cliffs, N. J.: Prentice-Hall, 1963.

Utterback, J. M., and W. J. Abernathy. "A Dynamic Model of Process Innovation by Firms." *Omega,* Vol. 3, No. 6, 1975.

Wheelwright, S. C., and S. Makridakis. *Forecasting Methods for Management.* New York: John Wiley & Sons, 1973.

Yourdon, E. "CALL/360 Costs." *Datamation,* November 1970, pp. 22–28.

Chapter 6

Distribution and Facility Location

The output of productive systems must be distributed to wholesale and/or retail outlets, consumers, clients, and patients. For products, the distribution system is transportation oriented, interrelated with the location of both producing and warehousing facilities, and designed for inventoriable items rather than those produced to order. The first major section of this chapter focuses on the nature of distribution systems, the number and location of producing plants, and the number and location of intermediate warehouses. The second major section discusses the distribution of medical services, public services, e.g., police and fire protection, and consumer outlets. The problem of distributing services entails locating the service facility in relation to users.

PRODUCT DISTRIBUTION SYSTEMS

The cost of the physical flow of product to consumer is of great economic significance; for the U.S., it is estimated at more than $100 billion per year. Since, for some products, the distribution cost may equal or considerably exceed production costs, management is keenly interested in the design and operation of distribution systems.

Distribution systems begin with producing facilities, and proceed, through a network of intermediate warehouses and stock points, to the ultimate user or consumer. The basic problem is numbers and locations of the plants and distribution warehouses. The economic criterion is to maximize contribution; however, this commonly is reduced to the objective of minimizing distribution costs. These costs also may involve production costs and system inventory costs. In addition to economic factors, other criteria must be met by systems designs—for example, achieving acceptable delivery times, and considering preferred routings and individual customers.

The Linear Programming Distribution Model

As a setting for an illustrative problem, let us assume the distribution situation indicated by Figure 1. Three factories, located in Chicago, Detroit, and Atlanta, produce identical products. The organization's distribution system has five major distribution points that serve market areas in Milwaukee, Cincinnati, Des Moines, Buffalo, and New York City. The three factories have capacities that determine the availability of product, and the market demand in the five major areas determines the requirements to be met. The problem, then, is to determine how to allocate the available product from the three factory locations to the five distribution points so that market demand is met and the cost of distribution for the entire system is minimized.*

Data for our illustrative problem are shown in Table 1. There are 19,000 units available at the Detroit plant, 28,000 at Chicago, and 25,000 at Atlanta, a total of 72,000. Similarly, demands in the five market areas are indicated in the bottom row of the table and also total 72,000 units, although this is not a necessary requirement for solution. Recall that these figures of units available and required are commonly termed the *rim conditions*. Table 1 also shows the distribution costs per 1,000 units for all combinations of factories and distribution points. These figures are shown in the small boxes. For example, the distribution cost between the Chicago plant and the Milwaukee distribution point is $34 per 1,000 units. For convenience, we have labeled the plants *A*, *B*, *C*, and the distribution points, *V*, *W*, *X*, *Y*, and *Z*.

An Optimal Solution. The distribution problem posed by Table 1 has an optimal solution, that has been generated by linear programming distribution

* This is the same illustrative problem used to discuss the development of the details of the linear programming distribution methodology in Appendix C.

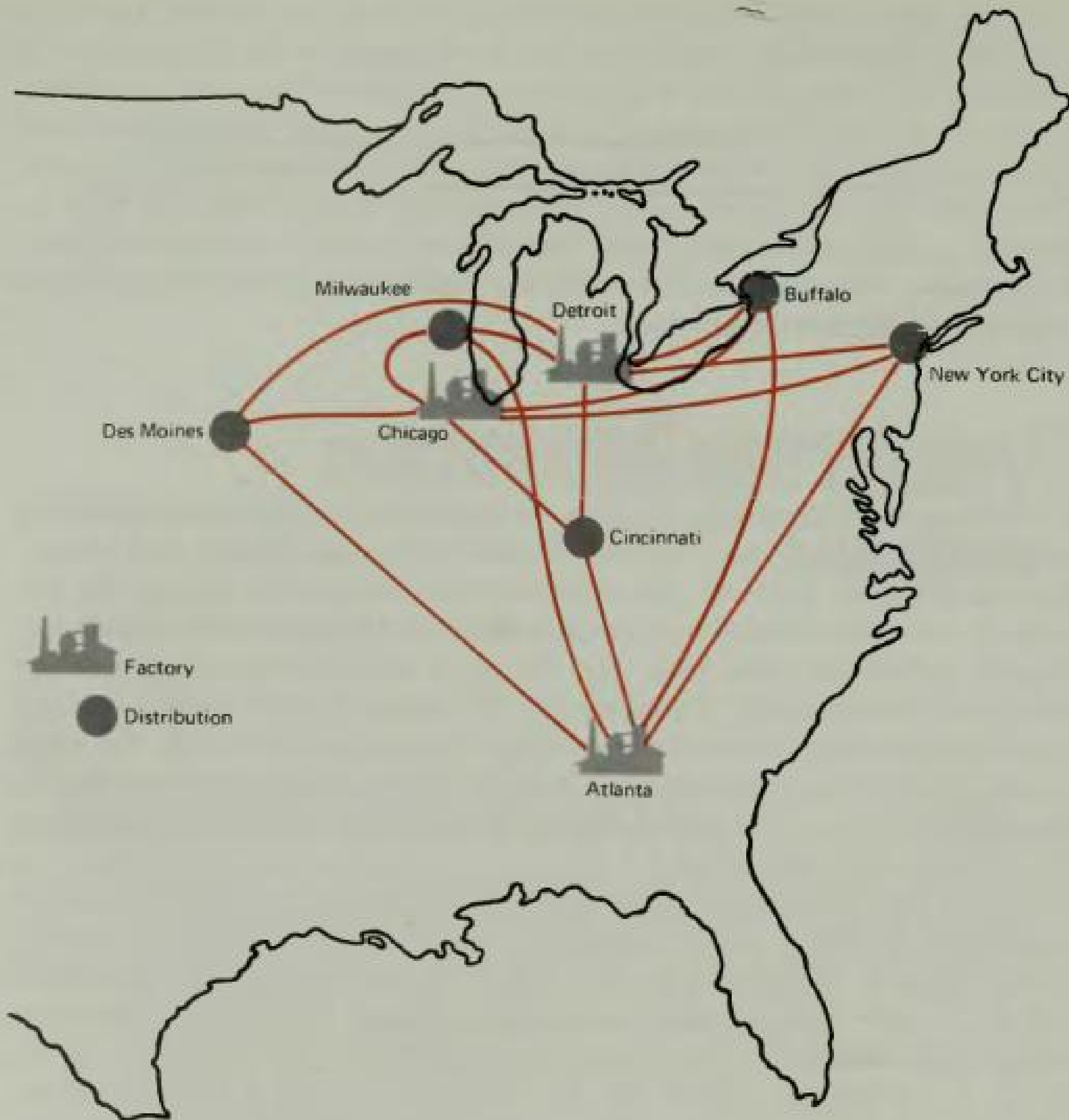


Figure 1. Geographical locations of factories and distribution points

methods (see Appendix C for a review of methods). and is shown in Table 2. The total distribution cost required by the optimal solution is $2,986. The solution is optimal because reevaluation of all open squares shows that no further improvement is possible. The evaluations of the open squares are shown in Table 2 by the small figures in the lower left hand corners of the squares. Note that all open squares show that further shifts in assignments either would increase distribution costs, or not change them. Standard computer codes are widely available for larger scale problems.

1,1,1,1,1, 10 zeros

TABLE 1

SUMMARY OF QUANTITIES OF PRODUCT AVAILABLE AND REQUIRED, AND DISTRIBUTION COSTS PER THOUSAND

(Distribution or Transportation Matrix)

From Factories / To Dist. Points	Milwaukee (V)	Cincinnati (W)	Des Moines (X)	Buffalo (Y)	New York City (Z)	Available at Factories 1000s ↓
Detroit (A)	42	42	44	40	44	19
Chicago (B)	34	42	40	46	48	28
Atlanta (C)	46	44	42	48	46	25
Required at Dist. Points 1000s →	11	13	7	17	24	72

TABLE 2

AN OPTIMAL SOLUTION

(Evaluation of all squares without assignments in this table results in no further improvement)

From / To	Milwaukee (V)	Cincinnati (W)	Des Moines (X)	Buffalo (Y)	New York City (Z)	Available 1000s ↓
Detroit (A)	42 +8	42 2	44 +4	40 17	44 0	19
Chicago (B)	34 11	42 10	40 7	46 +6	48 +4	28
Atlanta (C)	46 +10	44 1	42 0	48 +6	46 24	25
Required 1000s →	11	13	7	17	24	72

Total distribution cost = $2986.

TABLE 3
ALTERNATIVE BASIC OPTIMUM SOLUTION

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	$\boxed{42}$ +8	$\boxed{42}$ 2	$\boxed{44}$ +4	$\boxed{40}$ 17	$\boxed{44}$ 0	19
B	$\boxed{34}$ 11	$\boxed{42}$ 11	$\boxed{40}$ 6	$\boxed{46}$ +6	$\boxed{48}$ +4	28
C	$\boxed{46}$ +10	$\boxed{44}$ 0	$\boxed{42}$ 1	$\boxed{48}$ +6	$\boxed{46}$ 24	25
Required 1000s →	11	13	7	17	24	72

Alternate Optimal Solutions. The fact that open squares CX and *AZ* in Table 2 have zero evaluations is important, and gives us flexibility in determining the final plan of action. These zero evaluations allow us to generate other solutions that have the same total distribution cost as the optimal solution generated in Table 2. For example, since open square CX has a zero evaluation, we may make the shifts in allocations that its closed path indicates, and generate the alternative basic optimal solution by shifting assignments to it from square *CW*, as shown in Table 3. Another alternative optimum solution could be generated: since square *AZ* has a zero evaluation, we can shift assignments to it, as shown in Table 4.

But we are not finished yet. We can generate literally dozens of other optimal solutions from each of the basic solutions. Let us go back to our original optimal solution, shown in Table 2. Recall that we derived a second optimal solution, shown in Table 3, by shifting 1,000 units to square CX which had a zero evaluation. But it was not necessary for us to shift the entire 1,000 units to CX. We could have shifted only 500 units, or 400, or 300, or any fractional amount of the 1,000 units. All the basic alternate optimum solutions also could be varied in this way.

Where we have optimum solutions containing open squares with zero evaluations, we have great flexibility in distribution at minimum cost. Where fractional units are permitted, we have a tremendous number of alternate optimal solutions. This often may enable us to satisfy subjective factors in the problem and still retain a minimum distribution cost.

We will comment on some variations of the basic problem, since they are of

TABLE 4

SECOND ALTERNATE BASIC OPTIMUM SOLUTION

From \ To	*V*	*W*	*X*	*Y*	*Z*	Available 1000s ↓
A	42 +8	42 0	44 +4	40 17	44 2	19
B	34 11	42 10	40 7	46 +6	48 +4	28
C	46 +10	44 3	42 0	48 +6	46 22	25
Required 1000s →	11	13	7	17	24	72

interest in the problem of distribution. Three such variations follow: the effect of unequal supply and demand; the effect of different production costs at factory sites; and the effect of different selling prices in the market areas.

Unequal Supply and Demand. Now, suppose that supply exceeds demand, as shown in Table 5. As before, the total available is 72,000 units; however, the aggregate demand at the five distribution points totals only 67,000 units. This

TABLE 5

DISTRIBUTION MATRIX WITH SUPPLY EXCEEDING DEMAND, OPTIMUM SOLUTION

From \ To	*V*	*W*	*X*	*Y*	*Z*	Dummy	Available 1000s ↓
A	42	42 2	44	40 15	44 2	0	19
B	34 11	42 11	40 6	46	48	0	28
C	46	44	42	48	46 20	0 5	25
Required 1000s →	11	13	6	15	22	5	72

situation can be handled by creating a dummy distribution point to receive the extra 5,000 units. The nonexistent distribution point is assigned zero distribution costs, as shown, since the product never will be shipped. The optimal solution then assigns 67,000 of the 72,000 available units in the most economical way to the five real distribution points, and assigns the balance to the dummy department. Table 5 shows an optimal distribution plan for this situation.

When demand exceeds supply, we can resort to a modification of the same technique. We create a dummy factory to take up the slack. Again, zero distribution costs are assigned for the dummy factory, since the product never will be shipped. The solution then assigns the available product to the distribution points in the most economical way, indicating which distribution points should receive "short" shipments to minimize that total distribution costs.

Differential Production Costs. Suppose that the three factories have differing production costs due to differing labor costs, equipment, layout, and raw material sources. The problem involves allocating output to markets in order to minimize combined production plus distribution costs. This situation is shown in Table 6, where production costs per thousand are $50, $62, and $54 for plants *A*, *B* and *C*, respectively. The cost values in Table 6 are determined simply by adding $50 to all the distribution costs in row *A* of Table 1; $62 to all the distribution costs in row *B*; and $54 to all the distribution costs in row *C*. The solution, shown in Table 6, is an optimal allocation of the output from the three plants to the five distribution points, minimizing combined production-plus-distribution costs.

Note, however, that the distribution plan shown in Table 6 is exactly the same as the one shown in Table 2, which took only distribution costs into account. In addition, the evaluations of open squares, indicated by the small figures in the lower left hand corners of the squares, are exactly the same. Therefore, all the alternate optimal solutions discussed previously also can be generated. Of course, the distribution costs for the plan are exactly the same as in the solution of Table 2. Adding a constant to all the costs in a row has not changed the distribution plan. This is a fundamental point in the solution of linear programs. Its significance for distribution systems is that we can deal only with the distribution costs in this situation, in which all costs are assumed to be linear.

Differential Selling Prices. Suppose that the prices at which the product can be sold in the five different market areas also vary at $100, $105, $110, $115, and $120 in market areas *V*, *W*, *X*, *Y*, and *Z*, respectively. Now we can construct a contribution matrix, such as that shown in Table 7, in which each value in the

TABLE 6

OPTIMUM DISTRIBUTION PLAN WHEN PRODUCTION COSTS OF
$A = \$50$, $B = \$62$, and $C = \$54$ PER 1000 UNITS
ARE ADDED TO DISTRIBUTION COSTS

From \ To	*V*	*W*	*X*	*Y*	*Z*	Available 1000s ↓
A	92 +8	92 2	94 +4	90 17	94 0	19
B	96 11	104 10	102 7	108 +6	110 +4	28
C	100 +10	98 1	96 0	102 +6	100 24	25
Required 1000s →	11	13	7	17	24	72

Production costs:
A, 50 × 19 = $ 950
B, 62 × 28 = 1736
C, 54 × 25 = 1350
4036
Distribution costs 2986
Production and distribution costs $7022

little boxes represents the revenue, less the variable costs of production plus distribution. For example, the contribution values in column *V* of Table 7 are the revenues of $100 per thousand applicable to market area *V*, less the production plus distribution costs in that column. Thus the contribution for *AV* is $\$100 - 92 = \$8$.

Since we are dealing with contribution values rather than costs, we wish to generate a distribution plan that maximizes contribution. The distribution plan shown in Table 7 does just that. However, the distribution plan of Table 7 is exactly the same as the plan shown in Tables 2 and 6. Note also that the small figures in the lower left hand corners of the squares are identical, except that the signs have been reversed (since we are dealing with a problem involving maximizing contribution rather than minimizing cost). Of course, the total distribution cost has not changed. The general point is the same as the one made in connection with Table 7: adding a constant to the columns of the costs in the distribution matrix does not change the distribution plan. Even though the

TABLE 7

OPTIMUM DISTRIBUTION PLAN WHEN PRICES IN MARKET AREA OF $V = \$100$, $W = \$105$, $X = \$110$, $Y = \$115$, AND $Z = \$120$ ARE ACCOUNTED FOR*

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	8	13 (2)	16	25 (17)	26	19
B	4 (11)	1 (10)	8 (7)	7	10	28
C	0	7 (1)	14	13	20 (24)	25
Required 1000s →	11	13	7	17	24	72

Maximum contribution = \$1048
Production cost = 4036
Distribution cost = 2986
Total revenue = \$8070

*contribution is maximized

prices in market areas are different, the optimal distribution plan is the same when we assume linear costs and prices. Therefore, we can deal with just the costs of distribution in developing the best distribution plan.

There are other important technical aspects of the linear programming distribution model; however, the general conceptual framework is of considerable importance in generating excellent distribution plans that minimize objective costs. As we mentioned before, there are other objective and subjective factors that often enter the problem. Goal programming has been proposed as a methodology for handling some of them.

Goal Programming for Distribution Problems with Multiple Objectives

As we have said, the distribution problem has a multiple criteria set. The classic linear programming model provides a basis for minimizing distribution cost; however, a framework is needed in which to make trade-offs with other goals and objectives. In the distribution problem, some of the mangerial objectives, in addition to cost minimization, might be: minimizing system inventories; meeting the requirements of union contracts; special treatment of individual customers; stabilizing employment; balancing load among supplying plants; minimizing transportation hazards; and meeting transportation schedule contracts.

Lee and Moore [1973] present a goal programming model for distribution problems that provides managers with a rigorous framework to trade off some goals at the expense of others. Although the general linear programming model is used, Lee and Moore take advantage of the slack variables and minimize the deviations between goals that are set within the system of new problem constraints. As part of the process of constructing the model, all managerial goals are identified and arranged in order of priority. The solution procedure then successively seeks the achievement of goals in their order of priority. Since higher priority goals are constraints that cannot be violated, they may be achieved at the expense of lower priority goals.

LOCATIONAL CHOICE

Location is an important part of the design of a productive system, partly because the design integrally involves the distribution of the product or service, and partly because some details of physical layout may depend on location.

The location of facilities involves a commitment of resources to a longer-range plan. Thus, based on forecasts and predictions of the size and location of markets, the trends and goals for the organization are of great significance. Given these predictions, we allocate basic facilities and capacities for production and distribution that commonly account for the largest fraction of an organization's assets. These capital assets are enormous in manufacturing organizations, and even in service organizations the commitment of resources may be very large. But location and distribution are not important simply because of the concern over asset management. The location and distribution plans represents the basic strategy for gaining access to markets, and may have significant impact on revenue, costs, service levels to customers and clients, etc.

Location is not immediately obvious as a dominant factor in an enterprise's success or failure. Indeed, location is not uniformly important for all kinds of activities. For manufacturing organizations, decentralization within industries means that many good locations exist. A unique location that clearly is superior to all others may not exist. For example, in the U.S., television and radio plants are found in New England, the Middle Atlantic states, various parts of the Midwest, the South, and on the West Coast. Plastics raw material plants are located in Massachusetts, New York, New Jersey, Pennsylvania, Ohio, Michigan, and Tennessee. Plastics molding plants are dispersed throughout industrial America. Furniture is manufactured in North Carolina, New York, Illinois, Indiana, Virginia, Pennsylvania, California, Michigan, Ohio, Massachusetts, Wisconsin, Texas, and other states. The steel industry, which we tend to think of as centralized, actually is dispersed widely. The steel industry has plants in: eighteen states that produce pig iron; twenty-seven states that produce steel ingots; twenty-eight states that produce hot rolled products. Thus it appears, in practice, that there are likely to be a number of equally good locations, or that the location methods used could not discriminate among alternate locations.

General technological constraints commonly eliminate most of the possible locations; or a technological requirement may dominate and then, as Beckman [1968] states, that activity is oriented toward the technical requirement. For example, mining is raw material oriented; beer is water oriented; aluminum reduction is energy oriented; and service activities, including sales, are consumer or client oriented. If some technological requirement—such as raw material location, water, energy, etc.—does not dominate, then manufacturing industries may be transportation oriented, since raw materials must be transported to the plant and products distributed from it.

The intended criterion for choice of location is profit maximization for economic activities. If the prices of products are uniform in all locations, then the criterion becomes the minimization of relevant costs. If the costs of all inputs are independent of location but product prices vary, then the criterion for locational choice becomes maximum revenue. In such instances, locations will gravitate toward the location of consumers, and the general effect will be to disperse or decentralize facilities. If all prices and costs are independent of location, then the choice will be guided by: proximity to potential customers or clients; proximity to similar and competing organizations; proximity to centers of economic activity in general; and the personal choices of owners and managers.

INDUSTRIAL PLANT LOCATIONS

In most plant location models, the criterion is to minimize the sum of all costs affected by location. Some items of cost, such as freight, may be higher for city A and lower for city B; however, other costs (e.g., power) may have the reverse pattern. The best location is the one that minimizes costs on balance.

In attempting to ~~minimize costs, however,~~ we must consider not only today's costs but long-run costs as well. Therefore, we must try to predict the influence of some of the intangible factors that may influence future costs. Such factors as the attitude of city officials and townspeople toward a new factory site in their city may be an indication of future tax assessments. Poor local transportation facilities may mean future company expenditures to counterbalance this disadvantage. A short labor supply may cause labor rates to be bid up beyond rates measured during a location survey. The type of labor available may indicate future training expenditures. Thus, while a comparative cost analysis of various locations may point toward one community, an appraisal of intangible factors may lie behind a decision to select another. The result is an excellent example of a managerial decision requiring a multiple criterion function, in which trade-offs must be made between the various values and criteria [Buffa and Dyer, 1977; Easton, 1973].

The general problem indicating the nature of trade-offs that are required is illustrated by Table 8. Table 8a shows the results of comparative cost analyses for six alternate sites; Site 1 has the lowest projected monthly cost: $28,237. However, Table 8b lists fourteen subjective factors that management emphasized in this particular location study. The relative importance of these factors is not immediately obvious, nor is it obvious how they should be related to the objective costs. A formal methodology is needed to provide the basis for managerial decisions.

A Plant Location Model

where did data come from?

A model that attempts to deal with the multidimensional location problem was developed by Brown and Gibson [1972]. This model classifies criteria affecting location according to the model structure, quantifies the criteria, and balances or tradesoff among criteria.

Classification of Criteria. The model deals with any individual list of criteria set by the analyst, but classifies them uniformly as follows:

1. *Critical.* A criterion is critical if its nature may preclude the location of a plant at a particular site, regardless of other conditions that might exist. For example, a water oriented enterprise, such as a beer company, would not consider a site where water shortage was a possibility; an energy oriented enterprise, such as aluminum smelting, would not consider sites where low cost and plentiful electrical energy were not available. Critical factors have the effect of eliminating certain sites from consideration.

2. *Objective.* Objective criteria can be evaluated in monetary terms, such as

TABLE 8
(a) OBJECTIVE FACTOR COSTS, AND
(b) SUBJECTIVE FACTORS FOR SIX SITES

(a)

Site	*Material*	*Marketing*	*Utilities*	*Labor*	*Building*	*Taxes*	*Total Objective-Factor Cost (OFC)*
1	$1,079	$1,316	$ 9,460	$12,773	$514	$3,095	$28,237
2	945	1,485	11,563	11,249	563	3,470	29,275
3	490	1,467	12,768	10,422	539	3,580	29,266
4	979	1,600	10,548	12,159	490	3,755	29,531
5	925	1,263	10,898	12,333	612	3,701	29,732
6	1,507	1,950	11,628	12,244	612	3,393	31,334

(b)

Availability of transportation	Union activities	Community services	Competition
			Complementary industries
Industrial sites	Recreation facilities	Employee transportation facilities	
			Availability of labor
Climate	Housing		
		Cost of living	
Educational facilities	Future growth		

restricted plant building · govt regulat— water supply

Adapted with permission from: P. A. Brown and D. F. Gibson, "A Quantified Model for Facility Site Selection—Application to a Multiplant location Problem," *AIIE Transactions* 4(March 1972): 1–10. Copyright American Institute of Industrial Engineers, Inc. 25 Technology Park/Atlanta, Norcross, Georgia 30071.

labor, raw materials, utilities, taxes, etc. A factor can be both objective and critical; for example, the adequacy of labor would be a critical factor, while labor cost would be an objective factor.

3. *Subjective*. Subjective criteria are characterized by a qualitative type of measurement. For example, the nature of union relationships and activity may be evaluated, but no monetary equivalent can be established. Again, criteria can be classified as both critical and subjective.

Model Structure. For each site i, a location measure, LM_i, is defined which reflects the relative values for each criterion. The critical factor measure is the sum of the products of the individual critical factor indexes for site i with respect to critical factor j. Since the critical factor index for each site is either zero or one, depending on whether the site has an adequacy of the factor or not, if any critical factor index is zero then the critical factor measure and the overall location measure LM_i are also zero. Site i therefore would be eliminated from consideration.

The objective criteria are converted to dimensionless indexes to establish comparability between objective and subjective criteria. As a result, the site with the minimum cost has the largest objective factor measure; the relationships of total costs between sites is retained; and the sum of the objective factor is 1.

The subjective factor measure for each site is influenced by the relative weight of each subjective factor, and the weight of site i is relative to all other sites for each of the subjective factors. Preference theory is used to assign weights to subjective factors in a consistent and systematic manner. The application of preference theory is an extremely important part of obtaining relative evaluations of subjective factors, and the methodology ensures comparability with objective factors. The procedure involves comparing subjective factors two at a time: if the first factor is preferred over the second, then the numerical value of 1 is assigned to the first factor and 0 to the second, and vice versa for the opposite result. If neither factor is preferred, a rating of 1 is given to both. Procedures are included for higher order rankings as well. As with objective factors, the ratings are normalized, so that the sum of subjective weightings for a given site adds to 1.

Finally, the objective factor decision weight, X, must be determined. This factors establishes the relative importance of the objective and subjective factors in the overall location problem. The decision commonly is based on action by a management committee, reflecting policies, past data, and an integration of a wide variety of subjective factors. The determination of X could be subjected logically to a Delphi process.

Once all the data inputs have been established, the figure of merit, LM_i, can be computed for each site, and the site that receives the largest LM_i is selected. Brown and Gibson extend the model to multiplant location, and present a computed example of the evaluation of six sites, involving capacity constraints, for a specific industry. Sensitivity analyses are shown to indicate how decision would change when the objective factor decision weight, X, is varied from 0 to 1.0. The entire procedure is programmed for electronic computation, using a 0/1 programming alogrithm that is capable of treating problems as large as 150 variables and 50 constraints.

Capital Expenditure–Volume Effects

Another important variable among alternate locations is the relationship between fixed and variable expenses. The fixed investment costs can differ considerably, depending on local construction and land costs, and variations in the particular site selected (such as availability of railroad spurs into the property, and adequacy of power and gas lines). Therefore, the concept of breakeven analysis could be used to contrast the objective cost factors of individual locations. Figure 2 shows such a contrast between the fixed and variable components for nine locations. Of course, such a breakeven analysis is valid only for volumes near the design capacity, since large changes in volume would entail differences in capital investment as well as possible differences in variable costs. Note that although City *J* offers attractive fixed costs, the operating costs make it the most expensive location for the volume contemplated. When viewed in this way, City *D* appears to be the most economical location.

MULTIPLANT LOCATION

As we mentioned before, multiplant location is influenced by existing locations as well as by the kinds of economic factors that we have discussed already. Each location considered must be placed in economic perspective with the existing plants and market areas. The objective is to select the new location that minimizes the total production–distribution cost. This aim is somewhat different from the location analysis for a single plant, because each location requires a different allocation of capacity to markets from the several plants in order to minimize overall costs. The formal problem can be placed in a linear programming framework and solved in a distribution matrix, according to the general methods of Appendix C.

An Example

Let us examine the problem conditions of the Good-Wear Shoe Company case. The company manufactures a line of inexpensive women's shoes in two plants, Detroit and Chicago, and now distributes to five main distribution centers—Milwaukee, Cleveland, Cincinnati, Buffalo, and Atlanta—from which the shoes are shipped to retail shoe stores. The fifth center, Atlanta, has been added recently to serve the Southeast, an area in which the company has been expanding sales effort. To meet increased demand, the company has decided to build a new plant with a capacity of 25,000 pairs per week.

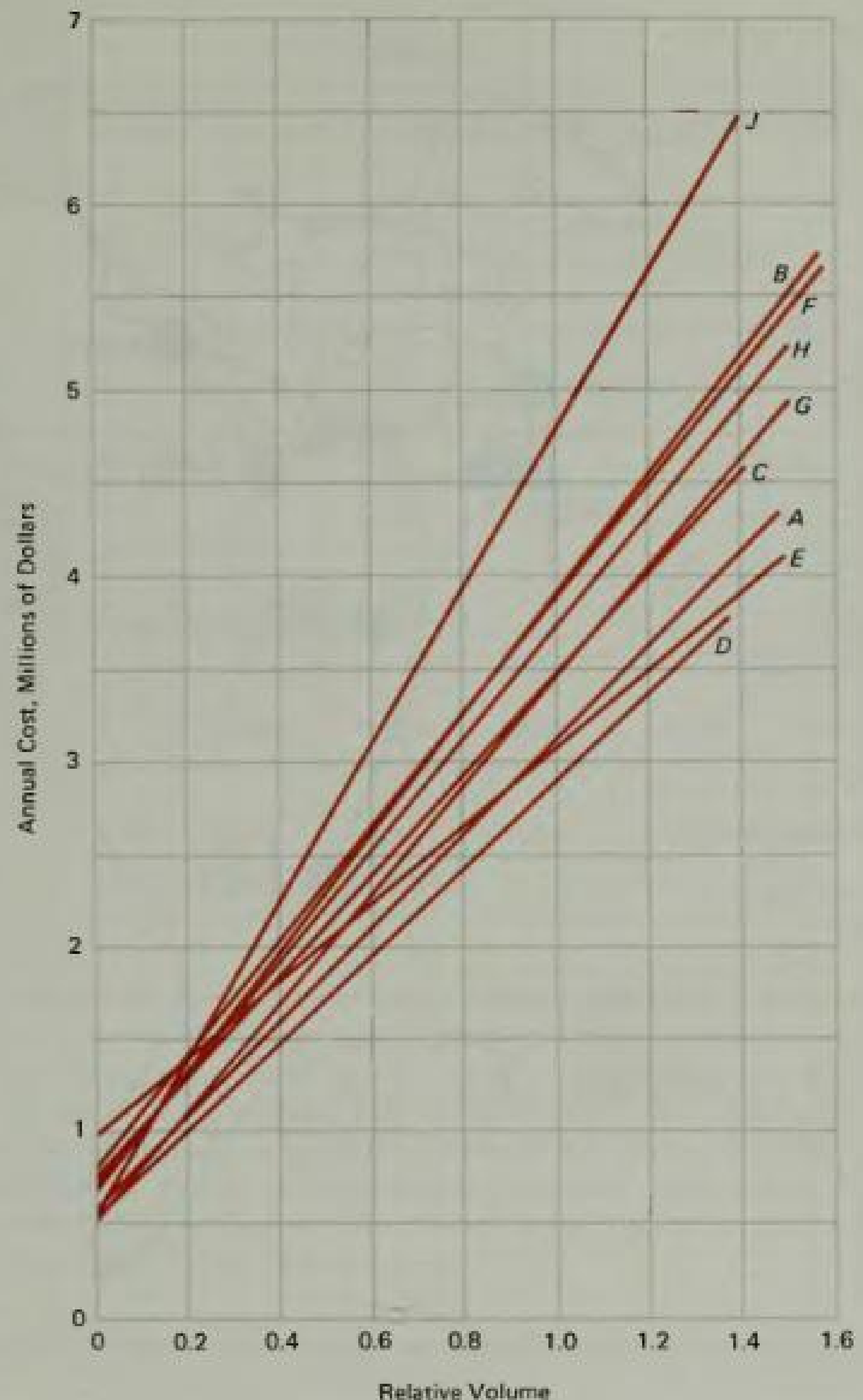


Figure 2. Fixed and variable cost relationships for the nine locations of Figure 2. Breakeven costs between pairs of alternatives occur where their cost-volume lines cross

General surveys have narrowed the choice to three locations: Cincinnati, Cleveland, and Atlanta (see Figure 3). The production and distribution costs, as well as the plant capacities and distribution demands, are shown in Table 9. The proposed plant capacity of 25,000 pairs weekly reflects average forecast demand in the several market areas and allows for some expected growth of sales. Distribution costs include freight,

Figure 3. Geographical locations of factories, distribution centers, and potential new factory locations

handling, and warehousing costs. As expected, the production costs are low in the Atlanta plant, but distribution costs are relatively high compared to the other two locations.

The important question now is, Which location will yield the lowest cost for the enterprise in combination with the existing plants and distribution centers? To determine this, we solve three distribution matrixes, one for each combination. Capacities and demands are as given, and the costs in each cell are the production plus distribution costs for each factory-distribution center combination. Figure 4 shows the resulting three optimum matrixes, with the total cost for each.

TABLE 9
PRODUCTION COSTS, DISTRIBUTION COSTS, PLANT CAPACITIES, AND MARKET DEMANDS FOR THE GOOD-WEAR SHOE COMPANY

From Plants / To Distribution Centers	Distribution Costs Per Pair, Handling, Warehousing, and Freight					Forecast Weekly Market Demand Pairs
	Existing Plants		Proposed Plant Locations			
	Detroit	Chicago	Cincinnati	Cleveland	Atlanta	
Milwaukee	$0.42	$0.32	$0.46	$0.44	$0.48	10,000
Cleveland	0.36	0.44	0.37	0.30	0.45	15,000
Cincinnati	0.41	0.42	0.30	0.37	0.43	16,000
Buffalo	0.38	0.48	0.42	0.38	0.46	19,000
Atlanta	0.50	0.49	0.43	0.45	0.27	12,000
Normal weekly plant capacity, pairs	27,000	20,000	25,000	25,000	25,000	
Unit production cost	$2.70	$2.68	$2.64	$2.69	$2.62	

The solutions to the formal linear programming problems show that the Atlanta plant is a slightly better location in terms of variable costs. Atlanta is favored again by lower land and construction costs. Finally, if we consider future possible expansion of markets in the South without going into a discussion of other intangible factors, the Atlanta location seems to have an edge in terms of both present and future costs. The problem solution, as shown, is fairly simple; however, other restricting conditions could have been added, such as inventory levels in the various locations, as well as the location of future distribution centers.

Locational Dynamics for Multiplants. The decision to build the Atlanta plant was based on current costs, demand breakdowns, and an assessment of the future. But suppose that the balance of these factors changes. Then the allocation of capacity to markets should change also, in order to yield a minimum total cost for whatever conditions exist. Thus, location analysis is continuous and necessary.

Suppose, for example, that after the Atlanta plant was built, our shoe company experienced a decline in demand because of increased competition from the low-cost Italian imports. Assume that, instead of a total demand of 72,000 units per week (as hoped and planned for), there is a demand of only 56,000 units, divided among the market areas as follows: Milwaukee, 9,000; Cleveland, 13,000; Cincinnati, 11,000; Buffalo, 15,000; Atlanta, 8,000.

From plants / To dist. centers	Detroit	Chicago	Cincinnati	Demand, 1000s ↓
Milwaukee	3.12	3.00, 10	3.10	10
Cleveland	3.06, 8	3.12	3.01, 7	15
Cincinnati	3.11	3.10	2.94, 16	16
Buffalo	3.08, 19	3.16	3.06	19
Atlanta	3.20	3.17, 10	3.07, 2	12
Capacity, 1000s →	27	20	25	72

Production cost = $192,500
Distribution cost = 26,450
Total = $218,950

From plants / To dist. centers	Detroit	Chicago	Cleveland	Demand, 1000s ↓
Milwaukee	3.12	3.00, 10	3.13	10
Cleveland	3.06	3.12	2.99, 15	15
Cincinnati	3.11, 8	3.10	3.06, 8	16
Buffalo	3.08, 19	3.16	3.07	19
Atlanta	3.20	3.17, 10	3.14, 2	12
Capacity, 1000s →	27	20	25	72

Production cost = $193,750
Distribution cost = 26,960
Total = $220,710

Figure 4. Optimum production-distribution solutions for three proposed locations for the added plant

To dist. centers \ From plants	Detroit	Chicago	Atlanta	Demand, 1000s ↓
Milwaukee	3.12	3.00 (10)	3.10	10
Cleveland	3.06 (15)	3.12	3.07	15
Cincinnati	3.11	3.10 (10)	3.05 (6)	16
Buffalo	3.08 (12)	3.16	3.08 (7)	19
Atlanta	3.20	3.17	2.89 (12)	12
Capacity, 1000s →	27	20	25	72

Production cost = $192,000
Distribution cost = 26,400
Total = $218,400

Figure 4. (continued)

The Atlanta and Detroit plants can meet this demand without the Chicago plant by resorting to overtime work. In fact, any two of the three plants can meet the demand if overtime work is used. Therefore, these alternatives, as well as the alternative of continuing to operate all three plants at partial capacities, can be used. Which alternative is best, and how much difference does it make? To determine the answers, we must know, in addition to the data that we have already, the cost of overtime work in each plant, the additional capacity available through overtime, and the cost reductions available by shutting down one plant. For each plant these data are as follows:

	Detroit	Chicago	Atlanta
Production cost at overtime	$ 3.37/unit	$ 3.33/unit	$ 3.27/unit
Additional overtime capacity, units	7,000/wk	5,000/wk	6,000/wk
Fixed costs:			
When operating	$12,000/wk	$9,000/wk	$13,000/wk
When shut down	5,000/wk	4,000/wk	6,000/wk

Now we can set up a distribution matrix for each of the four alternatives and determine the total costs for each plan, as well as the appropriate fixed costs.

To dist. centers \ From plants	Detroit	Chicago	Atlanta	Demand, 1000s ↓
Milwaukee	3.12	3.00 (9)	3.10	9
Cleveland	3.06 (13)	3.12	3.07	13
Cincinnati	3.11	3.10	3.05 (11)	11
Buffalo	3.08 (14)	3.16	3.08 (1)	15
Atlanta	3.20	3.17	2.89 (8)	8
Unused Capacity	0	0 (11)	0 (5)	16
Capacity, 1000s →	27	20	25	72

Variable cost = $169,650
Fixed cost:
Detroit, 12,000
Chicago, 9,000
Atlanta, 13,000 34,000
Total = $203,650

(a) All plants operating

Figure 5. Optimum production-distribution solutions for four alternatives of plant operations: (b), (c), and (d) have other equally optimal solutions

This is done and the optimum distribution solution for each plan, including the appropriate fixed costs, is shown in Figure 5. Note that to keep the problem within the linear programming framework, we regard the overtime capacity as a separate source of supply. In actual shipment there would be no need to separate units produced on overtime. The most advantageous action is to shut down the Chicago plant and depend on the Detroit and Atlanta plants, with 4,000 units per week produced on overtime in the Atlanta plant. This alternative is $1,500 per week better than the next best solution, and $2,250 per week better than the worst.

If the situation were even more serious, we might want to consider disposing of a plant. A similar analysis could be made to determine which plant should go. In such a case, all fixed costs attributed to the discarded plant would be eliminated, and the capital recovered from the sale would have to be evaluated.

To dist. centers / From plants	Chicago	Atlanta	Chic.-O.T.	At.-O.T.	Demand, 1000s ↓
Milwaukee	3.00 9	3.10	3.65	3.75	9
Cleveland	3.12	3.07 8	3.77 5	3.72	13
Cincinnati	3.10 11	3.05	3.75	3.70	11
Buffalo	3.16	3.08 9	3.81	3.73 6	15
Atlanta	3.17	2.89 8	3.82	3.54	8
Capacity, 1000s →	20	25	5	6	56

Variable cost = $177,730
Fixed cost:
Detroit, 5,000
Chicago, 9,000
Atlanta, 13,000 27,000
Total = $204,730

(*b*) Detroit plant shut down

Figure 5. (continued)

Foreign Locations

The lure usually held out for manufacturers to locate facilities in foreign countries has been the relatively cheap labor available in some areas. While rapidly increasing wages in many foreign countries have changed this situation within recent years, this argument still can be made in various foreign areas.

For a particular manufacturer, the important question is, *Is a net advantage available in a foreign location?* There are several important reasons why there may not be. Wage levels themselves are not the important parameter; rather, *labor costs* will determine the advantage or disadvantage; wages can be high and labor costs can be simultaneously low. The equating factor is *productivity.* While the American worker is paid much more per hour, the relatively large capital investment expended multiplies his/her efforts through special tools, mechanization, and automation.

Of course, the temptation is to assume that we can couple the advantage in lower wages with high productivity by using the same levels of mechanization and mangerial practice abroad. The difference in basic production economics in

To dist. centers / From plants	Detroit	Atlanta	Det.-O.T.	At.-O.T.	Demand, 1000s ↓
Milwaukee	3.12	3.10 (6)	3.79	3.75 (3)	9
Cleveland	3.06 (13)	3.07	3.73	3.72	13
Cincinnati	3.11	3.05 (11)	3.78	3.70	11
Buffalo	3.08 (14)	3.08	3.75	3.73 (1)	15
Atlanta	3.20	2.89 (8)	3.87	3.54	8
Unused capacity	0	0	0 (7)	0 (2)	9
Capacity, 1000s →	27	25	7	6	65

Variable cost = $173,150
Fixed cost:
Detroit, 12,000
Chicago, 4,000
Atlanta, 13,000 29,000
Total = $202,150

(c) Chicago plant shut down

Figure 5. (continued)

these two contrasting situations must be noted. Since labor is cheap relative to equipment in some foreign locations, we may find it wise to use relatively more labor and less expensive machinery; the resulting productivity and final labor costs thus would be more in line with those usually achieved in the foreign environment. The most economical manufacturing methods and techniques are not necessarily those with the greatest possible mechanization, but those that, for a given situation, strike a balance between the costs of labor and the costs of machinery.

There are many costs in addition to labor to consider. If there is a net labor cost advantage, will it be counterblanced by higher costs of materials, fuel and power, equipment, credit, etc.? TAXES, Foreign country laws

Studies of production costs in the United States and abroad, involving companies with both domestic and foreign operations, have indicated considerable variability in the relative advantage or disadvantage of foreign locations. Ap-

To dist. centers \ From plants	Detroit	Chicago	Det.-O.T.	Chic.-O.T.	Demand 1000s ↓
Milwaukee	3.12	3.00 9	3.79	3.65	9
Cleveland	3.06 9	3.12	3.73 4	3.77	13
Cincinnati	3.11 3	3.10 8	3.78	3.75	11
Buffalo	3.08 15	3.16	3.75	3.81	15
Atlanta	3.20	3.17 3	3.87	3.82 5	8
Unused capacity	0	0	0 3	0	3
Capacity, 1000s →	27	20	7	5	59

Variable cost = $178,400
Fixed cost:
Detroit, 12,000
Chicago, 9,000
Atlanta 6,000 27,000
Total = $205,400

(d) Atlanta plant shut down

Figure 5. (continued)

parently there are some products, industries, or companies that are favored by the structure of foreign costs; however, these same conditions are unfavorable for others. Products in the industries that have a relatively high labor content seemed to have lower costs abroad. On the other hand, industries whose cost structure is dominated by materials, energy, and capital had higher costs abroad.

WAREHOUSE LOCATION

While plant location often involves the personal preferences of the owners and managers, especially if the location also dictates their residences, warehouse location more often yields to rational analysis. A number of approaches to the problem currently are being taken, including linear programming, simulation, heuristic procedures, and the branch-and-bound technique. For example, simulation models have been developed and used by the Nestle Company [Gerson and Maffei, 1968], H. J. Heinz [Shycon and Maffei, 1960], and Ralston Purina [Markland, 1973]. Heuristic models have been developed [Keuhn and Hamburger, 1963] and applied at the B. F. Goodrich Company [Khumawala, 1972]. Finally, the branch-and-bound technique has been applied at ESSO [Effroymson and Ray, 1966].

Even with a single producing plant, the problem is not simple, for the entire distribution system design is involved. If the organization is committed to the site of the production plant, then it can be concerned with designing a distribution system to link its plant with its markets. Some of the remaining problems for resolution are:

1. Whether or not to use field warehouses and if so, how many and where to locate them.
2. What modes of transportation to use.
3. What inventories will be required to operate the system, including those at the plant, in transit, and in the warehouses.

When markets are concentrated, warehouse locations tend to be determined at or near the major points of concentration. Boundaries between warehouse territories theoretically are determined to equalize the distribution cost from either warehouse at the boundary. As a practical matter, natural boundaries often occur in the form of rivers, mountains, state lines, etc.

The practical problems increase substantially when multiple plants producing multiple products are servicing multiple warehouses. The products produced by the several plants may be overlapping or independent. Should each plant supply customers (or warehouses) directly, or should products not made in all plants be cross-shipped among plants so that orders can be filled as a unit? What territory should be served by a given plant for products made in more than one plant? For products made in more than one plant, how should production schedules be dovetailed to serve the entire market?

ESSO—The Branch-and-Bound Technique

Effroymson and Ray [1966] developed a model that later was improved by Khumawala [1972]. It involves a procedure that uses branch-and-bound methods and linear programming to produce optimal solutions with reasonable computing time. ESSO applied the location system to several location problems that included four plants, fifty warehouses, and two hundred customer zones. The procedure involves applying rules for including or excluding warehouse locations, depending on whether or not their competitive savings cover their fixed costs of operation. Linear programming is used at points in the procedure to compute lower bound costs. By following out the branches and computing upper and lower bound costs, warehouse locations can be either definitely included in the optimum solution or excluded, leading to the final optimum solution. An example of the procedure, together with a case history, is given in Atkins and Shriver [1968].

Ralston Purina—An Application of Simulation

Markland [1973] applied a computer simulation methodology in evaluating field warehouse location configurations and inventory levels for the Ralston Purina Company, and reported typical results for the Midwest region. The basic results for distribution in the ten Midwestern states indicated that only three of the five existing warehouses were required in an optimal solution, and an 85 percent inventory service level minimized distribution cost.

The basic structure of product flow is shown by Figure 6, which indicates the nature of the multilevel, multiproduct distribution system involving plant warehouses, field warehouses, wholesalers, and retail grocers. Shipments from the five warehouses may go to other plant warehouses, to field warehouses or to wholesalers; and inventories are maintained at these three levels. Also shipments may go from any of the five field warehouses to any of the twenty-nine demand analyses areas representing the wholesale level, or to any of the other field warehouses.

Model Structure. Markland modeled the distribution system on the basic format of systems dynamics as a dynamic feedback control system in which product flow is the main control variable. Product flow and inventory level equations were written to represent all flow combinations and inventory levels at plant and field warehouses. Constraints on maximum inventory levels at plant and field warehouses were established, as well as constraints on maximum plant production capacity for all products.

The costs associated with product distribution involved transportation costs, warehousing costs, and penalty costs. Both transportation and warehousing activities were represented in the model as nonlinear costs. Transportation costs reflected shipment weight, mode of shipment, and distance. Warehousing costs involved a fixed and a variable charge, based on company cost data.

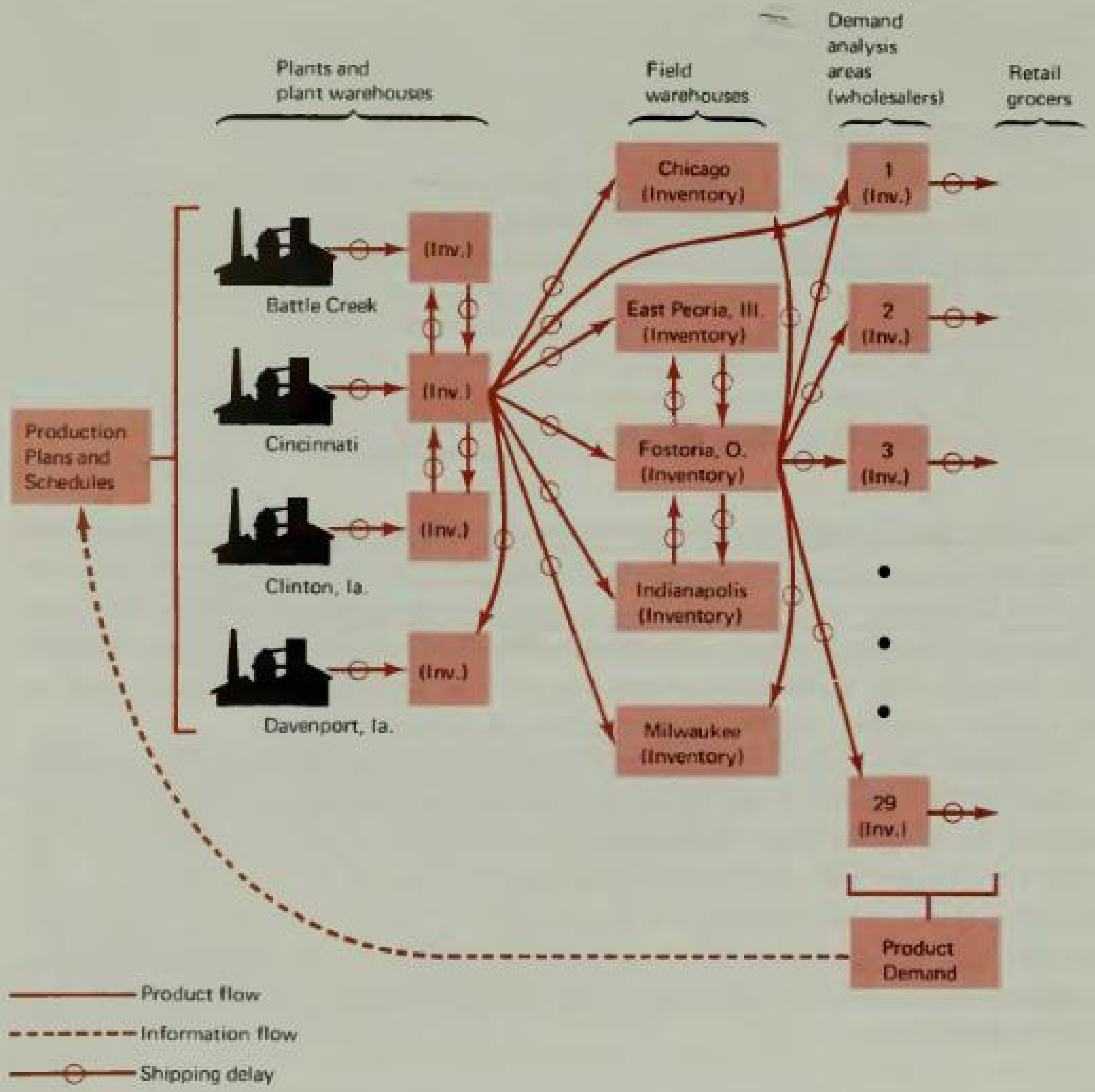


Figure 6. Ralston Purina distribution system. Only typical flow lines are shown

Simulation. The system was driven by a product demand simulator and the determination by customers of a shipment source. The shipment source was determined by the following heuristic:

> *Facilities Search Heuristic:* Search possible facilities in order of increasing distance from the end customer placing the order. If order is to a field warehouse, search all field warehouses first, then search all manufacturing facility warehouses. Do the opposite if the order is to a manufacturing facility warehouse. If the order is satisfied by a field warehouse (manufacturing facility warehouse) which is not the closest to the end-customer, compute a cost of order shifting which represents the customer's loss in goodwill from having his shipment delayed. If the order cannot be satisfied at any facility, create a back order and compute a back ordering cost which represents the

customer's loss in goodwill fron his demand not being satisfied. Utilize the same process for the situation in which the order can be partially satisfied at a facility, i.e., move from facility to facility, in order of increasing distance- until the order is completely satisfied or a back order is required [Markland, 1973].

Once the shipment source was established, the shipment mode was simulated from probability distributions, transportation cost computed, and inventory levels adjusted. The process then was repeated for all products, customer orders, and time periods.

Sample Results. The system was simulated for six different field warehouse configurations, in which the number of field warehouses was varied from zero to the existing five, and for different inventory service levels. Thirty-two field-warehouse location patterns that involved combinations of warehouses were tested, using a procedure of dropping warehouses from the existing pattern. Table 10 shows sample results, that indicate that one of the combinations of three field warehouses minimized distribution costs. Transportation costs in Table 10 dominated the choice, since the maximum penalty costs also occurred with the three-warehouse configuration. Note that Ralston Purina apparently saved $132,000 per year by consolidating field warehouses from five to three [(959 − 926) 1,000 × 4 = $132,000]. The policy of eliminating intermediate warehouses would increase costs considerably [(991 − 959) 1,000 × 4 =

TABLE 10

DISTRIBUTION COST SUMMARY (THREE-MONTH PERIOD), ALTERNATIVE WAREHOUSE LOCATIONS ($000)

	Warehouse Configuration					
Type of Cost	Five Field Warehouses	Four** Field Warehouses	Three** Field Warehouses	Two** Field Warehouses	One** Field Warehouses	Zero Field Warehouses
Transportation costs*						
Mfg. facility - customer	$ 514	$ 530	$ 525	$ 567	$ 574	$ 665
Mfg. facility - field warehouse	36	30	20	17	10	0
Field warehouse - customer	60	51	42	20	28	0
Subtotal	610	611	567	604	612	665
Warehousing costs						
Field warehouses	30	25	16	13	6	0
Mfg. facility warehouses	303	304	310	313	316	320
Subtotal	333	329	326	326	322	320
Penalty costs						
Order-shifting costs	12	8	9	7	5	22
Back ordering costs	4	2	4	4	3	4
Subtotal	16	10	13	11	8	26
Total distribution cost	$ 959	$ 950	$ 926	$ 941	$ 942	$ 991

*Transportation costs associated with interplant and interwarehouse shipments are not shown, for the sake of brevity. They were of small magnitude, and did not vary significantly between the various simulation runs.

**Minimum cost combination for the particular warehouse configuration.

SOURCE: R. E. Markland, "Analyzing Geographically Discrete Warehousing Networks by Computer Simulation," Decision Sciences 4(April 1973):216-236.

TABLE 11

DISTRIBUTION COST SUMMARY (THREE-MONTH PERIOD), ALTERNATIVE INVENTORY LEVELS ($000)

Type of Cost	Inventory Level										
	100%	95%	90%	85%**	80%	75%	70%	65%	60%	55%	50%
Transportation costs*											
Mfg. facility–customer	$ 460	660	544	543	559	575	613	638	694	736	761
Mfg. facility–field warehouse	147	45	40	36	41	36	38	39	41	43	44
Field warehouse–customer	235	64	50	58	58	58	61	71	60	50	52
Subtotal	842	769	684	637	658	669	713	748	795	829	857
Warehousing costs											
Field warehouse	318	28	27	26	26	25	25	24	24	24	23
Mfg. facility warehouses	71	303	304	305	306	307	308	309	310	310	312
Subtotal	389	331	331	331	332	332	333	333	334	334	335
Penalty costs											
Order-shifting costs	0	11	9	11	12	11	12	13	14	15	16
Back ordering costs	0	3	3	3	3	3	3	3	3	4	4
Subtotal	0	14	12	14	15	14	15	16	17	19	20
Total distribution cost	$1231	1114	1027	982	1005	1015	1061	1097	1146	1182	1212

*Transportation costs associated with interplant and interwarehouse shipments are not shown for the sake of brevity. They were of small magnitude, and did not vary significantly between the various simulation runs.
**Minimum cost inventory level.

SOURCE: R. E. Markland, "Analyzing Geographically Discrete Warehousing Networks by Computer Simulation," Decision Sciences 4(April 1973): 216–236.

$128,000 per year], and would be $260,000 per year more expensive than the optimal policy of using only three field warehouses.

Table 11 indicates that an 85 percent inventory level minimizes distribution cost. Again, the decision is virtually made by transportation costs. To understand the significance of Table 11, we must interpret 100 percent inventory service to mean that a customer order placed on a given facility is supplied immediately from inventory by that facility, with no back-ordering or shifting of the order to other facilities, as shown by the cost of Table 11 under the 100 percent inventory level column.

Note that the large cost reductions that come from lower inventory service levels largely are due to lower transportation costs. Why would a lower service level affect transportation cost? Presumably, a customer order would not necessarily be placed at a facility that could supply at the most advantageous transportation costs. When management can manipulate these orders, and supply from the lowest transportation cost source, the gains can be very great. The offset costs of such a policy are the loss of goodwill resulting from order-shifting and back ordering. But the net gains apparently are overwhelming. The 100 percent inventory level policy, compared to the 85 percent policy, results in an annual cost difference of $996,000.

Obviously, if the sample results for the Midwest states are representative, the cost reductions available to Ralston Purina are extremely large when extended to their national distribution system. Whether that system is considered as one large distribution system or as a set of independent regions (such as the Midwest region in the example) will have effects—particularly near the marginal areas at the borderlines between regions.

LOCATION OF REGIONAL HEALTH SERVICES

While industrial plant and warehouse location are influenced strongly by distribution costs, the location of services is oriented toward the location of users. Medical facilities must be placed within the reach of people who need these services. Emergency units need to be located so that they can provide a certain minimum response time. Retail outlets seek out locations that can maximize their revenue. Thus, the fundamental differences in the nature of the services location problem have necessitated different techniques than have been used for industrial plant and warehouse location.

The objective of health care planners is often stated as allocating facilities to locations such that primary health care demanded by the population is maximized. The implementation of such an objective depends on how one weights and trades off several criteria, such as overall utilization, average travel distance by users, distance per capita, and so on. In order to make rational location plans, demographic data is needed that characterizes aspects of user behavior. Abernathy and Hershey [1972] developed a location model that provides optimum locations within a region for different criteria, thus giving decision makers a basis for trade-off analysis in their determination of actual locations.

Model Description

The model assumes a defined medical service area, such as is shown in Figure 7, in which the numbered geographic areas represent census blocks. The xy axes provide a grid for specifying the location of populations and health care facilities. Three cities are located within the area of Figure 7, and demographic data are available concerning the population's behavior in using medical facilities. The demographic data are divided into strata that exhibit relatively homogeneous patterns of medical facility use, particularly with respect to the decrease in the propensity to seek care that results as distance to a health care facility increases.

Based on the stratified data, equations are developed that describe the utilization of a facility in relation to distance for all strata in each census block, and

Figure 7. A hypothetical medical service area with 32 census blocks and three cities. City populations are (approximately) $A = 17{,}000$, $B = 9{,}000$, and $C = 13{,}000$. Distances on x–y axes are in miles.

SOURCE: W. J. Abernathy and J. C. Hershey, "A Spatial-Allocation Model for Regional Health-Services Planning," *Operations Research* 20(May–June 1972): 629–642.

the probability of choosing a particular center, relative to the distance to the nearest center. Total demand is determined by summing the demand over all centers, strata, and census blocks. The model determines the (xy) coordinates for health centers in such a way that some criterion is optimized. In the example problem, four criteria were used:

1. *Maximum utilization*. Maximize the total number of visits to centers.
2. *Minimize distance per capita*. Minimize the average distance per capita to the closest center.
3. *Minimize distance per visit*. Minimize the average per visit travel distance to the nearest center.
4. *Minimize percent degradation in utilization*. Minimize the average reduction in the number of visits that individuals in the community make as a

TABLE 12
LOCATION COORDINATES IN MILES FOR THREE CRITERIA AND DIFFERENT NUMBERS OF CENTERS*

Center number	Criterion					
	(1) *Maximize utilization*		(2) *Minimize distance per capita*		(3) *Minimize distance per encounter*	
	x	y	x	y	x	y
I With 1 center						
1	21.00	−3.00	0.64	1.20	−8.70	10.10
II With 2 centers						
1	21.4	−3.7	17.6	−3.30	18.50	−3.30
2	−9.89	10.4	9.89	10.4	−9.90	10.40
III With 3 centers						
1	22.40	−3.1	21.52	−2.78	22.30	−3.20
2	−10.16	10.40	−10.20	10.40	−10.20	10.40
3	3.63	−2.75	3.60	−2.80	3.60	−2.80
IV With 4 centers						
1	22.40	−3.14	22.00	−3.50	21.23	−3.08
2	−10.20	10.40	−10.10	10.30	−9.80	10.40
3	3.59	−2.78	2.69	−4.80	3.61	−2.70
4	11.32	−2.25	3.76	3.04	−11.35	3.00
V With 5 centers**						
1	22.40	−3.10				
2	−9.72	10.61				
3	3.24	−3.19				
4	−11.62	3.24				
5	11.04	−2.00				

*See Figures 7 and 8 for locations of coordinates.
**Determined only for the first criterion.

SOURCE: W. J. Abernathy, and J. C. Hershey, "A Spatial-Allocation Model for Regional Health-Services Planning," *Operations Research* 20(May–June 1972): 629–642.

percentage of the visits made, if the individuals are in immediate proximity to a center.

The system was optimized using computer search methods that search a multidimensional criterion surface to find the optimum point. Any of the preceding criteria, as well as others, could be used.

An Example. Abernathy and Hershey provide the results of a study for the medical service region of Figure 7. Data were given for the key variables in four strata, together with the coordinate locations of population centers by stratum. Optimum locations were generated using the search methodology for one, two, three, and four centers for each of three criteria independently. Table 12 shows the coordinate locations of centers for three different criteria. Figure 8 shows the

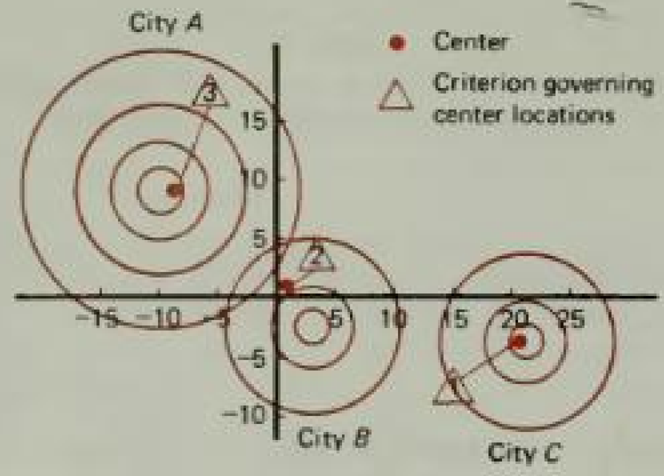


Figure 8. Location of one center based on three different criteria

SOURCE: W. J. Abernathy and J. C. Hershey, "A Spatial-Allocation Model for Regional Health-Services Planning," *Operations Research* 20(May–June 1972): 629–642.

locations of one center in relation to the grid and the three cities for the three criteria.

If only one health center is allocated to the region, its optimum location depends importantly on which criterion is used, because of the strongly different behavior patterns in different geographic areas, as shown in Figure 8. When Criterion 1 (maximize utilization) is used, the center is located near the center of City *C* because this area contains a large number of individuals for whom distance is a strong barrier. This location choice, however, tends to increase overall distance per capita. This criterion favors individuals whose use of the health center is curtailed most by large travel distance. However, it shifts the transportation costs to those who are less sensitive to distance. Criterion 2 (minimum distance per capita) results in a location near the population-distance centroid for the region; this happens to be near City *B*. Criterion 3 (minimum distance per encounter) results in a location near the center of the largest city, City *A*. This is regarded as a local criterion, since it does not consistently represent needs throughout the region, and results in the lowest overall utilization.

Abernathy and Hershey provide the decision maker with a set of graphs that relate the differential effects of the three criteria for one to four centers. These graphs can be used to help the decision maker decide on trade-offs among criteria. Each of the three graphs present the value of each criterion when it serves as the basis for optimizing the location of from one to four centers, and providing a comparison of the criterion with the degradation in the same criterion when each of the other two criteria is optimized. The decision maker can compare the cost implications of suggested numbers, sizes, and locations of centers with the marginal benefits of a given level of service in those instances where benefits are measured in terms of the three different criteria.

LOCATION OF EMERGENCY UNITS

The nature of the location of emergency units, such as ambulances, fire stations and police services, is somewhat more dynamic than that of industrial plants, warehouses, and medical facilities. Different demands for service and response time requirements create the need for adaptation, as units—called or dispatched to perform service—leave their previous locations unprotected. Since personnel costs constitute as much as 90 to 98 percent for emergency services [Chaiken and Larson, 1972], a commitment to fixed locations is obviously less justified.

Fire stations usually are situated at fixed locations when fire fighting services are dispatched, while ambulances and police cars are more mobile and may be dispatched while cruising, or from any location. As Chaiken and Larson point out, however, the distinction between fixed and mobile locations breaks down in periods of high demand, since the units may be dispatched directly from one incident to the next.

For emergency units, the location problem might more correctly be termed an *allocation* or spreadout *deployment* problem. It is the strategy by which the services are made available to meet some standard of performance, such as reaction time. Therefore, bound up within the location or deployment problem are: the total capacity of the system; the location or patrol patterns; priorities for different kinds of calls and the circumstances under which calls may be backlogged; decisions as to which units actually should be dispatched in response to a call; and the circumstances under which relocation or redeployment takes place. The major objective of an emergency system design is to reduce to a low level the possibility that a true emergency call will need to be backlogged because all the units are busy.

Design of Response Areas. The location problem is closely tied in with the area for which the emergency unit has primary responsibility. Several objectives may be involved in the actual design of these response areas including minimum response time, work load balancing, demographic homogeneity, and administrative requirements. Rules of thumb have involved the use of square or circular patterns to minimize travel time to the scene of an emergency. Speed, however, may depend on specific streets, and if travel time rather than distance is taken as the criterion, the longer dimension of a "square" travel time response area would correspond with the higher speed streets. Closely related to the design of response areas are site selection for new and changed facilities, and the dynamic problems of prepositioning at the beginning of tours of duty and repositioning to change locations as load develops and some units become occupied.

Location of Fire Units. For fire units, prepositioning means determining the location of the fixed position fire stations themselves. Hendrick and Plane [1973] divided the fire station location problem into two main classes. First are those in which the call rate is high relative to fire protection capacity, as in New York City or Chicago. In these situations, queuing analysis and stochastic simulation are appropriate, and the concepts of prepositioning and repositioning represent important strategies. On the other hand, in cases where the call-capacity ratio is relatively low, as in Denver, Hendrick and Plane felt that a static approach, using mathematical programming, was justified.

Working with the Denver Fire Department, Hendrick and Plane used a linear programming formulation with a criterion that was designed to minimize the number of fire stations, subject to response time to specific fire hazards, as constraints. The linear programming model provided rich information on alternatives, which in turn was used by decision makers—in conjunction with other objective and subjective information—to make final location choices.

Relocation of Fire Units. Kolesar and Walker [1974] designed a fire company relocation algorithm for use in New York City, whose problem, typical of large cities, is that more than one serious fire may be in progress simultaneously. On the average, ten such incidents occur daily in New York City, and a five-alarm fire in Manhattan could deplete the area of half its fire fighting units, resulting in a sharp degradation of fire protection. Under such circumstances, common practice is to relocate the remaining units in selected "empty" firehouses in order to anticipate further alarms with a new and better deployment. Using a computer based system, Kolesar and Walker developed a relocation method that effected minimum total expected response times.

Suppose that two fires simultaneously require the services of the seven fire companies shown in Figure 9. One region is uncovered. The question is how to redeploy in order to minimize expected response time. The model assumes that the call rate will be random. The "square root law," applied to approximate the expected response-distance, states that the expected response-distance is proportional to the square root of the area served by a fire company. Kolesar and Blum [1973] empirically validated the square root relationship.

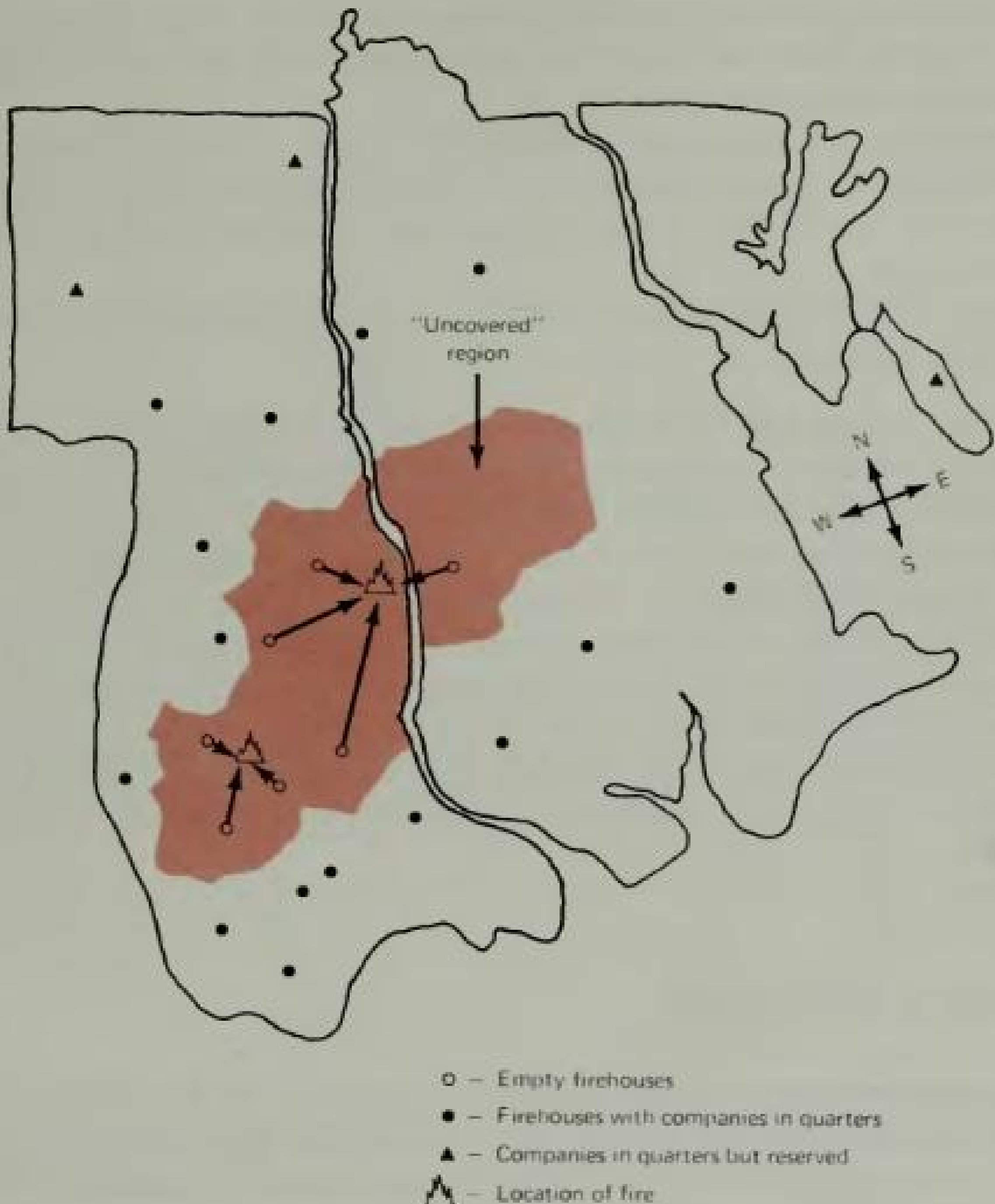


Figure 9. A sample relocation problem

SOURCE: P. Kolesar and W. E. Walker, "An Algorithm for the Dynamic Relocation of Fire Companies," *Operations Research* 22(March–April 1974): 249–274.

The algorithm for relocation developed by Kolesar and Walker performs with the following main steps:

1. Determine the need for relocation. An uncovered response area is detected by a program called "trigger". This program comes into play whenever a fire requires the use of at least three engines and two ladders. An uncovered area is one that falls below the minimum standard of having at least one company within x minutes of every alarm box.

2. Determine which of the empty firehouses should be filled. A heuristic rule is used to select the house that is associated with the largest number of uncovered response areas.
3. Determine the companies available to relocate. A heuristic rule chooses available companies with the lowest relocation costs. A feasibility check is made to be sure that no area becomes uncovered as a result of a potential relocation.
4. Solve a linear programming assignment problem to minimize total relocation distance.

Now let us return to the situation indicated by Figure 9, which shows the locations of two fires, the locations of the seven ladder companies working at the fires, and the region left uncovered. Step 1 shows nine uncovered response areas. Step 2 indicates that there are two solutions (each considering four of the seven empty houses), that provide a minimum coverage leaving no area uncovered but still require that a minimum number of companies move. In step 3 the program selects the companies that are to relocate. In Figure 10, the least cost assignment is indicated by the colored arrows; the least travel-distance solution produced by step 4 is shown by the black arrows. The solution produced by step 4 results in a 22 percent reduction in travel distance, and only a 9 percent increase in cost.

Location of Ambulances

The ambulance location problem generally is similar to the fire station location problem: the response time depends on the state of the system at the time that a call for service occurs. However, an ambulance normally responds with only one unit, and usually is occupied with that call for a relatively short service time rather than for several hours, as might be the case with fire companies. The number and location of hospitals will affect the length of time that an ambulance is tied up. The longer the ambulance is tied up giving service, the more likely it is that another call will be received while the unit is busy. Since the mean call rate is typically a function of the time of day, the size of the ambulance fleet in service also varies during the day. In Los Angeles, typical response times are 3.5 minutes, and typical ambulance service times are 17 minutes [Fitzsimmons, 1970].

Fitzsimmons [1970, 1973] developed a computer simulation model to study optimum ambulance service in the city of Los Angeles. Based on studies of past actual loads for each hour of the day, a specific mean hourly rate could be determined in proportion to load experience through the day. Previous statisti-

Figure 10. Solutions to the sample relocation problem

SOURCE: P. Kolesar and W. E. Walker, "An Algorithm for the Dynamic Relocation of Fire Companies," *Operations Research* 22(March–April 1974): 249–274.

cal studies showed that the Poisson distribution was a good description of call rates for the mean values. Finally, the simulation study determined an exact time for an incident by sampling over the hour from a uniform distribution. These components for determining load are shown in Figure 11 as input to the generation of an incident.

The generation of an incident is processed according to the general logic of the

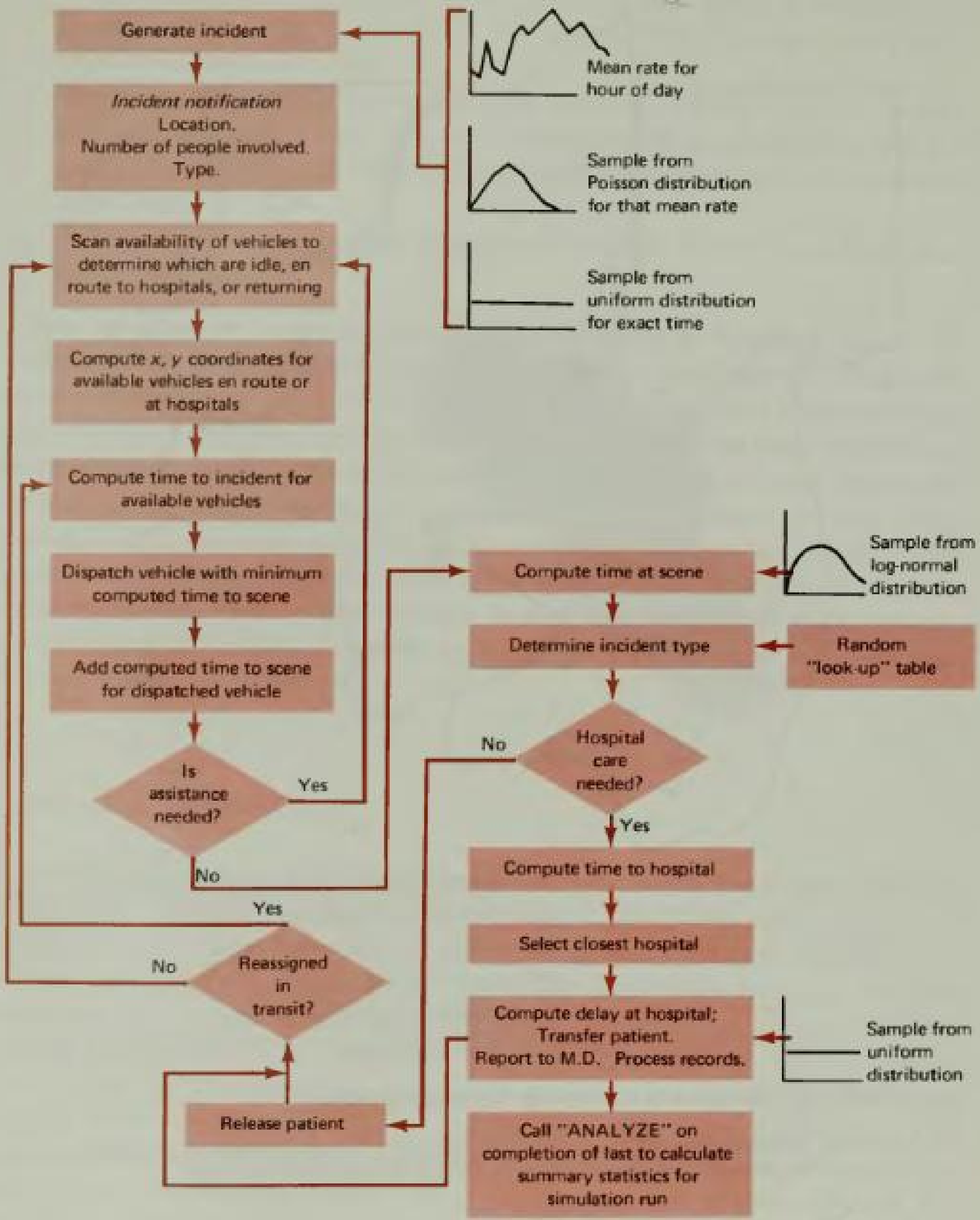


Figure 11. Summary flow chart for simulation of emergency medical system

flow chart shown in Figure 11. Travel time is computed as the sum of x and y distance-times to correspond to the usual urban layout plan. Additional Monte Carlo sampling is required to compute the time needed for an ambulance to arrive at the scene, the type of incident, and the delay at the hospital. While each vehicle has a home base, the simulator is designed to accommodate a mobile system (which can be dispatched en route) since vehicles can be reassigned in transit. The simulator included capability for both ambulances and helicopters. The model was validated in terms of equivalent analytical models for simple cases, and by means of actual data, for a portion of the existing Los Angeles system.

Use of the Emergency Medical Simulation Model

Various alternatives, in terms of number and location of ambulances and number and location of hospitals, were evaluated mainly in terms of response time.

Number and Location of Ambulances. Fitzsimmons evaluated the effect on response time of having one to ten ambulances in the system for single and dispersed home stations. The single home stations were at a hospital. In general, the response time for a single station leveled off at about three ambulances in the system, as shown in Figure 12; however, response time for dispersed deployment continued to decline. Waiting time for the single station fell to near zero with three ambulances in the system.

A series of experimental runs evaluated hypotheses concerning location deployments. Dispersed deployment dominated the single station alternative for all system performance criteria. Furthermore, optimal locations improved mean travel time to the scene by about 12 percent, compared to existing locations. Finally, it was found that optimal deployment was a function of incident or call rate.

Number and Location of Hospitals. A series of runs was made in which the number of hospitals was varied from one to four, using a single ambulance that was located centrally. Mean time to the hospital was reduced considerably when hospitals were added to the system, but response time and waiting time were only slightly reduced, and ambulance utilization declined about 2 percent, as shown in Figure 13. Mean time to the scene actually increased slightly. Optimally locating hospitals reduced mean waiting time only slightly, because of reduced ambulance utilization.

The model was applied to the problem of establishing locations for fourteen ambulances in the city of Los Angeles. The ambulance system in Los Angeles is under the jurisdiction of the fire department, and the problem was to select fourteen out of thirty-four sites that could physically accommodate an ambu-

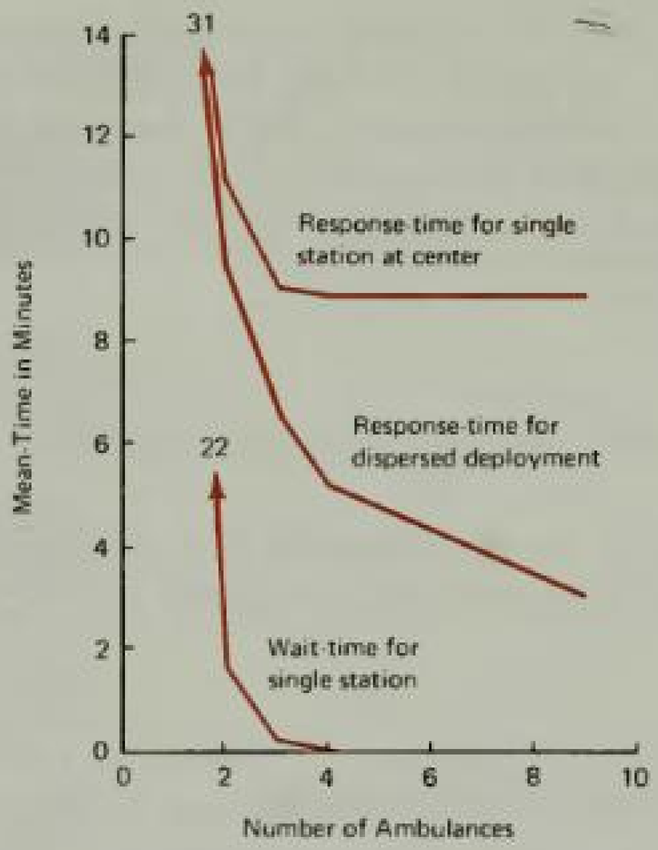


Figure 12. Response times for single ambulance station, and dispersed locations for uniform distribution of calls within the geographic area

SOURCE: J. A. Fitzsimmons, "Emergency Medical Systems: A Simulation Study and Computerized Method for Deployment of Ambulances" (Ph.D. diss. University of California, Los Angeles, 1970).

lance unit. The fire department had made preliminary decisions on location, with the object of locating an ambulance within 2.5 miles or every point in the city, or a five minute response time at thirty miles per hour. Application of the model resulted in the relocation of nine ambulances; mean response time was reduced by almost 9 percent, and the probability of a response time greater than six minutes was improved by 33 percent. The new deployment also resulted in a more balanced work load, due to shorter response times. It was found that ambulances had a 10 percent average utilization rate. As in all productive systems with large demand variation, good service creates considerable idle time, thus making it likely that an ambulance unit will be available when an emergency occurs.

Another approach to ambulance location, using a branch-and-bound procedure, was developed by Swoveland, Uyeno, Vertinsky, and Vickson [1973]. This approach was applied to locating ambulances for the city of Vancouver, Canada.

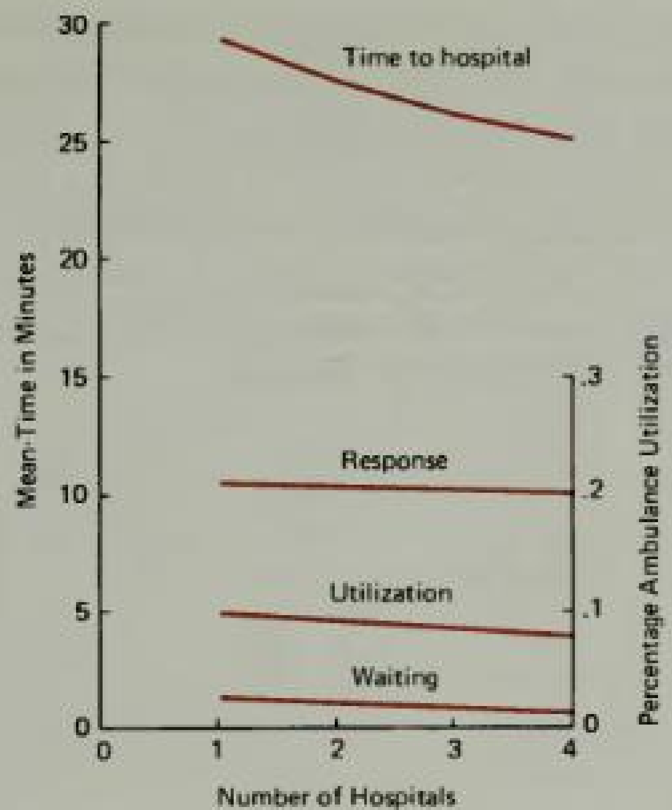


Figure 13. Effect of the number of hospitals in the system on response time

SOURCE: J. A. Fitzsimmons, "Emergency Medical Systems: A Simulation Study and Comized Method for Deployment of Ambulances" (Ph.D. diss., University of California, Los Angeles, 1970).

LOCATION OF RETAIL OUTLETS

Early models of retail outlet location were based on what was called a gravity model. The models involved the hypothesis that two cities attract retail trade from an intermediate town approximately in direct proportion to the populations of the two cities, and in inverse proportion to the square of the distances from the two cities to the intermediate town. While the gravity model had severe limitations, the gravity (or attraction) concepts have carried over into models that are nearly as simple but inject some aspects of assumed consumer behavior.

Huff [1962] developed a model in which the utility of a given shopping center is directly proportional to the ratio S/T^{λ}; that is, the ratio of the size of the shopping center to the time it takes for a consumer to travel there from his/her base. Size is measured in terms of square footage of selling space. The travel time is modified by λ, which is estimated empirically to reflect the effect of travel time on various kinds of shopping trips. Huff cited empirical evidence that these two factors exert an influence on a consumer's choice of a shopping center, and may be the only variables needed to predict such behavior.

Based on the utilities calculated for all competitive centers, one computes the probability of a consumer at a given origin traveling to a given shopping center. Multiplying this probability by the number of consumers in a defined area gives us an approximation of the expected number of consumers from a given area who are likely to travel to a given shopping center for a given type of shopping trip. The basic statement of the model, then, is:

$$E_{ij} = P_{ij}C_i = \frac{\dfrac{S_j}{T_{ij}^{\lambda}}}{\displaystyle\sum_{j=1}^{n} \frac{S_j}{T_{ij}^{\lambda}}} \cdot C_i$$

where,

E_{ij} = the expected number of consumers at i that are likely to travel to shopping center j.

C_i = the number of consumers at i.

P_{ij} = the probability of a consumer at a given point of origin i traveling to a given shopping center j.

S_j = the size of a shopping center j.

T_{ij} = the travel-time involved in getting from a consumer's travel base i to shopping center j.

λ = a parameter that is to be estimated empirically to reflect the effect of travel time on various kinds of shopping trips.

We shall not attempt to cover the reasoning by which we reach the conclusion that consumer behavior is measured adequately by the size and travel time variables. Rather, we will simply report the results of an empirical validation.

Empirical Validation

Huff selected a suburban community within metropolitan Los Angeles that contained three rather distinct neighborhoods. These neighborhoods were the points of consumer origin. Questionnaires were used to gather data from house-

TABLE 13

SIZES OF THE SELECTED SHOPPING CENTERS

Shopping Center (j)	*Gross Size* [*sq. ft. of selling area*] (S_j)
J_1	239,000
J_2	236,000
J_3	326,000
J_4	97,000
J_5	1,250,000
J_6	281,000
J_7	228,000
J_8	326,000
J_9	203,000
J_{10}	222,000
J_{11}	502,000
J_{12}	425,000
J_{13}	134,000
J_{14}	121,000

SOURCE: D. L. Huff, "Determination of Intra-Urban Retail Trade Areas" (University of California, Los Angeles: Graduate School of Management, 1962).

TABLE 14

PROXIMITY OF NEIGHBORHOODS TO SELECTED SHOPPING CENTERS

Shopping Center (j)	*Travel Time* [*minutes*] *from Neighborhood 1* (T_{1j})	*Travel Time* [*minutes*] *from Neighborhood 2* (T_{2j})	*Travel Time* [*minutes*] *from Neighborhood 3* (T_{3j})
J_1	2.8	3.6	4.2
J_2	11.1	6.8	8.6
J_3	13.3	17.0	14.3
J_4	15.0	14.7	8.6
J_5	15.4	16.1	20.5
J_6	14.0	17.8	15.2
J_7	17.8	17.4	11.5
J_8	21.7	21.4	15.5
J_9	22.9	22.7	27.0
J_{10}	20.0	17.8	27.1
J_{11}	15.7	19.2	17.4
J_{12}	27.6	27.7	25.8
J_{13}	10.8	7.7	5.2
J_{14}	10.4	9.4	16.1

SOURCE: D. L. Huff, "Determination of Intra-Urban Retail Trade Areas" (University of California, Los Angeles: Graduate School of Management, 1962).

TABLE 15

COMPARISON OF OBSERVED AND EXPECTED NUMBER OF CONSUMERS FROM EACH OF THE THREE NEIGHBORHOODS WHO LAST MADE A FURNITURE PURCHASE AT ONE OF THE SPECIFIED SHOPPING CENTERS*
(UNDIFFERENTIATED FOR AGE OR INCOME)†
(DISTANCE MEASURED IN TRAVEL TIME MINUTES)

Shopping Center	*Neighborhood 1*		*Neighborhood 2*		*Neighborhood 3*	
	Observed	*Expected*	*Observed*	*Expected*	*Observed*	*Expected*
J_1	51	51.66	68	65.83	80	78.21
J_2	0	1.50	4	16.72	1	7.43
J_3	0	1.30	0	3.27	1	1.95
J_4	0	0.00	0	1.33	0	3.05
J_5	3	3.43	24	14.06	11	2.30
J_6	6	0.98	6	2.55	12	1.38
J_7	0	0.00	3	2.17	3	2.78
J_8	0	0.00	1	2.00	0	15.0
J_9	0	0.00	0	1.10	0	0.00
J_{10}	2	0.00	16	2.02	4	0.00
J_{11}	0	0.00	0	3.88	1	1.58
J_{12}	0	1.31	0	1.50	1	0.00
J_{12}	0	0.91	0	7.28	8	21.82
J_{14}	0	0.91	6	4.29	0	0.00
Total	62	62.00	128	128.00	122	122.00

	Neighborhood 1	Neighborhood 2	Neighborhood 3
*r =	0.99	0.94	0.96
λ =	2.542	2.115	3.247
†Avg. adult age	37.15	35.52	38.10
Avg. stated income	6130	7246	6611

SOURCE: D. L. Huff, "Determination of Intra-Urban Retail Trade Areas" (University of California, Los Angeles: Graduate School of Management, 1962).

holders in the three neighborhoods. These people were asked questions about which shopping center they had last patronized in order to make several kinds of purchases, and which they normally patronized, as well as questions about family income etc. Information also was gathered on fourteen shopping centers within a twenty mile radius from each of the neighborhoods. Tables 13 and 14 indicate the data on gross size of the fourteen shopping centers, and the travel times to each of the shopping centers from each of the three neighborhoods, respectively. Estimates of the value of λ employing successive approximation were obtained for each of the different kinds of shopping trips.

Comparisons of observed and expected numbers of consumers going to each of the fourteen shopping centers then were generated. Table 15 involves furniture purchases. Note that the correlation coefficient between observed and expected values for each of the three neighborhoods is quite high.

Use of the Model

Based on the model, we can estimate the number of consumers coming to a given shopping area for a given type of purchase, as indicated previously. Then, using survey data to determine average household income in the area and the average family budget figures for various kinds of purchases (for example, furniture), we can determine the annual sales potential for a shopping center by multiplying each of the budget figures by the expected number of consumers for that type of purchase. Thus, the model produces data on alternate locations concerning the sales potential. The presumption is that a location would be chosen that would maximize potential revenue. Given the location in a shopping center, the problem becomes one of finding a suitable site, and a host of intangible criteria and values would enter into the criterion.

Huff [1966] also computerized the model so that a search could be made over a large geographic area in order to find the best retail location for a given situation.

Summary

An important aspect of a productive system design is the subsystem that distributes the products and services to users. The key question of where to locate physical facilities is an important part of that problem.

The conceptual framework of linear programming transportation methods helps us understand the allocation nature of the distribution of physical product from sources to destinations, within the constraints of the amounts available and demand requirements. While the solutions generated by linear programming are for optimum distribution cost, often there are alternate optimum solutions. In addition, near optimum solutions cost only slightly more. Thus flexibility creates actual shipment schedules that allow management to consider other values and criteria. Goal programming for distribution problems with multiple objectives provides a rigorous framework for trading off objective and subjective criteria.

The emphasis in industrial plant location is on minimizing costs. However, we are speaking of longer-run costs: many intangible factors may influence future costs. The Brown and Gibson model provides a framework for the integration of objective and subjective factors, using preference theory to assign weights to factors in a consistent and systematic manner. The breakeven concept examines the capital expenditure-volume effects, and displays the range of output volume for which a given solution may be effective.

Multiplant location is influenced by existing locations, since each location considered must be placed in economic perspective with the existing plants and market areas. Each alternate location that is considered results in a different allocation of capacity to markets; our objective is to minimize costs for the system as a whole. Locational dynamics provides a basis for balancing costs among producing plants as demand shifts in different market areas. In some instances, an appropriate economic decision might be to close a plant and enlarge capacity in the remaining plants through the use of overtime work.

The warehouse location problem has been the focus of much recent research. Unlike plant location, where the myriad objective, subjective, and critical factors present a complex problem involving multiple objectives, in the warehouse location problem the cost criterion dominates. Although the problem has certainly not been easy to solve, it has yielded progressively to quantitative analysis in three main lines of investigation: mathematical programming, heuristic and simulation approaches, and branch-and-bound methods.

While the industrial location problem is distribution oriented, the location of facilities for service operations is oriented toward the location of customers, clients, or users. Service facilities tend to be more local in nature, to bring the service to be rendered to users.

Abernathy and Hershey applied computer search methodology to optimize the location of regional health services in relation to demographic data concerning the behavior of the using population, and compared results for criteria of maximum utilization, minimum distance per capita, and minimum distance per encounter. The data are summarized to allow decision makers to make trade-offs between the different criteria.

The location of emergency units is presented in the context of a deployment problem. Bound up in the problem is the system capacity, patrol patterns, priorities for different kinds of calls, and the choice of which units actually are dispatched in response to a call. Relocation—that is, the repositioning of units as some units are assigned (as when a major fire breaks out)—is a significant problem. Fitzsimmons developed a computerized ambulance location program (based on a queuing model), and used a computer search problem to optimize mean response time for alternate locations.

Finally, the location of retail units is focused on the problem of maximizing revenue. The basic model involves the hypothesis that two cities attract retail trade from an intermediate town approximately in direct proportion to the population of the two cities and in inverse proportion to the square of the distances of the two cities to the intermediate town, modified by some aspects of consumer behavior.

Review Questions

1. In a linear programming distribution problem, how do we know whether or not we have alternate optimal solutions? Of what importance are alternate optimal solutions to a manager?

2. How can the existence of unequal supply and demand be handled in the framework of distribution methods of linear programming?

3. What is the effect on an optimal distribution cost plan of taking into account differing production costs that may occur in plants? differing selling prices that may occur in market areas?

4. What is goal programming, and how can it be used in distribution problems?

5. What is the meaning of the statement, "The location of breweries is water oriented"?

6. What is the resultant economic locational choice criterion, when, from location to location:
   a. Prices and costs vary.
   b. Prices are constant, but costs vary.
   c. Costs are constant, but prices vary.
   d. Prices and costs are constant.

7. Suppose that you are considering two alternate locations for an industrial plant. There are no discernible differences in revenues, and studies indicate that material, labor, and overhead costs are virtually identical. On what other bases would you make comparisons in order to provide information for the decision process?

8. In the context of the Brown–Gibson plant location model, define:
   a. Critical criteria.
   b. Objective criteria.
   c. Subjective criteria.
   d. Location measure.
   e. Objective factor decision weight.

9. Under what conditions might a plant location be most economical for one projected volume but not most economical if the projected volume is any larger?

10. How is the problem of locating an *additional* plant, where one or more plants already exist, different from locating or relocating a single plant?

11. In Question 3, we raise the issue of the effect on an optimal distribution cost plan of taking differing production costs into account. Now examine Figure 4, in which differing production costs were taken into account. Reconcile your answer to Question 3 with the results of Figure 4.

12. What is meant by the term *locational dynamics*?

13. In the example of four alternatives of plant operations summarized by Figure 5, suppose that the fixed costs, when the plants were shut down, were $6,000 per week for all plants. What, then, is the best operating plan?

14. Why is it not always advantageous for a manufacturer to locate plants in foreign countries where wages are somewhat lower? Can the manufacturer's labor costs actually be higher in some foreign situations?

15. Contrast the nature of the warehouse location and the plant location problems? What approaches have been taken with the warehouse location problem?

16. Explain the concept of *order-shifting* in Markland's facilities search heuristic in the Ralston Purina warehouse location study.

17. In the Ralston Purina warehouse location study, the number of warehouses was consolidated from five to three, based on the cost summary in Table 10. What is the source of the cost savings?

18. In the Ralston Purina study, why does transportation cost decline as the inventory service level is reduced from 100 percent to 85 percent? (See Table 11.)

19. In what ways is the Ralston Purina approach to warehouse location likely to represent a suboptimum solution?

20. Compare the problem of locating regional health facilities with the typical industrial plant location problem. In what ways are the two problems similar? different?

21. Under what conditions might the location of emergency units, such as fire stations, be regarded as a dynamic relocation- and redeployment problem? a static location problem?

22. Kolesar and Blum developed a relocation algorithm for fire units. Suppose that two fires break out simultaneously. How would the algorithm function in relocating the remaining units?

23. In the model for locating retail outlets, what factors are used to predict consumer behavior?

24. In the model used for locating retail outlets, how is the probability that a given customer will travel to a given shopping center calculated? What aspects of consumer behavior are implied by this calculation?

Problems

1. The Cosgrove Abrasives Company manufactures and distributes industrial abrasives, and currently is reviewing its distribution system. The products are manufactured and stored at factory warehouses in three locations, and the company ships to four independent distributors. The company is

planning for the upcoming period. The amounts available at each of the factory warehouses (W), and the demands from each of the independent distributors (D), are as follows:

Available		Demands	
W_1	300 units	D_1	200 units
W_2	200 units	D_2	100 units
W_3	400 units	D_3	450 units
	900 units	D_4	250 units
			1000 units

In dollars per unit, the shipping costs among all combinations of factory warehouses and distributors are as follows:

From \ To	D_1	D_2	D_3	D_4
W_1	5	2	6	7
W_2	3	5	4	6
W_3	4	5	2	3

When the problem was placed in the format of linear programming, using distribution methods, the optimum solution, at a total distribution cost of \$2,950, was as follows:

Ship From	To Distributor	Units
W_1	D_2	100
	D_3	200
W_2	D_1	200
W_3	D_3	250
	D_4	150
		900

The company president objected to the plan, even though it was pointed out that this plan achieved the minimum possible distribution cost. His objections were that the plan did not meet goals that were not included in the linear programming objective function. After considerable discussion, the president's "other" goals were summarized in the following list, in priority order:

P_1. Guarantee delivery of the entire order from D_4.

P_2. Ship at least 100 units over the W_3-D_1 route to satisfy his private agreement with a teamster's official.

P_3. Allow some back ordering, but meet no less than 80 percent of demand at each distributor.

P_4. If necessary, let total distribution cost increase slightly (say 10 percent) from the minimum of $2,950.

P_5. Minimize shipping over the W_2-D_4 route because of hazards.

P_6. Balance the percentage of demand met between D_1 and D_3; i.e., treat them alike.

P_7. As the president put it, "Last, but certainly not least important, minimize total distribution cost."

a. How well does the optimum distribution plan meet the president's objectives?

b. The distribution manager reexamined his *LP* solution and found that there were alternative optimum distribution plans. Not wishing to yield on minimum distribution cost as his prime goal, he pointed out that either of the following two alternative distribution plans also would yield a minimum total distribution cost of $2,950:

Plan 2

Ship From	To Distributor	Units
W_1	D_1	200
	D_2	100
W_2	D_3	200
W_3	D_3	250
	D_4	150
		900

Plan 3

Ship From	To Distributor	Units
W_1	D_2	100
	D_3	50
	D_4	150
W_2	D_1	200
W_3	D_3	400
		900

How well do Plans 2 and 3 meet the president's goals?

2. The president of Cosgrove Abrasives is a hard man, and refuses to go along with any of the three optimum distribution plans presented. He tells the distribution manager, "That's your optimum, not mine—I define the problem differently than you. If you want to use a model, find one that fits my definition of the problem!" After doing some research (i.e., reading this chapter), he discovered the article on goal programming by Lee and Moore [1973]. He recomputed the problem, using the manager's list of goals in the president's priority order, and the result was the following distribution plan which cost $3,360 to implement. How well does this plan meet the president's goals?

Ship From	To Distributor	Units
W_1	D_2	100
	D_4	200
W_2	D_1	90
	D_3	110
W_3	D_1	100
	D_3	250
	D_4	50
		900

3. The Good-Wear Shoe Company, discussed in the text, installed its Atlanta plant. Once that producing location proved successful and materials were imported for the existing markets, management turned its attention to the exploration of new markets. It found ready opportunities by expanding its sales efforts toward the West, South, and Southwest. Although it had been supplying these markets from existing distribution centers, current volume in the new locations raised the question of the advisability of a new warehouse location. Three possible locations were suggested because of market concentrations: Denver, Houston, and New Orleans. Data on capacities, demands, and costs are given in Table 16. Based on these data, which warehouse location should be chosen? What additional criteria might be invoked to help management make a choice? How should the decision be made whether to build the new warehouse or to continue supplying from the existing warehouses?

TABLE 16
PRODUCTION COSTS, DISTRIBUTION COSTS, PLANT CAPACITIES, AND MARKET DEMANDS FOR THE GOOD-WEAR SHOE COMPANY

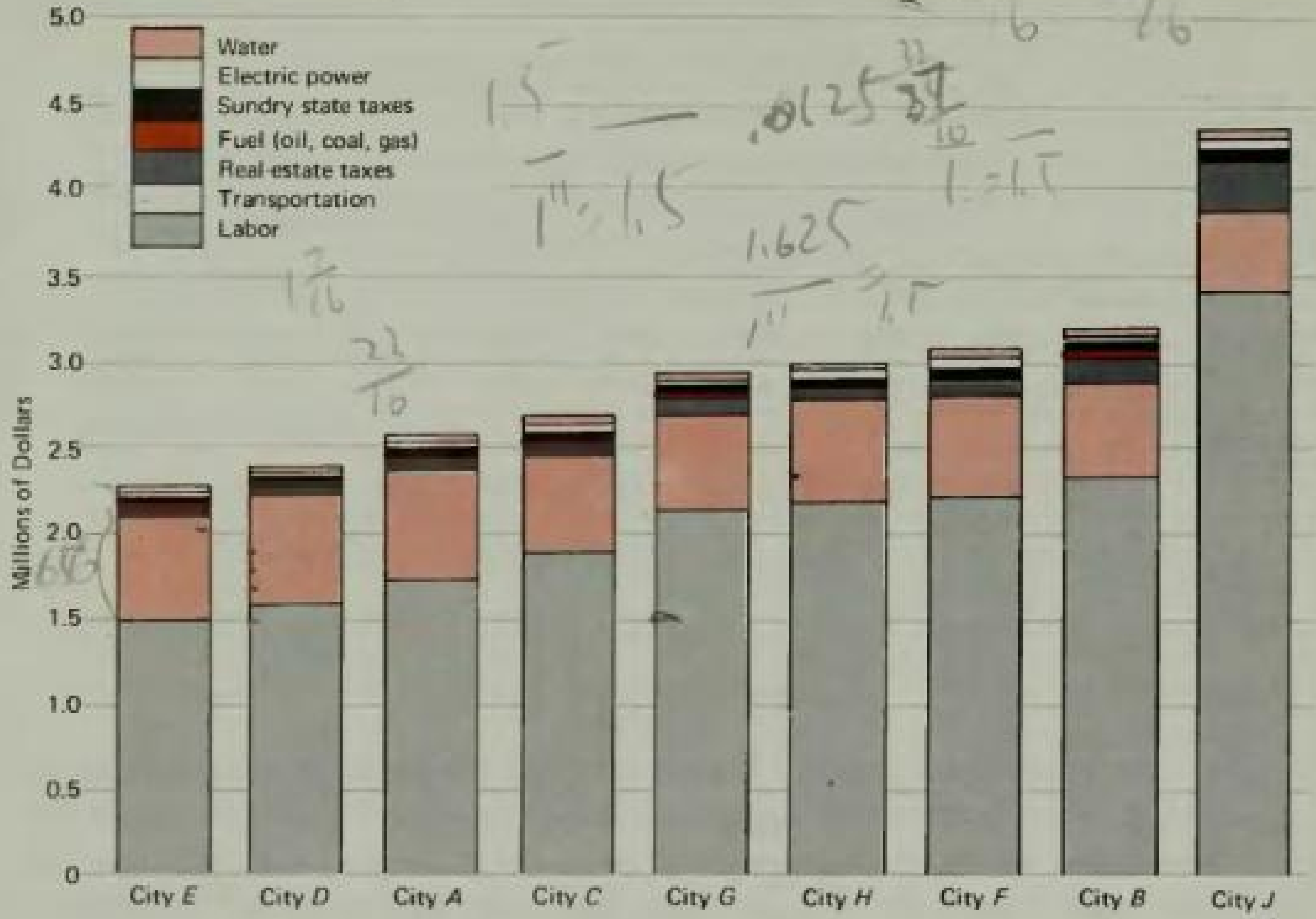
To Distribution Center / From Plants	Distribution Costs per Pair, Handling, Warehousing, and Freight								Normal Weekly Capacity Pairs	Unit Production Cost
	Existing Warehouses					Proposed New Warehouses				
	Atlanta	Buffalo	Cincinnati	Cleveland	Milwaukee	Denver	Houston	New Orleans		
Atlanta	\$0.27	\$0.46	\$0.43	\$0.45	\$0.48	\$0.65	\$0.58	\$0.55	25,000	\$2.62
Chicago	0.49	0.42	0.42	0.44	0.32	0.50	0.54	0.60	20,000	2.68
Detroit	0.50	0.38	0.41	0.36	0.42	0.55	0.60	0.65	27,000	2.70
Forecast weekly market demand, pairs	8,000	15,000	11,000	13,000	9,000	16,000	16,000	16,000		

p.157f

4. A manufacturer, Company X is considering the problem of relocating its plant. Nine locations are being considered actively. Figure 14*a* shows the estimated annual costs to operate the plant in each of the locations. As shown in Figure 14*b*, the nine sites have also been rated as "excellent," "plentiful," "very good," "good," "adequate," or "fair" for six subjective factors. The seventh factor, union activity, is rated "active," "significant," "moderate," or "negligible." In addition, Table 17 shows general and specific information concerning the nine cities.
   a. Given the objective and subjective data about the nine alternate sites, formulate a model for reconciling the objective and subjective factors.
      Hint: think in terms of rating scales for the subjective factors, and the relative weight to be placed on objective versus subjective factors.
   b. Given the rationale (model) developed in Problem 4a, which city would be selected?

highest score mean ip

5. Additional information was constructed concerning the estimated annual fixed costs of operation for Company X in Problem 4. These costs were considered important in relation to the annual operating costs given in Figure 14*a*, especially if breakeven between sites were to occur near the contemplated plant design volume. Which city appears most advantageous as a site when considered in terms of a breakeven analysis?



(a)

Factor	City E	City D	City A	City C	City G	City H	City F	City B	City J
Labor supply	adeq.	adeq.	plent.	plent.	adeq.	adeq.	plent.	plent.	plent.
Type of labor	good	good	excel.	excel.	excel.	excel.	excel.	excel.	excel.
Attitude	good	good	v.g.	v.g.	good	good	v.g.	v.g.	good
Appearance	fair	fair	good	good	excel.	fair	good	good	good
Transportation	good	good	v.g.	good	v.g.	good	v.g.	v.g.	v.g.
Recreation	good	v.g.	v.g.	v.g.	v.g.	good	v.g.	v.g.	v.g.
Union activity	sign.	sign.	neg.	neg.	mod.	sign.	sign.	mod.	act.

(b)

Figure 14. Summary of Company X's study of nine alternative locations: comparison of other factors influencing choice of plant site. (Reprinted by special permission of National Industrial Conference Board "Studies in Business Policy," no. 61 (New York: National Industrial Conference Board, 1953).)

bonds float on this city, not others

TABLE 19

COMPANY X: COMPARISON OF PROPOSED PLANT LOCATIONS

City	Population: 1960 Census	Population: Est. 1970	Population: Percent Minority	Bonded Indebtedness: Total ×1000	Bonded Indebtedness: Per Capita	Bank Deposits: Total ×1000	Bank Deposits: Per Capita	Taxes: For Plant (1)	Taxes: State Personal Income	Taxes: General Sales Taxes	Municipal Airports	Hospitals	No. of Beds	Churches	Schools	Hotels	No. of Rooms
City A	14,505	20,260	12.5	$ 277	$13.80	$16,385	$816.74	$ 75,000	Yes	None	Yes	2	119	25	8	4	295
City B	25,129	25,000	8.6	1,377	60.75	24,011	931.00	111,830	None	Yes	Yes	1	108	16	8	2	50
City C	8,734	10,500	15.0	None	0.00	10,000	1,023.50	48,100	Yes	None	No	1	75	10	6	1	120
City D	11,723	14,862	19.0	11	0.75	17,276	1,010.50	89,000	Yes	Yes	Yes	1	90	21	22	3	100
City E	7,250	8,000	17.0	600	75.00	8,288	1,102.60	440	None	Yes	No	2	69	8	5	1	100
City F	72,013	78,000	7.6	1,433	18.40	82,434	1,071.00	86,822	None	None	Yes	2	423	96	37	14	1,100
City G	21,960	23,100	7.8	272	14.00	20,000	797.11	110,000	None	Yes	No	1	112	22	8	8	187
City H	12,514	18,007	1.0	48	2.40	14,472	1,000.00	81,180	None	None	No	1	70	21	10	1	105
City I	109,822	115,000	4.0	5,874	44.50	168,755	1,761.00	202,146	None	None	Yes	4	611	78	47	8	

City	Electric Power Cost (2)	Water Cost (3)	Gas Rate (4)	Coal Rate (5)	Oil Rate (6)	No. of Present Industries (7)	Percent Unionized	Labor Rates: Unskilled M	Unskilled F	Semiskilled M	Semiskilled F	Skilled M	Skilled F
City A	$4,287	$419.65	41¢	$ 8.13	3¢	8	25%	4.50 to 4.75	4.45 to 4.60	4.76 to 5.00	4.60 to 4.75	5.00 to 5.00	4.75 to 5.00
City B	6,005	460.00	31¢	10.44	10.5¢	5	40%	5.00 to 5.25	5.00 to 4.85	5.40 to 5.25	4.90 to 5.25	5.25 to 5.75	
City C	8,066	444.00	41¢	8.00	3¢	1	0	4.00 to 4.90	4.50 to 4.55	4.90 to 5.00	4.65 to 4.75	5.25 to 5.50	4.80 to 5.00
City D	6,600	274.00	4¢	6.65 to 7.50	—	3	67%	4.50	4.00	4.80	4.70	5.00	—
City E	6,542	242.20	8¢	10.65	11.4¢	3	87%	4.00	4.45	4.75	4.65	5.00	4.75
City F	7,000	399.28	72¢	9.60	10¢	30	67%	5.00 to 5.05	4.75 to 4.80	5.25 to 5.35	4.00 to 4.75	5.30 to 5.50	5.00 to 5.05
City G	6,522	360.00	41¢	10.00	12.5¢	8	100%	4.90 to 5.00	4.75 to 4.90	4.90 to 5.25	4.85 to 5.05	5.25 to 5.80	4.95 to 5.45
City H	6,000	475.00	31¢	7.50	10¢	8	44%	4.75 to 4.90	4.90 to 4.75	4.91 to 5.31	4.85 to 4.90	5.65 to 5.80	None
City I	8,130	375.00	7¢	12.63	7.5¢	16	16%	3.608 av. (unspecified)		3.718 av. (unspecified)		5.475 to 5.60*	—

Adapted from National Industrial Conference Board

(1) Computed total tax on hypothesis of plant
(2) Per 200,000 kwh per month (1000 kw demand, 60% power factor)
(3) Cost for 1,000,000 gal [illegible] per month
(4) Per 1000 cu ft [illegible]
(5) Per ton [illegible] delivered
(6) Per gal. No. [illegible] fuel oil delivered
(7) Employing 100 people or more
*Same rate for males and females

City	Annual Fixed Cost, Thousands
A	$ 500
B	700
C	750
D	500
E	1,000
F	800
G	500
H	700
J	500

6. Table 18 gives data on the fixed and variable costs of operation for three possible warehouse locations in Detroit, Philadelphia, and Dallas, supplying four customer zones located in San Francisco, Chicago, Houston, and New York. In trying to decide on the best distribution system, can you find any warehouse locations that must be retained in the system because the *smallest* cost saving resulting from their use is greater than their fixed costs?

Hint: for Philadelphia shipping to New York, for example, there is a saving of 1 compared to the next best source of supply, Detroit, i.e., 5–4. But

TABLE 18

FIXED AND VARIABLE DISTRIBUTION COSTS FROM THREE POSSIBLE WAREHOUSE LOCATIONS TO FOUR CUSTOMER ZONES

Warehouse	Annual Warehouse Fixed Cost	Variable Distribution Costs from Warehouses to Customer Zones				
		C_1 San Francisco	C_2 Chicago	C_3 Houston	C_4 New York	
W_1 Detroit	3.0	9	4	6	5	24
W_2 Philadelphia	3.5	12	6	10	4	32
W_3 Dallas	2.5	8	6	3	12	29
		29	16	19	21	8 5

5=D

29-24=5>3

32-29=3>2.5

the fixed cost for the Philadelphia warehouse is 3.5, so the minimum savings are not greater than the fixed cost needed to keep Philadelphia open. This does not mean that Philadelphia should be closed; nor does it mean that it must be retained. Try the other combinations, always starting with the lowest cost supplier for a given customer zone.

7. Referring again to the data in Table 18, could any of the three warehouses be eliminated from consideration by using the following reasoning: Is there a warehouse that is retained in the system because of the rule stated in Problem 6, and because the *maximum* saving made by opening an additional warehouse is less than the fixed cost for keeping that warehouse open?

   Hint: Start with a warehouse that is kept open because of the rule in Problem 6, and assume that the others are closed. Now open one of them and compute the savings that result. That is, the warehouse just opened will have advantageous routes to supply certain zones. Add up the savings and compare them with the fixed costs for that warehouse. If the savings exceed the fixed costs, then that warehouse cannot be eliminated from consideration. But if the savings are less than the fixed costs, it can be eliminated from further consideration. The reason is that, on balance, the warehouse cannot compete with the warehouse that originally was fixed open.

References

Abernathy, W. J., and J. C. Hershey. "A Spatial-Allocation Model for Regional Health-Services Planning." *Operations Research* 20(May–June 1972): 629–642.

Atkins, R. J., and R. H. Shriver. "New Approach to Facilities Location." *Harvard Business Review*, May–June 1968, pp. 70–79.

Beckman, M. *Location Theory*. New York: Random House, 1968.

Brown, P. A., and D. F. Gibson. "A Quantified Model for Facility Site Selection—Application to a Multiplant Location Problem." *AIIE Transactions* 4(March 1972): 1–10.

Buffa, E. S. *Operations Management: Problems and Models.* 3d ed. New York: John Wiley & Sons, 1972. Chapters 10 and 11.

______, and A. E. Bogardy, "When Should a Company Manufacture Abroad." *California Management Review* 2(1960).

______, and J. S. Dyer. *Management Science/Operations Research: Model Formulation and Solution Methods.* Santa Barbara, Ca.: Wiley/Hamilton, 1977.

Chaiken, J. M., and R. C. Larson. "Methods for Allocation Urban Emergency Units: A Survey." *Management Science* 19. Part 2 (December 1972): 110–130.

Easton, A. *Complex Managerial Decisions Involving Multiple Objectives.* New York: John Wiley & Sons, 1973. Plant-location example, p. 288.

Effroymson, M. A., and T. A. Ray. "A Branch-Bound Algorithm for Plant Location." *Operations Research* 14(May–June 1966): 361–368.

Fitzsimmons, J. A. "Emergency Medical Systems: A Simulation Study and Computerized Method for Deployment of Ambulances. Ph.D. dissertation, University of California–Los Angeles, 1970.

______. "A Methodology for Emergency Ambulance Deployment." *Management Science* 19(February 1973): 627–636.

Geoffrion, A. M. "A Guide to Computer-Assisted Methods for Distribution Systems Planning." *Sloan Management Review* 16(Winter 1975): 17–41.

Gerson, M., and R. B. Maffei. "Technical Characteristics of Distribution Simulators." *Management Science* 10(1968).

Gibson, D. F., and J. Rodenberg. "Location Models for the Forest Products Industry and Other Applications." *AIIE Transactions* 7(June 1975): 143–152.

Helly, W. *Urban Systems Models.* New York: Academic Press, 1975. Chapter 7.

Hendrick, T. E., and D. R. Plane. "An Application of Operations Management for Municipal Fire Station Location." Working paper, University of Colorado, December 1973.

______. "Operations Research and Computer Systems for Fire Department Management." *IMSA Signal Magazine*, May–June 1973, p. 20.

Huff, D. L. "A Programmed Solution for Approximating an Optimum Retail Location." *Land Economics* 42(August 1966): 293–303.

______. "Determination of Intra-Urban Retail Trade Areas." University of California, Los Angeles, Graduate School of Management, 1962.

Keuhn, A. A., and M. J. Hamburger. "A Heuristic Program for Locating Warehouses." *Management Science* 19(July 1963): 643–666.

Khumawala, B. M. "An Efficient Branch and Bound Algorithm for the Warehouse Location Problem." *Management Science* 18(August 1972): 718–731.

______, and D. C. Whybark. "A Comparison of Some Recent Warehouse Location Techniques." *The Logistics Review* 7(1971).

Kolesar, P., and E. H. Blum. "Square Root Laws for Fire Engine Response Distances." *Management Science* 19(August 1973): 1368–1378.

______, and W. E. Walker. "An Algorithm for the Dynamic Relocation of Fire Companies." *Operations Research* 22(March –April 1974): 249–274.

Larson, R. C. *Urban Police Patrol Analysis*. Cambridge, Mass.: M.I.T. Press, 1972.

Lee, S. M., and L. J. Moore. "Optimizing Transportation Problems with Multiple Objectives." *AIIE Transactions* 5(December 1973): 333–338.

Magee, J. F. *Industrial Logistics*. New York: McGraw-Hill, 1968.

Markland, R. E. "Analyzing Geographically Discrete Warehousing Networks by Computer Simulation." *Decision Sciences* 4(April 1973): 216–236.

______. "Analyzing Multi-Commodity Distribution Networks Having Milling-in-Transit Features." *Management Science* 21(August 1975): 1405–1416.

Markland, R. E., and R. J. Newett. "Production-Distribution Planning in a Large-Scale Commodity Processing Network." *Decision Sciences* 7(October 1976).

McAllister, D. M. "Local Recreation Facility Location." In *Regional Environmental Management: Selected Proceedings of the National Conference*, edited by L. E. Coate and P. A. Bonner. New York: John Wiley & Sons, 1975.

Reed, Jr., R. *Plant Location, Layout, and Maintenance*. Homewood, Ill.: Richard D. Irwin, 1967.

Schollhammer, H. *Locational Strategies of Multinational Firms*. Los Angeles: Center for International Business, Los Angeles World Trade Center, 1974.

Shycon, H. N., and R. B. Maffei. "Simulation Tool for Better Distribution." *Harvard Business Review*, November–December 1960.

Whitman, E. S., and W. J. Schmidt. *Plant Relocation: A Case History of a Move*. New York: American Management Association, 1966.

Swoveland, C.; D. Ueno; I. Vertinsky; and R. Vickson. "Ambulance Location: A Probabalistic Enumeration Approach," *Management Science* 20. Part 2 (December 1973): 686–698.

Chapter 7

Processes and Job Design

As we mentioned previously, there are strategic decisions—involving the design of products and/or services and the location of the system—which are an integral part of the overall productive system design. The core of that productive system, however, is located in the complex of technology and people where the productive process is centered—in factories, hospitals, banks, offices, etc. We already noted that the entire design process contains interdependent components, and that the products, services, and locations are partially influenced by the core productive process, and vice versa. Although we shall not forget about this interaction in Chapters 7 and 8, this chapter will focus on process planning, and job design. Since these productive systems involve the adroit blending of technology and people to form a core productive system, they often are called *sociotechnical* systems. In the following chapter, we will consider some of the special design problems encountered in relating the processes and people in the form of facility layout.

The goal of the entire design process is to develop a rationale for the organization of the work to be done, and to relate this rationale to machines and technology in terms of work sharing between worker and machines, work flow, and physical facilities. The layout, which, on the surface, shows the spatial relationships, illustrates the physical integration of these factors. Whether or not the layout permits an effective design from points of view other than work flow and physical efficiency depends on the effectiveness of process planning and job design, and how technology and people are molded into a system.

Now we can consider the really fundamental alternatives of division of labor versus broad spectrum jobs—grouping tasks into jobs at fairly homogenous skill levels versus vertically integrating tasks into jobs. It is at this point that jobs can be created that either satisfy and fulfill workers, or dehumanize them.

We shall consider process planning, job design, and layout as an integral whole in the attempt to avoid the known effects that result when these elements are dealt with as separate, independent concepts. The known effects are that process planning (that is, technology and layout) has been thought of as the independent variable and that people and job designs have been thought of as the dependent variables. In that kind of framework, job designs were viewed as the results of process or technology planning. Currently developing concepts and practices consider the two components jointly to produce designs that satisfy the needs of both kinds of variables.

CRITERIA AND VALUES

From the time of the Industrial Revolution to the present, the main pressures that have influenced the design of processes and jobs are productivity improvement and economic optimization. These pressures have been associated with specialization. Adam Smith stated the advantages of division of labor as the guiding principle, and managers have applied this principle progressively over time. The automotive-type assembly lines epitomize the result, although the general division of labor principle has been applied throughout history and currently is being applied in office work as well as in other productive systems.

In 1776, Adam Smith enumerated the three important advantages of division of labor: (1) the development of a skill or a dexterity when a single task was performed repetitively; (2) a saving of the time normally lost in changing from one activity to the next; and (3) the invention of machines or tools that normally seemed to follow when workers specialized their efforts on tasks of restricted scope. To these three advantages, Charles Babbage [1832] added a fourth: the principle of limiting skills as a basis for wage payment. Babbage noted that: (1) wages paid were dictated by the most difficult or rarest skill required by the jobs; (2) division of labor enabled skills to be made homogenous within jobs more easily; and (3) for each job, one could purchase exactly the amount of skill needed. The result would be a lower total labor cost.

Scoville [1969], observing the current scene, notes that the specific technology involved determines the extent to which division of labor *can* be pursued, and that market size determines how far it *will* be pursued. Present day technology indeed has fostered specialization, and the size of markets for many products and services is enormous, often international, in scope. In short, from the time of Adam Smith to the present, the economic criterion has dominated. As a criterion, the economic effects of division of labor had no stopping rule. More finely divided jobs were assumed to be better because they were thought to increase productivity, and other factors were assumed to be correlated with productivity, including the workers' motivation (especially when incentive pay schemes were used).

Beginning in the early 1930s, however, another criterion was proposed as a counterbalance—job satisfaction. Results of the famous Hawthorne experiments indicated that workers responded to other factors in the work situation. In the late 1940s, the value of the job satisfaction criterion developed from a morale building program at IBM [Walker, 1950]. The term *job enlargement* was coined to describe the process of reversing the continuing trend toward specialization. Practical applications of job enlargement were written up in the literature. These applications described improvements in productivity and quality levels that resulted from combining operations to create jobs of broader scope. While job enlargement concepts did not specify any stopping rule either, they did provide a counterbalancing criterion.

In commenting on job satisfaction research, Davis [1972] points to several *values* held by the organizations that applied the division of labor criterion throughout the industrial era. He points out these widely held beliefs: that the worker could be viewed as an operating unit, and as such, could be adjusted and changed by training and incentives to suit the needs of the organization; that workers were viewed as spare parts and therefore were interchangeable in work assignments; that labor was thought of as a commodity to be bought and sold; that materialism, in its narrow sense of achieving material comfort, justified the means required to achieve it; and that many managers regarded jobs as isolated events in the lives of individuals—a noncareer.

Davis, however, is critical of most job enlargement and job satisfaction studies because they almost invariably accept the technology as given, and merely attempt to maximize satisfaction within technological constraints. He concludes that, throughout the industrial era, as a basis for process planning and job design, technology predominately has determined job content. This even holds true for most of the applications that consider job satisfaction as a criterion.

How Technology Can Determine Job Content

Consider a complex assembly process that has been broken down into a series of operations so that the product can be produced on an assembly line. The line is designed to meet certain capacity requirements; for example, an output of 480 units per 8 hour shift or 1 per minute. Output, then, dictates the maximum content of each operation, which can take no longer than 1 minute each. Furthermore, a certain sequence of assembly is required: Operation 1 takes the elements that come first that require 1 minute or less; Operation 2 the second 1-minute group; etc. Of course, we usually have some flexibility as to sequence; by rearranging the sequence, we end up with job content in each operation that largely has been determined by output requirements and the assembly line design framework.

In other situations, the process, machine, physical layout, time requirements, and traditions are likely to play a dominant role in determining job content. Each resulting job or operation can be analyzed from several viewpoints, such as motion patterns required, tools used, environmental impact, etc. But how would the resulting design compare with other basic alternatives of job content?

THE SOCIOTECHNICAL SYSTEM APPROACH

The essence of the current sociotechnical philosophy of designing processes and jobs is that constraints are imposed by (1) technology, limiting the possible arrangements of processes and jobs, and by (2) job satisfaction and social-system needs. The circle marked "technological constraints" in Figure 1 indicates that all job designs within the circle represent feasible solutions from a technological point of view, and that all designs outside the circle are infeasible. Similarly, the circle marked "social system constraints" in Figure 1 indicates that all job designs within that circle represent feasible solutions from the job satisfaction point of view, and that all designs outside its circle are infeasible. Within the shaded area of overlap between the two circles, however, we have a solution space that meets the constraints of both technology and the social system. The shaded area defines the only solutions that can be regarded as feasible in sociotechnical terms. Our objective is to optimize jointly the economic- and social system variables; that is, find the best possible solution to process and job designs within the feasible shaded solution space.

By looking within the joint feasible solution space, Scoville [1969] developed

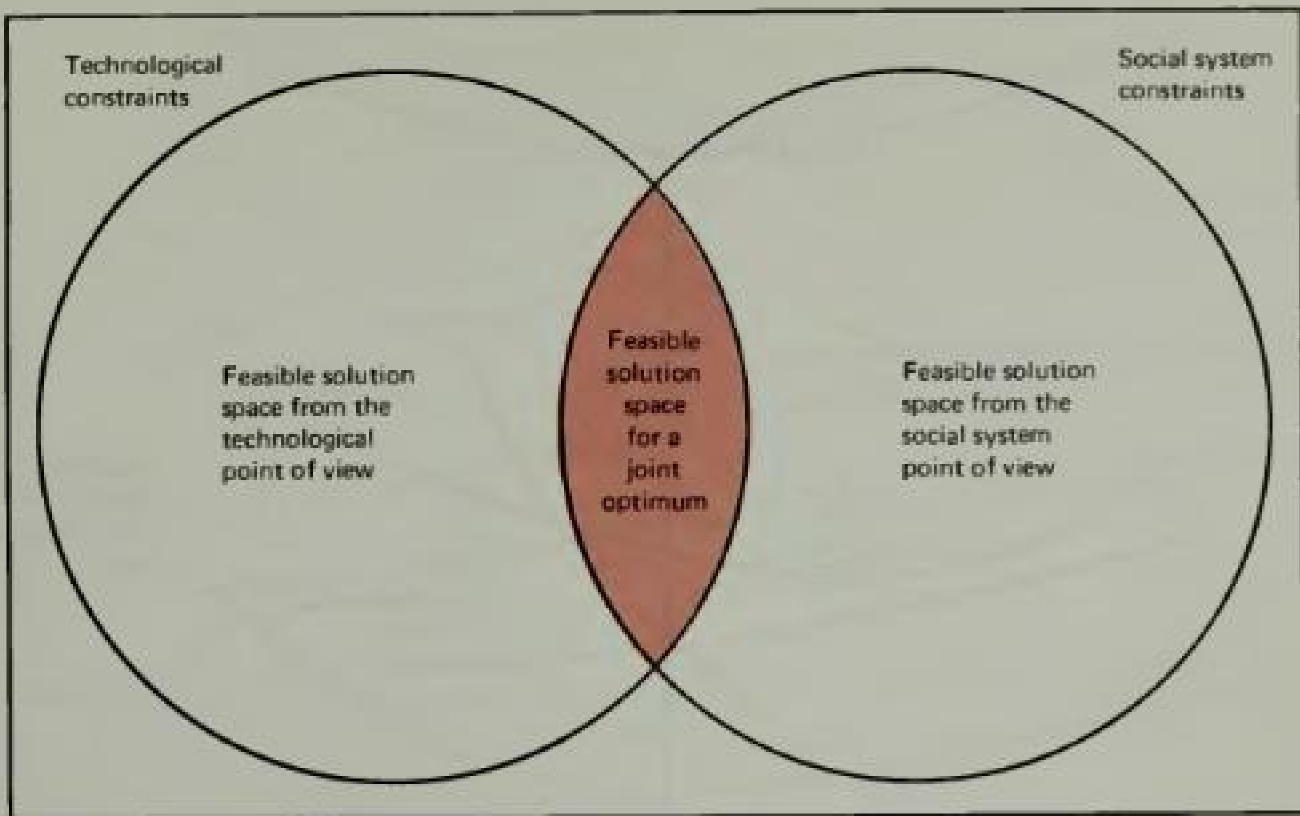


Figure 1. Feasible solution space within which a jointly optimum solution may be found, represented by the shaded area common to both of the defined feasible solution spaces for the technological and social systems

mathematical models that examine the question of job breadth separately from the points of view of managers and workers. Figure 2 shows a graphic form of the manager's model. Scoville argues for the shape of the curves rather than for any specific numerical solution.

The tasks and duties required can be reshuffled in many ways to form the continuum of narrow versus broad job-designs (represented by horizontal axes in Figure 2). For example, in auto assembly, jobs either can be divided finely (as is true on conventional auto assembly lines), or, at the opposite extreme, one worker on a team can assemble the entire vehicle. This kind of job enlargement can be termed *horizontal*. Alternately, a *vertical* enlargement occurs where jobs incorporate varying degrees of quality control, maintenance, repair, supply, and even supervisory functions.

The curves in Figure 2 are rationalized as follows:

1. The wage-cost curve reflects low productivity for both very narrow and very broad jobs, with a maximum occuring someplace in the middle range. Training costs go up as the scope of jobs increases, while turnover costs are most important for narrow jobs. Thus, the wage cost curve declines to a minimum and increases thereafter as job breadth increases.

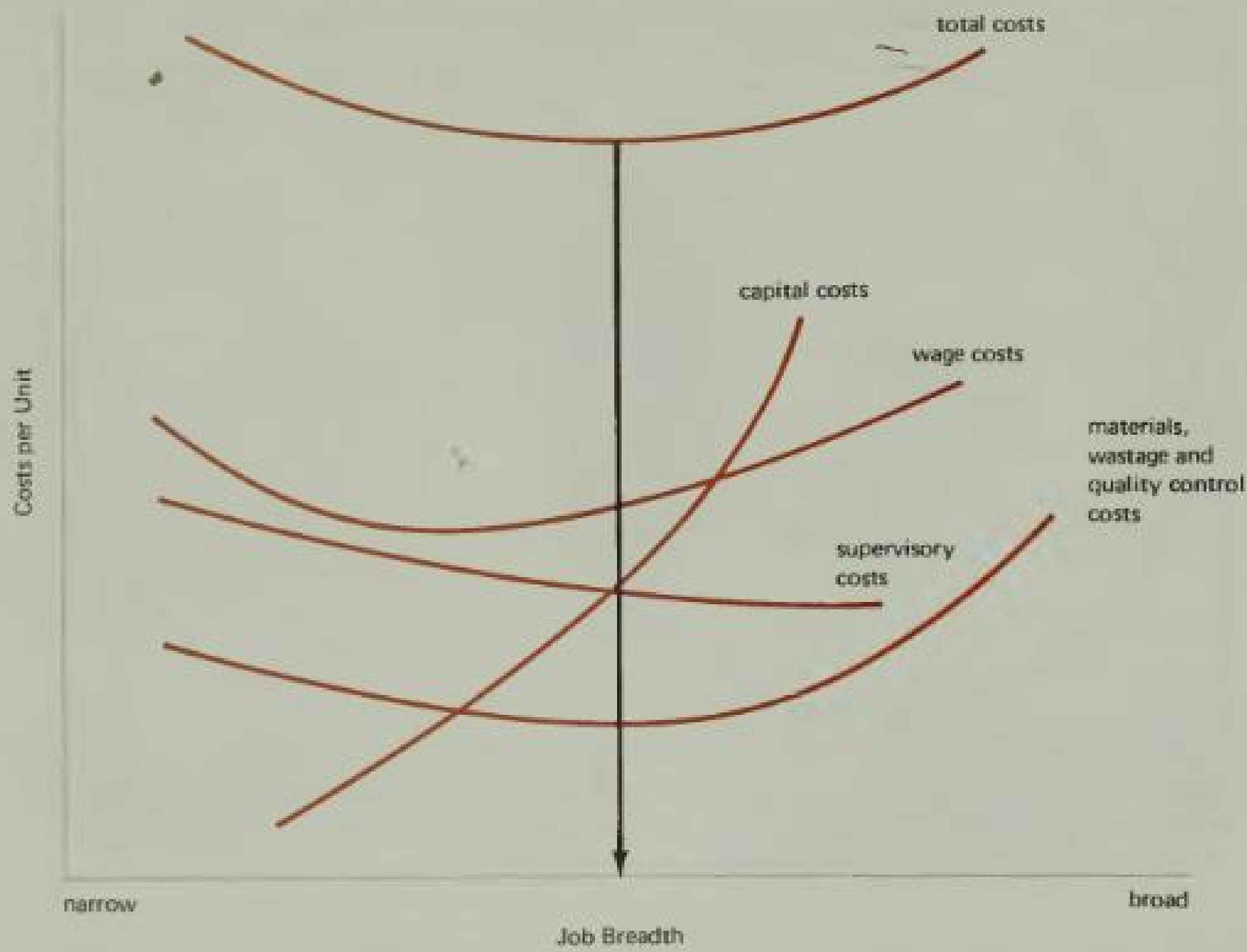


Figure 2. Manager's model of job breadth

SOURCE: J. G. Scoville, "A Theory of Jobs and Training," *Industrial Relations* 9(1969): 36–53.

2. Material, scrap, and quality control costs are high for narrow jobs, due to lack of motivation and penalties, and also are high for broad jobs because the advantages of division of labor are lost.
3. Supervisory costs decline with broader jobs because that function is progressively incorporated with jobs through job enlargement.
4. Capital costs per worker rise on the assumption that capital-labor cost ratios more than offset the inventory cost of goods in process.

The total cost curve in Figure 2 is the sum of the individual cost component curves, and reflects an optimum point at which managers would choose to operate. As with all cost allocation problems in which each pure strategy involves a cost, the joint optimum must lie someplace between the extremes. The pure strategies cannot represent the optimum, since the best solution necessarily results from a balance of costs.

Scoville's model of job breadth from the point of view of workers is shown graphically in Figure 3. The wage-productivity and employment probabilities

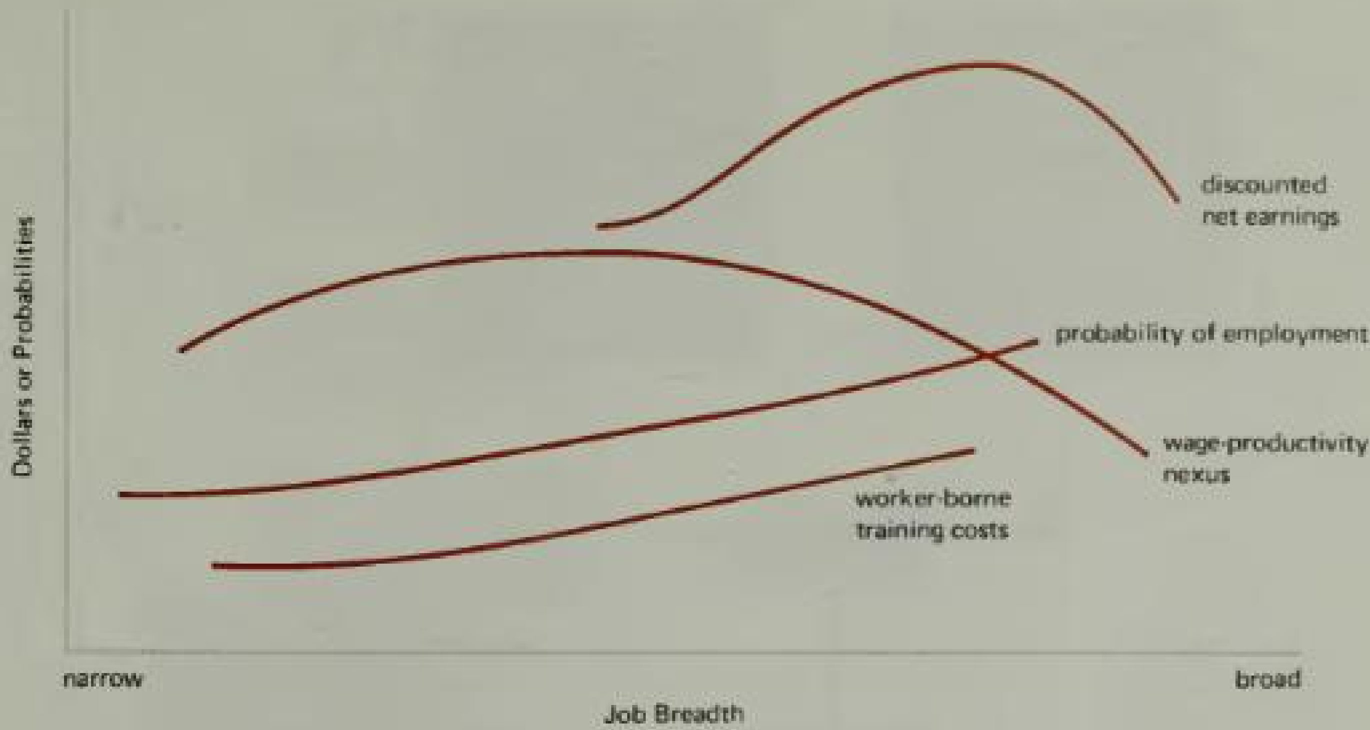


Figure 3. Worker's model of job breadth

SOURCE: J. G. Scoville, "A Theory of Jobs and Training," *Industrial Relations* 9(1969): 36–53.

curves are multiplied to produce the discounted expected earnings curve. If we then subtract the worker-borne training costs, we obtain a net economic benefits curve which has a maximum near the broad end of the job breadth spectrum. Therefore, optimum job breadth, from the workers' point of view, is likely to be nearer the broad end of the spectrum, compared to the optimum from the managers' point of view.

Thus Scoville's models indicate that there are optimum job designs from an economic point of view. However, since the factors that enter the two models are differnet, managers and workers are not likely to agree entirely on how work should be organized.

Therefore, our rationale for process and job design is an equilibrium model. In this model the balance of forces between labor and management is likely to produce an organization of work someplace within the joint feasible region of Figure 1. Figure 4 indicates that job content is a result of such a process: the constraints define a joint solution space, and the criteria and pressures determine where within the solution space we find a given work pattern. In the rationale of Figure 4, given job content, the detailed methods for actually performing work are determined in a second phase that draws on psychological and physiological data, work flow principles, etc. We shall discuss these concepts later in the chapter.

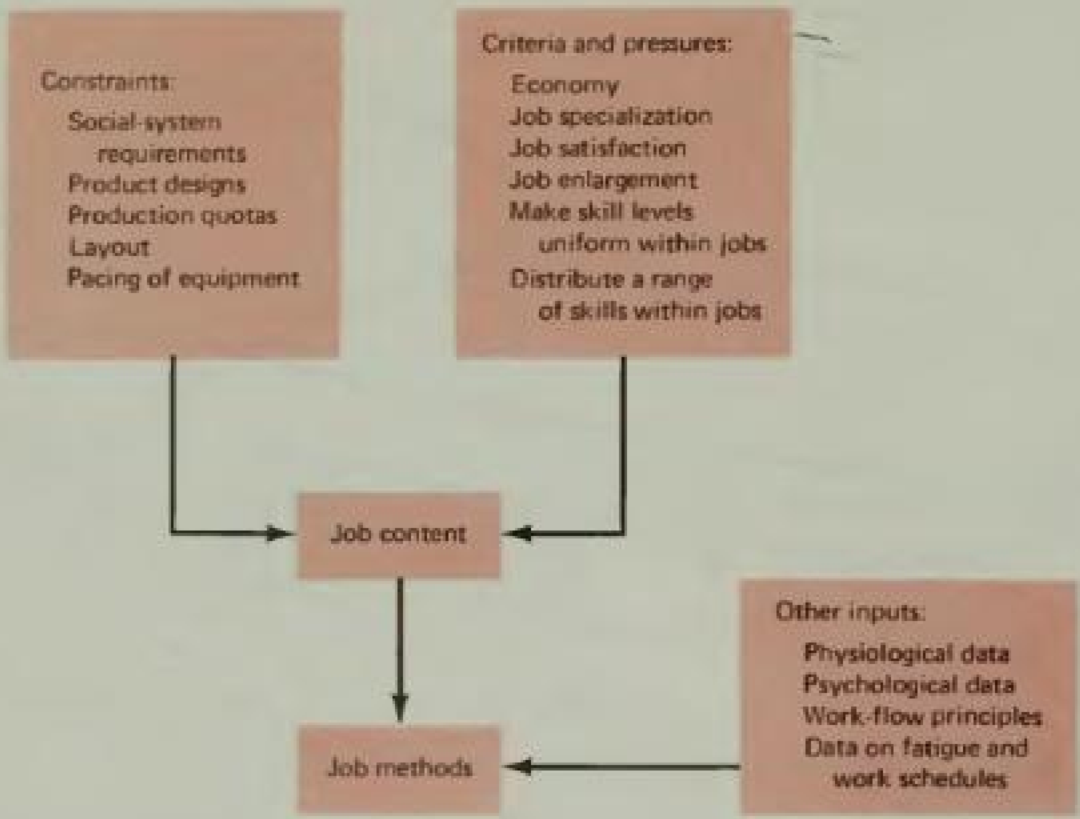


Figure 4. Relationship of constraints, criteria, and other pressures in determining job content. Inputs to job methods design.

The following sections discuss two views of the organization of work. The first is essentially a technological view, that considers the role of humans within technology. The second is the sociotechnical view, that revolves around job enlargement and the concepts of autonomous work groups.

THE TECHNOLOGICAL VIEW OF PROCESS PLANNING AND JOB DESIGN

While the general methods we shall describe were developed in manufacturing systems, they have been adapted and widely used in many other situations (e.g., offices, banks, hospitals, etc.). Thus, although we will take our examples from manufacturing settings, the methods are not restricted to only such settings.

Figure 5 shows the overall development of processing plans in a manufacturing situation. Process planning takes as its input the drawings or other specifications that might indicate *what* is to be made, and also the forecasts, orders, or contracts that indicate *how many* are to be made. The drawings then are analyzed to determine the overall scope of the project. If it is a complex assembled product, considerable effort may go into "exploding" the product into its components of parts and subassemblies. This overall planning may take the

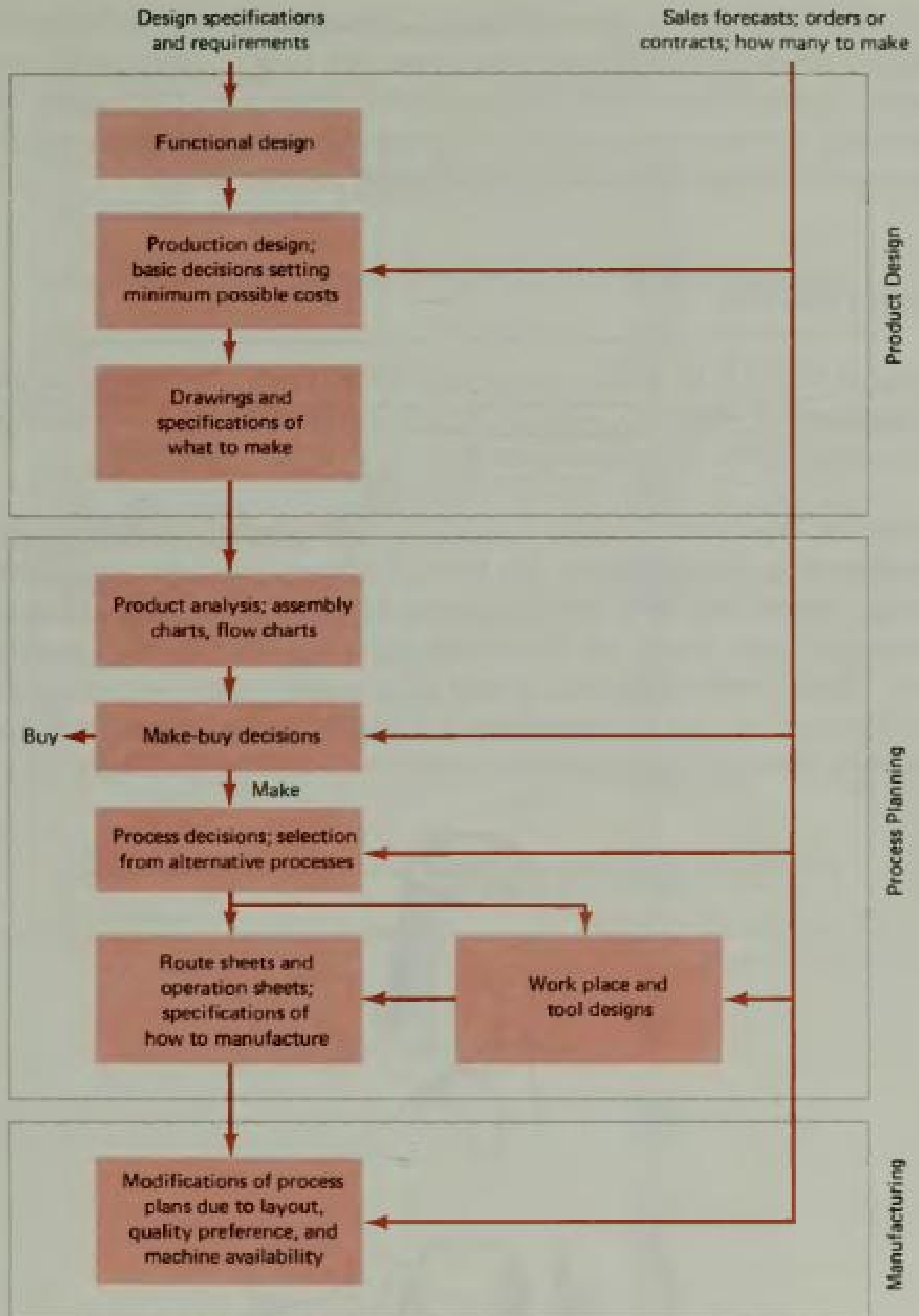


Figure 5. Diagram of the overall development of processing plans

form of special drawings that show the relationship of special parts, cutaway models, and assembly diagrams. At this point, preliminary decisions may be made about subassembly groupings to determine which parts to make and which to buy, as well as to determine the general level of tooling expenditure.

Then, for each part, a detailed routing through the system is developed. Technical knowledge of processes, machines, and their capabilities is required, as well as a knowledge of costs and production economics. Ordinarily, a range of processing alternatives is available. The selection may be influenced strongly by the overall volume and projected stability of product design.

Product Analysis

The product that is to be manufactured is analyzed from a technological point of view to determine what processes are required. We will take as an example the twelve part capacitor shown in Figure 6.

Assembly or "Gozinto" Charts. Schematic and graphic models commonly are developed to help visualize the flow of material and the relationship of parts (e.g., where they flow into the assembly process, which parts make up subassemblies, and where the purchased parts are used in the assembly sequence). Thus, for the capacitor, a first step might be the preparation of an "assembly chart" or, as it is often called, a "Gozinto" (goes into) chart. Figure 7 is an assembly chart for the capacitor. Notice how clearly the chart shows the

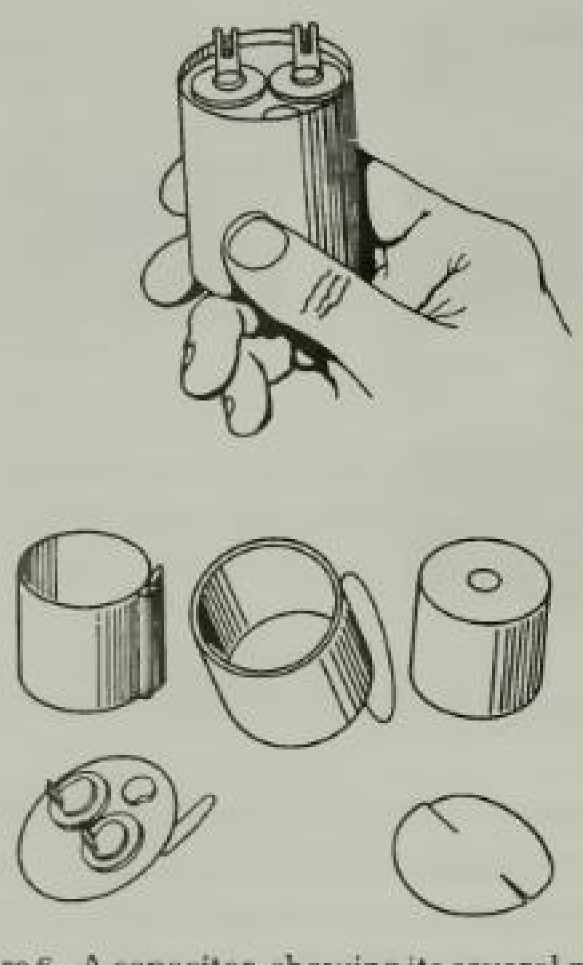


Figure 6. A capacitor, showing its several parts

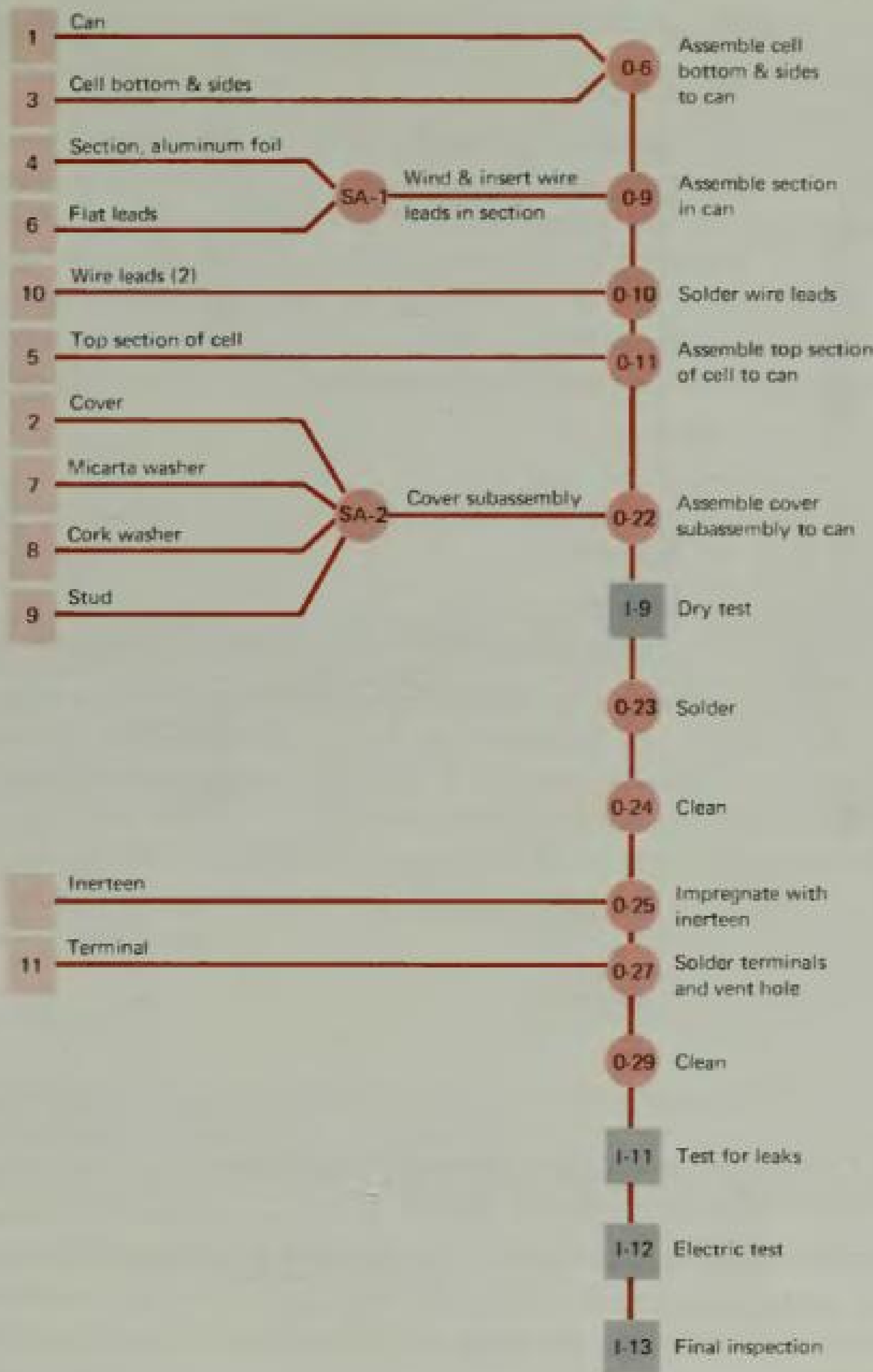


Figure 7. Assembly or "Gozinto" chart for capacitor of Figure 6

relationship of parts, the sequence of assembly, and which groups of parts make up subassemblies. Figure 7 summarizes the technology required to assemble the capacitor.

The assembly chart can be useful in making preliminary plans regarding

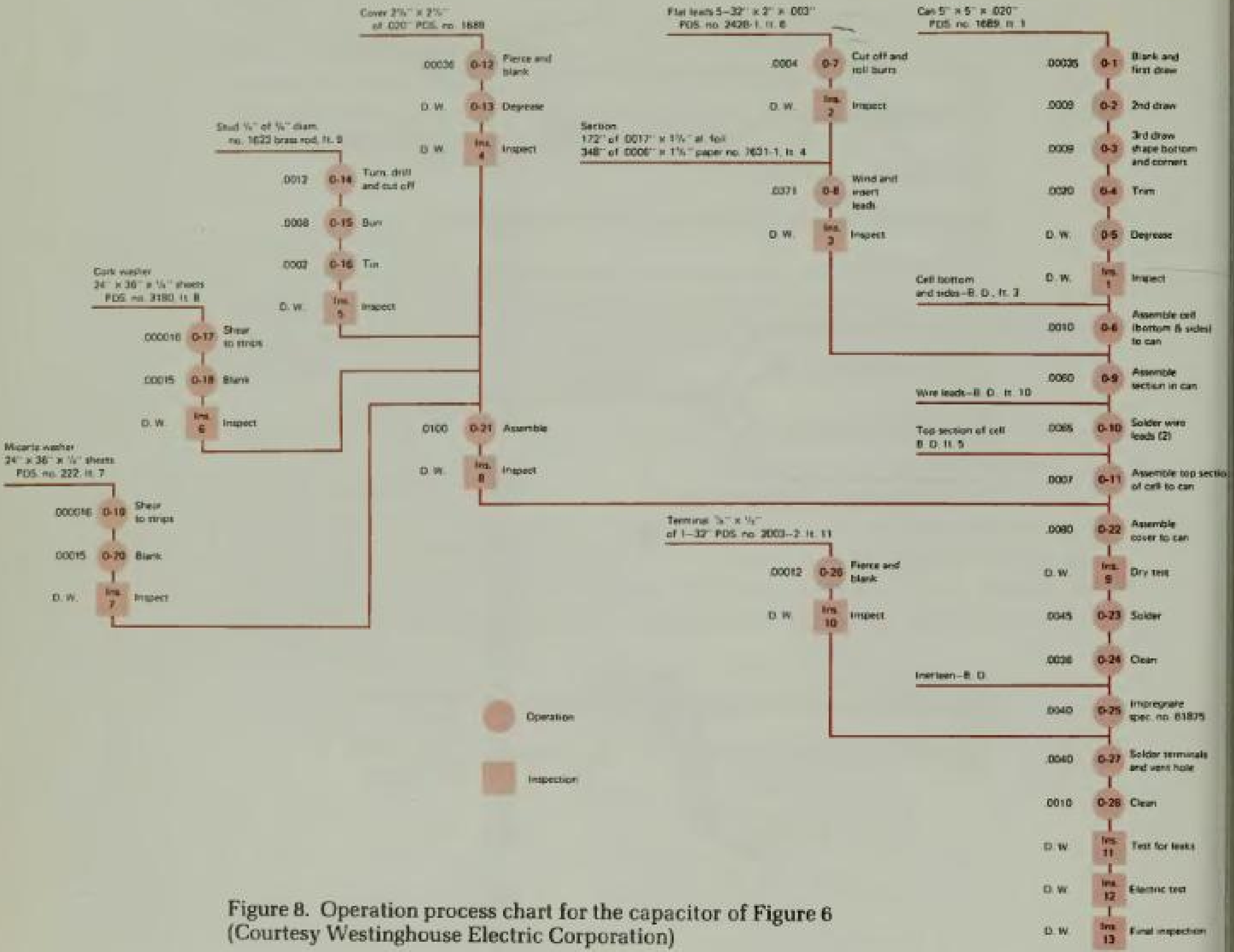


Figure 8. Operation process chart for the capacitor of Figure 6
(Courtesy Westinghouse Electric Corporation)

probable subassemblies (alternatives usually exist; for instance, the cover subassembly could be assembled separately) and appropriate general methods of manufacture. For example, where in the process might production lines be appropriate?

Operation Process Charts. If the product already is engineered, we have complete drawings and specifications of the parts and their dimensions, tolerances, and materials to be used. From the specifications and the forecast, we can develop a plan of "how to manufacture." Decisions must be made concerning which parts to purchase and which to manufacture in-house. The engineering drawings specify the locations, sizes, and tolerances for holes to be drilled, surfaces to be finished, etc., for each part. With this information, together with estimates of quantities to be produced and of manufacturing processes, we can specify the most economical equipment, processes, and sequences of processes.

The resulting operation process chart for the capacitor is shown in Figure 8, where circles represent operations to be performed and squares indicate required inspection points. Figure 8 is essentially a summary of all required operations and inspections for the capacitor. It is a general plan for manufacture. While the focus of Figure 8 is on the technological processing required, it is obvious that the jobs to be performed by workers also have been specified.

ANALYSIS OF MAN-MACHINE RELATIONSHIPS

Given the product analysis and the required technological processing that results from Figure 8 or similar analyses, individual job designs become the focus. The concepts and methods used have developed over a long period, beginning with scientific management era led by Frederick W. Taylor. The professional designers of processes and jobs have been industrial engineers in industry. In the post World War II period, psychologists and physiologists have contributed concepts and methods concerning the human role in man-machine systems that have reinforced the technological view.

Man vs. Machine

In the technologists' view, a human being has certain physiological, psychological, and sociological characteristics that define both his/her capabilities and limitations in the work situation. These characteristics are not thought of as fixed quantities, but rather as distributions that reflect individual variations.

In performing work, man's functions are envisioned in terms of three general classifications:

1. Receiving information through the various sense organs, i.e., eyes, ears, sense of touch, etc.
2. Making decisions based on information received and information stored in the memory of the individual.
3. Taking action based on decisions. In some instances, the decision phase may be virtually automatic because of learned responses, as in a highly repetitive task. In others, the decisions may involve an order of reasoning, and the result may be complex.

Note that the general structure of a closed-loop automated system is parallel in concept. Where is the difference? Are automated machines like men? Yes, in this model of man in the system, machines and men are alike in certain important respects. Both have sensors, stored information, comparators, decision makers, effectors, and feedback loops comparable to the automatic system for maintaining a given output shown in Figure 8 of Chapter 3. The difference is in the human's tremendous range of capabilities and in the limitations that human physiological and sociological characteristics impose. Thus, machines are much more specialized in the kinds and range of tasks they can perform. Machines perform tasks as faithful servants, reacting mainly to physical factors. For example, bearings may wear out because of a dusty environment, but man reacts to the psychological and sociological environment as well as the physical environment.

Although there are few really objective guides to the allocation of tasks to humans and machines on other than an economic basis, a subjective list of the kinds of tasks most appropriate for humans and for machines is given by McCormick [1970].

Human beings appear to surpass existing machines in their ability to:

1. Detect small amounts of light and sound.
2. Receive and organize patterns of light and sound.
3. Improvise and use flexible procedures.
4. Store large amounts of information for long periods and recall relevant facts at the appropriate time.
5. Reason inductively.
6. Exercise judgment.
7. Develop concepts and create methods.

Existing machines appear to surpass humans in their ability to:

1. Respond quickly to control signals.
2. Apply great force smoothly and precisely.
3. Perform repetitive and routine tasks.
4. Store information briefly and then erase it completely.
5. Perform rapid computations.
6. Perform many different functions simultaneously.

Of course, these lists of relative advantage for humans and machines raise a question: Why don't business, industry and government use humans and machines according to these guides? We all have observed that humans are used extensively to perform the tasks given in the list for machines. The answer lies in the balance of costs for a given situation. Both labor and machines cost money; when the balance of costs favors machines, conversions normally have been made.

Computers

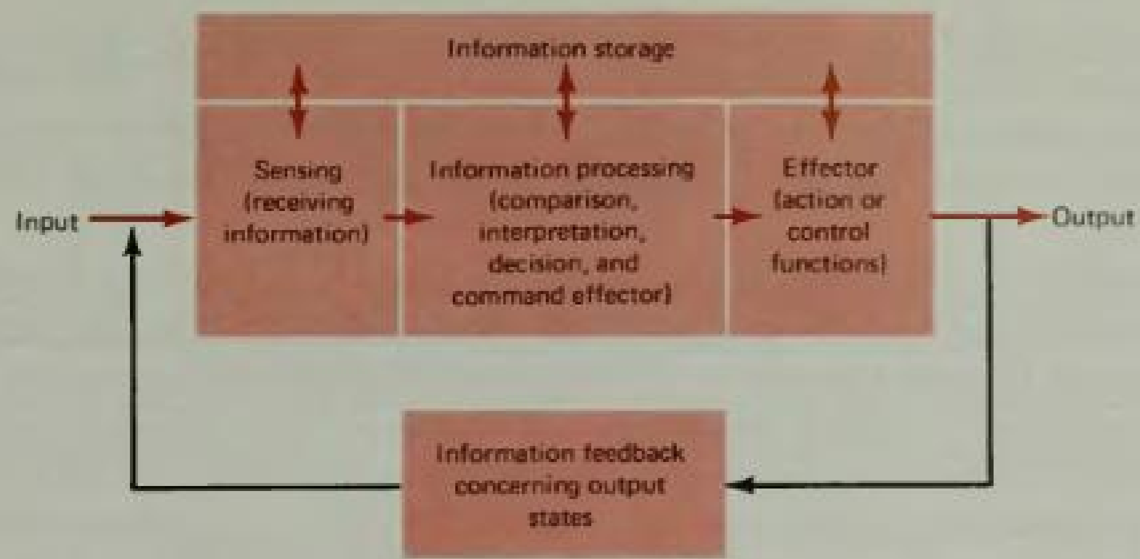


Figure 9. Functions performed by human or machine components of human-machine systems

Conceptual Framework for Man-Machine Systems

We have noted that humans and machines can be thought of as performing similar functions in accomplishing work tasks, although each has comparative advantages. The functions they perform are represented in Figure 9, and generally are comparable to those of the closed-loop feedback system.

Information is received by the *sensing function*. Human sensing is accomplished through the various sense organs of eyes, ears, sense of touch, etc. Machine sensing can parallel human sensing through electronic or mechanical devices, though it is usually much more specific or single-purpose in nature than the broadly capable human senses.

Information Storage for man takes place in the human memory or by access to records. Machine information storage can occur through magnetic tape or drum, punched cards, cams and templates, etc.

The function of *information processing* and *decision* takes sensed and/or stored information and produces a decision by some simple or complex process. The processing could be as simple as a choice between two alternatives, depending on input data; or it could be very complex, involving deduction, analysis, or computing to produce a decision for which a command is issued to the effector.

The *effector or action function* occurs as a result of decisions and command, and may involve human- or machine triggering of control mechanisms, or the communication of decisions. Control mechanisms, in turn would cause something physical to happen (e.g., moving the hands or arms, starting a motor, increasing or decreasing the depth of a cut on a machine tool, etc.).

Input and output are related to the raw material or the thing being processed. The output represents some transformation of the input, in line with our previous discussions of systems in general. The processes themselves may be of any type; chemical, changes in shape or form, assembly, transport, clerical, and so on.

Information feedback concerning the output states is an essential ingredient, for it provides the basis for control. Feedback operates to control the simplest hand motion through the senses and the nervous system. For machines, feedback concerning the output states provides the basis for machine adjustment. Automatic machines couple the feedback information directly so that adjustments are automatic (closed-loop automation). When machine adjustments are only periodic, based on information feedback, the loop is still closed, but not on a continuous and automatic basis.

Types of Man-Machine Systems

We shall use the module of the functions performed by man or machine shown in Figure 9 to discuss the basic structure for three typical systems: manual, semiautomatic, or mechanical and automatic systems. Figure 10 uses the module of Figure 9 to show the structure of the three types of systems in schematic form.

Manual systems involve humans using only mechanical aids or hand tools. The person supplies the power required and acts as controller of the process; the tools and mechanical aids help multiply his efforts. The basic module of Figure 9 describes the functions where the human directly transforms input into output, as shown in Figure 10*a*. In addition, we must envision the manual system as operating in some working environment that may have an impact on both the human and the output.

Semiautomatic systems involve the human mainly as a controller of the process, as indicated in Figure 10*b*. The human interacts with the machine by sensing information about the process, interpreting it, and using a set of controls that may start and stop the machine and possibly make intermediate adjustments. Power is normally supplied by the machine. Figure 11 shows the general cycle of activity of a semiautomatic type man-machine system imbedded in the working environment. Some combinations of the manual and semiautomatic systems also supply the human with some of the system power—perhaps in loading the machine, or in some activities that may involve the human while the machine goes through its cycle. Common examples of semiautomatic systems are the machine tools that frequently are used in the mechanical industries.

Automatic systems presumably do not require a human, since all the functions of sensing, information processing and decision making, and action are

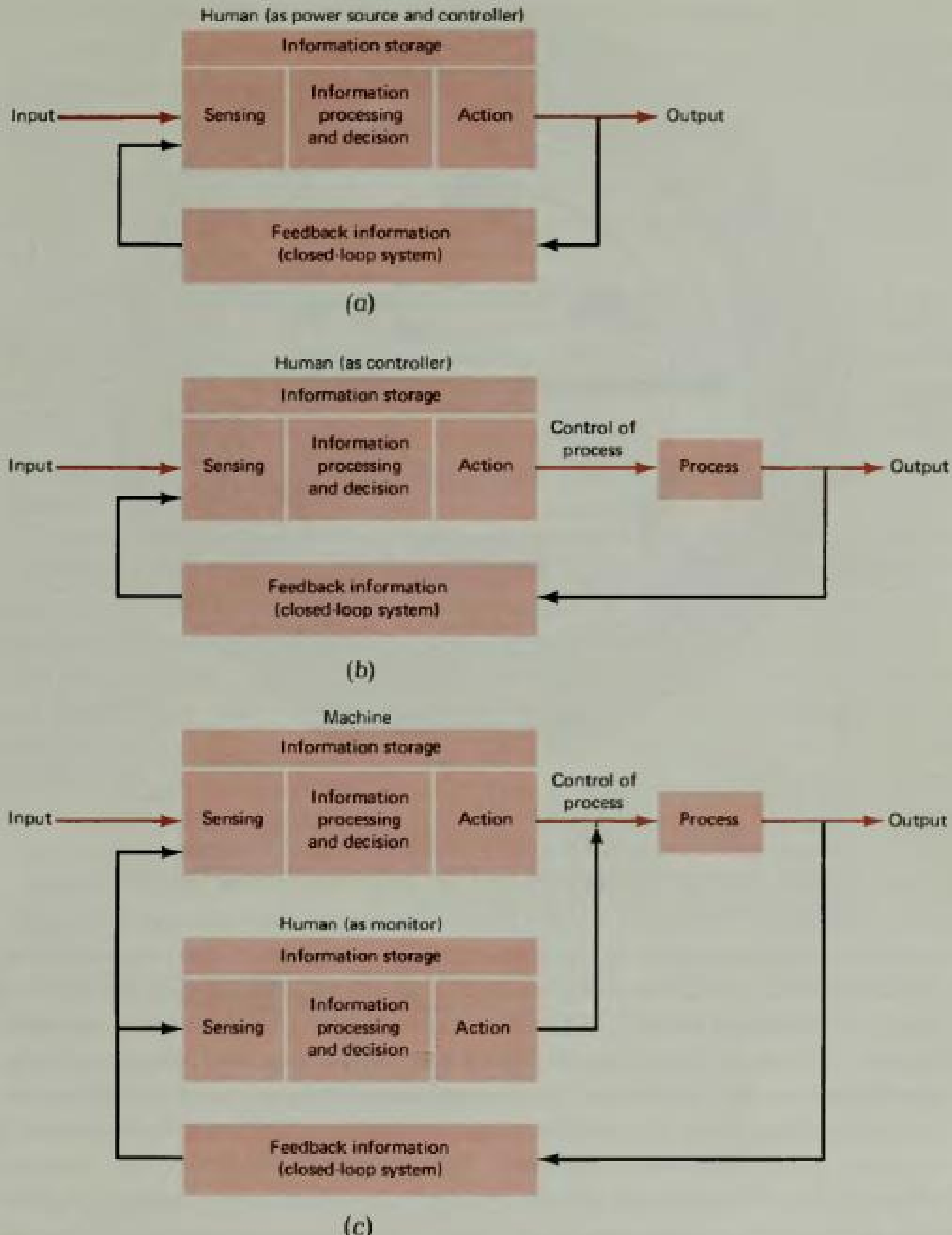


Figure 10. Schematic diagram of human and machine functions in three types of systems (From *Human Factors Engineering* by E. J. McCormick. Copyright 1970. Used with permission of McGraw-Hill Book Company.)

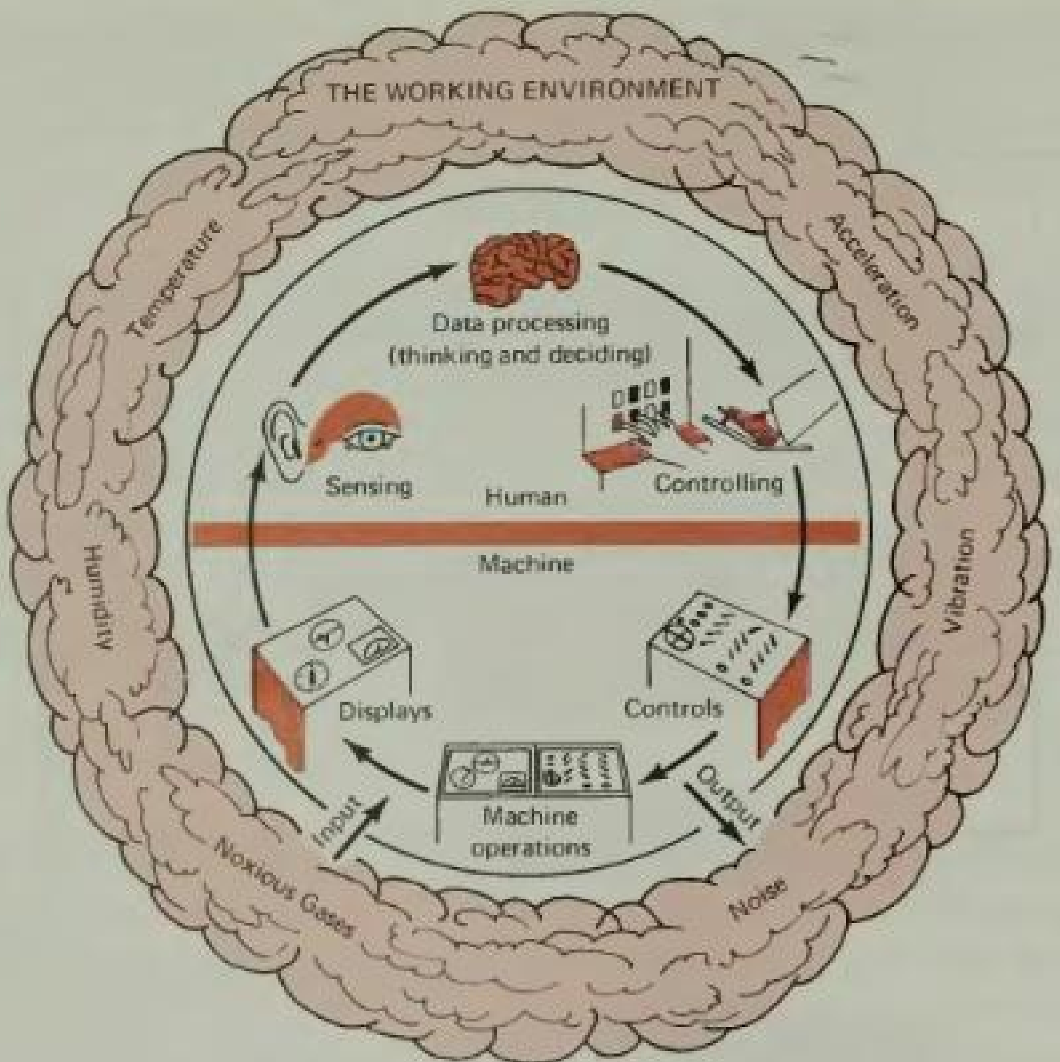


Figure 11. Simplified model of a human-machine system for semiautomatic operations (From A. Chapanis, *Man-Machine Engineering*, Copywright ©1965 by Wadsworth Publishing Company, Belmont, California. Reprinted by permission of the publisher, Brooks/Cole Publishing Company, Monterey, California.)

performed by the machine. Such a system would need to be fully programmed in order to sense and take required action for all possible contingencies. Even if such machines could be designed, automation at such a level is not economically justified. Therefore, Figure 10c indicates the human role as a monitor to help control the process. In this role, the human periodically or continuously maintains surveillance over the process through displays that indicate the state of the crucial parameters of the process.

Figure 11, representing the general cycle of activity for semiautomatic operations, is of central value, for it involves all the relationships that industrial engineers attempt to consider in their designs. In Figure 11, the human is receiving and interpreting information about the process from information displays, and manipulating mechanisms that control the processes. The human also is maintaining a general surveillance over the process. The machine responds to these control actions to convert input to output. Thus, we have represented all the kinds of human functions for the three types of man-machine systems discussed. For example, the manipulative activity represents the dominant features of manual systems; the information input, sensing, and controlling activities represent both semiautomatic and automatic systems; and the information input and general monitoring or surveillance activities represent automatic systems. The impact of the environment and the physical relationships of the human to work flow and work place arrangements apply to all man-machine systems.

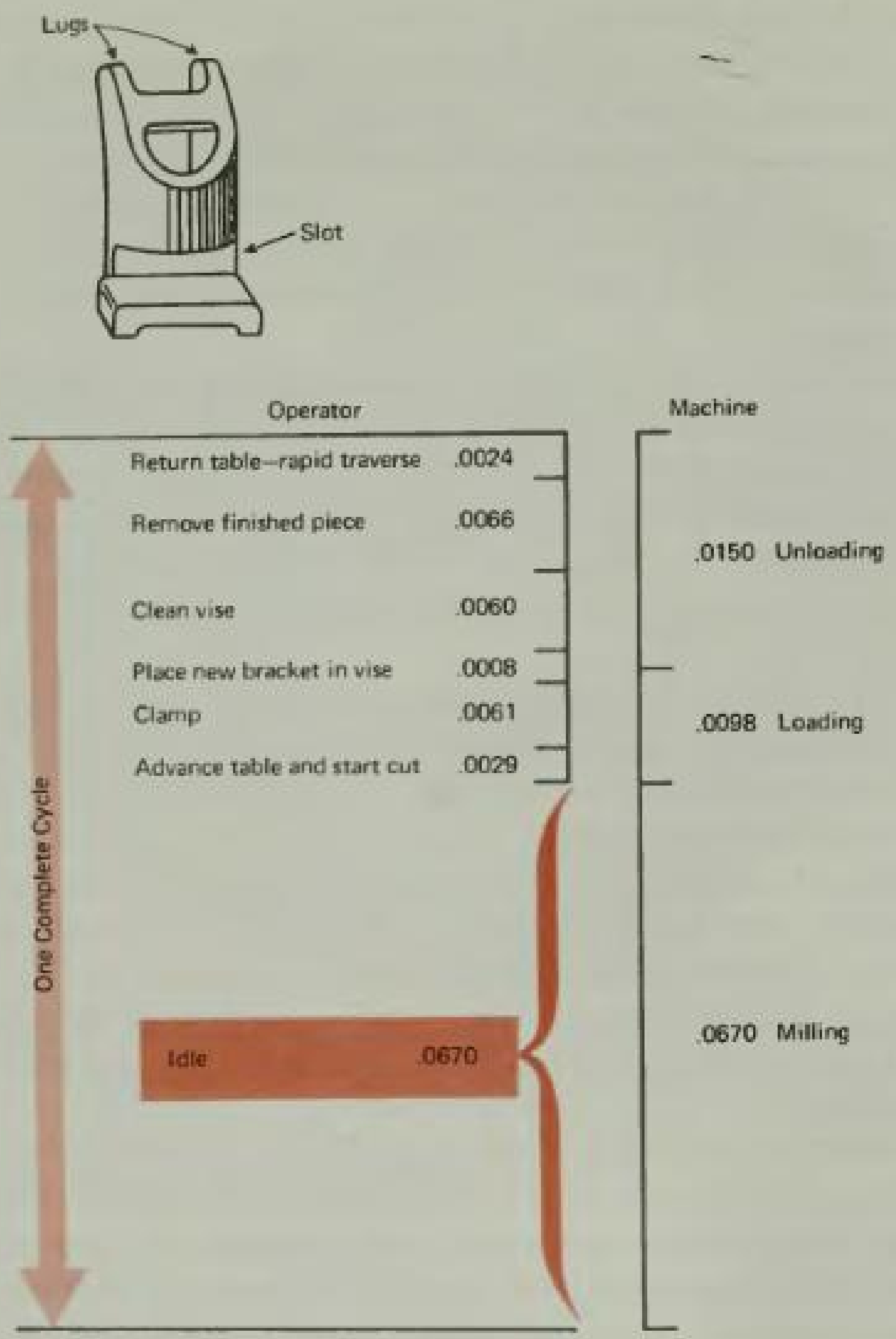
In analyzing manual activity and man-machine cycles, various types of graphic models are used, such as: operation charts; micro-motion or "simo" charts, which plot the simultaneous activities of the two hands to the thousandth of a minute; man-machine or multiple-activity charts; etc. These graphic models are used as design aids, in conjunction with the principles of motion economy, to develop superior designs from the point of view of the effective use of the human.

The *activity chart* is an example of graphic models used to analyze operations in terms of the time required to perform major manual and machine elements or activities. Relationships between man and machine or between crew members then can be examined when the activities are represented on a time scale. Let us take the milling of a slot in a bracket as an example. Figure 12 shows the bracket and the activity chart for the milling of the slot. The major elements of the repetitive work of the human and the machine have been plotted side by side on a time scale. In this instance, the times are recorded in decimal hours.* Frequently, the major objective of the analysis is to maximize both human and machine utilization.

In the example of the milling of the bracket slot, we see that the machine is utilized 100 percent, since it always is being loaded, unloaded, or actually taking a cut. At no time is it idle. Nevertheless, the machine effectiveness could be increased by improving the manual methods of loading and unloading, thus improving the machine output per unit of time. The techniques for accomplishing this require a detailed study of the manual activity by other graphic means. On the other hand, the human is idle 73 percent of the cycle while he or she waits for the milling machine to complete its cut. This general situation is common in many kinds of machine and process operations.

The question is what to do with the human's idle time in such a situation. Perhaps the first consideration is whether the operator is really idle, since some types of machine operations require operator vigilance and surveillance during

*Time values commonly are derived through the well established methods of work measurement. The resulting output standards also are used as a basis for incentive wage payment plans. For methods, see Barnes [1968] and Nadler [1970].



Time Recap for One Cycle

Operator occupied	.0248	.0670	Machine running
		.0248	Machine preparation
Operator idle	.0670	.0000	Machine idle
Total	.0918	.0918	Total

Figure 12. Activity chart of "mill slot in bracket" (Courtesy Westinghouse Electric Corporation)

the machine cycle. An attempt to utilize such time could affect quality adversely. Often, however, this is purely idle time. When this is true in repetitive operations of considerable volume it may be possible to have the human operate two or more machines. For the bracket slot milling operation, the operator could handle three machines doing the identical operation without introducing any idle machine time. Idle machine time develops beyond that number, and an economy study would be required to determine whether idle human- or machine time would be preferable.

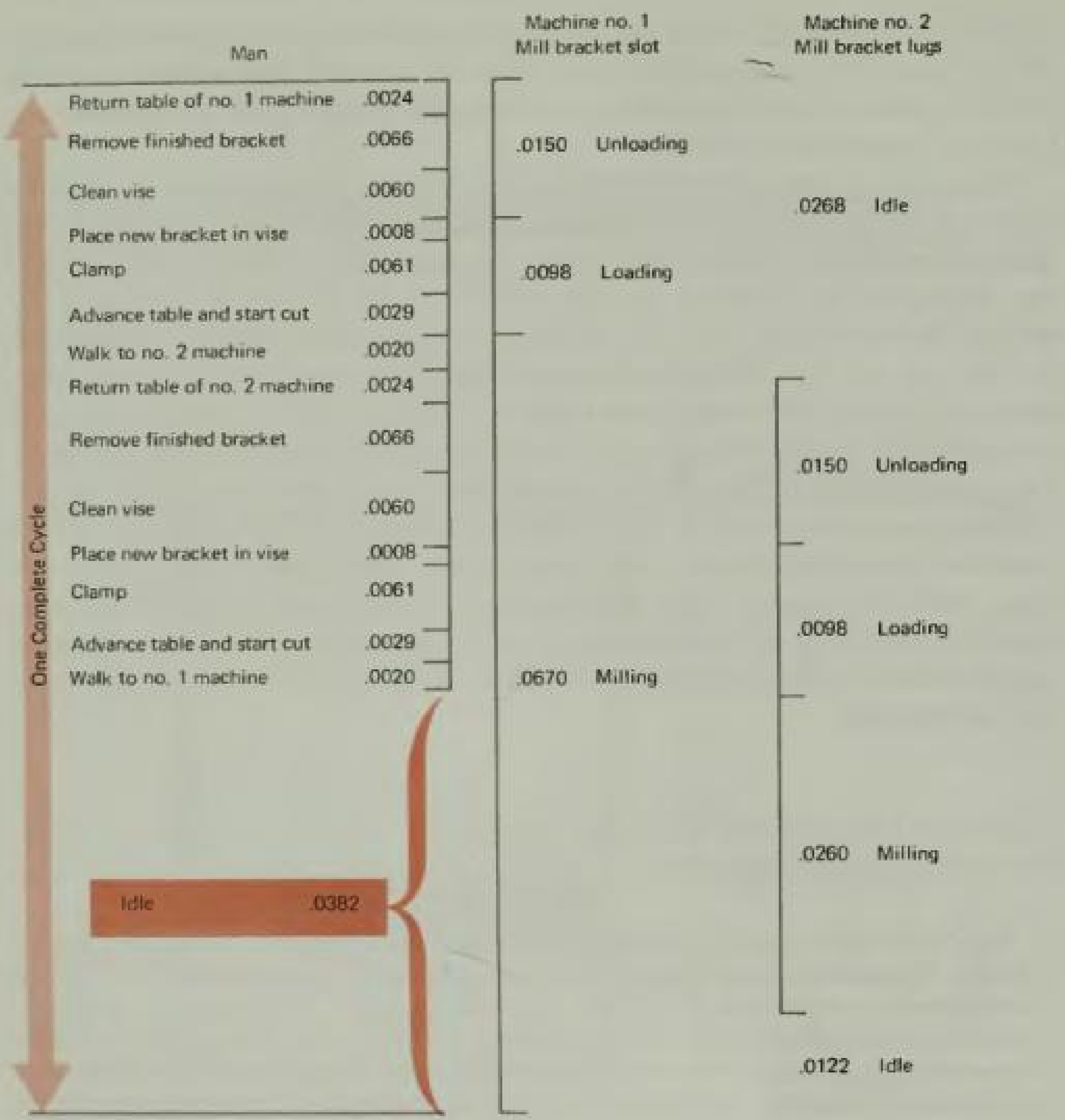
Where the volume of activity in question does not justify multiple machine operation, examination of the operation process chart for that part might reveal other elements that could be performed during the idle time. As an example, the operation that followed the slot milling on the bracket happened to be milling the lugs. Figure 13 is the resulting composite operation of milling both the slot and the lugs. Note that the man's idle time now has been reduced to about 30 percent of the total cycle. However, in so structuring the job, about 45 percent of idle machine time has been introduced into the lug milling operation. The question again is whether idle human- or idle machine time is preferable.

The industrial engineer's general methods in designing processes and jobs are presented in detail in [Barnes, 1968; Nadler, 1970; Niebel, 1974]. [Chapanis, 1965; Fogel, 1963; McCormick, 1970; and Woodson, 1966] present detailed concepts and data concerning information input, visual and auditory displays, analysis of manual control activity, speed and accuracy of motor responses, and the physical environment.

THE SOCIOTECHNICAL VIEW OF PROCESS PLANNING AND JOB DESIGN

> We are living in a period of rapidly increasing, wrenching social and technological change. Organizations and job structures are undergoing change in response to these developments and because of our increased knowledge of the behavior of organizations and individuals. Western industrial society is in transition from one historical era to another, and the environmental characteristics of the emerging era will lead to crisis and massive dislocation unless there is some form of adaption. The structures of most purposive organizations, based as they are on concepts of the industrial era, are becoming increasingly disfunctional because they stand at the confluence of changes involving technology, social values, the economic environment, and the practice of management [Davis and Taylor, 1972].

To the sociotechnical school, the concepts and methods of the technological view are mechanistic and view the human as a machine or, worse, as just a link in a machine. Sociotechnical advocates feel that, without question, the central focus of the technological approach is technology itself, that technology is taken as given without exploring the full range of the possible alternatives. Davis states, "When we examine the industrial era's approach to organization, 'Scientific Management'—which in reality is the machine theory of organization—we find some deeply disturbing omissions. No clear objectives concerning roles for men



Time Recap for One Cycle

		.0670	Machine running	.0280	Machine running
Man occupied	.0536	.0248	Machine preparation	.0248	Machine preparation
Man idle	.0382	.0000	Machine idle	.0390	Machine idle
Total	.0918	.0918	Total	.0918	Total

Figure 13. Activity chart for the composite operation, "mill slot and lugs of bracket" (Courtesy Westinghouse Electric Corporation)

as men are visible, although objectives are clearly defined for men as machines [Davis, 1972].

Briefly, sociotechnical theory rests on two essential premises. The first is that there is a joint system operating, a sociotechnical system, and that "joint optimization" is appropriate (as indicated by the Venn diagram in Figure 1). The second premise is that every sociotechnical system is embedded within an environment that is influenced by a culture and its values, and by a set of generally accepted practices where there are certain roles for organizations, groups, and people.

Englestad [1972] developed a set of psychological job requirements that translate the empirical evidence suggesting that workers prefer tasks that have a substantial degree of wholeness (that is, that show a strong Gestalt), and that allow the individual to have control over the materials and the processes involved. Englestad's psychological job requirements are as follows:

1. The need for the job content to be reasonably demanding in terms other than sheer endurance, yet to provide at least a minimum of variety (not necessarily novelty).
2. The need for the opportunity to learn on the job (which implies standards and knowledge of results) and go on learning. Neither too much nor too little opportunity must be offered.
3. The individual's need for some minimum area of decision making.
4. The need for some minimum degree of social support and recognition in the work place.
5. The need to be able to relate what one does and what one produces to one's social life.
6. The need to feel that the job leads to some sort of desirable future.

Thus, the principles of process and job design may be summarized as: the application of the concept of joint optimization between technological and social-system values; and the idea of wholeness and self control, as elaborated in the list of psychological job requirements. These are the principles that guide the overall organization of work in specific situations, resulting in the establishment of job content. The determination of job methods, in the sociotechnical view, results from a concept of the semiautonomous work group which, by and large, creates its own methods.

To date, wholeness has been achieved through job enlargement and enrichment. The sociotechnologists feel that even the following examples fall short, because the technology was taken as given, even though job satisfaction was a major criterion.

JOB ENLARGEMENT AND ENRICHMENT

The term *job enlargement* originated as a result of a morale building program at the IBM Corporation. Their company president felt deeply that the then current modes of organizing work left very little to hold the interest of, and give satis-

TABLE 1
NUMBER OF OPERATIONS PERFORMED CORRELATED WITH DEGREE OF JOB INTEREST, PRESENT JOB

Operations Performed	*Very or Fairly Interesting*	*Not Very or Not at All Interesting*	*Total*
1	19	38	57
2–5	28	36	64
5 or more	41	18	59
	88	92	180

SOURCE: C. R. Walker and R. H. Guest, *The Man on the Assembly Line* (Cambridge, Mass.: Harvard University Press, 1952)

faction to, the worker. A conscious program of enriching jobs in variety, interest and significance was established. Since the IBM experience, a number of case studies and research works have indicated that benefits can result from organizing work in such a way that jobs of broader scope result. One survey [Walker and Guest, 1952] attempted to correlate degree of job interest with the number of operations performed. They asked 180 assembly line workers this question: "Would you say your job was interesting, fairly interesting, not too interesting, or not at all interesting?" The results are tabulated in Table 1. There are significant differences between the interest-noninterest categories in relation to the number of operations performed. This essentially lends credence to the hypothesis that jobs of broader scope lend interest and satisfaction to the worker.

The approaches to job enlargement and enrichment have included both *horizontal* and *vertical* integration of tasks and responsibilities. Horizontal integration involves a broadening of the tasks performed, ~~though the new set~~ of tasks may be basically similar in nature to the old set. For example, instead of performing only a small portion of the assembly of an item, the ultimate of horizontal integration might require a single worker to assemble the entire unit. Each worker who formerly performed only a portion of the assembly in assembly line fashion would be involed in a somewhat longer cycle time. The rationale is that horizontal integration provides a feeling of closure and having completed something. It also provides greater flexibility in scheduling since there is reduced interdependence between workers.

Vertical integration involves the addition of responsibilities, such as elements of quality control, or planning and scheduling of the basic tasks being performed. Herzberg [1968] provides a list of seven types of vertical integration in jobs and associated motivators in Table 2.

TABLE 2

TYPES OF VERTICAL INTEGRATION IN JOBS

Type of Vertical Integration	*Motivators Involved*
A. Removing some controls while retaining accountability	Responsibility and personal achievement
B. Increasing the accountability of individuals for own work	Responsibility and recognition
C. Giving a person a complete natural unit of work (module, division, area, and so on)	Responsibility, achievement, and recognition
D. Granting additional authority to an employee in his/her activity; job freedom	Responsibility, achievement, and recognition
E. Making periodic reports directly available to the worker him or herself rather than to the supervisor	Internal recognition
F. Introducing new and more difficult tasks not previously handled	Growth and learning
G. Assigning individuals specific or specialized tasks, enabling workers to become experts	Responsibility, growth, and advancement

SOURCE: F. Herzberg, "One More Time: How Do You Motivate Employees?" *Harvard Business Review*, January–February 1968, p. 59.

Job Methods Design and Semiautonomous Work Groups

It always has been assumed that professional job designers should be able to design superior methods of work because of their knowledge of work flow, psychological and physiological capabilities of humans, machines, and tool design. This accepted principle began to be questioned when the human problem of resistance to change was recognized. We long have been aware of the problems of introducing change in an existing situation. At the risk of oversimplifying, the approach to the problem has been to involve in the change process the people who would be affected by the change.

Many case studies have shown that involvement seemed to produce a "low-threat" situation, whereas the more traditional mode of introducing the change —through the authority system of the organization—was felt to be a "high-threat" technique. High-threat techniques produce resistance to change. Such resistance can manifest itself in many ways: through noncooperation, poor motivation, and even active opposition or sabotage. A superior job design—from the point of view of work flow, tools, and psychophysiological data—can be masked by the effects produced through resistance to the change.

The low-threat approach tends to minimize opposition, and it has been shown that under certain circumstances, a work group can be a driving force for promoting change. Full participation in the decisions that affect the work group seems to be called for. In its full meaning, this includes the design of the procedures and methods by which the work of the group is carried out. Following are reports of case studies and experiments on work designs that were produced by semiautonomous work groups.

The Hovey and Beard Company. This company manufactured a line of wooden toys. One part of the process involved spray painting partially assembled toys, after which the toys were hung on moving hooks that carried them through a drying oven. The operation, staffed entirely by women, was plagued with absenteeism, high turnover, and low morale. Each woman at her paint booth would take a toy from the tray beside her, position it in a fixture, and spray on the color according to the required pattern. She then would release the toy and hang it on the conveyor hook. The rate at which the hooks moved had been calculated so that each woman, once fully trained, would be able to hang a painted toy on each hook before it passed beyond her reach.

The women who worked in the paint room were on a group incentive plan, that tied their earnings to the production of the entire group. Since the operation was new, they received a learning allowance that decreased by regular amounts each month. The learning allowance was scheduled to fall to zero in six months, since it was expected that the women could meet standard output or more by that time. By the second month of the training period, trouble had developed. The women had progressed more slowly than had been anticipated, and it appeared that their production level would stabilize somewhat below the planned level. Some women complained about the speed that was expected of them, and a few of them quit. There was evidence of resistance to the new situation.

Through the counsel of a consultant, the foreman finally decided to bring the women together for more general discussions of working conditions. After two meetings in which relations between the work group and the foreman were somewhat improved, a third meeting produced the suggestion that control of the conveyor speed be turned over to the work group. The women explained that they felt that they could keep up with the speed of the conveyor, but that they could not work at that pace all day long. They wished to be able to adjust the speed of the belt, depending on how they felt.

After consultation, the foreman had a control with a dial marked, "low, medium, and fast" installed at the booth of the group leader, who could adjust the speed of the conveyor anywhere between the lower and upper limits that had been set. The women were delighted and spent many lunch hours deciding how the speed should be varied from hour to hour throughout the day. Within a week, a pattern had emerged: the first half-hour of the shift was run on what the women called "medium speed" (a dial setting slightly above the point marked "medium"). The next two-and-a-half hours were run at high speed, and the half-hour before lunch and the half-hour after lunch were run at low speed. The rest of the afternoon was run at high speed, with the exception of the last forty-five minutes of the shift, which were run at medium speed.

In view of the women's report of satisfaction and ease in their work, it is interesting to note that the original speed was slightly below medium on the dial of the new control. The average speed at which the women were running the belt was on the high side of the dial. Few, if any, empty hooks entered the drying oven, and inspection showed no increase of rejects from the paint room. Production increased, and within three weeks the women were operating at 30 to 50 percent above the level that had been expected according to the original design.

As a semiautonomous group, participation in the critical decision had opened the way to an acceptance of a higher average speed than had been expected. Although this was only one element in the overall design of the system, it was an extremely important element. The interesting thing about this case is that, based on research knowledge about the effective design of production lines, neither the original solution nor the work group's solution represents ideal design. Formal analysis, using a waiting line model [Hunt, 1956] and experimental research, indicate that rigid pacing at any speed represents a relatively poor solution to serialized operations [Conrad, 1955; Conrad and Hille, 1955].

The best design would eliminate rigid pacing completely, replacing it with a system where the conveyor provided only transportation between the several operations, allowing queues of material to build up between each operation in the sequence. The rationale is simply that the existence of queues provides greater flexibility that makes it possible for operators to average their slow cycles against their fast cycles in their work time distribution patterns. The result is more effective utilization of labor on the line. Had the women had the opportunity to participate more fully in the design of the system, and had they been given some information about the system with which they were dealing, it might have been possible to derive an even better system than the one that finally resulted.

Coal Mining in England. Trist and coworkers report a study about coal mining in England, that took place over a long period of time, and was reported periodically in the literature [Trist and Bamforth, 1951; Emery and Trist, 1960; Trist et.al., 1963]. The study contrasts two alternate ways of organizing work in a coal mining operation, technology and other conditions remaining the same. The conventional approach had been to divide the labor force of forty-one men into fourteen segregated task groups, and to give them specialized tasks to perform.

In the conventional organization, activities were divided into seven specialized tasks. Each was carried out by a different task group, and mine output depended on completion of a working cycle. The cycle consisted of preparing an area for coal extraction, using machinery to dig the coal out of the face, and removing the coal with the aid of conveyors. each of the tasks had to be completed, in sequence and on schedule, over three working shifts. On each of the shifts, one or more task groups could perform their work only if the preceding tasks had been completed. The filling tasks for coal removal were the most onerous, and frequently were not completed, resulting in a delay in the work cycle and a reduction of output. Each worker was paid an incentive to perform his task, without reference to the other tasks of workers in his or other work groups. The result of the conventional system was the development of isolated work groups, each with its own customs, agreements with management, and pay arrangements focused on its own interests. Coordination between men and groups on different shifts, and control of work, had to be provided entirely by management.

In the new, composite design, all the required roles were allocated to members internally, by the work group itself. The basic objective of the new design was to maintain continuity for achieving a complete work cycle (which commonly extended over more than one shift). Integration of the objectives of the crews in the three shifts was aided by: (1) setting goals for the performance of the entire cycle; (2) making payments to the group as a whole for the completion of all the tasks required in the cycle; and (3) offering an incentive payment for output. The wage payment scheme placed responsibility on the group as a whole for all operations, and generated the interrelationship of individuals performing different tasks. The nature of the working relations and payment scheme led to the spontaneous development of worker interchangeability according to need. This, in turn, required the development of multiskilled workers and, in essence, an extremely flexible work force.

Many other case studies have reported the results of various aspects of job enlargement and semiautonomous work groups: Harwood Manufacturing Company [Cartwright and Zander, 1960]; home appliance manufacturing [Connant and Kilbridge, 1965]; pharmaceutical appliances [Davis and Canter, 1956]; maintenance craftsmen in a chemical plant [Davis and Werling, 1960]; the paper and pulp industry in Norway [Englestad, 1972]; textile weaving in India [Rice, 1953]; petroleum refining [Susman, 1972; Taylor, 1972]; and the IBM Corporation [Walker, 1950].

Summary

In the early stages of industrial development, technological factors dominated the design of processes and jobs. In essence, jobs were the result of the technological process planning, and human factors were regarded more as limitations. Beginning as early as 1920, however, people have insisted that human factors be given greater consideration, and the conceptual framework that includes human factors has developed progressively, culminating in the sociotechnical approach. The evidence seems to indicate that consideration of human factors sometimes can result in increased productivity. Even so, the weight placed on human values will continue to be emphasized, even if productivity were to decrease. The sociotechnical framework provides a model in which these conflicting criteria can be examined and attempts can be made to optimize within the feasible region of Figure 1.

The sociotechnical model is a general one that can incorporate various social criteria related to environmental protection, energy conservation, consumer protection, and so on. The conceptual framework provides a stage on which the decision maker may play out the results of alternatives and trade off the conflicting criteria. The determination of these trade-offs is one of the major tasks that will face today's operations management student in the design, implementation, and operation of productive systems in his/her professional future.

While we need not hold a funeral for the "division of labor" concept, it is obvious from the case studies on job enlargement and enrichment that practice has overshot the optimum level of specialization. Scoville's models of job breadth form a framework for considering the economic consequences of specialization. From both the manager's and the worker's point of view, there is an optimum specialization, though the two optima may not be identical.

To this point, we have not considered processes and spatial relationships among and between workers. The flow of work through a productive facility needs to be given careful thought in order to minimize costs. In addition, sociotechnical system designs depend in part on the actual physical layout; thus, intercommunication can occur both for business and social reasons. The basic organization of work, that we have been discussing, leads into a discussion of the nature of facility layout.

Review Questions

1. In Scoville's view, what are the factors that limit the extent of division of labor in practice?
2. In Davis's view, what are the values held by organizations that applied the division of labor criterion throughout the industrial era?
3. Taking an assembly line as an example, what are the factors that determine the nature and content of jobs that result?
4. What is the sociotechnical systems approach to job design?
5. What is the nature of the manager's model of job breadth? of the worker's model of job breadth? Do they result in the same kinds of job designs?
6. What is the technological view of process planning and job design?
7. What are the tools of product analysis? How are job designs affected by their use?
8. In the technologist's view, what are the human functions in performing work? In which kinds of functions does the human have superiority over machines, and vice versa?
9. What is the human role in the models for manual, semiautomatic, and automatic systems?
10. What is the sociotechnologist's view of process planning and job design?
11. What are Englestad's psychological job requirements? Are they compatible with the technological view of job design?
12. Given the principles of division of labor and counterbalancing concepts

of job enlargement, is there any way of deriving some middle ground concept representing optimal job design?

13. Summarize the results of the IBM experience with job enlargement.

14. What is a semiautonomous work group? What evidence supports the superiority of this approach? the inferiority?

15. Summarize the results of the Hovey and Beard Company's experience with semiautonomous work groups. In what kinds of situations do you think that this approach is valid?

16. Summarize the results of the studies of coal mining in England. In what kinds of situations do you feel that the approach taken there could not be transfered with the expectation of similar results?

Problems

1. Figure 14 shows a pipe valve with its various parts, together with an assembled view.* Construct an assembly chart for this product.

2. Construct an operation process chart for the pipe valve based on the Assembly-Fabrication information in Table 3 concerning processing requirements. The machines are available for use on the parts indicated, as well as for other parts.

3. Develop a summary of equipment requirements for the production of 200 pipe valves per hour. How are the equipment requirements affected by an increase in production rate to 350 valves per hour?

---

*The pipe valve problems are reproduced from A. L. Roberts, *Production Management Workbook* (New York: John Wiley & Sons, 1962), p. 22.

TABLE 3
DATA FOR PROBLEM 2

Assembly Operations	Operations	Machine	Output in Parts Per Hour
	1. Final assembly	Bench	30
	2. Clean	Solvent tank	100
	3. Inspect (pressure test)	Water test stand	50
	4. Pack in boxes	Bench	500
Fabrication Operations			
Body, Part #1			
Cast bronze	1. Cast	Bench mold	40
	2. Clean	Tumble barrel	150
	3. Machine-thread and face three surfaces	Turret lathe	25
Bushing, Part #2			
	1. Cast	Bench mold	80
	2. Clean	Tumble barrel	300
	3. Machine all inside and outside diameters	Turret lathe	60
Stem, Part #3			
⅜″ bar stock	1. Machine all surfaces and cut off	Automatic screw machine	200
Cap, Part #5			
¾″ hex bar stock	1. Machine all surfaces	Automatic screw machine	500
Handle, Part #6			
Cast bronze	1. Cast	Bench mold	50
	2. Clean	Tumble barrel	200
	3. Machine two surfaces	Turret lathe	80
	4. Broach square hole	Broach	80

Note: 100-percent inspection required for all fabricated parts prior to delivery to final assembly.

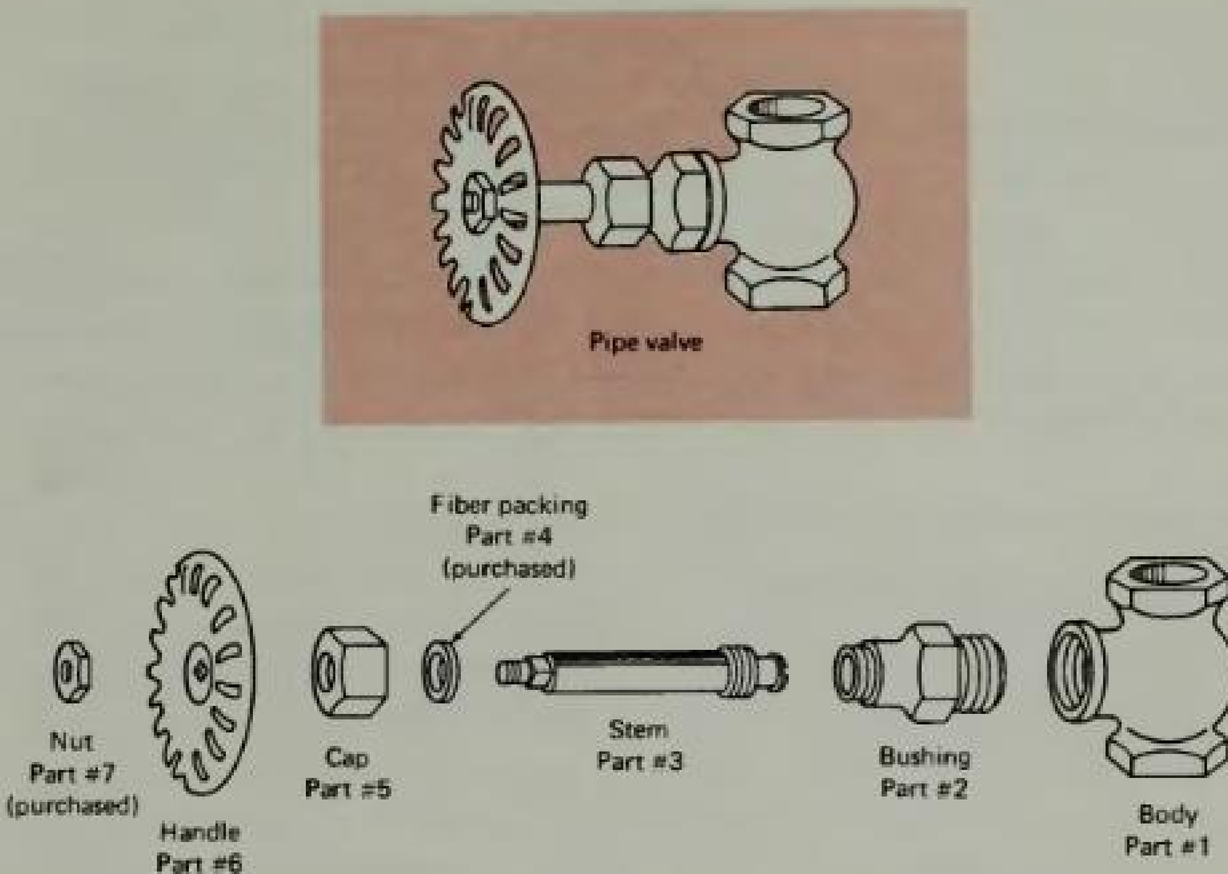


Figure 14. Pipe valve, assembled and exploded views

4. The Laudenbach Beer Company for a number of years has used pasteboard cases to pack six units of six-pack cans for shipment to wholesalers and retailers in its regional marketing area in eleven western states. The total volume of beer shipped from its two plants in Los Angeles and San Francisco was 36 million cans (1 million cases) last year, and this volume is expected to be maintained and probably increased by 15 percent over the next five years.

The Laudenbach Company always has purchased its pasteboard cases in "knocked-down" form, inprinted with the company name and other information. At each of the two plants, the company has box folding equipment in its production lines which ~~set up the box blanks.~~ This requires a gluing operation for the bottom flaps as well as the forming of the case, and is done automatically except for: maintaining a supply of the case blanks in the machine supply hopper; maintaining glue supply; and general surveillance of the machine and its operation. per minute for

A new machine has come to the attention of the company's production analyst; it will perform the case making operation, feeding case blanks to the case forming machine. Printing requirements would be handled separately ~~on sheets of paper~~ that would be glued to the case as a part of the case making operation. The existence of the machine has raised the issue

TABLE 4
COST DATA FOR THE LAUDENBACH COMPANY

	Cost per Hundred		
	Purchase Blanks	One Case Machine Los Angeles Plant	One Case Machine in Each Plant
Purchased blanks	$22.39	—	—
Glue	0.70	$ 0.70	$ 0.70
Labor	3.02	3.90	3.90
Freight, blanks to plants	0.95	—	—
Insurance	0.09	0.15	0.15
Maintenance	0.12	0.15	0.15
Pasteboard (raw material)	—	8.20	8.20
Printed labels	—	0.30	0.30
Scrap	—	0.20	0.20
Floor space	0.30	0.40	0.75
Trucking blanks to San Francisco	—	1.50	—
Totals	$24.57	$15.50	$14.35

Other data:	
Installed cost of case machines	$40,000 each
Expected economic life	8 years
Tax depreciation term (straight-line)	10 years
Incremental tax rate	50 percent
Interest rate	15 percent
Overhead factor at current output rates (25 percent variable)	$10/1000
Cases required per year	1,000,000

of whether the case blanks should be purchased or made in-plant, as well as some side issues regarding the company's general make-buy policies. The new machines cost $40,000 each, installed, and have a capacity considerably in excess of the needs at each of the two plants. Approximatately twelve hours production per day from the case machine will satisfy two-shift requirements in the San Francisco plant. The alternatives being considered are: the continuance of the current practice of purchasing case blanks: the installation of one case making machine at the Los Angeles plant, with trucking of the blanks to the San Francisco plant; and the installation of a case making machine in each of the two plants. In the Los Angeles location, space for the new machine can be found by a minor relayout of the packaging area. This relayout is estimated to cost $500 for the relocation of one machine and rerouting of a small section of conveyor. Suitable space exists in the San Francisco location. Cost data are as shown in Table 4.

The Laudenbach Company has a reputation in the community as a steady employer, and with its vendors as a stable customer. Suppliers commonly quote a better price to Laudenbach than they would to customers whose future business might be uncertain because of business fluctuations. On being presented with the production analyst's report, Laudenback executives were hesitant, fearing a breakdown in relations with suppliers because rapid obsolescence of the case making machine was anticipated. A review of the history of such machines showed that the designs had changed radically in the last few years, although there was no indication that the present machine design was not reliable in performance.

Quick obsolescence

A final issue, discussed at length by the company officials, centered on the nature of the new operation. The Laudenbach Company has been exclusively a beer company, not a paper company. "Isn't it smarter to stay out of case making and stick to beer making? We know beer making, and our reputation rests on that fact. If we go into case making, perhaps we should also consider integration back one more step and make the pasteboard too. If we are going to consider backward integration, there probably are other possibilities more closely allied to beer making that we should consider. Let the case makers have this business of ours. They are specialists and so are we."

In the light of all factors, what analysis would you make and what action do you recommend?

5. A young couple has started a launderette. Their business, The Blue Monday Launderette, offers the following services:

a. Washers for use by customers.
b. Drying service for customers' washes.
c. Dry cleaning and finished laundry.
d. Washing and fluff drying.

The husband performs the "fluff dry" work (washing, extracting, and fluff drying). The wife handles all other matters, such as dealing with customers and aiding people who do their own washing. She also folds and bundles the "fluff-dry" work during slack periods in the day.

The "fluff-dry" business has been unusually heavy, and the couple has been forced to work evenings to get work out on time. It is their contention that five washers, if kept busy, can do the "fluff dry" work. But there are so many other things to be done, such as semidrying the wet clothes in the extractor and loading and unloading the dryer, that one person simply cannot keep the washers loaded. The wife does not have time to assist the husband during normal working hours.

The following times are required to perform the various tasks:

	Time (minutes)
Washer (five available for "fluff dry")	
Load soiled clothes and soap, set water temperature, start machine	2
Running-time (automatically stops)	40
Unload wet clothes into cart	2
Extractor (one available for "fluff dry"; each holds only one washer load)	
Load wet clothes, start machine	2
Running time (automatically stops)	5
Unload semidry clothes into cart	2
Dryer (two available for "fluff dry"; each holds only one washer load)	
Untangle and load clothes, start machine	3
Running time (automatically stops)	20
Unload dry clothes into cart	2
Miscellaneous	
Travel times between equipment	Negligible

a. With the data given, construct an activity chart for the best method of coordinating the work of one man, five washers, one extractor, and two dryers.

b. What is the overall cycle time (time difference between identical points in the process) such as loading washer No. 1 on consecutive loads?

6. The Precision Aerospace Manufacturing Company (PAMCo) is considering the installation of its first numerically controlled machine tool. The machine being considered is a highly flexible numerically controlled (N/C) milling machine capable of handling a wide variety of machining tasks. The machine could be purchased for $150,000. Since the machine would be PAMCo's first N/C machine, the manufacturing engineer held that it was justified as *manufacturing research* and need not meet the same rigid justification requirements normally established for conventional equipment. This view found small support when presented to the manufacturing manager.

The frustrated engineer, now determined to make his point, decided to dig deeply to show the far reaching effects that numerical control methods

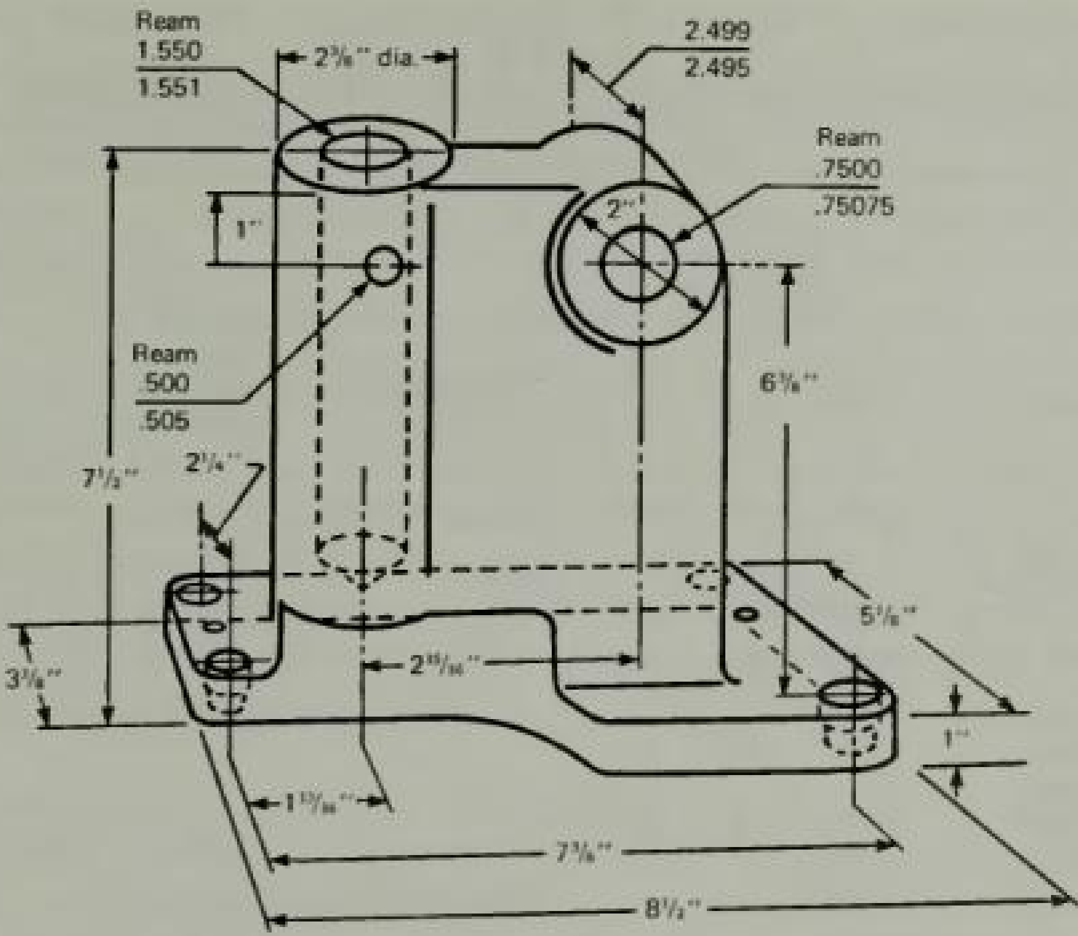


Figure 15. Casting to be machined by N/C methods at PAMCo

could have on manufacturing costs. He felt that the direct cost reductions were relatively obvious, but that the indirect cost reductions were not properly understood and were perhaps more important than the direct cost reduction for N/C. To make the contrast a specific one, the engineer selected one particular part which he felt was typical (see Figure 15) and based all calculations on that part.

Direct Cost Reductions

The engineer obtained actual data for the manufacture of the part by conventional means and supplemented these data with estimates for part programming and the expected production cycle for the N/C machine. Table 5 is a summary of the time requirements for the two contrasting technologies. The engineer then projected his figures to an annual basis by assuming two-shift operation (4,000 hours per year) for the N/C machine and the equivalent worker-machine hours required by conventional methods, $\frac{0.895}{0.405} \times (4{,}000) = 8{,}840$ hours. The reduction in work hours by use of the N/C machine was 4,850 hours per year. He then took the average hourly

TABLE 5
ESTIMATED TIME REQUIREMENTS, CONVENTIONAL VERSUS N/C MACHINING

Conventional Machining				*N/C Machining*			
	Hours per 20-Part Lot				*Hours per 20-Part Lot*		
Operation	*Setup*	*Run*	*Total*	*Operation*	*Part Programming and Setup*	*Run*	*Total*
Mill boss	0.70	0.80	1.50	All operations	1.70	6.40	8.10
Drill, Team, and center bore boss hole	1.70	4.20	5.90				
Straddle mill 2 bosses	0.90	0.75	1.65				
Drill, hole, and ream 1½ hole	1.20	4.00	5.20				
Drill, bore, and ream ½" hole	1.05	2.60	3.65				
Totals	5.55	12.35	17.90	Totals	1.70	6.40	8.10
Total time per part, hours: 0.895				Total time per part, hours: 0.405			

machine shop wage of $5.50 per hour. To this he added 30 percent for fringe benefits to cover retirement, vacations and holidays, social security, insurance, and so on. In addition, the factory overhead was computed at the rate of 25 percent of direct labor cost, so the resulting effective hourly labor cost was $8.52 per hour. The computed direct labor cost reduction was 4,850 × $8.52 = $41,332 per year.

Indirect Cost Reductions

Some of the indirect effects the engineer found were in the category of tooling costs, inventories, intraplant material handling, and inspection.

Tooling Costs. On investigation, the engineer found two kinds of tooling cost reductions. The first was a one-time savings that would occur initially as new jobs were tooled up for the N/C machine to the point where the machine was operating at full load. He felt that these were not annual savings in tooling costs but a reduction in investment required; therefore, this saving should be deducted from the investment required for the N/C machine. The average tooling cost per job by conventional machining had been \$1300; \$600 was projected for the average N/C job. The engineer figured that fifty new jobs would load the new machine, so an additional investment in tools of \$35,000 was prevented, in his view. The net installed cost of the N/C machine was therefore \$115,000.

In addition, he figured that there was an annual tool cost reduction that stemmed from two sources: the annual turnover of jobs for the machine, each new job requiring the smaller tool cost typical of N/C methods; and the annual reduction in the costs of storing, transporting, and repairing conventional tools. He found from company records that about 25 percent of the jobs turned over each year, which would result in an annual tool cost reduction of \$8,750. On investigating the tool repair and carrying costs, the engineer found them to be 20 percent of the value of tooling. He felt that legitimate cost reductions to include were the inventory carrying and repair costs on the one time tool cost saving of \$35,000, or an additional annual cost reduction of \$7,000.

Inventory Cost-Reductions. The engineer reasoned that the combination of many operations into one, as indicated in Table 5 eliminates the dead-time between operations for handling and storage. This results in a shorter overall time that the job spends in process. In addition, the typically smaller setup time (cost) leads to shorter, more frequent runs, that result in smaller finished goods inventories.

Another factor leading to smaller finished goods inventories was the short lead time for N/C processes that would make possible a smaller safety or "buffer stock requirement. Finally, typical inventory losses due to obsolescence should result from the shorter turnaround or lead times. Because of the short lead times, parts formerly made to stock could be made to order, thus reducing substantially the risk of obsolescence.

On checking with the production control department, the engineer found that a typical job was in process for three months. Experience in plants where N/C has been tried seemed to average two months. In addition, the value of in-process inventory would be smaller because of the smaller direct costs indicated in Table 5. Thus, the amount of in-process inventory would be 33 percent less and the value of an item in inventory would be 24 percent less, assuming that valuation for inventory purposes would be average material cost of \$5 per piece plus value added by manufacture. Thus, the total value of in-process inventory should be reduced by 51 percent.

Calculating present equivalent in-process inventory levels, the engineer came up with a figure of \$11,800, based on the average number of parts in process and their average value. The reduction of in-process inventory was therefore $0.51 \times 11{,}800 = \$6{,}018$. Previous studies had shown that the cost of carrying in-process inventory was 25 percent, so that cost reduction was therefore $1,505. The engineer did not attempt to appraise the cost reductions due to effects on finished goods inventories and risk of obsolescence.

Intraplant Material Handling Cost Reductions. The combining of operations due to N/C methods completely eliminates the usual material handling activities normally required between operations. Previous studies in the Industrial Engineering Department indicated that the average material handling trip cost $0.75. If we assume the part under study to be the typical one, five operations are combined into one. There will be about 500 lots per year, so $4 \times 500 = 2{,}000$ trips per year saved. The annual cost-reduction that results is $0.75 \times 2{,}000 = \$1{,}500$.

For some of the parts manufactured, there was the possibility of considerable damage in transit; in some instances, special handling and packaging were used to avoid damage. A reduction in handling would, of course, also reflect itself in a reduction in these costs, but no records existed to help estimate the magnitude of the saving. Besides, the engineer felt that he must be getting close to the needed total cost reduction to justify the new N/C machine. Before stopping, however, he investigated one more possibly fruitful area: inspection.

Inspection Cost Reduction. The engineer reasoned that fewer inspections would be required if operations were combined, that less frequent inspection would be necessary, and that necessary inspection often would be less detailed. Since less human error is involved in the operation of N/C machines, the inspector is, in effect, checking the correctness of the part program by carefully checking the first part produced. The result is that good control often can be maintained by first piece inspection only. Since, in programming a part, all dimensions are in reference to some point, the inspection of the spacing of a series of holes may be accomplished by inspecting the first and last holes only. If these two holes are correctly positioned, those in between will be too.

Not being sure how to estimate these inspection savings, the engineer surveyed the N/C literature and found that a rule of thumb placed N/C inspection costs at one-half those for conventional machining. He estimated that approximately $3,500 per year was being spent for inspection

TABLE 6
SUMMARY OF DIRECT AND INDIRECT COST REDUCTIONS FOR AN N/C MILLING MACHINE COSTING $150,000 INSTALLED

Direct cost reduction:	$41,332
Labor	
Indirect cost reduction:	
Tool cost (annual turnover of jobs)	8,750
Carrying charges on one-time tool cost-reduction	7,000
Inventory	1,505
Material handling	1,500
Inspection	1,750
Total annual cost reduction	$61,827
One-time reduction in tooling investment	$35,000

labor on a volume of parts that was approximately equivalent to what the N/C machine could handle. He therefore set the cost reduction due to inspection at $1,750.

Summary of Cost Figures

On completing his survey the engineer summarized his cost study, as shown in Table 6. He completed a new report summarizing the advantages of N/C machines, together with his recommendation that the machine be purchased immediately. Then he presented it to the manufacturing manager.

a. Evaluate the engineer's cost study.
b. Should the manufacturing manager approve the recommendation?

7. *Job design field study.* Find some set of activities on your college or university campus, or in the community, that may be regarded as some reasonable whole, producing some kind of definable output. Examples might be the Registrar's Office, the Financial Aids Office, the Department or School Office; or in the community, a food service (fast or otherwise), supermarket, bank, post office, or a small manufacturing plant.
   a. Analyze what currently is being done. What are the inputs and outputs, what are the jobs, how do they interact with each other, etc.? Make flow charts and layout sketches to summarize flow and data.
   b. How do the processes, tools, equipment, and layout contribute to the definition of job content; that is, actually dictate the scope of jobs?

c. Generate alternative ways that the work to be accomplished could be organized, divided up among the employees, etc. Let your imagination run freely. If the work is currently finely divided, see if the scope of jobs logically could be enlarged. If the jobs are currently very broad in scope, think in terms of specialization.

d. Evaluate the alternatives and present reasons why your recommended design has important advantages and should be adopted. What are the disadvantages of your recommended design?

References

Barnes, R. M. *Motion and Time Study: Design and Measurement of Work* 6th ed. New York: John Wiley & Sons, 1968.

Beer, M., and E. F. Huse. "Systems Approach to Organization Development." *Applied Behavioral Science* 8(January 1972): 79–109.

Buffa, E. S. "Pacing Effects in Production Lines." *Journal of Industrial Engineering* 12(1961).

———, "Toward a Unified Concept of Job Design." *Journal of Industrial Engineering* 11(1960): 346–351.

Cartwright, D., and A. Zander, eds. *Group Dynamics: Research and Theory.* Evanston, Ill.: Row, Peterson, 1960.

Capanis, A. *Man-Machine Engineering.* Belmont, Ca.: Wadsworth Publishing Company, 1965.

Conant, E. H., and M. D. Kilbridge. "An Interdisciplinary Analysis of Job Enlargement: Technology, Costs and Behavioral Implications." *Industrial and Labor Relations Review* 18(October 1965): 377.

Conrad, R. "Setting the Pace." London: Medical Research Council, APU 232–55, Applied Psychology Research Unit, 1955.

———, and D. A. Hille. "Comparison of Paced and Unpaced Performance at a Packing Task." *Occupational Psychology* 24(1955): 15–28.

Davis, L. E. "The Coming Crisis for Production Management." *International Journal of Production Research* 9(1971): 65–82.*

———. "Job Satisfaction Research: The Post-Industrial View." *Industrial Relations* 10(1971): 176–193.*

———, and R. R. Canter. "Job Design Research." *The Journal of Industrial Engineering* 12(November–December 1956).

———, and A. B. Cherns, eds. *Quality of Working Life: Problems Prospects, and State of the Art*. Vols. 1 and 2. Glencoe, Ill.: The Free Press, 1975. Particularly Vol. 1, Chapters 1 and 12.

———, and J. C. Taylor, eds. *Design of Jobs*. Middlesex, England: Penguin Books, 1972.

———, and E. L. Trist. "Improving the Quality of Work Life: Experience of the Socio-Technical Approach." In *Work in America*. Cambridge, Mass.: MIT Press, 1972.

———, and R. Werling. "Job Design Factors." *Occupational Psychology* 24 (1960): 109–132.

Elliot, J. D. "Increasing Office Productivity Through Job Enlargement." *The Human Side of Office Manager's Job*. Office Management Series no. 134. New York: American Management Association, 1953, pp. 5–15.

Emery, F. E., and E. L. Trist, "Socio-Technical Systems." In *Management Science, Models and Techniques*, C. W. Churchman and M. Verhulst, eds. London: Pergamon Press, 1960.

Englestad, P. H. "Socio-Technical Approach to Problems of Process Control." In L. E. Davis and J. C. Taylor, eds. *Design of Jobs*. Middlesex, England: Penguin Books, 1972.

*Asterisk indicates that this article *also* can be found in. Davis, L. E., and J. C. Taylor, eds. *Design of Jobs*. Middlesex, England: Penguin Books, 1972.

Fogel, L. J. *Biotechnology: Concepts and Applications.* Englewood Cliffs, N.J.: Prentice-Hall 1963.

Guest, R. H. "Job Enlargement: A Revolution in Job Design." *Personnel Administration* 20(March–April 1957): 13–15.

Hackman, J. R., and E. E. Lawler III. "Employee Reaction to Job Characteristics." *Journal of Applied Psychology* 55(1971): 265–286.

Herzberg, F. "One More Time: How Do You Motivate Employees?" *Harvard Business Review*, January–February 1968.

Hunt, G. C. "Sequential Arrays of Waiting Lines." *Operations Research* 4(1956): 674.

Marks, A. R. N. "An Investigation of Modifications of Job Design in an Industrial Situation and their effects on Measures of Economic Productivity." Ph.D. dissertation, University of California, Berkeley, November 1954.

McCormick, E. J. *Human Factors Engineering.* 3d ed. New York: McGraw-Hill, 1970.

Millen, R. A. "*A Generalized Framework for the Evaluation of Work Systems.*" Ph.D. Diss., University of California, Los Angeles, 1976.

Nadler, G. *Work Design.* Rev. ed. Homewood, Ill.: Richard D. Irwin, 1970.

Niebel, B. W., and A. B. Draper. *Product Design and Process Engineering.* New York: McGraw-Hill, 1974.

Rice, A. K. "Productivity and Social Organization in an Indian Weaving Shed." *Human Relations* 6(November 1953): 297.

Schoen, D. R. "Managing Technological Innovation." *Harvard Business Review*, May–June 1969.

Scoville, J. G. "A Theory of Jobs and Training." *Industrial Relations* 9(1969): 36–53.

Susman, G. I. "The Impact of Automation on Work Group Atuonomy." *Human Relations* 23(1970): 567–577.*

Taylor, J. C. "Some Effects of Technology in Organizational Change." *Human Relations* 24(1971):105–123.*

Trist, E. L., and K. W. Bamforth. "Some Social and Psychological Consequences of the Longwall Method of Coal Mining." *Human Relations* 4(February 1951): 3–38.

———; G. W. Higgin; H. Murray; and A. B. Pollack. *Organizational Choice*. London: Tavistock Publications, 1963.

Turner, A. N., and P. R. Lawrence. *Industrial Jobs and the Worker*. Boston: Harvard Business School, 1965.

Walker, C. R. "The Problem of the Repetitive Job." *Harvard Business Review* 28(May 1950): 54–58.

———, and R. H. Guest. *The Man on the Assembly Line*. Cambridge, Mass.: Harvard University Press, 1952.

Woodson, W. E. *Human Engineering Guide for Equipment Designers*. 2nd ed., Berkeley: University of California Press, 1966.

Woodward, J. *Industrial Organization: Theory and Practice*. London: Oxford University Press, 1965.

Chapter 8

Layout of Physical Facilities

Layout is the integrating phase of the design of a productive system. The basic objective of layout is to develop a system that meets requirements of capacity and quality in the most economical way. Here, the specifications of what to produce, how it is to be produced, and what the quantity should be (forecasts, orders, or contracts) become the basis for developing an integrated system. This integrated system must provide for: machines, work places, and storage in the capacities required so that feasible schedules can be determined for the various parts, products, and services; a transportation or communication system; and auxiliary services for production and for personnel (such as medical facilities and cafeterias).

Because of the dynamic character of our economy, the design of this integrated production machine must retain an appropriate degree of flexibility to provide for future changes in: the design of products and services; volume of activity and mixes; and for advancing production technology. Both the site and the building should make it possible to expand operations in a way that dovetails with existing operations. Certain financial and physical restrictions are a normal part of the layout problem. The physical restrictions may be due to the site (its size, shape, and orientation in relation to roads, railroads, and utilities) or to local ordinances or state laws that specify building restrictions and safety codes. In the redesign or relayout of facilities, the existing building imposes severe restrictions.

These general statements about the layout problem indicate something of its complexity. Almost all the factors that enter the problem tend to interact. For example, providing flexibility affects the nature of processes and capacities, which, in turn, interact with short- and long-run costs. Material transportation methods affect not only transportation costs but also the amount of handling at machines and work places. The physical arrangement and relative location of work centers are important in determining transportation costs and direct labor costs. Storage locations and capacities interact with transportation costs and delay times.

We could continue describing the interdependent nature of the factors involved in the layout problem, but our point is that optimizing the design of such a system is largely an art. The value of principles, rules, and guides in determining suboptimal solutions for components of the large problem have been demonstrated. For example, queuing theory can provide a basis for optimal design capacities of service centers such as tool cribs, maintenance crews, supermarket check-out counters, bank tellers, etc. Human engineering data can help optimize work place design; operation sequence analysis can help determine the best relative work center locations (discussed later in this chapter). And principles of production economics can help us select the most economical processes, designs, handling systems, etc.

However, there is no overall general theory that enables us to relate the multitude of influencing factors in a composite optimal design. Rather, the development of a good layout is the result of a sequence of major decisions on such questions as location, design capacity, and general modes of processing. These decisions are followed by a multitude of less significant, but important, decisions related to the selection and placement of equipment, allocation of space, basic flow patterns, etc.

Several components of system design have been discussed separately, both because we have special knkowledge about them and because they are important enough to be treated distinctly. Processes, job and method design, production design of the product or service, process planning, and facility location are examples of such components. These, however, tend to blend with the overall activity of layout during the actual development of a physical facility design.

THE CAPACITY DECISION

In contemplating a new physical facility design or the redesign or expansion of an existing system, a high-level decision on the design capacity is required. It is not enough simply to look at annual activity of various products and services because the activity or sales pattern may reveal seasonal fluctuations. Do we design for the peak or for some average level? If we choose to follow the demand curve, we shall minimize inventory risks. but we shall have a fluctuating labor force and our facility will have idle capacity, except during peak activity. If we adopt some intermediate level, we shall tend to stabilize employment levels and utilize facilities better, but we will have to build inventories to meet the sales peak. Which plan will minimize the combined cost of inventories, investment costs, and labor turnover costs? This is a highly significant problem in economic analysis. It can be approached as a programming problem in which production capacity is allocated to operating periods, minimizing the combined plant investment costs, inventory costs, and turnover cost. Investment costs would be approximated by a nonlinear function of capacity.

Future Capacity

Then there is the question of forecasts. Do we build for a capacity that matches our present experience, or do we attempt to build for some forecast level of one, five, or ten years hence? Can we afford to build for more capacity than we need now? We must remember that, because capacity is bought in "chunks," successive units of capacity are not equally expensive. Moreover, it commonly is true that, at any one level there will be idle capacity in certain equipment classifications. Therefore, moving to the next level does not entail the purchase of equipment items where we already have idle capacity. When capacity is built to match some forecast of future needs, it is common to buy equipment for current needs only, and to provide space in the building and layout for the additional equipment when it is needed. Thus, planning is for a future capacity, but extra overhead is carried for the building space only. As the capacity is needed, machines can be integrated into the system without need for relayout.

The question really relates to the provision for extra space to match a future forecast. Some additional space probably can be justified when we consider that space added later is more expensive per square foot because existing walls must be removed, doors cut through to the new space, etc. ~~Relayout costs which would be required to integrate the new space into the production system are more important than the cost of the additional space itself.~~ If the new space is not integrated into the system but becomes a "thumb" stuck onto the existing layout, we would pay extra costs daily in the form of higher material handling costs. These extra costs of adding future space must be balanced against the incremental costs of building more space now and carrying it as added overhead until it is needed.

Subcontracting and Multiple Shift Effects on Capacity

Another aspect of the capacity problem is the question of how to meet capacity needs. The investment-capacity ratio can be altered by the amount of subcontracting and by the intensity with which we use facilities; that is, one, two, or three shifts. The economics of make-versus-buy are especially significant when new capital equipment is involved, as they would be with the design of a new facility, for now we have no idle capacity behind which "make" can hide. Rather, we are setting the future pattern, so overall make-buy policies need to be reviewed.

Determining the number of shifts appropriate for a given organization is not a simple question. If we use two shifts instead of one, investment costs are not halved because, as we noted before, increments of capacity are not equally expensive. Many other costs also are involved. Wage premiums of 10 to 15 percent are common for second shifts, and many people question whether productivity and scrap ratios are as good on multiple shifts. Also, multiple shifts ordinarily increase supervision costs. There is no single answer to the question of shift operation, since the relative importance of building and equipment investment costs and labor costs varies from industry to industry. Economic analysis is required for each situation. In general, industries with very heavy investments in buildings and equipment per worker, such as steel, chemical, and oil refining industries, find multiple shifts more economical; those with moderate and low investments per worker find that wage premiums more than counterbalance the investment savings of multiple shifts.

Translating Capacity into Workable Units

What is the meaning of the term *capacity*? In the steel industry one thinks in terms of tons of steel per day, week, or month; an automobile manufacturer in terms of cars; a hospital in terms of patient-days; a post office in terms of tons of mail per day by class of mail; etc. How about a job shop machining company? Its products are so different that capacity in terms of the output of finished products is quite meaningless. Here, capacity must be expressed in more universal terms. The units commonly used are *available hours* in various machine classifications per day, week, or month. This output potential is a good general measure of capacity for plant design because we can convert fairly easily to a physical-capacity equivalent—the number of *machines* required. The number of machines required in various classifications is the data that we finally want to work with in developing a layout. We need to translate everything into physical units of capacity. In doing this, we must be careful to make allowances for two factors which reduce the utilization of equipment: the facility efficiency factor and the scrap factor.

Through the *facility efficiency factor* we recognize that, because of scheduling delays, machine breakdowns, preventive maintenance, etc., a portion of the available hours cannot be used. Facility efficiency factors vary according to the type of equipment and the company; they range generally from 0.50 to 0.95. Thus, if 100 outboard motors per week translated into the need for 550 milling machine hours per week we would need the equivalent of $550/0.80 = 688$ hours if our efficiency factor was 0.80, since we would expect about 138 machine-hours to be unavailable to us.

Through the *scrap factor* we recognize that, for any real production process, we shall produce some bad parts or products. When we decided to design a plant to build 100 outboard motors per week, obviously we were thinking of good motors that were free of defective parts. But some of our capacity will be used up producing scrap, and so we must allow for this event too. If we expected 3 percent scrap in our milling operations, 688 hours of available machine-hours must be increased to $688/0.97 = 709$ hours. Now, if we expect to work 75 hours per week on two shifts, we need $709/75 = 9.45$ machines. As we noted, physical capacity comes in "chunks," so we must provide for either 9 or 10 milling machines. If we decide on 10 machines, we should expect to have some idle capacity. If we decide to squeeze by with 9 machines, we should expect some bottlenecks now and then, which we may try to make up for by overtime work. Thus, a capacity of 100 outboard motors translates into an equivalent milling machine capacity of 9 or 10 machines.

In summary, the capacity question involves important decisions that will determine total investment and future operating costs. These factors are determined by the selection of output levels in relation to seasonal activity levels; the determination of how many shifts are most economical; the decision of what proportion of the total effort to subcontract; and the determination of how much excess or "growth" capacity is economical. The organization's financial strength and ability always present restrictions in determining how close we may approach ideal capacity conditions at a given time.

BASIC LAYOUT TYPES

In general, layouts can be classified either as process oriented or product oriented. In a *process layout*, equipment of the same functional type is grouped together. Thus we would have lathes grouped together, milling machines

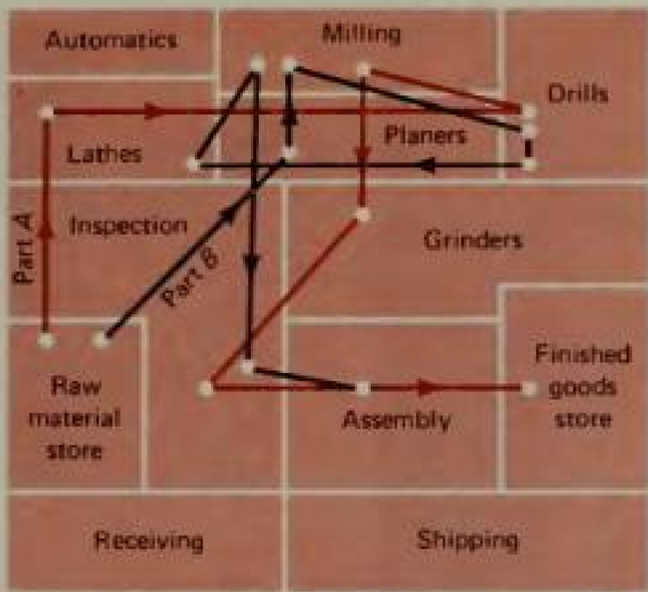


Figure 1. Process or functional layout. Machines are arranged in functional groups. Parts take various routings as dictated by their design requirements. Illustrative routes for two parts, *A* and *B*, are shown. Parts are moved from operation to operation in batches or lots and stored temporarily at each work station to await their turn

grouped together, inspection occurring in one place, all assembly occurring in one place, etc., as shown in Figure 1. Process layout closely follows our model for intermittent productive systems, discussed in Chapter 2.

Figure 2 shows the typical arrangement for product or line layout, which follows our model for continuous systems. The name *product layout* is used because the basic organization of the layout is dictated by the part or product. Equipment is arranged according to the sequence in which it is used for a given part or product, following the processing sequences. If similar equipment is required for both parts *A* and *B*, it normally would be duplicated in the two lines of Figure 2, even though the equipment is not fully utilized for either part. Flow patterns for line layout are fixed by the nature of the layout and the type of material handling equipment commonly used.

As we shall see, these concepts of flow that were developed in manufacturing can apply to service and nonmanufacturing situations as well.

PROCESS LAYOUT FOR INTERMITTENT SYSTEMS

Our discussion of descriptive models of production in Chapter 2 explained that intermittent systems are so classified because the flow is intermittent. Processing units are organized by function on the assumption that the functional skills and expertise are available in that service center or facility. Therefore, an X-ray department of a hospital normally offers such a broad range of skills and knowledge that a physician can call for virtually any kind of X-ray and expect it to be performed. Of course, a given X-ray department may be limited by the lack of certain equipment or by personnel training and skill limitations.

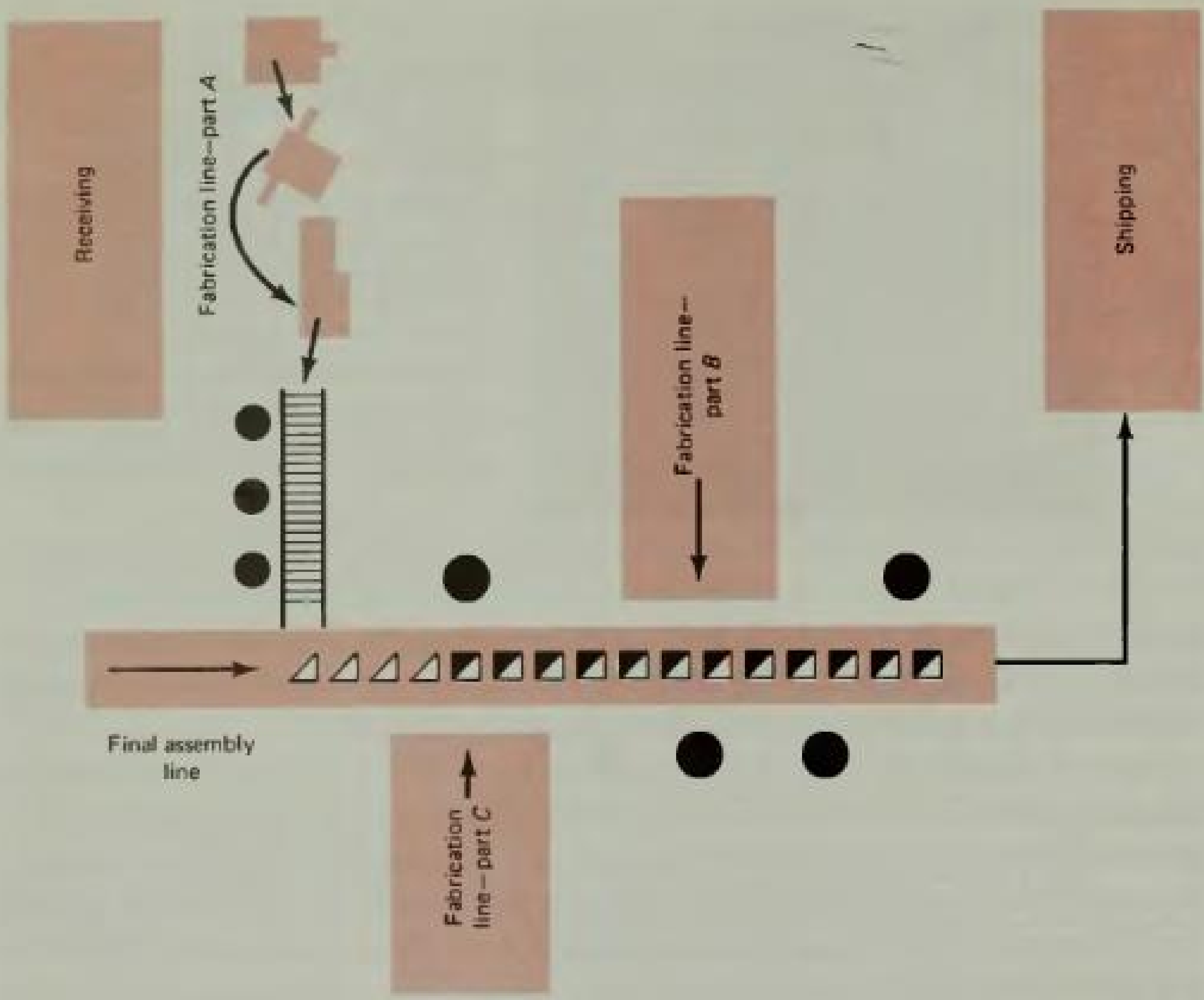


Figure 2. Product or line layout. Machines and equipment are arranged according to the sequence of operations required to fabricate and assemble. Machines and workers are specialized in the performance of specific operations, and parts approach continuous movement

Many examples of intermittent systems can be found in practice; for instance, in manufacturing, hospital and medical clinics, large offices, municipal services, and libraries. In every situation, the work is organized according to the function that is performed. The machine shop industry is one of the most common examples, and the name and much of our knowledge of intermittent systems results from the study of such manufacturing systems. Table 1 gives a summary of typical departments or service centers that occur in several generic types of intermittent systems.

In all the generic types of intermittent systems, the item being processed (part, product, information, person) normally goes through a processing sequence; but the work to be done and the sequence of processing vary, depending on the objective. At each service center, the specification of what is to be accomplished

TABLE 1
TYPICAL DEPARTMENTS OR SERVICE CENTERS
FOR VARIOUS GENERIC TYPES OF INTERMITTENT SYSTEMS

Generic system	*Typical Departments or Service Centers*
Machine shop	Receive, Stores, Drill, Lathe, Mill, Grind, Heat-Treat, Inspection, Assembly, Ship
Hospital	Receiving, Emergency, Wards, Intensive-Care, Maternity, Surgery, Laboratory, X-ray, Administration, Cashier, etc.
Medical clinic	Initial Processing, External Examination, Eye, Ear, Nose and Throat, X-ray and Fluoroscope, Blood Tests, Electrocardiagraph and Electroencephalograph, Laboratory, Dental, Final Processing.
Engineering office	Filing, Blueprint, Product Support, Structural Design, Electrical Design, Hydraulic Design, Production Liaison, Detailing and Checking, Secretarial Pool.
Municipal offices	Police Dept., Jail, Court, Judge's Chambers, License Bureau, Treasurer's Office, Welfare Office, Health Dept., Public Works and Sanitation, Engineer's Office, Recreation Dept., Mayor's Office, Town Council Chambers.

determines the details of processing and the time required. For each service center, we have the general conditions of a waiting line (queuing) system, with variable arrival and processing rates. The detailed structure of the activities within the service center could follow any of the four basic structures of waiting line situations shown in Figure 1 of Appendix D. When we view an intermittent system as a whole, we can visualize it as a network of queues with variable paths or routes through the system, depending on the details of processing requirements. Obviously, an important problem in the design of such systems is the relative location of the service centers or departments.

Decision to Organize for Intermittent Flow

The nature of intermittent systems indicates that we are dealing with situations in which routes through the system must vary depending on requirements, and in which the processing required at each type of operation also may vary.

To obtain reasonable utilization of personnel and equipment in intermittent flow situations, we assemble the skills and machines to perform a given function in one place and route the items being processed to the functional centers. If we tried to specialize according to the processing requirements of each type of order in production line fashion, we would have to duplicate many kinds of skills and equipment. The utilization for each order might be very low unless the volume for that type were very large. Thus, when flexibility is the basic system requirement, the intermittent mode is likely to be the most economical. The flexibility required may be of several types: flexibility of routes through the system; flexibility in the volume of each order; and flexibility in the design or processing requirements of the item being processed.

Other advantages of the intermittent design become apparent when the intermittent concept is compared to the continuous flow or production line concept. The jobs that result from an intermittent organization are likely to be broader in scope and require more job knowledge. One is an expert, say, in some field of work, whether it is heat-treating, medical laboratory work, structural design, or city welfare (as shown in Table 1). Even though the functional mode implies a degree of specialization, it is specialization within a field of activity, and the variety within that field can be considerable. A pride in workmanship has been traditional in this form of organization of work by trades, crafts, and relatively broad specialities. In short, job satisfaction criteria seem relatively easier to meet in these situations than when specialization results in highly repetitive activity.

THE RELATIVE LOCATION OF FACILITIES PROBLEM

Once we decide to organize productive facilities on a functional basis, the central design problem is how to relate the individual functional units to one another. In a machine shop, should the lathe department be located adjacent to the mill department? In a hospital, should the emergency room be located adjacent to intensive care? In an engineering office, should product support be located adjacent to electrical design? In municipal offices, should the welfare and health department offices be adjacent to each other? Although these are general questions for which we do not expect specific answers, the locations will depend on the need for one facility to be adjacent relative to the need for other pairs of facilities to be adjacent. We must allocate locations based on relative gains and losses for alternatives, and seek to minimize some measure of the cost of having facilities nonadjacent.

Criteria

The problem of allocating locations to departments depends on some criterion, or measure, of effective relative locations. While the details of the criterion may vary depending on the specific system, we always are attempting to measure the interdepartmental interactions required by the nature of the system. How much business is carried on between departments, and how do we measure it? In manufacturing systems, material must be handled from department to department; in offices, people normally walk between locations to do business and communicate; and in hospitals, patients must be moved and/or nurses and other personnel must walk from one location to another. Table 2 summarizes criteria for four systems.

By their very nature, intermittent systems have no fixed path, and therefore we must aggregate for all paths and seek a combination of relative locations that minimizes the criterion. While this location combination may be poor for some paths through the system, in the aggregate it will be the best set of locations. Inevitably, we either are trying to minimize (1) system interdepartmental transactions costs, or (2) the walking time of patients, customers, or clients. Some approaches to the problem try to approximate these objective criteria by rating the degree to which it is essential that each pair of departments or facilities be located close together [Lee and Moore, 1967; Seehof and Evans, 1967].

TABLE 2
CRITERIA FOR DETERMINING THE RELATIVE LOCATION OF FACILITIES IN INTERMITTENT SYSTEMS

Generic System	*Criterion*
Manufacturing	Interdepartment material handling cost
Hospital	Personnel walking cost between departments
Medical Clinic	Walking time of patients between departments
Offices	Personnel walking cost between areas and equipment, or face-to-face contacts between individuals

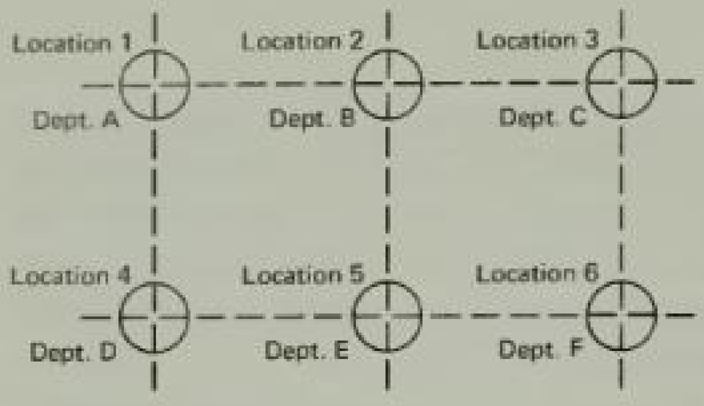


Figure 3. There are $6! = 720$ arrangements of the six process-areas in the six locations of the grid

Complexity of the Relative Location Problem

A relatively simple example will help us discover the nature of the problem that we need to solve. Figure 3 shows, in schematic form, six process areas arranged on a grid. If any of the six departments can be located in any of the six alternate locations, there are $6! = 720$ possible arrangements, of which ninety are different in terms of their effects on the cost of interdepartmental transactions. For the trivial problem of Figure 3, then, we could consider enumerating the different location-combinations, comparing aggregate costs, and selecting the combination with minimum cost, although this task would be very tedious. However, the number of combinations to evaluate goes up rapidly as we increase the number of departments. For just 9 departments on a 3 x 3 grid, we have more than 45,000 combinations. For 20 departments arranged on a 4 x 5 grid, we have 608×10^{15} combinations. Therefore, we must rule out enumeration as a practical approach to the problem.

Operations Sequence Analysis

An early graphical approach to the problem led to a powerful computerized model, which we shall also examine. Operations sequence analysis maintains close contact with the nature of the problem and is useful for relatively small problems.

For example, a private industrial clinic performs services under contract to a number of business and industrial firms. These services include medical examinations for new employees, as well as annual physical examinations for all employees. The clinic performs eight types of examination sequences, depending on the details of the individual contracts. Many people who have been examined have complained about excessive walking because of the clinic layout. The director of the clinic wants to know what a good solution would look like.

Table 3 represents a one-month sample of the flow between the eleven departments, aggregating over all types of examinations. Table 3 is called a load summary, and represents the flow among all combinations of departments.

The graphical approach to finding a solution places the information contained in the load summary of Table 3 in an equivalent schematic diagram, in which circles represent the departments or functional service centers. Connecting lines are labeled to indicate the intensity of travel or transactions between centers. Figure 4 is a first solution, and is obtained merely by placing the work centers on a grid, following the logic of the pattern indicated by Table 3. When all connecting lines have been drawn on the diagram and labeled, we have an initial

TABLE 3

LOAD SUMMARY. SAMPLE OF NUMBER OF TRIPS PER MONTH BETWEEN ALL COMBINATIONS OF DEPARTMENTS FOR AN INDUSTRIAL MEDICAL CLINIC

From / Departments		To: 1 Initial Processing	2 Eye Examination	3 Ear, Nose, and Throat	4 X-ray and fluoroscope	5 Blood Tests	6 Blood Pressure Check	7 Respiratory Check	8 Electrographic	9 Laboratory	10 Dental Examination	11 Final Processing
Initial processing	1		600									
Eye examination	2			400	100			100				
Ear, nose, and throat	3				350	50						
X-ray and fluoroscope	4						100	450				
Blood tests	5							50				
Blood pressure check	6				100					150	100	
Respiratory check	7						50		450	100		
Electrographic	8						200			250		
Laboratory	9										500	
Dental examination	10											600
Final processing	11											

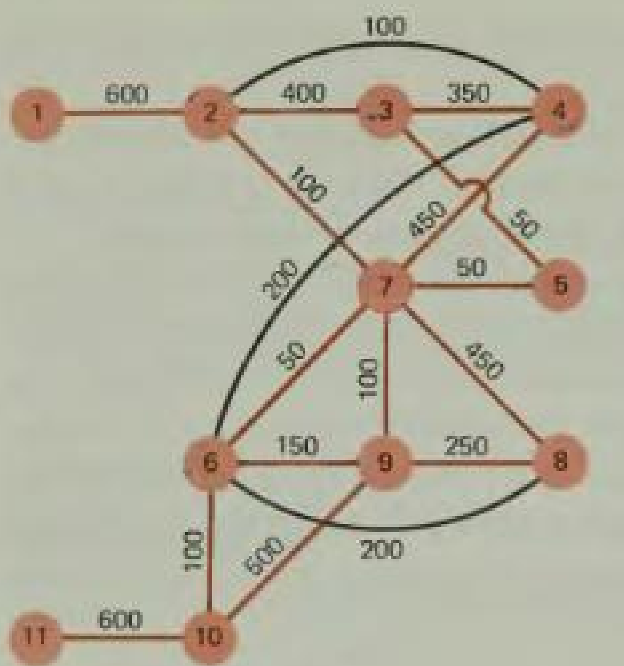


Figure 4. Initial graphic solution developed from load summary of Table 3. It can be seen by inspection that 4 can be moved to the position between 2 and 6 to eliminate 300 non-adjacent loads. The positions of 8 and 9 are improved by replacing 9 by 8 and moving 9 to the position just below 8

SOURCE: E. S. Buffa, "Sequence Analysis for Functional Layouts," *Journal of Industrial Engineering* 6(March–April 1955).

solution that may be improved by inspecting the effect of changes in location. When an advantageous change is found, the diagram is altered. For example, in Figure 4 we see immediately that work center 4 has a total of 300 trips to or from work centers that are not adjacent, i.e., 2 and 6. If 4 is moved to the location between 2 and 6, all loads to and from 4 become adjacent.

Further inspection shows that 200 nonadjacent trips occur between work centers 6 and 8. Is an advantageous shift possible? Yes: by moving 9 down and placing 8 in the position vacated by 9, the number of nonadjacent trips is reduced from 200 to 100. Figure 5 shows the diagram with the changes incorporated.

Further inspection reveals no further obvious advantageous shifts in location, so we adopt Figure 5, which has a $2 \times 100 = 200$ trip-distance rating, as our final

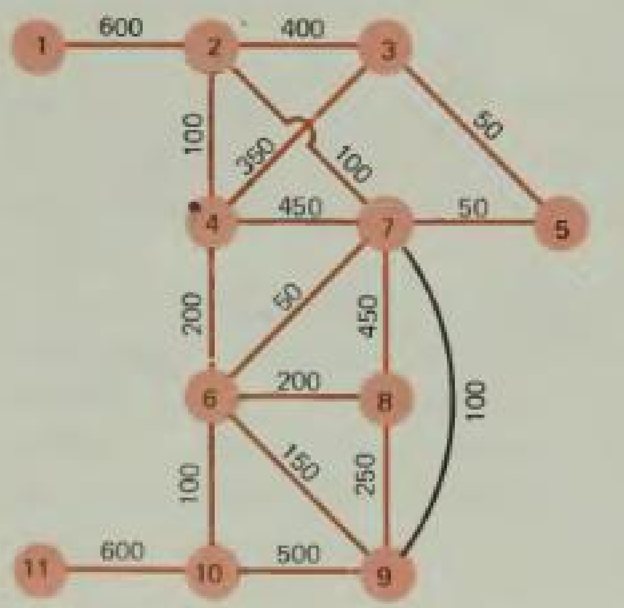


Figure 5. Schematic diagram incorporating changes suggested by Figure 4. Solution is not necessarily optimal but no further location changes seem to yield improvement

SOURCE: E. S. Buffa, "Sequence Analysis for Functional Layouts," *Journal of Industrial Engineering* 6(March–April 1955).

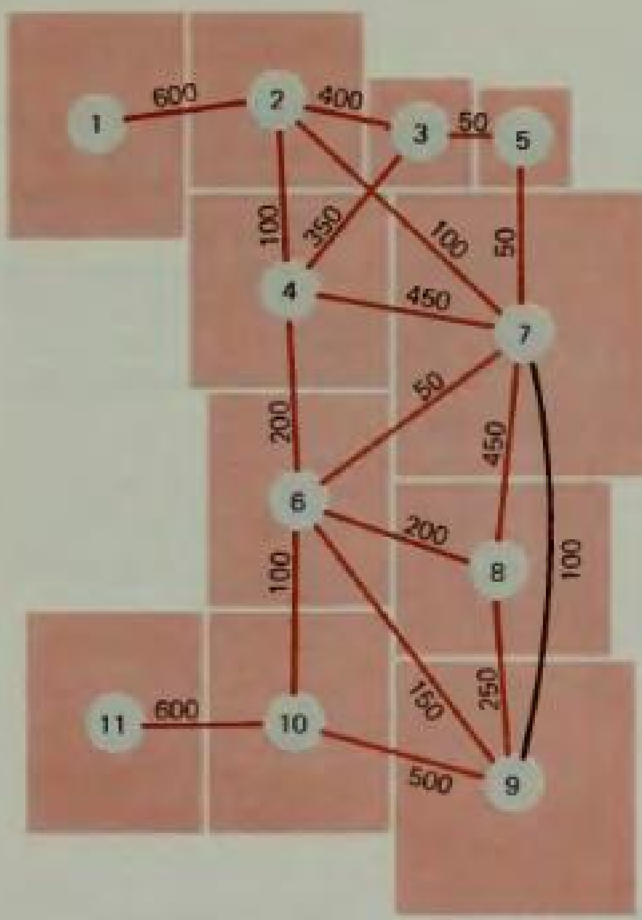


Figure 6. Initial block diagram. Estimated department areas are substituted for circles in the schematic diagram of Figure 5. Block templates are used for estimated area requirements for the various work centers

SOURCE: E. S. Buffa, "Sequence Analysis for Functional Layouts," *Journal of Industrial Engineering* 6(March–April 1955).

schematic diagram. For larger problems, the grid distance becomes an important part of the measure of effectiveness, because work centers might be separated by two, three, or four grid units. Figure 5 is not an optimum solution in the mathematical sense; however, it represents a good solution because most of the work centers are adjacent to the other work centers involving interdepartmental flow.

The Block Diagram. Now that we know how the medical clinic departments should be located in relation to one another, we can use the final schematic diagram as a basis for developing a block diagram in which the physical areas required by the work centers take the same relative locations. Estimates of the areas required by each department are required.

The block diagram itself is developed by substituting estimated areas for the small circles in the schematic diagram. Initially, this can be done with block templates to find an arrangement that is compatible with both the flow pattern of the schematic diagram and the various size requirements for departments. Figure 6 shows such an initial block diagram for our example. We can see that the essential character of the schematic diagram is retained. However, Figure 6 obviously does not yet represent a practical solution. A slight variation of the shapes of departments will enable us to fit the system into a rectangular configuration and meet possible shape and dimension restrictions that may be

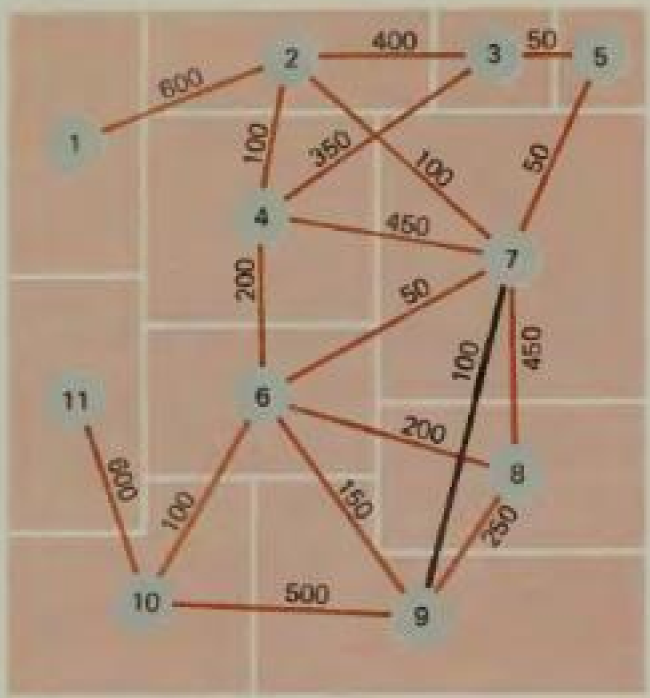


Figure 7. Block diagram which takes account of rectangular building-shape and other possible restrictions of shape and dimension imposed by site, but still retains approximate department area requirements and idealized flow pattern

SOURCE: E. S. Buffa, "Sequence Analysis for Functional Layouts," *Journal of Industrial Engineering* 6(March-April 1955).

imposed by the site or by an existing building. Figure 7 shows such a block diagram.

The final block diagram represented by Figure 7 becomes an input to layout for aisles, equipment, work place design, and other important details. The detailed layout phase would undoubtedly require minor shifts in space allocation and shape; however, the basic relationships among department locations can be retained.

CRAFT (COMPUTERIZED RELATIVE ALLOCATION OF FACILITIES)

The graphical approach to the determination of the relative location of departments has obvious limitations. The technique depends on the individual analyst's insight in order to evolve better schematic diagrams and, as the number of activities centers increases, the technique breaks down rapidly. Practical problems in facility location often involve twenty or more activity centers, and this number is already at the limit for feasible use of the operation sequence analysis technique. To overcome this limitation, a computerized relative allocation of facilities technique (CRAFT) was developed [Armour and Buffa, 1963], which easily handles up to forty activity centers and has other important advantages.

The CRAFT Program

The CRAFT program takes as input data matrixes of interdepartmental flow and material handling cost (or some other measure of the need for work centers to be close to each other), together with a representation of a block layout. The block layout that is fed in may be the existing layout, or any arbitrary starting solution if a new facility is being developed. The program calculates department centers and an estimate of total interaction cost for the input layout. The governing heuristic algorithm then asks: What change in interaction cost would result if the locations of two departments were exchanged? Within the computer, the locations of the two departments or activity centers are exchanged, and the interaction costs are recomputed. Whether the result is an increase or a decrease, the difference is recorded in the computer's memory. The program then asks the same question for other combinations of two departments, and again records cost differences. It proceeds in this way through all combinations of exchanges.

When the cost differences for all such combinations have been computed, the program selects the exchange that would result in the largest reduction, makes the exchange in locations on the block diagram, and then prints out the new block layout, the new total interaction cost, the cost reduction just effected, and the departments involved in the exchange. The basic procedure then is repeated, generating a second, third, etc., improved block layout. Finally, when the procedure indicates that no further location exchanges that reduce interaction cost can be made, the final block layout is printed out. This becomes the basis for a detailed template layout of the facility.

Typical computer output for the block layout is shown in Figure 8. Departmental areas are identified by letter code, and each character represents a given number of square feet.

As presently constructed, the program has the capacity for handling forty departments or activity centers. Any departmental location can be fixed down simply by specifying in the instructions that the department (or departments) is not a candidate for exchange. This feature has great practical importance, because existing layouts cannot be completely rearranged. Fixed locations may develop when costly heavy equipment has been installed, or the location of receiving or shipping facilities may be determined by the location of roads or railroad spurs. Finally, the locations of some work groups often make it desirable to treat them as fixed points in the layout. A new development also makes it possible to include relocation costs as well as interaction costs in the computer program for relayout. [Hicks and Cowan, 1976].

Industrial Application of CRAFT

Computerized facilities layout programs have been most widely applied in industry. However, they have also been used in service and nonmanufacturing applications. To the author's knowledge, the CRAFT program itself has been used in four aircraft plant applications, two of the largest automobile companies, two computer manufacturers operations, a pharmaceutical manufacturer, a meat packer, a precision machine shop, a movie studio, and a hospital. Since the program is freely available and has been circulated widely, many other applications undoubtedly have been made.

A Precision Manufacturer in the Aerospace Industry. A particularly interesting application was made in the machine shop of a precision manufacturer in the aerospace industry [Buffa, Armour, and Vollmann, 1964]. The shop occupied about 42,000 square feet of floor area, and the majority of orders were for small precision parts in low quantities. Because of the small physical size of orders, most material handling among departments was accomplished by the machinists themselves, carrying orders to and from a central holding and dispatch area (Department K) in tote pans. The current layout had grown around the central holding and dispatch department.

In addition to the layout itself, one important question that management wished to evaluate was the validity of having material flow through the central holding area. Although the use of the central holding area offered a small plant the advantage of close control, it was obvious that the physical flow of material to and from the holding area entailed incremental material handling costs. The alternate policy was to dispatch orders directly from department to department, using an information system for control.

Another operating policy that management wished to evaluate was the use of machinists to accomplish the majority of interdepartmental material handling. Originally, when the plant was small, the machinists were used for this purpose because distances were short and it was felt that these short walks gave machinsts a break from usual routines. As time passed, however, the plant was enlarged and management felt that the use of specialized material handlers should be evaluated.

In summary, then, there were four basic conditions for which the determination of the best layout and the material handling costs associated was desirable:

1. Current practice: material flow through Department K. Machinists used for handling.

2. Material flow through Department K. Using material handlers.
3. Dispatch using information system. Machinists used for handling.
4. Dispatch using information system. Using specialized material handlers.

For purposes of analysis, twenty-two plant areas were designated as department centers. To determine interdepartmental flow, it was necessary to analyze approximately 1,600 shop orders (approximately an eight week sample) on which were indicated the routing required to fabricate the parts. Fortunately, these data were available on punched cards; therefore, it was relatively simple to develop a matrix that showed the number of loads flowing among all combinations of departments for the four different conditions. A similar matrix for the four conditions was developed to show the material handling cost in dollars per 100 feet of movement for combinations of departments for which flow occurred.

The CRAFT program then was used to evaluate both layout configuration and the four alternative policies for order dispatching and material handling. Table 4 summarizes the results. The far-right column indicates the percentage reduction in the cost of each alternative compared to the cost of the existing layout and current policies. Alternative 4, which uses an information system for dispatch and control and specialized material handlers, results in a 73 percent

TABLE 4
MATERIAL HANDLING COSTS FOR FIVE LAYOUTS

Alternatives	*Total 8-Week Material Handling Cost*	*Percentage Reduction from Existing Layout*
Existing layout—current policies	$3294.98	–
Cost for best layout under following conditions:		
1. Current practice: Material flow through Department K—machinists used for handling	2645.08	20%
2. Material flow through Department K—using material handlers	2402.97	27%
3. Dispatch utilizing information system—machinists used for handling	1186.89	64%
4. Dispatch utilizing information system—using specialized material handlers	900.93	73%

SOURCE: E. S. Buffa, G. C. Armour, and T. E. Vollmann, "Allocating Facilities With CRAFT," *Harvard Business Review*, March–April 1964.

reduction. When the eight-week figures are placed on an annual basis, the cost reduction potential is approximately $15,500. If the company demands a 10 percent return on plant investments of this kind and imposes a severe three-year payback for capital recovery and return, it could afford to spend over $38,000 for the plant relocations and alterations necessary for relayout.

Improvement in material handling cost is possible through all four policies. Alternate 1 indicates that a 20 percent material handling cost reduction results from a layout improvement with existing policies. An additional 7 percent improvement results from using specialized material handlers. The use of an information system apparently leads to the largest improvement. It is important to note, however, that we are dealing with an integrated system, and that interacting effects may have occurred among layout configuration, material handling method, and dispatch and control system.

Office Layout

The same general concepts about the relative location of facilities apply to the layout of large offices with functional groupings, although the measure of effective relative location may change. A number of applications of the CRAFT program in office situations have been reported.

Vollmann, Nugent, and Zartler [1968] applied a variant of the program to the office layout of an oil company in the southwestern United States. The locations of twenty-seven persons and seven significant pieces of equipment were involved in an office occupying 5,600 square feet. The criterion to be minimized was the aggregate of the product of cost-weighted trips by employees by the distance walked. Thus, the cost of a highly paid executive walking between centers has greater significance than lower paid employees when weighted in the criterion. A large number of alternatives were evaluated, including those involving personal preferences for location of certain individuals (for example, given that a certain individual merits a corner office by virtue of his position in the organization, how should the other individuals be located relative to him?).

Volgyesi [1970], a practicing architect, used the CRAFT program an an input to architectural design of an office. On the basis of surveys of the volume of interdepartmental interactions and an estimate of the cost per unit of these interactions, Volgyesi produced a relative priority for locating departments near one another in terms of a ranking scale. A ranking of "one" indicated no interaction between two departments, and "six" indicated a very high volume of interactions between two departments. These priority ratings formed the basis for the equivalent of the volume matrix. The program resulted in a nearly 16

TABLE 5
FUNCTIONAL DEPARTMENTS AND AREA REQUIREMENTS FOR A BRANCH CITY OFFICE

Department	Area requirements, square feet
A. Building and safety	1500
B. City attorney	900
C. Local councilman's office	1000
D. Contract administrators	1200
E. Engineering bureau—public works	3600
F. Planning department	1000
G. Police	3000
H. Recreation and parks	1000
I. Sanitation	2000
J. Street maintenance	1200
K. Traffic	800
L. General administration office	400
	17,600

percent improvement compared with proposals made, and suggested important alternatives for the location of some areas which had not been considered previously.

Municipal Offices. Some large cities, particularly those that are spread over a large geographic area, have established branch city offices to bring services closer to the people. In one such instance, the architect was considering two basic designs to accommodate the twelve major functional departments listed in Table 5. The two designs were a rectangular configuration, and a central patio around which the functional departments were arranged.

The rectangular design was, of course, cheaper to construct, but the decision makers felt that the central patio design offered significant advantages for the quality of the working environment, as well as an attractive physical design. It provided natural lighting for almost all offices, and gave individuals who walked to and from various offices the opportunity to go through the patio and be outdoors in an attractive setting. An additional question, then, was: How much added interaction cost was involved in the central patio design?

Sample data were gathered on face-to-face contacts initiated between offices, and preliminary architects' layouts of both designs were inputted into the CRAFT program. The program found improved solutions for both basic designs, reducing the interaction cost for the rectangular design by 16 percent and the patio design by 28 percent, compared with preliminary layouts. The best configurations found for the two designs are shown in Figure 8. The interaction cost of the patio design is 15 percent greater than for the rectangular design. The

Depts.

(a)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	G	G	G	G	B	B	A	A	A	A	A	A	L	L	L	L
2	G		G	B	B	B	B	A	A	A	A	E	E	E	F	F
3	G		G	B	B	B	B	A			A	E		E	F	F
4	G		G	G	E	E	E	E	E	E	E	E		E	F	F
5	G	G	G	G	E									E	F	F
6	G	G	C	C	E	E	E	E	E	E	E	E	E	E	F	F
7	G	G	C	C	C	C	J	J	J	I	I	I	K	K	K	K
8	G	G	C	C	C	C	J	J	J	I		I	K	K	K	K
9	G	G	D	D	D	D	J	J	I	I		I	I	H	H	H
10	G	G	D			D	J	J	I			I	I	H		H
11	G	G	D	D	D	D	J	J	I	I	I	I	H	H	H	H

TOTAL COST 13375.20 EST. COST REDUCTION 77.73

MOVEA C MOVEB D MOVEC ITERATION 5

(b)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	A	A	A	A	A	E	E	E	E	E	E	E	E	H	H
2	A				A	E	E	E	E	E	E	E	E	H	H
3	A	A	A	A	A	E	E	F	F	F	F	F	H	H	H
4	E	E	E	E	E	E	E	F	F	F	F	F	H	H	H
5	E			E	M	M	M	M	M	M	M	J	J	K	K
6	E	E		E	M						M	J	J	K	K
7	D	E	E	E	M						M	J	J	K	K
8	D	D	D	D	M						M	J	J	K	K
9	D	D		D	M						M	J	J	I	I
10	B	D	D	D	M						M	J	J	I	I
11	B	B	B	B	M	M	M	M	M	M	M	I	I	I	I
12	B	B	B	B	G	G	G	G	L	L	C	C	I		I
13	G	G	G	G	G			G	L	L	C	C	I		I
14	G							G	G	C	C	C	I		I
15	G	G	G	G	G	G	G	G	G	C	C	C	I	I	I

TOTAL COST 15354.11 EST. COST REDUCTION 458.80

MOVEA L MOVEB G MOVEC ITERATION 6

Figure 8. Alternate configurations for a municipal services branch: (a) Rectangular, and (b) Central patio design. Total interdepartmental interaction cost for the central patio design is 15 percent greater than the rectangular design

managerial decision needs the quantitative input. However, the criteria for decision are broad, and they include job satisfaction and architectural attractiveness.

Applications to Macro Systems

In larger-scale systems, the same general point of view on the relative location of facilities applies. Many manufacturing plants are enormous in scale, with several buildings covering many acres. The material handling cost between units located in separate buildings can be very large, justifying careful study of relative locations.

One such application took place at a movie studio [Buffa, Armour, and Vollmann, 1964], where the major functional departments—wardrobe, furniture storage, nursery (for dressing a stage with plant material), sound stages, etc.—were interconnected by a material handling system, which was basically a fleet of trucks.

In dressing a stage before shooting a scene, the director must specify the kinds of furnishings and other props desired. Because of the esthetic aspects required for this procedure, many trips may be needed to obtain exactly the right effects. After the scene has been shot, all the materials must be redistributed to their storage places for later reuse. In such a situation the material handling costs are immense, and the best relative location of various departments strongly depends on these costs. The CRAFT program indicated that it was possible to reduce one studio's material handling cost by $240,000 per year, through relocation of many of the activities involved. Since relocation on such a scale is expensive, many alternatives were computed, fixing the location of certain facilities (e.g., sound stages) in their present locations, and determining the best locations for movable functions in relation to the fixed-location facilities.

In another study of the facilities of a large, complex aerospace manufacturer, these same concepts were valuable in relocating departments within the system. The study focused on the relocation of special laboratory service (which had been decentralized) and raised the question of whether or not to centralize—and, if so, which location of several possibilities was most desirable. The results indicated that centralizing the service in one of the available locations was 14 percent less costly than the second best location.

Comparative Effectiveness of CRAFT

Other approaches to the relative location problem have been developed, some of which have been computerized. Wimmert [1958] developed a quantitative procedure. Hillier and Connors [1966] developed a quadratic assignment algorithm. In 1967, two computerized approaches were developed that use a rating system to approximate the relative need for pairs of departments to be adjacent or close to each other. ADELP was developed by Seehof and Evans [1967], and CORELAP was developed by Lee and Moore [1967]. Both programs use a scale that assigns numerical weights to the following descriptors: absolutely essential, essential, important, optional importance, unimportant, undesirable, and same department. Independent comparative studies have been carried out by Denholm and Brooks [1970], Ritzman [1972], Nugent, Vollmann, and Ruml [1968], and Zoller and Adendorff [1972].

Assumptions and Limitations of CRAFT

CRAFT is a macro layout model that emphasizes the large differences among alternatives. It assumes that a good final layout must be based on a block diagram that permits low cost aggregate flow. In taking the aggregate view, some of its assumptions should be understood so that results can be interpreted properly.

One of CRAFT's major assumptions is that flow among departments or work centers occurs between the centroids of the departments. When the department shape is approximately square, this is a fairly good assumption. On the other hand, when departments are long narrow rectangles or irregular shapes, the assumption is not a good one. Also, sometimes the minimizing algorithm will produce odd department shapes. The analyst must regard the CRAFT output as an input to further analysis. Department shapes often are regularized anyway to fit in with building limitations (e.g., columns) and to lay out aisles and other spaces in a practical manner.

The nature of the criterion used also involves significant assumptions. Whatever the criterion used—material handling costs, walking cost, and so on—the assumption is that the criterion varies linearly with distance. In addition, the very nature of layout studies places faith in past data on flow and its composition, although the layout itself is made to accommodate future flow patterns. Studies indicate, in some situations at least, that flow patterns are fairly stable over time. However, if they are dynamic, a layout based on CRAFT or any other technique will become obsolete. Finally, in some situations, handling costs may be joint with several products and can only be allocated arbitrarily.

The issues of assumptions and limitations are more throughly discussed by Vollmann and Buffa [1966].

Comparison with Visual Based Methods

Scriabin and Vergin [1975] conducted experiments in which computer models (including CRAFT) were put in competition with individuals making visual relative location arrangements of departments. Problem sizes were varied from five to twenty departments, and resulting mean material handling costs were compared with the best computer solutions. The conclusion was that the visual based methods were superior, or at least equal, to the computer based methods.

Unfortunately, the experimental structure failed to consider some of the most important problem characteristics. All departmental areas were assumed to be equal, as in the operations sequence analysis technique discussed earlier in this chapter. The computer techniques exhibit their power in the more complex situations representing real problems (in which the size of departments may vary significantly). In addition, the example problems used by Scriabin and Vergin involved flow patterns that exhibited moderate flow dominance. When flow dominance is prominent and areas are equal, relative location patterns are fairly obvious, and a visually based technique such as operations sequence analysis is appropriate [Vollmann, 1964].

Thus, a valid test of such a comparison must deal with unequal areas, different degrees of flow dominance (including cases where flow is not dominant), and problem sizes probably as large as thirty to forty departments. Nevertheless, even in cases where computer solutions are valid, visual examination of results may produce improvements. Practitioners would not simply accept computer solutions as final without examining them further [Buffa, 1976].

PROJECT SYSTEMS AND FIXED-POSITION ASSEMBLY

Large-scale projects are common in today's economy. In terms of flow, they are intermittent systems, and their physical layout deserves a special comment. Examples of familiar large-scale projects are the huge aerospace projects, such as the Polaris missile; aircraft assembly; ship building; or large construction projects. Some of these kinds of projects involve basically conventional manufacturing systems to produce components that go into the project itself, such as occurs in missile- and aircraft fabrication, and in the manufacture of some components that are assembled into ships, buildings, dams, and so on. The systems that manufacture these components basically will be typical of either intermittent or continuous manufacturing systems.

The heart of the project concept, however, lies in the assembly process, which is usually done on a fixed-position basis, either by necessity (as with buildings, dams, and bridges) or for reasons of economy (as with missiles, aircraft, ships, and other very large projects). In these fixed-position assembly situations, the equivalent of functional centers commonly are arranged around the unit being constructed in "staging areas." Some of the staging areas will be storage locations where the material or components will wait until needed in the process. Other areas may involve some degree of fabrication or prefabrication before final assembly on the unit.

Proximity of staging areas to the major unit may depend on frequency of use and travel time between the staging area and the unit. For example, in constructing a skyscraper, the heating and air conditioning unit need only be installed once; however, forming lumber and reinforcing steel will be used continuously throughout the rough construction stages. The general concept of the CRAFT location algorithm applies to project systems as well as to the kinds of situations already discussed. Since the project is commonly a one-time system, however, formal location procedures have not been used in practice, though the potential for use is valid.

PRODUCT LAYOUT OF CONTINUOUS SYSTEMS

When we speak of continuous systems, we normally think of manufacturing systems that are organized and physically laid out according to product. Whatever is being processed moves through a sequence of operations at rates that approach continuous movement (thus the name). Continuous movement is a relative term that distinguishes the character of such systems from batch processing, where, by contrast, movement is intermittent.

While most continuous systems are found in manufacturing, the idea of direct flow—where all the work to be done is subdivided into specialized operations and physically is arranged in the sequence of required operations—has been applied in some nonmanufacturing activities. For example, the cafeteria line is an application of the production line concept.

Decision to Organize for Continuous Flow

The managerial decision to organize the work on a product or line basis is a significant one. Some important requirements should be met, and there are some consequences from the attitudes of the work force that should be carefully weighed before implementing the decision. The aspects of division of labor versus job enlargement were discussed in Chapter 7, so we will simply mention that the resulting issues are an important consideration in the decision to organize for continuous flow. The materials presented in this chapter are based on the premise that we are involved in the design of a sociotechnical system, not a machine.

If we are to organize for continuous flow, the following requirements should be met:

1. Volume adequate for reasonable equipment utilization.
2. Reasonably stable product demand.
3. Product standardization.
4. Part interchangeability.
5. Continuous supply of material.

Each of these requirements needs to be qualified. (1) The concept of adequate volume presumes an economic analysis to determine the breakeven volume between line organization and alternatives. Reasonably good equipment utilization is associated with high volume. (2) Stable demand is required in terms of a minimum run that would cover at least the special tooling costs for a line. (3) Thus, stable demand is associated with product standardization. Engineering changes in product designs can be accommodated by production lines, but changes cannot be too frequent. We must have an "economical run" to cover to the costs of retooling and relayout that design changes may require. (4) Part interchangeability is required so that no special reworking or refitting of parts is needed during assembly. If parts are not interchangeable at assembly, the flow of work is disrupted because of imbalance. (5) Finally, when we have the high-volume, standardized-product situation described by the foregoing requirements, continuous supply of material is crucial. The lack of supply of a single part or item or raw material can force the entire process to be stopped, and the resulting downtime costs can be very large. e.g. Computer

Continuous flow concepts have found their greatest field of application in assembly rather than in fabrication. A moment's reflection makes obvious why this is true. Machine tools commonly have fixed machine cycles; this factor makes it difficult to achieve balance between successive operations. The result is poor equipment utilization and relatively high costs. In assembly operations where the work is more likely to be manual, balance is much easier to obtain because the total job can be divided into minute elements. If Operation 10 is too short while 16 is too long, part of the work of Operation 16 probably can be transferred to 10 (perhaps the tightening of a single bolt). Since very little equipment is involved at each operation anyway, utilization of equipment may not be of great importance.

When the conditions for continuous flow systems are met, significant advantages can result. The production cycle is speeded up because materials approach continuous movement. Since very little manual handling is required, the cost of material handling is low. In-process inventories are lower compared with batch processing because of the relatively fast manufacturing cycle. Because aisles are not used for material movement and in-process storage space is minimized, less total floor space is commonly required than for an equivalent intermittent system, even though more individual pieces of equipment may be required. Finally, the control of the flow of work (production control) is greatly simplified for continuous systems because routes become direct and mechanical. No detailed scheduling of work to individual work places and machines is required, since each operation is an integral part of the line. Scheduling the line as a whole automatically schedules the component operations.

Given the decision for designing a continuous flow system, the major problems are (1) deciding on a production rate for cycle time, and (2) subdividing the work so that smooth flow can result. The subdivided activities need to be balanced so that each operation has an equivalent capacity.

THE LINE BALANCE PROBLEM

Balance refers to the equality of capacity or output of each of the successive operations in the sequence of a line. If all are equal, we have perfect balance and expect smooth flow. If they are unequal, we know that the maximum possible output for the line as a whole will be dictated by the slowest operation in the sequence. This slow, or bottleneck, operation restricts the flow of parts on the line much as a half-closed valve restricts the flow of water, even though the pipes in the system may be capable of carrying twice as much water. Thus, when imbalance exists in a line, we have wasted capacity in all operations except the bottleneck operation.

Production Lines as Queuing Systems

One of four basic structures of waiting line models is the single channel, multiple phase case.* This structure is a sequence of several of our basic input-transformation-output modules in series, with the output of one phase or stage being the input to the next. If the line is not mechanically paced, the concept of such a structure parallels the actual flow that takes place in a production line.

$A \Rightarrow B \Rightarrow \infty$

*See Appendix D for a discussion of waiting line theory. Figure 1c of that appendix shows the structure of the single channel, multiple phase case.

As we noted previously, the important problem is to balance the stages so that each has the same amount of work to do. Actually, however, we are dealing with work-time distributions, and therefore an interplay between arrivals and service times at each stage will exist, as in any queuing system. This means that there will be some instances when items are waiting to be processed at each stage, and other instances when the waiting line is empty and the stage is idle. Therefore, the actual operation of a line complicates the problem of achieving balance. We often try to minimize the probability of having any stage idle by flooding the system with in-process inventory so that work usually will be waiting at each stage.

Returning to the problem of balance, to date waiting line theory has not been a fruitful approach in solving the balance problem, even though it gives useful insights into how a line functions. Since we have work-time distributions to deal with, the balance problem is a stochastic one. Most of the work on line balance, however, has simplified the basic problem, assuming the deterministic case where the service times are constant values, and that work always is available at each stage. More recently, research work has focused on stochastic line balancing [Reeve and Thomas, 1973].

Deterministic Form of Line Balancing

The following material assumes the deterministic form of the line balance problem. First we shall consider procedures for *assembly line* balancing. Later we shall see what differences might exist if we were dealing with a fabrication line.

In order to achieve balance to the best of our ability, we need to know the performance times for the smallest possible whole units of activity, such as tightening a bolt or making a solder joint. We also must know the flexibility in the sequence of these tasks or activities. There are, of course, certain limitations on the sequence of the tasks. For example, a washer must go on before the nut; wires must be joined physically before they can be soldered; a hole must be drilled before it can be reamed, and reamed before it can be tapped. On the other hand, the sequence may be irrelevant; for example, the order in which a series of nuts is put on. This sequence flexibility is important in helping us to specify the groups of elements making up operations, or stations, for the line that achieves the best balance.

~~Let us use~~ an example to see the nature of the problem: Figure 9 shows the cylinder subassembly for a typical small air compressor; the parts are named and numbered. By examining the assembly, we readily can see the sequence restrictions that we would have to observe. When assembling the cylinder head to

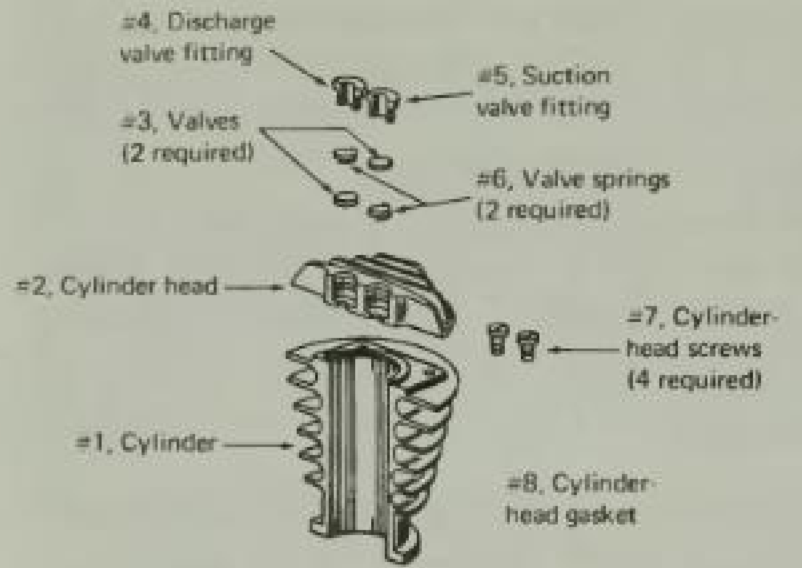


Figure 9. Cylinder subassembly for a typical air compressor

the cylinder, we would have to position the cylinder head gasket (part 8) first. Also, when assembling the discharge valve unit, we would have to put the valve itself (part 3) in first, followed by a valve spring (part 6), and finally by a discharge valve fitting (part 4). A similar procedure for the suction valve unit would be followed, but the sequence of the valve and the spring would be reversed.

These sequences must be observed, because the cylinder subassembly cannot be assembled correctly in any other way. On the other hand, it makes no difference whether the valve units are assembled to the cylinder head before or after the cylinder head is assembled to the cylinder. Similarly, which valve unit is assembled first is irrelevant. The cylinder head is joined to the cylinder by four screws. All four need not be assembled at the same time, and there is no required sequence for their assembly.

These task-sequence restrictions are summarized in Table 6 so we can use the result to advantage. In general, the assembly tasks listed in Table 6 are broken down into the smallest whole activity. Note, for example, that the screws and the valve fittings are positioned first and the threads engaged, so that tightening can take place separately, perhaps as a part of the next station or some subsequent station. For each task, we note in the right-hand column the task or tasks that must precede it. Thus, tasks *a*, *e*, and *i* can take any sequence, because no tasks need precede them. However, task *b* (position cylinder head on cylinder) must be preceded by task *a* (cylinder head gasket on cylinder). Task c-1 must be preceded by tasks *a* and *b*. The repetition of *a* is not absolutely necessary, since we know that *b* must be preceded by *a*. With this information, together with the task times also given in Table 6, we can construct the diagram in Figure 10.

TABLE 6

LIST OF ASSEMBLY TASKS SHOWING SEQUENCE RESTRICTIONS AND PERFORMANCE TIMES FOR THE CYLINDER SUBASSEMBLY OF FIGURE 7

Takt = 1/rate production rate

Task	*Performance Time, Seconds*	*Task Description*	*Task Must Follow Task Listed Below*
a	1.5	Position cylinder head gasket (No. 8) on cylinder (No. 1)	—
b	2.0	Position cylinder head (No. 2) on cylinder (No. 1)	a
c-1	3.2	Position cylinder head screw (No. 7) in hole and engage threads	b
c-2	3.2	Repeat	b
c-3	3.2	Repeat	b
c-4	3.2	Repeat	b
d-1	1.5	Tighten cylinder-head screw	c
d-2	1.5	Repeat	c
d-3	1.5	Repeat	c
d-4	1.5	Repeat	c
e	3.7	Position valve (No. 3) in bottom of discharge hole	—
f	2.6	Position valve spring (No. 6) on top of valve in discharge hole	e
g	3.2	Position discharge valve fitting (No. 4) in hole and engage threads	f
h	2.0	Tighten discharge valve fitting	g
i	3.1	Position second valve spring (No. 6) in bottom of suction hole	—
j	3.7	Position second valve (No. 3) on top of spring in suction hole	i
k	3.2	Position suction valve fitting (No. 5) in hole and engage threads	j
l	2.0	Tighten suction valve fitting	k

Figure 10 merely reflects in a graphical way the sequence requirements that we have determined. For convenience, the performance times are indicated beside the tasks. Now we can proceed with the grouping of tasks to obtain balance. But balance at what level? What is to be the capacity of our line? This is an important point, and one that makes a balancing problem difficult. If there were no restriction on capacity, the problem would be simple; we could take the lowest common multiple approach. For example, if we had three operations that required 3.2, 2.0, and 4.0 minutes, respectively, we could provide eight work places of the first, five of the second, and ten of the third; thus, the capacity of the line would be 150 units per hour at each of the operations, and a cycle time of 0.4 minutes. But capacity has been specified by balance rather than by the considations that we discussed previously.

$\frac{60}{3.2} = 18.75$ units

150

.4 for 1
.4 2nd
.4 3rd

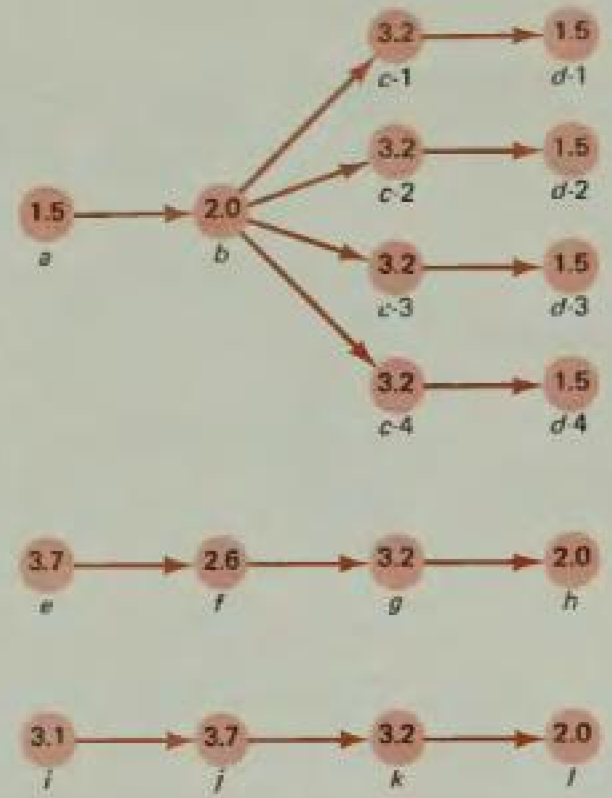


Figure 10. Diagrammatic representation of the sequence requirements shown in Table 6. Numbers indicate performance times of tasks

We must take the capacity of the line as given and develop good balance within that restriction. For illustrative purposes, let us assume that we must balance our line for a 10 second cycle. A completed unit would be produced by the line every 10 seconds. To meet this capacity requirement, no station could be assigned more than 10 seconds' worth of the tasks shown on the diagram of Figure 10. We then could group the tasks into station assignments. The total of all task times is 45.8 seconds. Therefore, with a 10 second cycle, five stations would be the minimum possible. Any solution that required more than five stations would increase direct labor costs. Figure 11 shows a solution that yields five stations. While this is a simple example, it illustrates the conceptual problems of line balancing.

Practical Balancing Methods. The general concepts of line balancing that we have discussed have been implemented through a number of practical methods for the large-scale problems in industry. These methods have ranged from linear programming models [Bowman, 1960] and dynamic programming [Held, Karp, and Sharesian, 1963] to heuristic methods [Helgeson and Birnie, 1961; Kilbridge and Wester, 1961; Tonge, 1960] and a computer based biased sampling technique [Arcus, 1966]. Perhaps of the greatest interest for large-scale problems are, COMSOAL: a Computer Method for Sequencing Operations for Assembly Lines, developed by Arcus; and a new computer model called MALB, developed by Dar-El [1973], based on an improvement of the Ranked Positional Weight Technique.

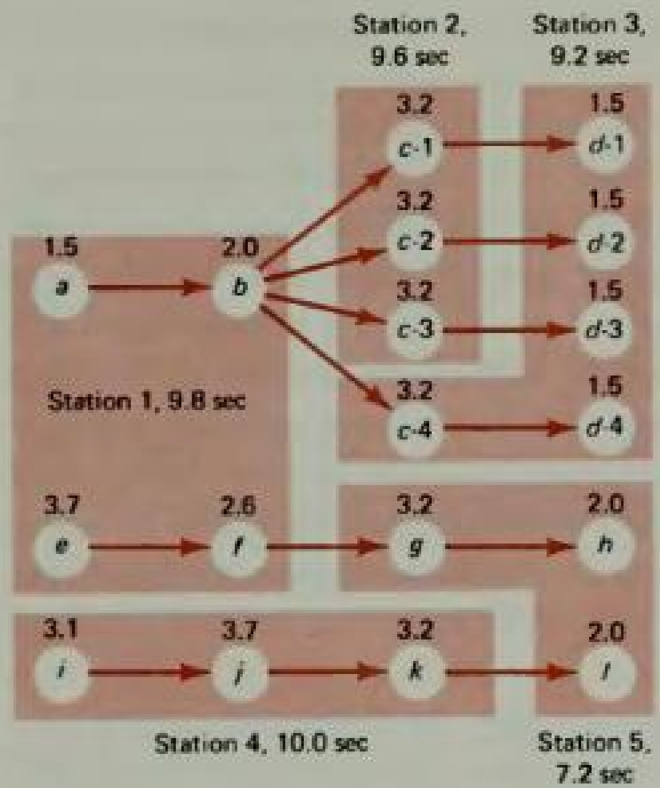


Figure 11. A solution to the line balancing problem which requires no more than 10 seconds per station and does not violate the task sequence requirements

COMSOAL methodology is based on the rapid generation of feasible solutions by a computer routine. The probability of generating either optimal or very good solutions is finite, and depends on the sample size of feasible solutions generated. The nature of the computer program biases the generation of feasible solutions toward the better ones in order to save computer time. COMSOAL already has been implemented at Chrysler Corporation; Arcus has used it on a hypothetical line of 200 stations with 1,000 tasks, achieving the known optimal solution of zero idle-time.

Mastor [1970] evaluated a large number of different line balance models in a set of experiments involving a range of problem sizes (numbers of tasks to be assigned); line lengths (number of stations in the line); and order strengths (degree of sequence restriction imposed in the precedence relationships). He found that the dynamic programming formulation model was consistently the best performer. Although COMSOAL was a close second, the computer time required was much smaller. For large problems (70, 92, and 111 tasks), both techniques performed almost identically when the COMSOAL program was run for 999 sequences. Even for the large sample size on the 111 task problem, COMSOAL required only 56 seconds of computer time, compared with 94.3 seconds for the dynamic programming model.

Mastor's comparative study did not include the improvement on the Ranked Positional Weight Technique. While the Basic Ranked Positional Weight Technique did not perform particularly well, Dar-El [1973] compared the computerized improvement (MALB) with COMSOAL. The test problems ranged in size from 50 to 140 tasks. The comparative results indicate an advantage for MALB, in terms of both the resulting balance delay of idle time and computing time requirements.

Auxiliary Balancing Techniques. Other techniques are available to obtain balance in both the design and operation stages. If one station requires greater time than the others, a careful study of the activities, as well as the psychophysiological aspects of the tasks, may result in a reduction of time. Compensations for imbalance can be accomplished by assigning fast operators to the limiting operations. When some very fast operations cannot be combined into a single station (as might be true with machine operations), material banking before and after the fast operations may be required. These fast operations would be operated during only a small portion of the day; the machine, the operator, or both, could be used for other purposes.

Balancing Fabrication Lines. Conceptually, there is no difference in the balancing procedures for assembly and fabrication lines. However, the fixed machine cycles found in fabrication operations considerably limit freedom in achieving balance. It is often impractical to divide a machine operation into two or more suboperations in order to level out the time requirements at each station. This situation partially accounts for the fact that fabrication lines generally are economical only for very high volumes, since good balance is likely to be achieved only at high levels of output. Otherwise, equipment utilization is very low. If we try to use the idle capacity to machine other parts, material-handling costs increase because the equipment has been located so that it will minimize handling costs for the line product. The process layout, which minimizes overall handling costs, is usually more economical under these conditions.

Production Line Design and Job Satisfaction

As noted in Chapter 7, the production line concept epitomizes division of labor and specialization, and is the target of considerable criticism. The general line balance concepts just discussed do not seem to leave a door open for discussion of job design alternatives. The process starts with the determination of a cycle time that meets the output rate needs, and progresses toward the generation of stations (jobs) made up of tasks that meet the needs of the cycle and do not violate the technological constraints. Job satisfaction as a criterion seems to have no role in this process. The issue is: Are there alternatives that allow us to address the question of the degree of fractionation of jobs, while retaining the production line concept?

Multiple Stations and Parallel Lines. Given a capacity or production rate requirement based on forecasts of demand and market predictions, we can meet that requirement by a single line with cycle time *c*; or two parallel lines with cycle time 2*c*; or three parallel lines with cycle time 3*c*; etc. Buxey [1974] has developed extended line balance programs that enable us to use multiple stations. Of course, as the number of parallel lines increases, the scope of jobs increases; and, finally, we have complete horizontal enlargement. The point is that the alternatives do exist, even with the line organization of work.

In addition to increasing job scope, there are a number of advantages that result from the parallel line-multiple station concept. First, from the line balance point of view, balance may be easier to achieve because the larger cycle time offers a greater likelihood of attaining a good fit with low residual idle time. This is particularly true when some of the task times are nearly equal to the single-line cycle time. Furthermore, a multiple line design increases flexibility of operations enormously. Output gradations are available. That is, one can have one, two, three, etc., lines operating or not operating; and one can work overtime or undertime with all the line combinations. There are fewer dependent operations; for example, if there is a difficulty with an operation in Line 1, it may not affect Line 2. A machine breakdown in Line 1 need not stop the operation of Line 2.

From a human organization viewpoint, the parallel line-multiple station concept has all the advantages of job enlargement. Work groups can be smaller and more cohesive. A team spirit may be engendered by competition between teams on the bases of output, quality, safety, and other dimensions.

SERVICE FACILITIES

reread

Plant Services

Many plant services must fit into the overall layout. Because many of these activities are not part of the direct production activity of the enterprise, the idea has persisted that whatever space is left over is good enough for them. Actually, some of these activities (e.g., receiving, shipping, and warehousing) are in the direct material flow, and they process the product as do the production department. Others (e.g., maintenance facilities and tool cribs) do not work on the product but their activities do interact with production costs; therefore, their physical location and capacity deserve careful thought. The overall material flow patterns should be the major factors in determining the relative locations of receiving, shipping, storage, and warehousing areas. Our discussion of operation sequence analysis for process layouts included these functional areas. Since the material flow is more obvious for line layouts, the proper location for these areas is fairly well determined.

There is no obvious answer to the capacity question for receiving areas. In general, the problem is such that we do not have control over the rate at which materials come in. Since receipts of shipments from suppliers occur in a somewhat random pattern, a good design provides capacity that meets the reasonably expected peak loads for truck and rail docks, unloading crews, and temporary setdown areas for materials. Here again, waiting line theory may provide a guide for determining what these capacities should be. Of course, many other factors influence the details of the layout of receiving areas, such as climate, safety codes, handling equipment, dock heights, and the need to accommodate a variety of vehicles.

The location of tool cribs is important because of the travel time of high-priced mechanics to and from the area. Therefore, a study of the use frequency in relation to the physical layout of the production areas should determine a good location or locations. The tool storage problem is comparable to the material- and part storage problem in using space efficiently while making items available quickly and conveniently. The number of attendants required to serve the tool crib is another waiting line problem.

Maintenance facilities commonly are provided for buildings and grounds, plant utilities, and machinery and equipment. The capacity of maintenance for machinery and equipment again poses the problem of balancing the idle time of maintenance crews against the idle time of production workers and of losses of output capacity. Ordinarily, a considerable amount of idle capacity in equipment and crews is justifiable, as solutions to waiting line models of these types of problems would show.

flexible within constraints, central location for all main users

Employee Services

Present day personnel services cover a broad spectrum, including parking, cafeterias, medical services, credit unions, locker rooms, toilets and lavatories, and, often, recreational facilities. Obviously, providing for these services presents layout problems. In many instances, the location of these services does not have an effect on production costs, since the services are used after hours. In these instances, the layout problem concerns providing the space that is designed to perform the services in the amounts that are required. The activities must be studied to determine what must be done, and facilities must be provided accordingly.

For those services used during working hours, such as medical facilities, toilet facilities, and drinking fountains, the size of the facility and its location in relation to the users becomes important. Studies of travel distances to and from the service facility should be made in order to determine reasonable locations. Waiting line models again are useful in determining a balance between waiting times of employees and service capacity costs. In one large company that offered a broad medical service, the question of whether or not an additional doctor on the staff was warranted was answered by a waiting time study. The results of the study indicated that there was an average of fifteen employees in the waiting room during the eight-hour work day; assuming a 2,000 working-hour year and a modest average hourly wage of $4, this translates into $120,000 of waiting time per year. The study led to both an enlargement and decentralization of medical services.

Summary

Intermittent systems are those in which the flow is intermittent and processing units are organized by function. Such systems have general capability in some field of specialization.

Since intermittent systems can process a variety of work, routes through the system must vary, and the resulting problem of physical design and arrangement of work centers can be very complex. Thus, the relative location of departments sets the design for the system in an overall sense. The objective of the physical arrangement is to minimize some appropriate measure of the interaction among work centers. This may be material handling cost in manufacturing systems, cost of personnel walking between locations in hospitals, clinics, and offices, or simply the need to have face-to-face contact, as in some office situations.

Operations sequence analysis is a simple, graphic technique for handling small problems. Larger problems require a more powerful tool, such as the CRAFT program. As inputs, the CRAFT program requires (1) an initial spatial array, and (2) matrixes of the volume of interaction and interaction cost per volume unit, per unit distance. The algorithm functions by computing the changes in interaction cost that would result if locations were exchanged, making exchanges progressively for those switches with the greatest cost reduction potential. The program may be of value for both new design and redesign of existing facilities.

Large-scale projects are intermittent systems where the assembly process is usually accomplished in a fixed position. The CRAFT program concept for locating staging areas around the unit being constructed is valid, but has not been used.

Continuous system design is centered in line balancing, where the problems of system capacity and the capacity of individual operations and their balance are resolved. While the queuing model is conceptually valid in helping us understand what happens in the flow of items in a production line, it provides little help in solving the line balance problem.

The results of Mastor's comparative investigation underscore the effectiveness of the Held et al. and the Arcus models over a wide range of problem sizes, order of strengths, and line lengths. In addition, more recent work indicates that the MALB program must be included as a very effective method for large-scale problems.

Review Questions

1. In planning for a facility layout, the design capacity is an important decision, since capital investment requirements largely are determined by the decision, and the basic strategy for output scheduling also is set. Under what conditions should we design for peak demand needs? For average demand needs?

2. What are the alternate strategies for providing for future capacity in a facility layout?

3. What are the general economic considerations that result in the use of multiple shifts?

4. Define the terms *facility efficiency factor* and *scrap factor*. What are their functions in computing capacity needs?

5. Define the terms *process layout* and *product layout*.

6. Under what conditions would process layout be appropriate?

7. What is the relative location of facilities problem?

8. Consider the situation of the private industrial clinic used as an example for "operations sequence analysis." Recall that the clinic performs eight

types of examination sequences, depending on the details of individual contracts. Since there have been many complaints about excessive walking by those who have been examined, why not organize the entire clinic on the basis of the eight types of examination sequences in production-line fashion?

9. Table 2 summarizes types of criteria that might be used to determine the relative location of facilities in intermittent systems. Can you think of situations where criteria other than those listed might be important?

10. In the operations sequence analysis technique, what is the criterion for deciding whether or not departments should be located adjacent to one another?

11. In the operations sequence analysis technique, how do we translate the schematic diagram into a block diagram?

12. Explain the nature and functioning of the CRAFT algorithm. By what criterion would the algorithm exchange the location of departments in a factory layout? an office layout? a municipal office complex?

13. Explain the assumptions and limitations of the CRAFT layout model.

14. What is *fixed position assembly*? How does it relate to large-scale project systems?

15. What kinds of material handling systems do you think might be appropriate for process layouts?

16. In the decision to organize for continuous flow, why are each of the following factors requirements that should be met?
    a. Volume adequate for reasonable equipment utilization.
    b. Reasonably stable product demand.
    c. Product standardization.
    d. Part interchangeability.
    e. Continuous supply of material.

17. Define the meaning of each of the five requirements listed in Question 16; that is, what is their meaning with respect to the decision to organize for continuous flow?

18. Define the nature of the line balance problem.

19. What is the definition of a *task* in the line balance problem?

20. Describe the line balance problem as a queuing model. What simplifications are made in the deterministic form of line balancing?

21. What is the comparative performance of the practical methods of balance models discussed in this chapter?

22. How is *fine tuning* of line balance solutions achieved after installation?

23. How is it possible to introduce the question of appropriate job breadth into the process for designing and balancing production lines?

24. What kinds of material handling systems do you think might be appropriate for product or line layouts?

Problems

1. Assume a simple situation in which there are four departments to be arranged on a simple grid, as in Figure 3. Because of symmetry, most of the possible arrangements are not really different in their effect on aggregate interaction cost. There are $4! = 24$ possible arrangements; however, only three of them are fundamentally different. Enumerate these three arrangements.

2. How do unequal areas change the answer to Problem 1? Suppose, for example, that area requirements are:

Department	Area
A	$2 \times 4 = 8$
B	$2 \times 2 = 4$
C	$1 \times 2 = 2$
D	$1 \times 2 = 2$
	16

a. How many different arrangements are now possible? Enumerate them.
b. How many different arrangements are possible if the areas are as follows:

Department	Area
A	$1 \times 4 = 4$
B	$1 \times 4 = 4$
C	$1 \times 2 = 2$
D	$1 \times 2 = 2$
	12

3. Consider the schematic diagram shown in Figure 5. Can you improve it?

4. A manufacturing concern has four departments, and the flow between combinations of departments is as follows:

	To			
From	*A*	*B*	*C*	*D*
A		2		2
B	2		4	
C		3		1
D	2		1	

a. Using the operations sequence analysis technique, how should the departments be arranged?
b. Now suppose that the area requirements are as follows:
   Department *A*—3,600 sq. ft.
   Department *B*—2,400 sq. ft.
   Department *C*—2,400 sq. ft.
   Department *D*—1,600 sq. ft.

   Sketch the block diagram based on your answer in Question 4a.
c. Now assume that the four departments are located in two separate buildings that are 100 feet apart. The two buildings have floor areas of $60 \times 100 = 6{,}000$ square feet, housing departments *A* and *C*, and $40 \times 100 = 4{,}000$ square feet, housing departments *B* and *D*, respectively. What should be the space allocation to the four departments if the material handling costs are as follows:

From \ To	A	B	C	D
A		1	1	2
B	1		1	2
C	1	1		2
D	2	2	2	

5. An organization does job machining and assembly, and wishes to relayout its production facilities so that the relative location of departments reflects in a somewhat better way the average flow of parts through the plant. Table 7 shows an operations sequence summary for a sample of seven parts, with approximate area requirements for each of the thirteen machine or work centers. The numbers in the columns headed by each of the parts indicate the number of the work center to which the part goes next. Just below the sequence summary is a summary of production per month and the number of pieces handled at one time through the shop for each part.

TABLE 7

Machine or Work Center	Area, Square Feet	Work Center Number	Part						
			A	B	C	D	E	F	G
Saw	50	1		2	2				2
Centering	100	2		4	3				3
Milling machines	500	3	5	9	5	5		4	4
Lathes	600	4		5,7	7		5	10	5
Drills	300	5	8	3	4	11	7		6
Arbor press	100	6					11		7
Grinders	200	7		12	12		6		8
Shapers	200	8	9			3			9
Heat treat	150	9	11	4					10
Paint	100	10						11	11
Assembly bench	100	11	12	13	13	13	13	13	12
Inspection	50	12	13	11	11				13
Pack	100	13							
Production Summary									
Pieces per month			500	500	1600	1200	400	800	400
Pieces per load			3	100	40	40	100	100	2
Loads per month			250	5	40	30	4	8	200

a. Develop a load summary showing the number of loads per month going between all combinations of work centers.

b. Develop a schematic layout which minimizes nonadjacent loads. loads.

c. Develop a block diagram that reflects the approximate area requirements, and results in an overall rectangular shape.

6. A layout study was made in the engineering office of a large aerospace manufacturer. The study initially focused on the flow of work through the system, but meaningful cost data was difficult to generate on this basis. The search for realistic measures of effectiveness finally was narrowed down to the relative location of people in the organization as required by their face-to-face contacts with others in the organization. It was decided to collect data on face-to-face contacts initiated by each person for a

TABLE 8

NUMBER OF FACE-TO-FACE CONTACTS PER MONTH IN AN ENGINEERING OFFICE

From \ To	(A) Filing	(B) Supervision	(C) Blueprint	(D) Product Support	(E) Structural Design	(F) Electrical Design	(G) Hydraulic Design	(H) Production Liaison	(I) Detailing and Checking	(J) Secretarial Pool
A Filing		15				5		10		15
B Supervision	20		25	40	100	90	80	160	85	60
C Blueprint										
D Product Support	10	15					20	280		10
E Structural Design	50	20	600			40			340	50
F Electrical Design			475					160	270	60
G Hydraulic Design	10		460	20				140	320	45
H Production Liaison	20			200	160	190	240			680
I Detailing and Checking		210	690	40	190	240	80			20
J Secretarial Pool		25						15		

TABLE 9
AREA REQUIREMENTS AND AVERAGE WAGE RATES
FOR TEN GROUPS IN A LARGE ENGINEERING OFFICE

	Group	Area	Average Hourly Wage
A	Filing	20 × 15 = 300 sq. ft.	$2.25
B	Supervision	30 × 15 = 450 sq. ft.	5.00
C	Blueprinting	40 × 15 = 600 sq. ft.	2.10
D	Product support	25 × 20 = 500 sq. ft.	2.70
E	Structural design	65 × 25 = 1625 sq. ft.	4.50
F	Electrical design	25 × 35 = 875 sq. ft.	4.50
G	Hydraulic design	45 × 30 = 1350 sq. ft.	4.50
H	Production liaison	20 × 70 = 1400 sq. ft.	2.70
I	Detailing and checking	70 × 25 = 1750 sq. ft.	3.60
J	Secretarial pool	70 × 15 = 1050 sq. ft.	2.40
		90 × 110 = 9000 sq. ft.	

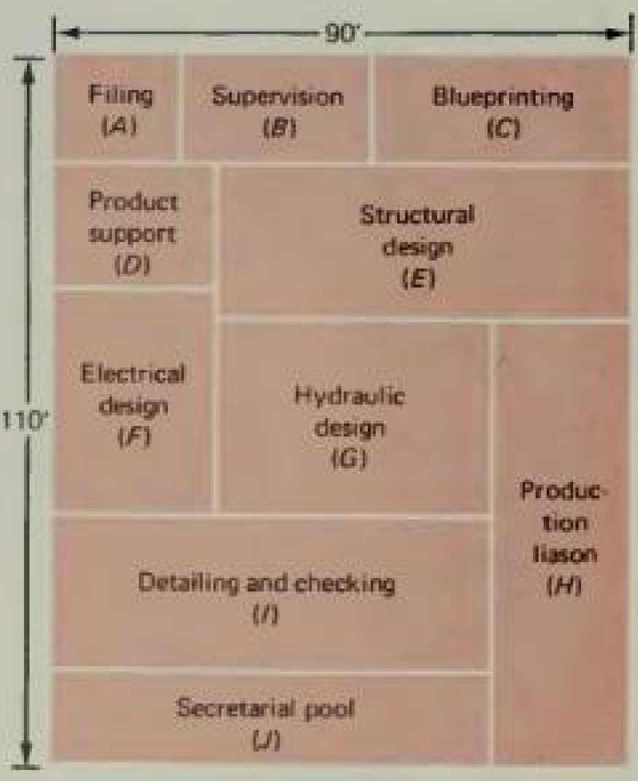


Figure 12. Existing block layout for ten groups within a large engineering office

TABLE 10
LIST OF TASKS, PERFORMANCE TIMES, AND SEQUENCE RESTRICTIONS FOR THE ASSEMBLY OF A PRODUCT

Task	*Performance Time, Minutes*	*Task Must Follow Task*
a	6	—
b	2	*a*
c	5	*a*
d	7	*a*
e	1	*a*
f	2	*b*
g	3	*b,c,d,e*
h	6	*f*
i	5	*g*
j	5	*h*
k	4	*i,j*

one-month period. These data were accumulated in the form of the matrix of Table 8.

The entire department was divided into ten groups or areas, as indicated in Table 8. Each cell value indicates the number of face-to-face contacts initiated by that group; for example, fifteen from *A* to *B*. Table 9 summarizes the area requirements for each group, as well as the average hourly wage paid in each group. Figure 12 shows the present block layout.

Prepare a new block layout within the constraints of the overall size of the layout shown in Figure 12.

7. Table 10 summarizes the required tasks, performance times, and sequence restrictions for an assembly. Construct a precedence diagram for the assembly.

References

Apple, J. M. and M. P. Deisenroth. "A Computerized Plant Layout Analysis and Evaluation Technique (PLANET)." *Proceedings*. American Institute of Industrial Engineers, 23rd Annual Conference and Convention, Anaheim, Calif., 1972.

Arcus, A. L. "COMSOAL: A Computer Method for Sequencing Operations for Assembly Lines." *International Journal of Production Research* (1966).

Armour, G. C., and E. S. Buffa. "A Heuristic Algorithm and Simulation Approach to Relative Location of Facilities." *Management Science* 9(1963): 294-309.

Atkinson, G. A., and R. J. Phillips. "Hospital Design: Factors Influencing the Choice of Shape." *Architects Journal*, April 1964.

Bowman, E. H. "Assembly Line Balancing by Linear Programming." *Operations Research* 8(1960): 385.

Buffa, E.S. "Communication to the Editor on a Paper by Scriabin and Vergin." *Management Science* 23(September 1976).

______. "The Additivity of Universal Standard Data Elements, I." *Journal of Industrial Engineering* 7(September-October 1956): 217-223.

______. "Pacing Effects in Production Lines." *Journal of Industrial Engineering* 7(November-December 1961).

______. "Sequence Analysis for Functional Layouts." *Journal of Industrial Engineering* 6(March-April 1955).

______; G. C. Armour; and T. E. Vollmann. "Allocating Facilities with CRAFT." *Harvard Business Review* 42(March-April 1964): 136-159.

______, and W. H. Taubert. *Production-Inventory Systems: Planning and Control*. Rev. ed. Homewood, Ill.: Richard D. Irwin, 1972. Chapter 10.

Buxey, G. M. "Assembly Line Balancing with Multiple Stations." *Management Science* 20(February 1974): 1010-1021.

Conrad, R. *Setting the Pace*. APU-232-55. London, England: Medical Research Council, Applied Psychology Research Unit, 1955.

______, and B. A. Hille. "Comparison of Paced and Unpaced Performance at a Packing Task." *Occupational Psychology* 29(January 1955): 15-28.

Dar-El, E. M. (Mansoor). "MALB—A Heuristic Technique for Balancing Large Scale Single-Model Assembly Lines." *AIIE Transactions* 5(December 1973).

______. "Solving Large Single-Model Assembly Line Problems—A Comparative Study." *AIIE Transactions* 7(1975): 302–310.

______, and R. F. Cother. "Assembly Line Sequencing for Model Mix. *International Journal for Production Research* 13(1975):463–477.

Denholm, D. H., and G. H. Brooks. "A Comparison of Three Computer Assisted Plant Layout Techniques." *Proceedings*. American Institute of Industrial Engineers, 21st Annual Conference and Convention, Cleveland, 1970.

Freeman, D. R., and J. V. Jucker. "The Line Balancing Problem." *Journal of Industrial Engineering* 18(1967): 361.

Francis, R. L., and J. A. White. *Facility Layout and Location: An Analytical Approach*. Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Fryer, J. S. "Organizational Segmentation and Labor Transfer Policies in Labor and Machine Limited Production Systems." *Decision Sciences* 7(October 1976).

Gutjahr, L., and G. L. Nemhauser. "An Algorithm for the Line Balancing Problem." *Management Science* 11(November 1964).

Held, M.; R. M. Karp; and R. Sharesian. "Assembly-Line Balancing: Dynamic Programming with Precedence Constraints." *Operations Research* 11(May–June 1963).

Helgeson, W. B., and D. P. Birnie. "Assembly Line Balancing Using the Ranked Positional Weight Technique." *Journal of Industrial Engineering* 12(November–December 1961).

Hicks, P. E., and T. E. Cowan. "CRAFT-M for Layout Rearrangement." *Industrial Engineering* 8(May 1976): 30–35.

Hillier, F. S., and M. M. Connors. "Quadratic Assignment Problem Algorithms and the Location of Indivisible Facilities." *Management Science* 13(September 1966): 42–57.

Hoffman, T. R. "Assembly Line Balancing with a Precedence Matrix." *Management Science* 9(July 1963).

Hunt, G. C. "Sequential Arrays of Waiting Lines." *Operations Research* 4(1956).

Ignall, E. J. "A Review of Assembly Line Balancing." *Journal of Industrial Engineering* 16(1965).

Jackson, J. R. "A Computing Procedure for a Line Balancing Problem." *Management Science* 2(April 1956).

Kaiman, L. "Computer Programs for Architects and Layout Planners." *Proceedings* American Institute of Industrial Engineers, 22nd Annual Conference and Convention, Boston, 1971.

Kala, R., and G. G. Hitchings. "The Effects of Performance Time Variance on a Balanced 4-Station Manual Assembly Line." *International Journal of Production Research* 6(October 1973).

Kilbridge, M. D. "The Balance Delay Problem." *Management Science* 8(October 1961).

Kilbridge, M. D., and L. Wester. "A Heuristic Method of Assembly Line Balancing." *Journal of Industrial Engineering* 12(July–August 1961).

———. "A Review of Analytical Systems of Line Balancing." *Operations Research* 10(September–October 1962).

Klein, M. "On Assembly Line Balancing." *Operations Research* 11(March–April 1963).

Lee, R. C., and J. M. Moore. "CORELAP-Computerized Relationship Layout Planning." *Journal of Industrial Engineering* 18(March 1967): 195–200.

Lew, P., and P. M. Brown. "Evaluation and Modification of CRAFT for an Architectual Methodology." *In G. T. Moore, ed., Emerging Methods in Environmental Design and Planning.* Cambridge, Mass.: MIT Press, 1970.

Mansoor, E. M. (Dar-El). "Assembly Line Balancing—An Improvement on the Ranked Positional Weight Technique." *Journal of Induatrial Engineering* 15(March–April 1964).

Mastor, A. A. "An Experimental Investigation and Comparative Evaluation of Production Line Balancing Techniques." *Management Science* 16(July 1970): 728–746.

Moodie, C. L., and H. H. Young. "A Heuristic Method of Assembly Line Balancing for Assumptions of Constant or Variable Work Element Times." *Journal of Industrial Engineering* 16(January–February 1965).

Moore, J. M. *Plant Layout and Design*. New York: Macmillan, 1962.

Mojena, R., T. E. Vollmann, and Y. Okamoto. "On Predicting Computational Time for a Branch and Bound Algorithm for the Assignment of Facilities." *Decision Sciences* 7(October 1976).

Muther, R., and K. McPherson. "Four Approaches to Computerized Layout Planning." *Industrial Engineering* 2(1970):39–42.

Nugent, C. E.; T. E. Vollmann; and J. Ruml. "An Experimental Comparison of Techniques for the Assignment of Facilities to Locations." *Operations Research* 16(January–February 1968): 150–173.

Reed, R. *Plant Layout*. Homewood, Ill.: Richard D. Irwin, 1961.

———. *Plant Location, Layout and Maintenance*. Homewood, Ill.: Richard D. Irwin, 1967.

Reeve, N. R., and W. H. Thomas. "Balancing Stochastic Assembly Lines." *AIIE Transactions* 5(September 1973).

Ritzman, L. P. "The Efficiency of Computer Algorithms for Plant Layout." *Management Science* 18(January 1972): 240–248.

Scriabin, M., and R. C. Vergin. "Comparisons of Computer Algorithms and Visual Based Methods for Plant Layout." *Management Science* 22(October 1975): 172–181.

Seehof, J. M., and W. O. Evans. "Automated Layout Design Program." *Journal of Industrial Engineering* 18(December 1967): 690–695.

Sommers, M. S., and J. B. Kernan. "A Behavioral Approach to Planning, Layout, and Display." *Journal of Retailing*, Winter 1965–1966, pp. 21–27.

Sury, R. J. "Aspects of Assembly Line Balancing." *International Journal of Production Research* 9(1971).

Thomopoulos, N. T. "Line Balancing–Sequencing for Mixed-Model Assembly." *Management Science* 14(October 1967).

______. "Mixed Model Line Balancing with Smoothed Station Assignments." *Management Science* 16(May 1970).

______. "Some Analytical Approaches to Assembly Line Problems." *Production Engineer*, July 1968.

______, and M. Lehman. "The Mixed Model Learning Curve." *AIIE Transactions* 1(June 1969).

Tonge, F. M. "Assembly Line Balancing Using Probabilistic Combinations of Heuristics." *Management Science* 11(May 1965).

______. *A Heuristic Program for Assembly Line Balancing*. Englewood Cliffs, N.J.: Prentice-Hall, 1961.

______. "Summary of A Heuristic Line Balancing Procedure." *Management Science* 7(October 1960).

Volgyesi, A. S. "Space-Age Approach to Space Allocation." Computer Decisions, May 1970, pp. 32–35.

______. "Toronto General: The Hospital That Computer Built." Computer Decisions, September 1969.

Vollmann, T. E. "*An Investigation of the Bases for the Relative Location of Facilities*." Ph.D. diss. University of California, Los Angeles, 1964.

______, and E. S. Buffa. "The Facilities Layout Problem in Perspective." *Management Science* 12(June 1966): 450–468.

______; C. E. Nugent; and R. L. Zartler. "A Computerized Model for Office Layout." *Journal of Industrial Engineering* 19(July 1968): 321–327.

Wester, L., and M. D. Kilbridge. "Heuristic Line Balancing: A Case." *Journal of Industrial Engineering* 13(May 1962).

Wimmert, R. J. "A Mathematical Method of Equipment Location." *Journal of Industrial Engineering* 9(November–December 1958).

Young, H. H. "Optimization Models for Production Lines." *Journal of Industrial Engineering* 18(January 1967).

Zoller, K., and K. Adendorff. "Layout Planning by Computer Simulation." *AIIE Transactions* 4(June 1972): 116–125.

Part IV

Operations Planning and Control

The chapters in Part 3 dealt with the design of productive systems. When such a system has been designed, we may presume that there has been a careful specification of the thing being produced (products or services), a determination of the system location, a specification of the nature and sequence of processes and jobs, and a physical design or layout that relates the processes into a unified physical system. All these problems are concerned with strategic decisions of longer-term significance.

We now turn out attention to the shorter-term, day-to-day, month-to-month kinds of decisions that bear on operations planning and control. Given the physical system, what kinds of plans are necessary for effective operation, what kinds of controls are necessary, and what are the criteria by which we select from among alternative plans and controls? Part 4 is concerned with such questions as: What policies and procedures will guide us in setting basic activity rates? Should we hire or lay off personnel, and in what numbers? When is using overtime justified instead of increasing the size of the work force? When should we take the risk of accumulating seasonal inventories to stabilize employment? How big should inventories be to sustain the process? What policies and procedures are appropriate for controlling inventories and reordering materials? What policies and procedures are necessary to schedule workers and equipment for effective operations? Should the utilization of workers and machines be maximized, or is there a value to idleness? How do we maintain the reliability of the productive system so that specified quantity and quality are produced? When is a preventive maintenance policy justified? What policies and procedures can be effective in controlling labor and other costs? Is it possible to control costs, or is it really the activities that must be controlled?

OPERATIONS PLANNING AND CONTROL SYSTEMS

Figure 1 presents the broad relationships of the plans and controls for the operations phase of a productive activity. Physically, the productive process takes as inputs labor, materials, equipment and physical facilities, and energy, and converts these inputs into useful outputs of goods and services. Above the productive process shown in Figure 1, we have outlined the planning processes in block form. The basis for short-term plans centers in forecasts, and forecasting thus will be the subject of the next chapter.

Using forecasts, we must make aggregate plans and schedules. These aggregate plans set the basic activity and manpower levels in the short term—they represent short-term capacity decisions. Such decisions involve determining whether to enlarge or contract the size of the labor force, how much to use overtime capacity, and, if we are dealing with an inventoriable product, whether to build or draw down inventories. Aggregate planning methods (see Chapter 10) become the basis for other plans, such as raw material ordering and inventories, equipment, and detailed manpower schedules.

Below the productive process in Figure 1 are the systems for controlling the quality, quantity, and costs for the plans made. As in all kinds of control processes, we need a way of monitoring the aspects over which we wish to establish control, and we need standards for comparison. The control system makes comparisons, interprets results, and takes action to readjust processes to conform to standards. Management attempts to achieve a system optimum by means of various types of controls. Inventories are significant in many systems, and Chapter 11 considers the functions and control models for inventories. Industrial planning, scheduling, and control systems are the subject of Chapter 12, and Chapter 13 discusses project planning and control systems. Some of the special problems of service systems are covered in Chapters 14 and 15, and, finally, the concepts and methods for maintaining system reliability are discussed in Chapter 16.

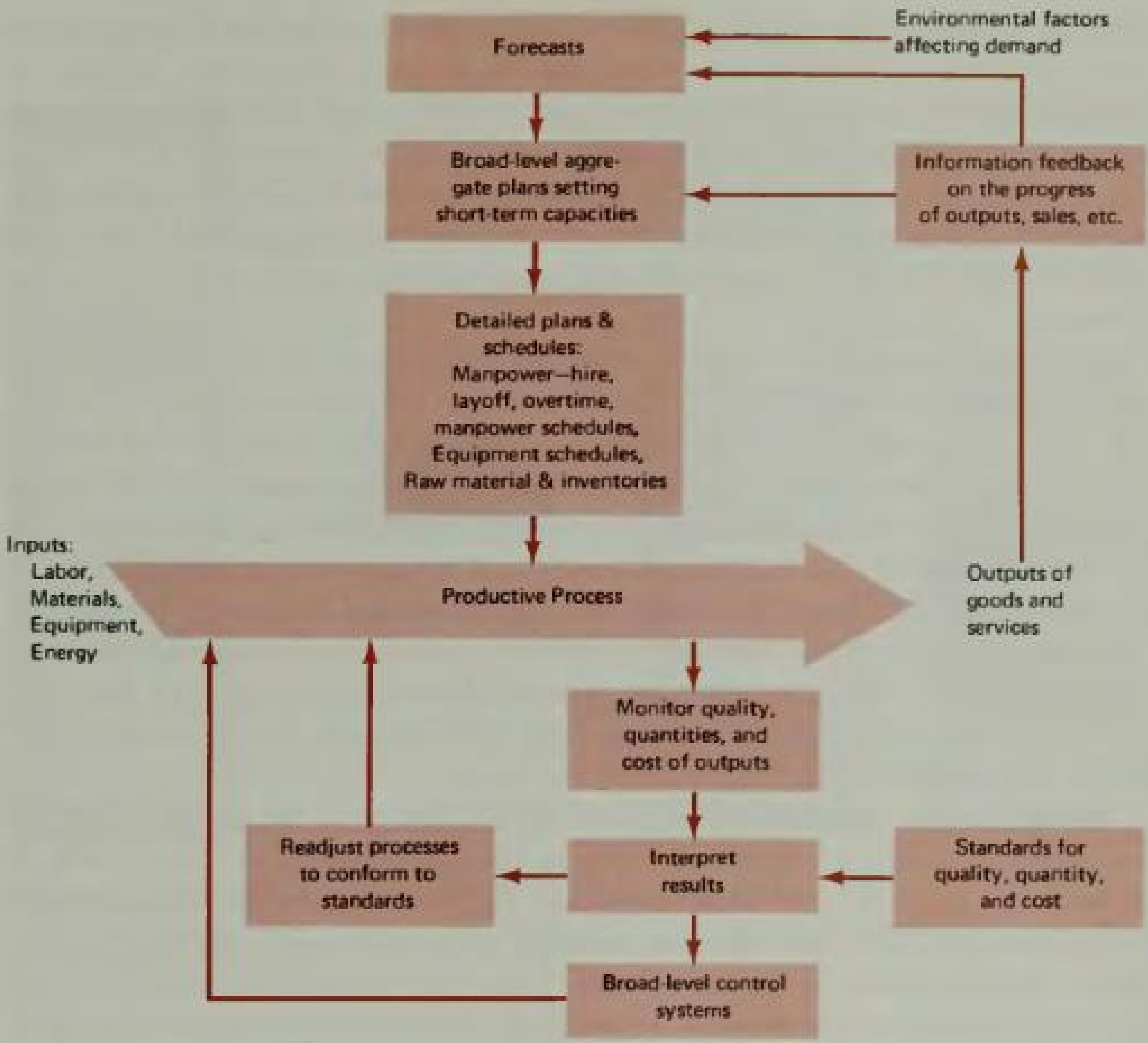


Figure 1. Operations planning and control systems

Classification of Systems

It is useful to establish a general classification of productive systems for the operations phase, rather than to continue discussing such systems in general, because the nature of the most important operating problems is quite different for different systems. We shall establish two bases of classification: (1) continuous versus intermittent systems, and (2) the output of inventoriable versus noninventoriable items.

Continuous verses Intermittent Systems. The continuous-intermittent basis of classification was used in discussing the models of productive systems in

Chapter 2 and throughout Part 3. Continuous systems are typified by production lines, continuous chemical processes, and, in general, production systems of enterprises that produce the high-volume standardized products for which our society is noted. There are also some nonmanufacturing systems that follow the continuous model, such as the cafeteria line and some large-scale office operations. The basis for the term *continuous*, however, is that the physical flow of whatever is being processed either is continuous or approaches continuous movement.

On the other hand, intermittent systems are those where the physical facilities must be flexible enough to handle a wide variety of products and sizes, or where the basic nature of the activity imposes changes of the design of the system outputs from time to time. In such instances, no single sequence pattern of processes is appropriate; therefore, the relative location of the operations must be a compromise that is best for all products. The emphasis is on flexibility of product design, processes, flow paths through the system, and so on. The system is termed *intermittent* because the flow is intermittent. Intermittent systems are characterized by custom- or job order type machine shops, batch type chemical operations, general offices, large-scale one-time projects, hospitals and other health care facilities, etc.

The basis of the continuous–intermittent classification is the general nature of the physical layout of facilities in the productive system. We now wish to broaden our conception of what constitutes the system to include the supply of raw materials at the input end of a manufacturing enterprise, and the distribution of finished goods at the output end. Thus, the term *production–inventory system* for the expanded concept includes the inventories throughout, and suggests a second basis for classification.

Systems for Inventoriable versus Noninventoriable Items. Figure 2 is a diagram of a production–inventory system for an inventoriable product. It emphasizes the broad flow characteristics of the system as a whole. In Figure 2, the manufacturing phase could be either continuous or intermittent. Of course, many intermittent manufacturing systems produce to inventory. Such systems may be arranged physically and scheduled internally as job shops. Common examples are the machine shops of the large automotive companies. Such shops are closed to job order from outside the organization, and may produce a set of products repetitively, in cycles. Equipment is time shared among many different products, and the products are produced to inventory because they are based on standard designs for which predictable markets exist. Such intermittent shops are called *closed job shops* because they are closed to outside job order.

Thus, we may have production–inventory systems in the inventoriable clas-

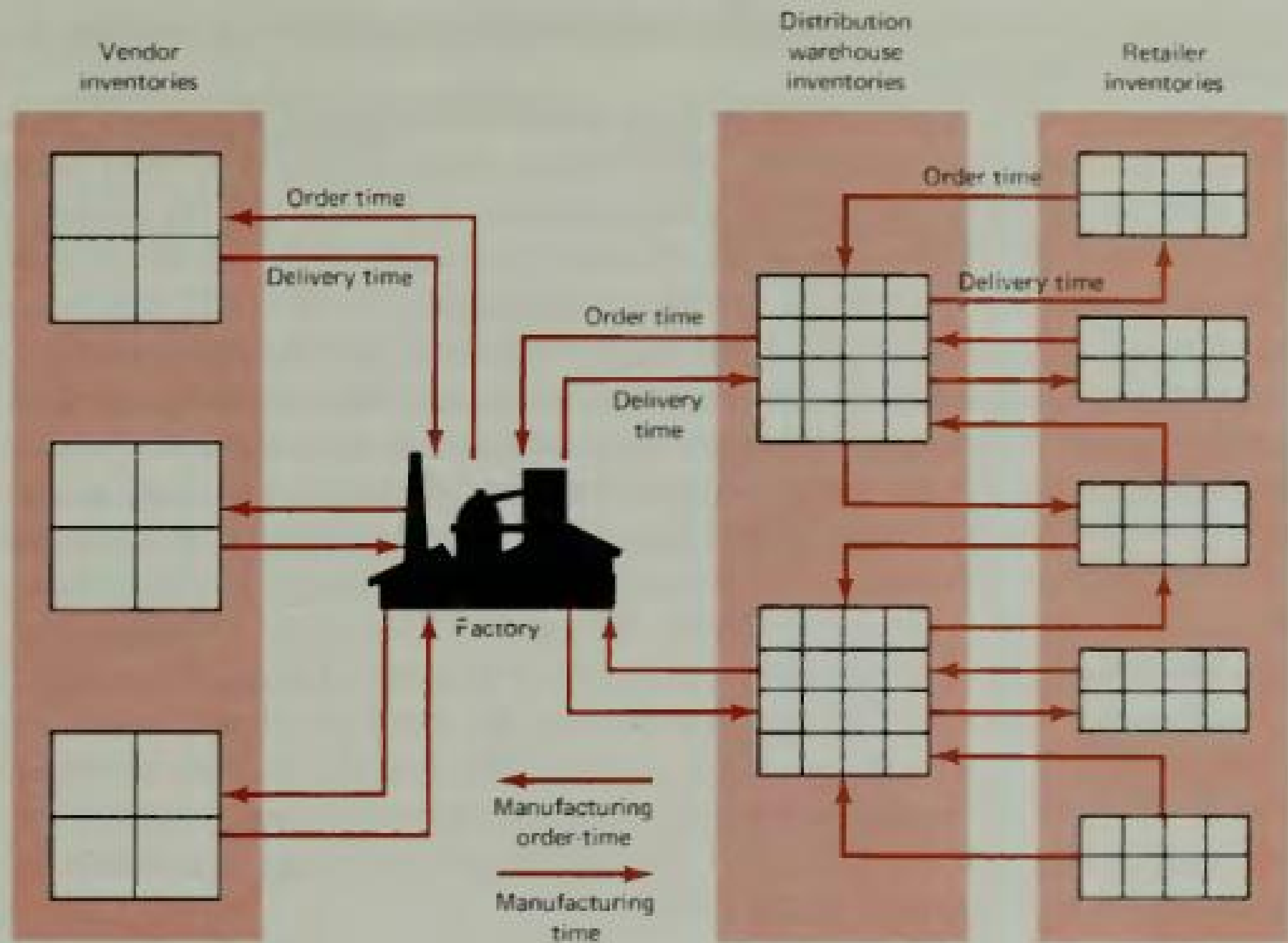


Figure 2. Supply-production-distribution system for an inventoriable product. Time lags are characteristic between and within each stage, and inventories serve vital functions at each stage

sification whose physical configuration may be continuous as well as intermittent. If we refer again to Figure 2, we find a commonly occurring enterprise specialization in the distribution of products. These enterprises specialize in the distribution of products and are called, simply, *distributors*. The portion of Figure 2 under managerial control for distribution systems focuses mainly on the inventories. The operations phase of such enterprises concerns the replenishment of inventories, the control of inventory levels, and shipment. Service rendered may be judged in terms of frequency of stock out or inability to supply the demands for products from inventories. Distribution systems are nearly pure inventory systems from an operations management viewpoint, and are of interest simply because they are common in society. Chapter 11 is devoted to inventories and inventory models. Thus, production-inventory systems for inventoriable items include systems in which production is continuous, systems in which production may be intermittent, and systems involving only distribution. It is important to group systems involving inventoriable items together because we can use inventories effectively in aggregate planning and programming, as discussed in Chapter 10.

Custom products and services require production-inventory systems that have no finished goods inventories. Nevertheless, many such systems may face inventory problems for raw materials and materials in process, and for supplies that may be used up as a part of the productive process. Open job shops and large-scale one-time projects are prominent examples. The open job shop produces to job order or contract. Since these orders or contracts may never be repeated, flexibility is a crucial requirement. Managerial attention is focused on scheduling and utilizing workers and machines to meet agreed due-dates and quality standards. The difference between the open job shop and the large-scale project in our classification is largely one of scale and complexity. There is no clear dividing line between the two systems, except that when a contract is large enough and complex enough to justify the special PERT-type planning and scheduling techniques discussed in Chapter 13, it is called a large-scale project.

Nonmanufacturing and service systems are the other prominent types of systems that produce no inventoriable output. The demand for their services must be met from current capacity, and inventories cannot be used as a means of buffering the fluctuations in demand. This fact has far reaching importance in planning and scheduling for such systems.

A Comparison. To summarize, the continuous-intermittent classification results in the following:

Continuous Systems	Intermittent Systems
Production-inventory systems for high-volume standardized products	Job shops (open and closed)
Distribution systems	Large-scale projects
Service and nonmanufacturing systems	Service and nonmanufacturing systems

This classification is most useful for designing the layout of physical facilities and for indicating the nature of detailed scheduling of both workers and machines.

The inventoriable-noninventoriable item classification results in the following:

Systems for Inventoriable Items	Systems for Noninventoriable Items
Continuous high-volume standardized product systems	Open job shop systems
Closed job shop systems	Large-scale projects
Distribution systems	Service and nonmanufacturing systems

This classification is most useful in determining the nature of appropriate aggregate planning and programming. The production–inventory-system concept also is useful in helping us focus on the system as a whole, thus helping us avoid the organization suboptimization that can result if the organization is segmented into producing and distributing functions.

Chapter 9

Demand Forecasting for Operations

Planning and control for operations requires a crucial input regarding the forecast of what we expect to happen. Its orientation is relatively current, since we need to make plans that range from a day-to-day to a year-to-year basis. Although the planning horizon depends on the particular application, it is relatively short; and we need forecasting methods that can be adapted to the needs of current, and often detailed, information.

Table 1 summarizes some of the most pertinent and useful forecasting methods under three main headings. In Chapter 5, we discussed the predictive methods in connection with predicting markets and volumes for new product introductions and longer-range predictions. The causal methods are most appropriate for short- and medium range forecasting, particularly as a basis for the aggregate planning methods that we shall discuss in Chapter 10. The time series models apply to shorter-range forecasts for operations, and are particularly valuable for inventory and production control. (See Chambers, Mullick, and Smith [1971] for an excellent survey of methods and applications.)

Requirements of Forecasts for Operations

The demand forecasting function serves many broad managerial purposes in both profit and nonprofit kinds of organizations. If it is to be useful for operations planning and control, demand forecast data must be available in a form

TABLE 1
METHODS OF PREDICTION AND FORECASTING

Method	*General Description*	*Application*	*Relative Cost*	*References**
Predictive methods:				
Delphi	Expert panel answers a series of questionnaires where the answers of each questionnaire are summarized and made available to the panel to aid in answering the next questionnaire.	Long-range predictions, new products and product development, market strategies, pricing and facility planning.	Medium-High	
Market surveys	Testing markets through questionnaires, panels, surveys, tests of trial products, analysis of time series.	Same as above.	High	
Historical analogy and life cycle analysis	Prediction based on analysis of and comparison with growth and development of similar products. Forecasting new product growth based on the S-curve of introduction, growth, and market saturation.	Same as above.	Medium	
Causal forecasting methods:				
Regression analysis	Forecasts of demand related to economic and competive factors which control or cause demand, through least squares regression equation.	Short- and medium range forecasting of existing products and services. Marketing strategies production and facility planning.	Medium	Benton; Chambers, Mullick, and Smith; Huang; Wheelwright and Makridakis; Parker and Segura
Econometric models	Based on a system of interdependent regression equations.	Same as above.	High	Chambers, Mullick, and Smith; Huang; Wheelwright and Makridakis
Time series forecasting models:				
Moving averages	Forecast based on projection from time series data smoothed by a moving average, taking account of trends and seasonal variations. Requires at least two years of historical data.	Short-range forecasts for operations such as inventory, scheduling, control pricing, timing special promotions.	Low	Brown, 1963
Exponential moving averages	Similar to moving averages, but averages weighted exponentially to give more recent data heavier weight. Well adapted to computer application and large numbers of items to be forecast. Requires at least two years of historical data.	Same as above.	Low	Benton, Box and Jenkins; Brown; Wheelwright and Makridakis

*References refer to items in reference list at the end of chapter.

that can be translated into demands for: specific items of material; time in specific equipment classifications; specific labor skills; etc. Therefore, forecasts of gross dollar demand, demand by customer or client classification, or demand by broad product or service classifications are of limited value for the planning and control of inventories, use of labor and equipment, and operations in general.

Since planning and control for operations necessarily must take place at several different levels, it is unlikely that one kind of forecast can serve at all levels. The immediate problems are controlling inventories, providing raw materials and labor required for current production programs, and planning the use of workers and machines on a day-to-day, week-to-week, month-to-month basis. However, because we must look somewhat further into the future to provide new capacity or a different kind of capacity, forecasts of different time spans are required as the basis for operating plans. These plans are: (1) plans for current operations and for the immediate future; (2) intermediate-range plans to provide for the required capacities of workers, materials, and equipment for the next one to twelve months; and (3) long-range plans for capacity, locations, changing product and service mix, and the exploitation of new products and services. We already discussed the longer-range plans in connection with productive system design. Our present interest is in forecasting needs that can serve intermediate-range plans and plans for current operations.

Components of Demand

Figure 1 shows the three-year record of the demand for a computing service. Briefly, the company involved derives income from the use of its proprietary programs, and the forecasting of usage can have an important bearing on many factors in planning operations. In general, this figure shows an increasing demand for the service, with considerable variation in demand. In attempting to construct forecasting models, we are interested in the kinds of variation.

Some of the variation in demand can be explained by a statistical model and some cannot. The inexplicable variations we call *random* variations. In forecasting demand for operations, we would like not to respond to what might be simply random variations. Some of the variations in Figure 1 certainly fall into the random category. Since many different customers use the company's program to suit their own schedules and needs, many of the seemingly odd increases or decreases from month to month come from this source. In using the programs, company *A* may decide to initiate a new planning cycle that may have

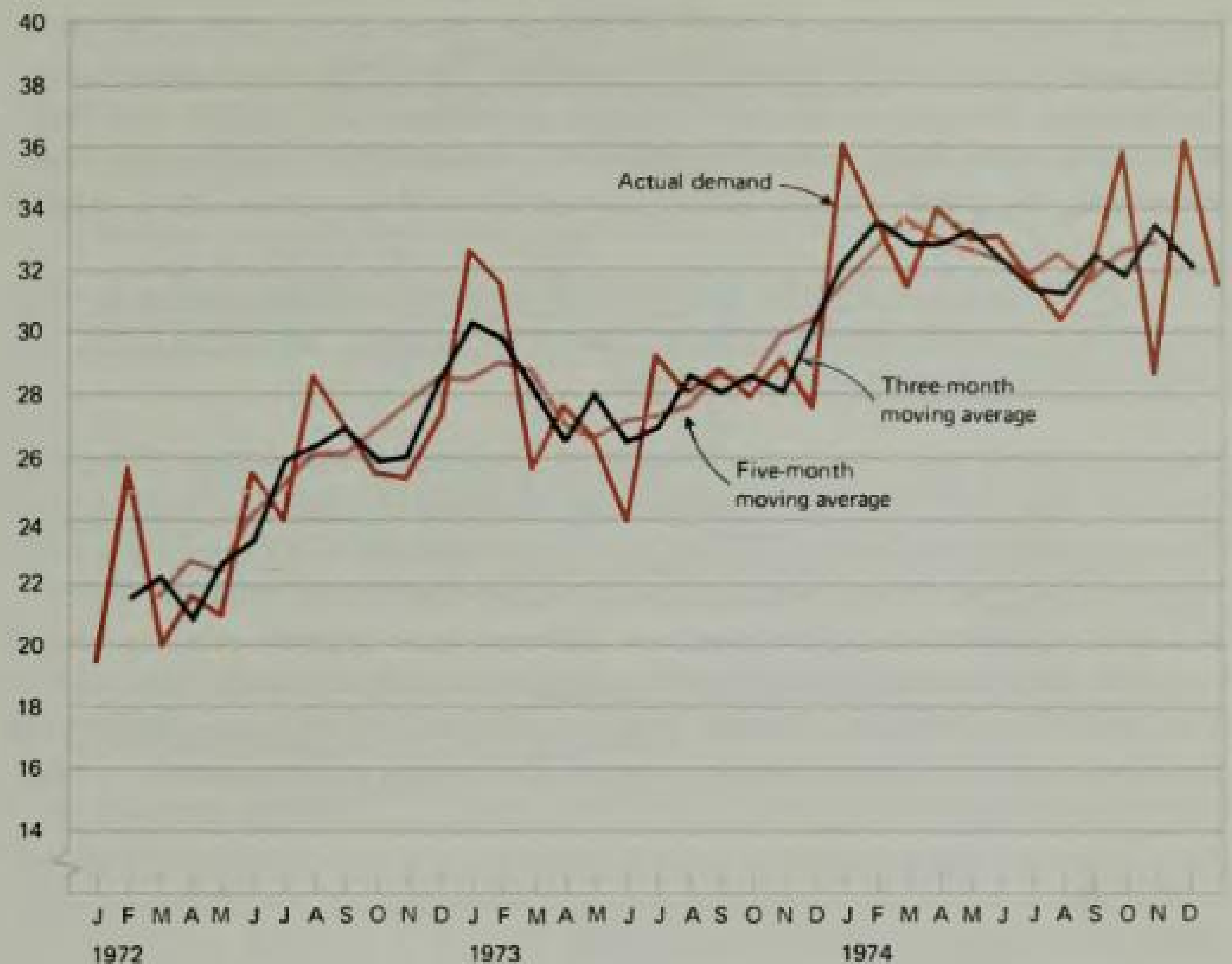


Figure 1. Actual monthly program usage with three- and five month moving averages shown

affected demand in August 1972, while Company *B* may shut down for two weeks in August for vacations. The aggregation of all users' needs produces this kind of unexplained, or random, variation. Trying to explain each month's demand is probably not worth the time and cost.

Although some of the variation in Figure 1 is not explicable, the aggregation of usage by all customers does produce a logical pattern. In spite of the random variations, there is some *average* demand for any particular period. For example, the average for the first quarter of 1972 is 21.51; for the first six months it is 22.03; and for the full year of 1972 it is 24.06. Each of these measures of average demand has some meaning, although they have limited usefulness in the form given.

As we already noted, Figure 1 obviously illustrates an upward *trend* in the demand for program usage; and the averages for the first quarter, first six months, and the entire year verify an upward trend during 1972. The same conclusion is not quite so clear in 1973, however. The comparable figures for the first quarter, first six months, and full year of 1973 are: 29.72, 27.80, and 28.03. Each comparable figure indicates a growth in demand from year to year; however, both the graph and the 1973 figures indicate that demand may exhibit *seasonal* variation.

The components of demand that we need to consider in forecasting models are random, average levels, trend, and seasonal. Cyclical variations related to the business cycle are beyond our scope.

Autocorrelation is an important concept in determining the nature of the raw demand data and the kind of forecasting model that might be appropriate. Autocorrelation measures the degree to which demands in different time periods are correlated. For example, a strong trend in demand data would result in a high correlation among demands in different time periods. Also, a strong seasonal pattern would produce intercorrelation among demands in Periods 1, 13, 25; 2, 14, 26; 3, 15, 27, and so on. Analyses of autocorrelation coefficients are therefore used to determine the best type of forecasting model to use.

TIME-SERIES FORECASTING METHODS

Moving Averages

When we look at the actual demand in Figure 1, the average demand for the entire three years is meaningless, since both trend and seasonal components are important factors. Therefore, we need a methodology that emphasizes the recent experience and estimates the trend effect. Furthermore, the estimate of demand for the upcoming period must take account of the expected seasonal variation. Finally, we wish not to be influenced by the random variations in demand when making an estimate for the upcoming period; or, preferably, we may wish to state the expected demand with probable limits of variation from the expected value due to random causes.

The common way to smooth the effects of random variations in demand is to estimate average demand by some kind of moving average. Table 2 gives some sample demand data taken from Figure 1 for the first six months. Both Figure 1 and Table 2 show that actual demand is quite variable. The three-month moving average, however, is much more stable because the demand for any one month

TABLE 2
ACTUAL DEMAND FOR PROGRAM USAGE AND
THREE- AND FIVE-MONTH MOVING AVERAGES

Single moving average

Date		Actuals	Three-month Moving Average	Five-month Moving Average
1977	Jan.	19.36	—	—
	Feb.	25.45	21.51	—
	Mar.	19.73	22.22	21.36
	Apr.	21.48	20.66	22.57
	May	20.77	22.56	22.24
	June	25.42	23.33	23.96
	July	23.79	25.85	25.03
	Aug.	28.35	26.31	—
	Sept.	26.80	—	—

receives only one-third weight. Extreme values are discounted; if they are simply random variations in demand, and if we gauge demand by the three-month moving average, we are not strongly influenced by them.

As indicated in Table 2, we can calculate the moving average by averaging the first *n* items of data into a single number (where *n* is the number of items of data to be averaged). We then increment the grouping by one data-item, as shown by the successive brackets in Table 2 for the three-month moving average. For example, the three-month moving average for February is 21.51, which represents the average of the actual demands for January, February, and March.

When choosing the size of *n*, we must remember that the larger the value of *n*, the greater the dampening effect. While a large *n* will show the overall trend, it can obliterate seasonal cycles. On the other hand, smaller values of *n* result in very little smoothing effect, and the moving average tends to resemble the original data.

These comparative effects are shown in Table 2, where both three- and five-month moving averages are computed. Figure 1 shows three- and five-month moving averages plotted in comparison with actual demand for monthly program usage. Note the smoothing effects and how the moving average lines reveal the trend and seasonal components in the data. The five-month moving average produces a greater smoothing effect, as we would expect. However, a conflict of objectives can be observed here. The five-month moving average discounts random effects more successfully, but the three-month moving average gives more weight to the most recent data. Since the program usage data in Figure 1 shows both trend and seasonal components, we have a keen interest in emphasizing the most current data in the moving average.

EXPONENTIALLY WEIGHTED MOVING AVERAGES

One effective and convenient method of accomplishing differential weighting and smoothing is by exponentially weighted moving averages. The simplest exponential smoothing model estimates a smoothed average demand for the upcoming period $\bar{F}_t$ by adding or subtracting a fraction (α) of the difference between actual current demand D_t and the last smoothed average $\bar{F}_{t-1}$. The new smoothed average $\bar{F}_t$ is then:

New smoothed average = old average + α(new demand–old average)

Or stated symbolically,

$$\bar{F}_t = \bar{F}_{t-1} + \alpha(D_t - \bar{F}_{t-1}) \tag{1}$$

The smoothing constant α is between 0 and 1 with commonly used values of 0.01 to 0.30. Equation 1 can be rearranged in a more convenient, and possibly more understandable, form, as follows:

New smoothed average = α(new demand) + $(1 - \alpha)$(old average)

Or stated symbolically,

$$\bar{F}_t = \alpha D_t + (1 - \alpha)\bar{F}_{t-1} \tag{2}$$

If $\alpha = 0.10$, then Equation 2 says that the smoothed average in the upcoming period $\bar{F}_t$ will be determined by adding 10 percent of the new actual demand information D_t, and 90 percent of the last smoothed average $\bar{F}_{t-1}$. Since the new demand figure D_t includes possible random variation, we are discounting 90 percent of those variations. Obviously, then, small values of α will have a stronger smoothing effect than large values. Conversely, large values of α will react more quickly to real changes (as well as random variations) in actual demand. The choice of α normally is guided by judgment, though studies could produce economically best or near-best values of α. Berry and Bliemel [1974] show how computer search techniques can be used to determine optimum values of α.

Equation 2 gives weight to all past actual demand data, although this is not obvious. This weighting occurs through the chain of periodic calculations to produce smoothed averages for each period. In Equation 2, for example, the term $\bar{F}_{t-1}$ was computed from

$$\bar{F}_{t-1} = \alpha D_{t-1} + (1 - \alpha)\bar{F}_{t-2}$$

which includes the previous actual demand D_{t-1}. The $\bar{F}_{t-2}$ term was calculated

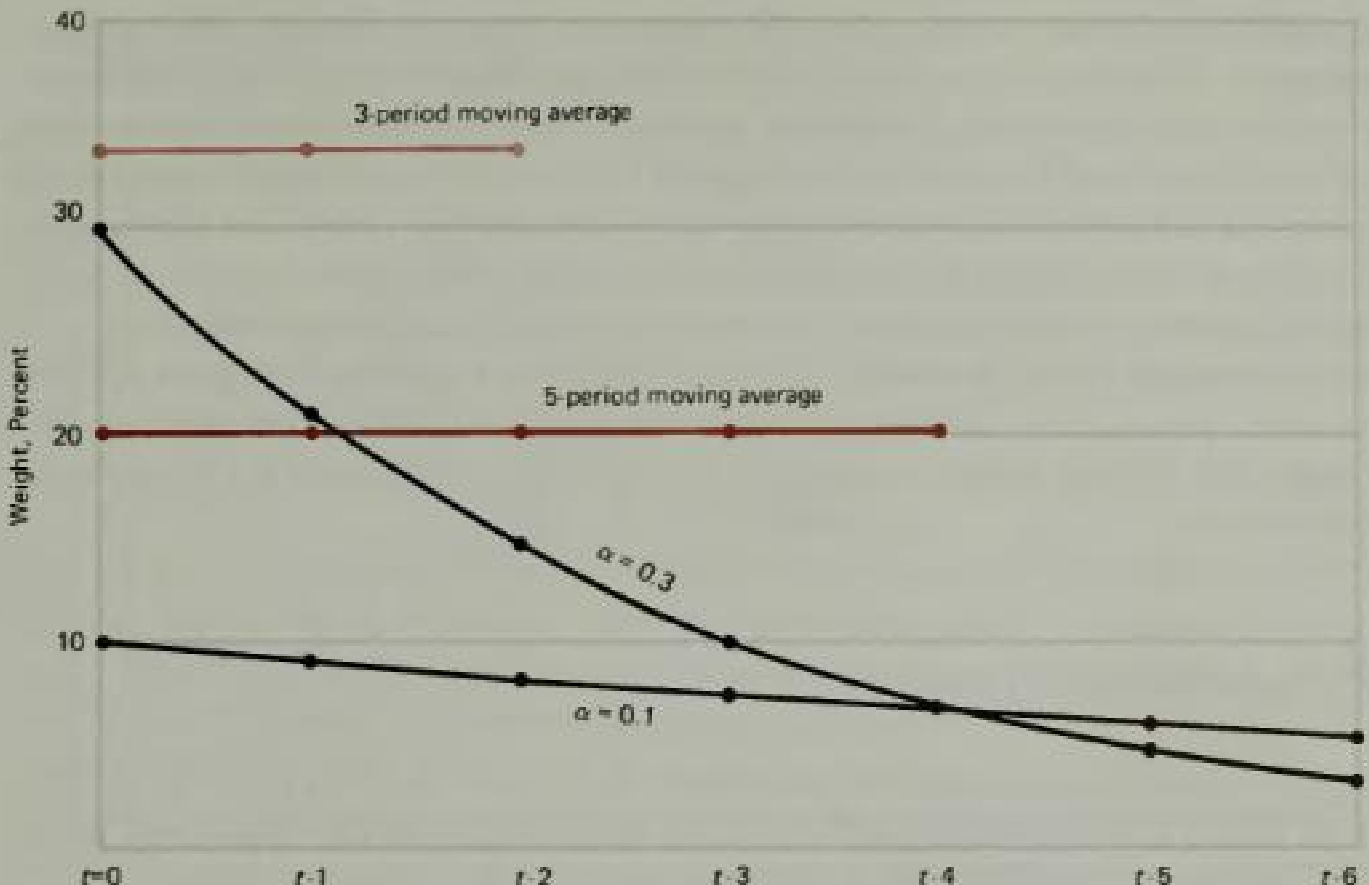


Figure 2. Comparative weightings given past data by 3- and 5- period moving averages and by exponentially weighted moving averages with $\alpha = 0.1$ and 0.3

in a similar way, which included D_{t-2}, and so on back to the beginning of the series. Therefore, the smoothed averages are based on a sequential process representing all previous actual demands.

Figure 2 shows comparative weightings given data by three- and five period moving averages, and by exponentially weighted moving averages with $\alpha = 0.1$ and 0.3. Note the effectiveness of the exponentially weighted averages in placing heavier weight on the most recent data. Another factor implicit in Figure 2 is that exponentially weighted data give a weight to all prior data, although the effect of old data will be small. Simple moving averages give weight only to the periods included in the average. It also is worth noting at this point that the exponential forecasting system is a direct application of the negative feedback concepts discussed in Chapter 3. Equation 2 is a feedback equation in which a fraction of the output is fed back with lag to stabilize the output. The feedback concept enables the forecast produced by exponential smoothing to be stabilized, and Chapter 3's general comments about the behavior of such systems apply directly here.

It is important to place the time periods for $\bar{F}_t$, D_t, and $\bar{F}_{t-1}$ in perspective, and to recognize that the so-called new smoothed average $\bar{F}_t$ is not an extrapolation beyond known demand data. Rather, it is the most current smoothed average. In a sense, then, it is not a forecast but a statement of current demand. As a parallel example, consider driving a car under three conditions of guidance: (1) with a view out the windshield; (2) with the windshield covered but with a view out the side windows; and (3) with the windshield and the side windows covered and a view out the back window, using the rear view mirror. Our situation roughly compares with being able to look out the side windows to see where we actually are. However, since there must be some time lag in compiling information on actual demand D_t, we are really looking out the *rear* side windows. However, no guesses, projections, or extrapolations are involved. Each updated figure is based only on past actual demand data, weighted to emphasize the most recent experience.

Extrapolation and Forecast

Since no trend or seasonality is included in the model, direct extrapolation from $\bar{F}_t$ to infer a forecast is justified. Therefore, the forecast for the upcoming period D^*_{t+1} is taken directly as the computed value of $\bar{F}_t$. (Starred symbols (*) will represent extrapolated or forecasted values.) Table 3 shows computations and forecasts for the first six months of the program usage data.

Trend Effects

If a trend were present in the data, Equation 2 would respond to it—but with a lag. However, the *apparent* trend for each period is simply the difference

TABLE 3
SAMPLE COMPUTATIONS FOR $\bar{F}_t$ AND THE FORECAST, D^*_{t+1}, FOR THE SIMPLE EXPONENTIAL SMOOTHING MODEL*

Date	Actuals	Smoothed Average $\bar{F}_t$	Forecast, D^*_{t+1}
Initial	–	23.0	–
1972, Jan.	19.36	22.27	–
Feb.	25.45	22.91	22.27
Mar.	19.73	22.27	22.91
Apr.	21.48	22.11	22.27
May	20.77	21.84	22.11
June	25.42	22.56	21.84
July	–	–	22.56

*Data for actuals from Table 2, $\alpha = 0.2$.

between the last two smoothed averages, $\bar{F}_t - \bar{F}_{t-1}$. This difference represents another series that can be estimated and smoothed by exponentially weighted averages, just as with average demand. Therefore, the new average trend adjustment $\bar{T}_t$ is

$$\bar{T}_t = \alpha(\bar{F}_t - \bar{F}_{t-1}) + (1 - \alpha)\bar{T}_{t-1} \tag{3}$$

The expected demand $E(D_t)$ for the current period, including trend adjustment is the smoothed average $\bar{F}_t$ computed in Equation 2 plus a fraction of the new average trend adjustment $\bar{T}_t$ computed in Equation 3:

$$E(D_t) = \bar{F}_t + \left(\frac{1 - \alpha}{\alpha}\right)\bar{T}_t \tag{4}$$

The term $(1 - \alpha)/\alpha$ corrects for lag in the trend adjustment $\bar{T}_t$ (see Brown [1959, pp. 192–196].* Note that the only facts and figures required to update a forecast are α, D_t, $\bar{F}_{t-1}$, and $\bar{T}_{t-1}$; yet all past actual demand data are weighted in the model, but with rapidly decreasing weights as we proceed backward in time. If a large number of items were to be forecast (as commonly happens), the entire process could be computerized, requiring minimum computer storage for each item. The only new data that would be supplied to the computing program each period would be the current actual demand D_t.

Extrapolation and Forecast. As with the no-trend model, Equation 4 involves no extrapolation beyond known demand data. To extrapolate from Equation 4 to forecast D^*_{t+1} requires that we add $\bar{T}_t$, the most recent average trend adjustment,

Predecessor

$$D^*_{t+1} = E(D_t) + \bar{T}_t = \bar{F}_t + \frac{1}{\alpha}\bar{T}_t \tag{5}$$

To forecast k periods in the future, we add $k\bar{T}_t$ to $E(D_t)$, or

$$D^*_{t+k} = E(D_t) + k\bar{T}_t = \bar{F}_t + \left(\frac{1}{\alpha} + k - 1\right)\bar{T}_t \tag{5a}$$

Table 4 shows sample computations of smoothed average, trend adjustments, and forecast for an exponential smoothing model with $\alpha = 0.2$. Figure 3 shows a graph of exponentially smoothed forecasts with and without trend adjustment for the data of the computer program usage. Note that the two forecasts are similar in form; however, since the actual data contain considerable trend, the forecast without trend adjustment lags behind and is lower than the forecast with trend adjustment. If the actual data were to stabilize, exhibiting no trend, the two forecasts would tend to converge.

More sophisticated exponential forecasting models have been developed in

*The lag factor for $\alpha = 0.1$ is $(1 - 0.1)/0.1 = 9$. This seems large, but it actually indicates how sluggish the forecasting system is when small smoothing constants are used. When $\alpha = 0.4$, the lag factor is $(1 - 0.4)/0.4 = 1.5$.

TABLE 4

SAMPLE COMPUTATIONS OF SMOOTHED AVERAGE, TREND ADJUSTMENTS, AND FORECAST FOR AN EXPONENTIAL SMOOTHING MODEL ($\alpha = 0.2$)

Date	Actuals, D_t	Smoothed Average, $\bar{F}_t = \alpha D_t + (1-\alpha)\bar{F}_{t-1}$	Current Apparent Trend, $\bar{F}_t - \bar{F}_{t-1}$	Average Trend Adjustment, $\bar{T}_t = \alpha(\bar{F}_t - \bar{F}_{t-1}) + (1-\alpha)\bar{T}_{t-1}$	Forecast for period $t+1$, $D^*_{t+1} = \bar{F}_t + \frac{1}{\alpha}\bar{T}_t$
Initial	–	23.00	–	1.000	–
1972, Jan.	19.36	22.27	−0.728	−0.146	–
Feb.	25.45	22.91	0.636	0.011	21.54
Mar.	19.73	22.27	−0.636	−0.119	22.96
Apr.	21.48	22.11	−0.158	−0.127	21.68
May	20.77	21.84	−0.269	−0.155	21.48
June	25.42	22.56	0.715	0.019	21.07
July	–	–	–	–	22.65

which the α parameters used for $\bar{F}_t$ in Equation 2 and $\bar{T}_t$ in Equation 3 are set individually to minimize measures of forecast errors. When D_t is subject to rapid change, these and other variations of the basic exponential smoothing models may provide a more stable response. See Brown [1963], Groff [1973], and Winters [1960].

Seasonal Adjustments

The basis for including a seasonal adjustment to an exponentially smoothed forecast model is to develop a base series that represents the seasonal cycle. The base series usually is constructed from last year's experience in some way. If the seasonal pattern is strong and relatively invariant, then the base series could be simply the period-by-period demand for last year, or an idealized cycle. If the peaks and valleys shift forward or backward slightly from year to year, then an averaging process may be used, such as a three-month moving average centered on the month for which the average is being determined, as shown in Table 2.

The general methods used are similar to the trend model, although they exhibit more complexity in order to handle the seasonal component.

See Brown [1959] and Buffa and Taubert [1972] for computed examples. Figure 4 shows an exponentially smoothed forecast with trend and seasonal adjustments for the program usage data.

Comparing the forecasts shown in Figures 3 and 4 for the usage data, it appears that the addition of trend adjustment, and both trend and seasonal

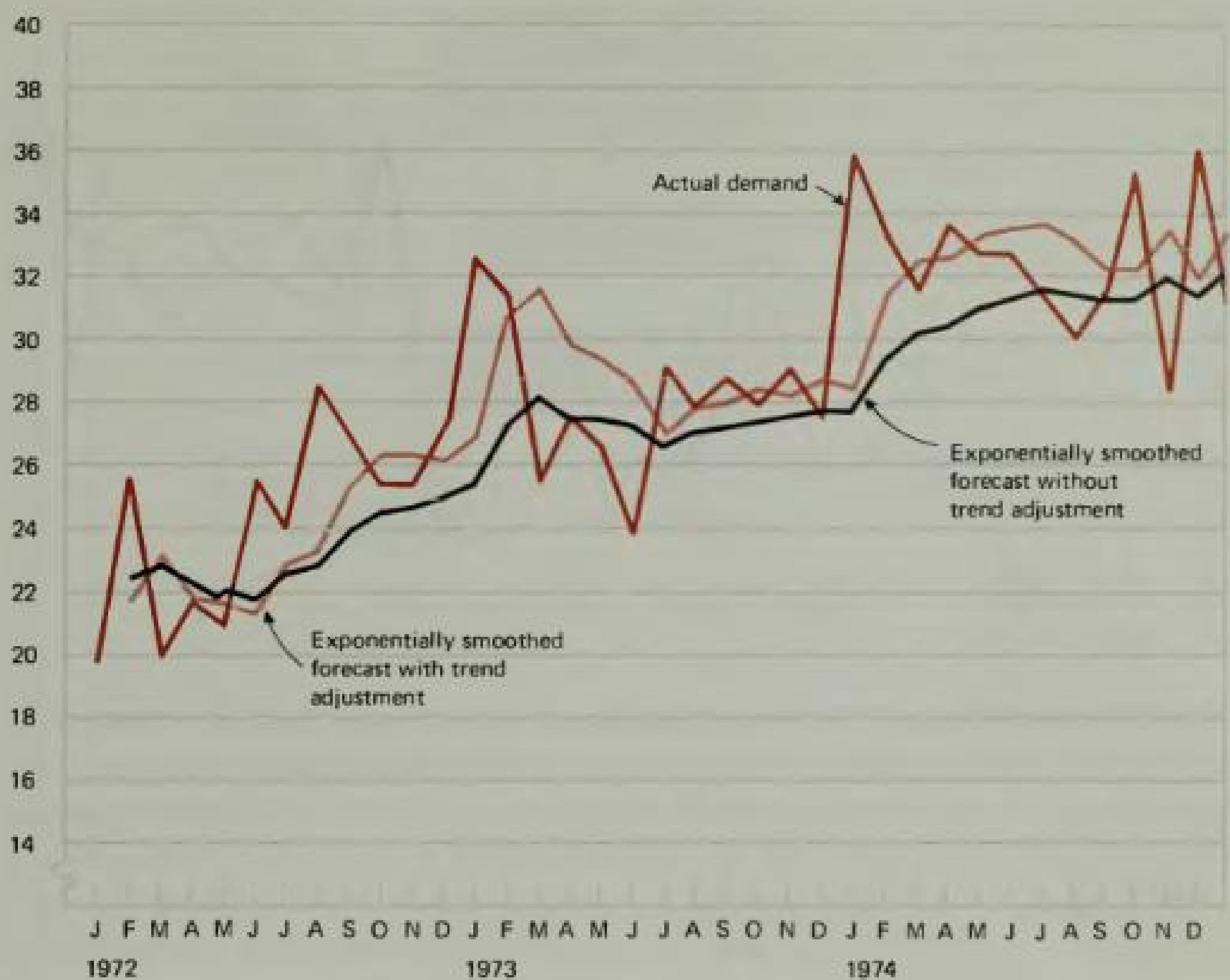


Figure 3. Forecasts of program usage by exponential smoothing models with and without trend adjustment

adjustments, progressively produce better forecasts. A common practical measure of the forecast accuracy is the Mean Absolute Deviation (MAD). MAD is simply the sum of the absolute deviations between actual demand and forecasts, divided by the number of observations. The MAD measures for the simple model, the model with trend, and the model with trend and seasonal adjustments, are 2.38, 2.36, and 1.74, respectively. These figures indicate that forecasting accuracy has been increased with the successive addition of the adjustments.

Adaptive Methods

As we have noted, it is common to use fairly small values of α in exponential smoothing systems in order to filter out random variations in demand. When actual demand rates increase or decrease gradually, such forecasting systems

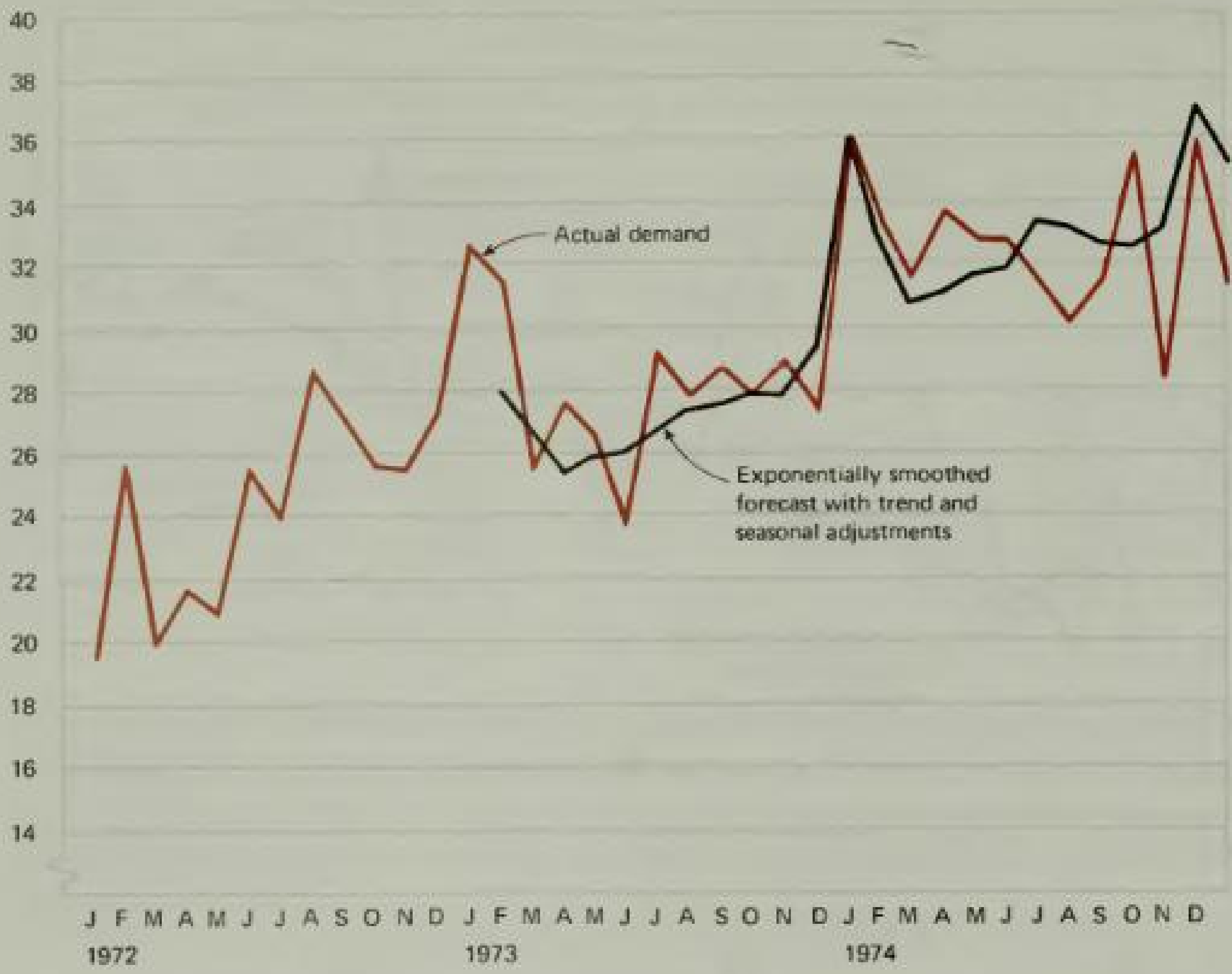


Figure 4. Forecast of program usage using an exponential smoothing model with trend and seasonal adjustments ($\alpha = 0.1$)

can track the changes rather well. If demand changes suddenly, however, a forecasting system using a small value of α will lag behind the actual change substantially. Thus, adaptive response systems have been proposed.

The basic idea of adaptive systems is indicated in Figure 5, where the forecast error is monitored and, based on preset rules, the weights assigned to past data are adjusted by changing the parameter α. The forecast then is computed using the new weights. If, for example, a step change in demand were to occur because of a radical change in the market, a large error would result. The large error signals that α should increase, giving greater weight to current demand. The forecast would then reflect the change in actual demand. When actual demand stabilized at the new level, adaptive systems reset the value of α to a lower level, which would filter out random variations effectively. A number of approaches to adaptive response systems have been proposed. Eilon and Elmaleh [1970] and Roberts and Reed [1969] have proposed systems in which the

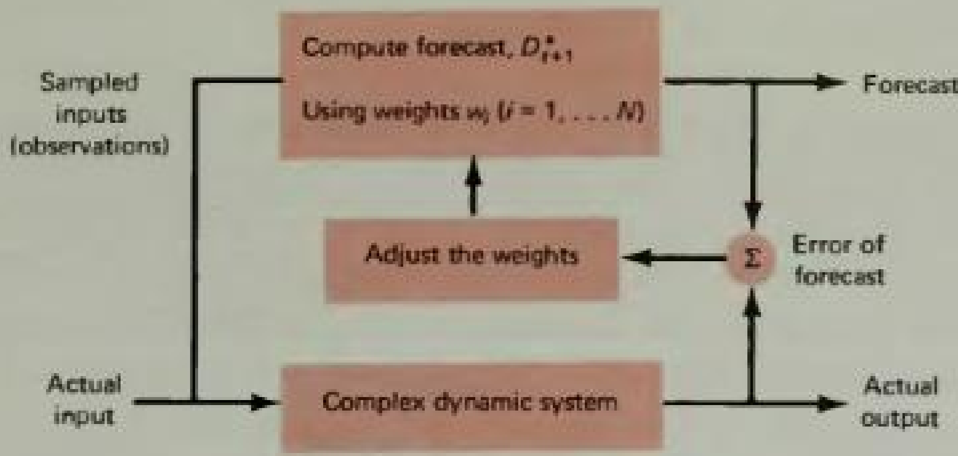


Figure 5. General structure of adaptive forecasting models, where the weights assigned to past actual demand data are adjusted depending on the forecast error

SOURCE: S. C. Wheelwright and S. Makridakis, *Forecasting Methods for Management* (New York: John Wiley & Sons, 1973)

parameter α is reset periodically. Trigg and Leach [1967] and Whybark [1972] have proposed continuous tracking systems that can reset the parameter each period. Montgomery [1970] has proposed adaptive control of the smoothing parameters by evolutionary operation, and Rao and Shapiro [1970] have proposed a system using evolutionary spectra. Whybark [1972] has made a comparative study of the performance of alternate adaptive systems.

CAUSAL FORECASTING METHODS

When we have enough historical data and experience, it may be possible to relate forecasts to the factors in the economy that *cause* the trends, seasonals, and fluctuations. Thus, if we can *measure* the causal factors, and we have determined their *relationships* to the demand for product or service of interest, then we can compute forecasts of considerable accuracy.

The factors that enter causal models are of every conceivable type—Gross National Product, disposable income, new marriages, housing starts, inventories, cost of living indexes—as well as predictions of dynamic factors and/or disturbances, such as strikes, actions of competitors, sales promotion campaigns, etc. The causal forecasting model expresses mathematical relationships between the causal factors and the demand for the item being forecast, and is indeed the most sophisticated of forecasting tools. As indicated in Table 1, there are two general types of causal models, and the costs range from medium to high for installation and operation.

Regression Analysis

Forecasting based on regression methods establishes a forecasting function called a regression *equation*. The regression equation expresses the series to be forecast, such as dollar sales or quantities sold, in terms of other series that presumably control the sales, or cause them to increase or decrease. The rationale can be general or specific. For example, in furniture sales, we might postulate that sales generally are related to disposable personal income—if disposable income is up, sales will increase, and if people generally have less money to spend, sales will go down. The empirical relationship is established through the regression equation. More specifically, we might postulate that furniture sales are controlled to some extent by the number of new marriages and/or the number of new housing starts. These both are specific indicators of possible demand for furniture.

Table 5 gives data on these three independent variables—housing starts, disposable income, and new marriages—and on sales of a hypothetical furniture company called the Cherryoak Company. We propose to build a relationship between the observed variables and company sales, where sales are dependent, or caused by, the observed variables. Therefore, sales are termed the dependent variable, and the observed variables are called the independent variables. The correlation coefficients between sales (S) and each of the independent variables are:

1. Disposable personal income (I)	0.805
2. Housing starts (H)	0.435
3. New marriages (M)	0.416

Since disposable income (I) correlates most strongly with company sales, let us use it as an example. Using regression analysis, we can determine the straight line that best fits the data expressing the relationship between sales (S) and disposable income (I). From statistics, we know that the regression equation represents a straight line that minimizes the square of the deviations from it, and sets the sum of the simple deviations to zero. The regression equation for the data of company sales (S) versus disposable income (I) is

$$S = 72.5 + 0.23I \tag{6}$$

where the coefficient, 72.5, is the y-axes intercept and the slope of the straight line is 0.23. Note that the form of the equation is in the standard format of the equation of a straight line, $y = a + bx$, where y is the dependent variable, x the independent variable, a the y intercept, and b the slope. In regression analysis, a

H = 3 variables

TABLE 5

DATA FOR 24 YEARS (1947–1970) USED IN PERFORMING REGRESSION ANALYSIS TO FORECAST 1971 SALES OF CHERRYOAK COMPANY

depend

Year	Housing Starts (H) [Thousands]	Disposable Personal Income (I) [$ Billions]	New Marriages (M) [Thousands]	Company Sales (S) [$Millions]	Time, T
1947	744	158.9	2.291	92.920	1
1948	942	169.5	1.991	122.440	2
1949	1.033	188.3	1.811	125.570	3
1950	1.138	187.2	1.580	110.460	4
1951	1.549	205.8	1.667	139.400	5
1952	1.211	224.9	1.595	154.020	6
1953	1.251	235.0	1.539	157.590	7
1954	1.225	247.9	1.546	152.230	8
1955	1.354	254.4	1.490	139.130	9
1956	1.475	274.4	1.531	156.330	10
1957	1.240	292.9	1.585	140.470	11
1958	1.157	308.5	1.518	128.240	12
1959	1.341	318.8	1.451	117.450	13
1960	1.531	337.7	1.494	132.640	14
1961	1.274	350.0	1.527	126.160	15
1962	1.327	364.4	1.547	116.990	16
1963	1.469	385.3	1.580	123.900	17
1964	1.615	404.6	1.654	141.320	18
1965	1.538	436.6	1.719	156.710	19
1966	1.488	469.1	1.789	171.930	20
1967	1.173	505.3	1.844	184.790	21
1968	1.299	546.3	1.913	202.700	22
1969	1.524	590.0	2.059	237.340	23
1970	1.479	629.6	2.132	254.930	24
1971		634.4		295.99	

SOURCE: *Statistical Abstract of the United States* (Washington, Bureau of the Census).

Note: Company sales and disposable per-capita income have been adjusted for the effect of inflation and appear in constant 1959 dollars.

SOURCE: G. G. C. Parker and E. L. Segura, "How to Get a Better Forecast," *Harvard Business Review*, March–April 1971.

and *b* are termed the regression coefficients and are the parameters that specify the equation.

The regression line is plotted in Figure 6, showing some specific points for selected years. These points illustrate the kinds of forecast errors that would have resulted if we had used this equation to forecast Cherryoak furniture sales. To use the regression equation to forecast sales, we simply insert the value of *I* and compute sales S. For example, if $I = 700$, then the forecaster could compute the value of S as $S = 72.5 + 0.23 \times 700 = \$233.5$ (million).

24 observations
4 variables

Cs = H + DPI + NMar.
3 independent variable

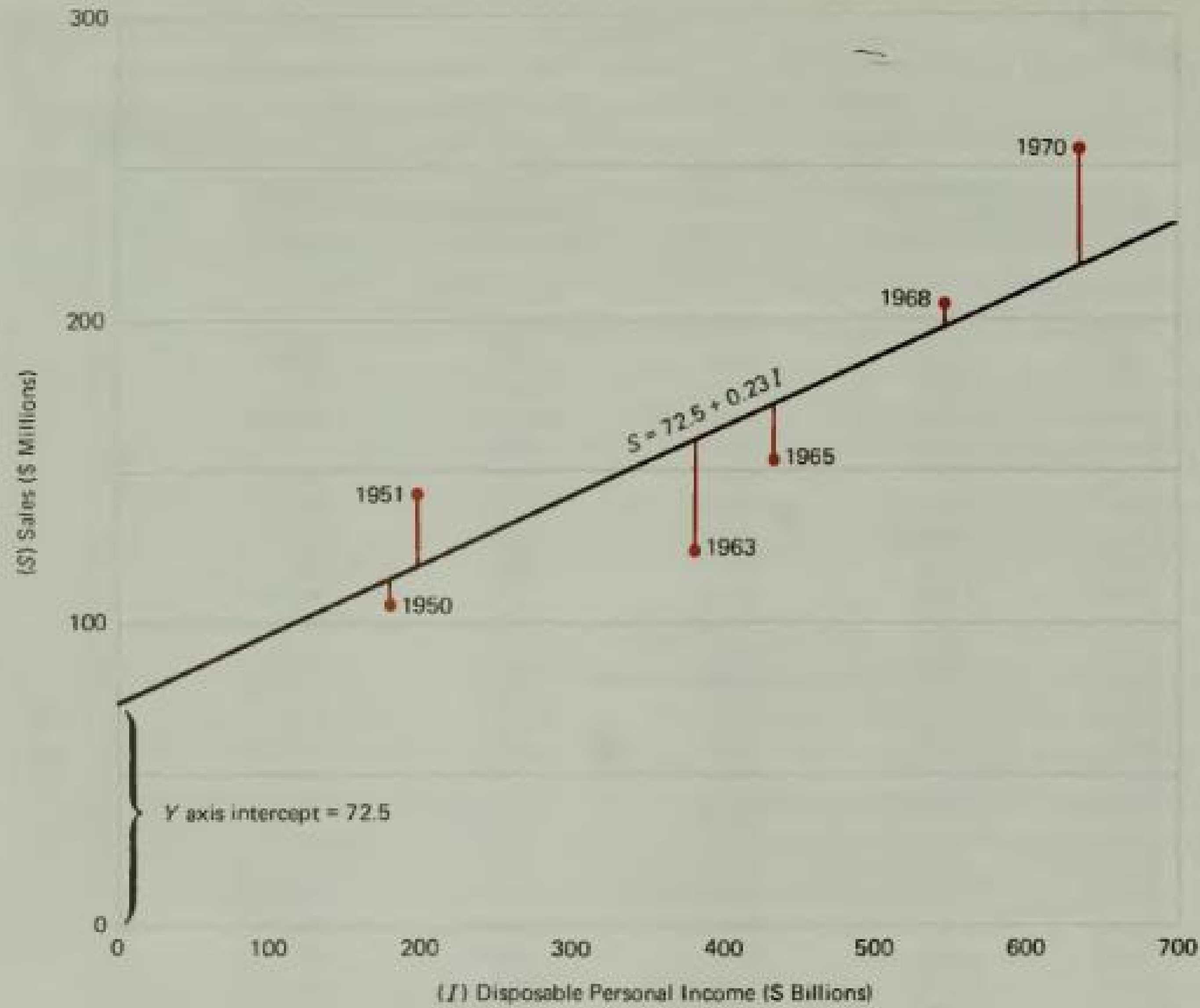


Figure 6. Simple regression line for sales dependent on disposable income. Data from Table 5.

Reliability of Forecast. A number of statistical tests can be performed to help determine the adequacy of the regression equation as a forecasting device. Data resulting from these statistical tests commonly are generated automatically in standard regression analysis computer programs. Our interest is particularly in the coefficient of determination and the standard error of estimate. In addition, there are important statistical tests concerning the significance of the regression coefficients; however, we will not discuss them.

The coefficient of determination is simply r^2, the correlation coefficient squared. For our example, $r^2 = 0.805^2 = 0.65$. The coefficient of determination states the proportion of the variation in the regression equation that is explained by the independent variable. For our equation, then, 65 percent of the variation in sales is controlled by variation in I, and 35 percent is unexplained. Thus, we

best fit

can expect large forecast errors if we use Equation 6. Apparently, other variables account for a substantial fraction of the changes in S that actually occur.

The standard error of estimate indicates the expected range of variation from the regression line of any forecast made. For example, the standard error of estimate for our data and Equation 6 is 38.7. Since we assume a normal distribution of sales for each value of I, this means that we can expect with some confidence that, two-thirds of the time, our estimate of S will be in the range of $\pm$ \$38.7 million. Therefore, if $I = 700$, then $S = 233.5$, as computed previously. However, with a standard error of estimate of 38.7, we are actually stating that two-thirds of the time we would expect the actual value of S to be in the range of 194.8 to 272.2—a rather broad range.

Obviously, we need to improve the forecasting ability of Equation 6. We can accomplish this by including other causal factors in the regression equation.

Multiple Regression. The general concepts of simple regression analysis can be extended to include the effects of several causal factors though multiple regression analysis. For the data of Table 5, Parker and Segura [1971] developed a multiple regression equation based on disposal income, time trend, gross sales lagged one year, and new housing starts lagged one year. The forecasting accuracy that resulted was indicated by an $r^2 = 0.95$, and a standard error of estimate of 9.7.

Requirements and Assumptions. A considerable historical record is necessary for regression analysis to have validity. As a rule of thumb, a five-year record is needed for one independent variable eight years are needed for two independent variables, and a longer history is needed for three or more independent variables. Of course, these data requirements often severely limit applications.

Furthermore, four very important assumptions in regression analysis should be met. (1) Linearity, which states that the dependent variable is related linearly to the independent variables. When the linear relationship does not hold, transformations often can be made that make it possible to meet the requirements of this assumption. (2) The variance of errors is constant. (3) Errors from period to period are independent of one another, or not autocorrelated. (4) The errors are distributed normally. The nature and importance of these assumptions is covered in greater detail in Benton [1972], Box and Jenkins, [1970], Huang [1970], and Wheelwright and Makridakis [1973]. Obviously, considerable knowledge of statistical methods is required to apply regression analysis appropriately.

Beyond ignoring one or more of the important assumptions, one of the great dangers in misapplying regression analysis is assuming that a good fit to historical data guarantees that the regression equation will be a good forecasting device. The regression equation itself should be an expression of a good causal theory, relating the factors in the regression model. In addition, we need to understand the potential importance of factors that are not included in the model. One of the differences between time series forecasting models and causal methods is that time series models accept increases or decreases in demand in an unbiased way, without investigating the reasons for the increase or decrease. On the other hand, causal methods demand an explanation for demand changes that occur, within the rationale of the forecasting system.

Econometric Forecasting Methods

In simplest terms, econometric forecasting methods are an extension of regression analysis to include a system of simultaneous regression equations. If, for example, we attempted to include the effect of price and advertising in our regression equation, then we would see the possibility of an interdependence; our own sales could affect these factors, as well as vice versa.

For example, let us assume that sales is a function of GNP, price, and advertising. In regression terms, we would assume that all three independent variables are exogenous to the system and thus are neither influenced by the level of sales itself nor by one another. This is a fair assumption as far as GNP is concerned. If, however, we consider price and advertising, the same assumption may not be valid. For example, if the per unit cost is of some quadratic form, a different level of sales will result in a different level of cost. Furthermore, advertising expenditures may well influence the price of the product, since production and selling costs influence the per unit price. The price, in turn, is influenced by the magnitude of sales, which also can influence the level of advertising. All this points to the interdependence of all four of the variables in our equation. When this interdependence has any strength, regression analysis cannot be used. If we want to be accurate, we must express this sales relationship by developing a system of four simultaneous equations that can deal with the interdependence directly.

To date, econometric models have been used largely in connection with relatively mature products for which a considerable historical record is available, and in industry and broad economic forecasts. For example, the Corning Glass Works developed econometric models to forecast TV-tube sales [Chambers, Mullick and Smith, 1971, 1974].

Summary

Short-term decisions focus on operating the system. While planning certainly is required, its horizon is normally a year or less. Planning is done at the broad aggregate level to make basic decisions that set the short-term capacities, and detailed plans for the use of productive resources, such as manpower and equipment schedules, and raw material orders.

Managerial control systems represent another focus of the operations phase of activity. First, we need standards of performance for cost, quality, and quantities of output. Then we need to monitor output in order to compare it with standards and interpret the results. These comparisons enable management to complete the control loop by readjusting processes to correct for any malfunctions.

The forecast is central to planning and control for operations. At the broadest operating horizon, forecasting is an important input to aggregate planning for the use of labor, facilities,and inventories. Forecasts by item are important as a basis for controlling inventories. Because of the possibility of a large number of items and forecast model complexity, computerized systems are common and often are required.

Forecasting models for operations are quantitatively oriented, and are based on time series analysis or causal methods. Time series analysis systems ordinarily have the shortest time horizons, and are most accurate for the short term. Moving averages, and exponential smoothing systems are in this category. When autocorrelation is present, we must take account of trend and seasonal factors; the exponential smoothing concepts have important application, partly because of the ease with which we can allocate greater weight to the most recent actual demand. Adaptive systems allow the resetting of parameters, either periodically or continuously, to adjust the weights of past data if demand changes sharply.

Causal methods relate forecasts to factors in the economy that cause the increases and decreases in demand. The presence of autocorrelation in demand data reduces the accuracy and value of causal methods; indeed, the absence of autocorrelation is an important assumption in the use of regression and econometric models. Though causal methods are more expensive to develop and use than time series models, they are particularly appropriate and accurate for medium- as well as short-range-horizon applications. Thus, they are particularly important for broad level aggregate planning.

Review Questions*

1. Discuss the broad outlines of operations planning and control. What is the role of forecasts? What kinds of plans must be made for day-to-day operation? What is the general structure of controls to ensure that actual outputs conform to plans?

2. What classification of productive systems is most useful in examining questions of system design? in examining questions of short-term planning and control?

3. What is a job shop? an open job shop? a closed job shop?

4. What common components of demand do we wish to consider in a forecasting system for operations?

5. What is autocorrelation? Does it facilitate or tend to defeat attempts to forecast?

6. What are the general smoothing effects of a small number of periods in a moving average (perhaps three to five periods)? What are the effects on the ability of a moving average to track a rapidly changing demand?

7. What are the general smoothing effects of values of α in the 0.01 to 0.20 range in exponential forecasting systems? How well do systems with low values of α track rapidly changing demand?

8. Show that exponential forecasting systems actually give weight to all past data.

9. How can we rationalize the fact that, to extrapolate or forecast using the basic model of Equation 2, we simply use the most current estimate of demand $\bar{F}_t$ and phase it forward by one period?

10. What is the function of the term $(1 - \alpha)/\alpha$ in Equation 4?

*Questions may also refer to the Part IV Introduction or Chapter 9.

11. If Equation 4 yields the most current estimate of demand, what is the simple rationale of Equation 5 used to forecast demand for period $t + 1$?

12. If we use a no-trend model to forecast demand where a trend actually exists, what is the nature of the error that results?

13. What is the function of the base series in exponential forecasting models that involve seasonal adjustments?

14. What is the general structure of the adaptive forecasting systems discussed? Which kinds of adaptive systems perform best?

15. Distinguish between the statistical methodology of causal methods and time series methods of forecasting.

16. As in the Cherryoak Company example used in the text, if we find a regression equation that accurately fits historical data, why not assume that it will be a good forecasting device? Why have an explanatory theory?

17. Define as measures of forecast reliability in regression analysis:
    a. The coefficient of determination.
    b. The standard error of the estimate.

18. What are the assumptions made in regression analysis?

19. How is econometric forecasting different from regression analysis?

smoo factor

Problems

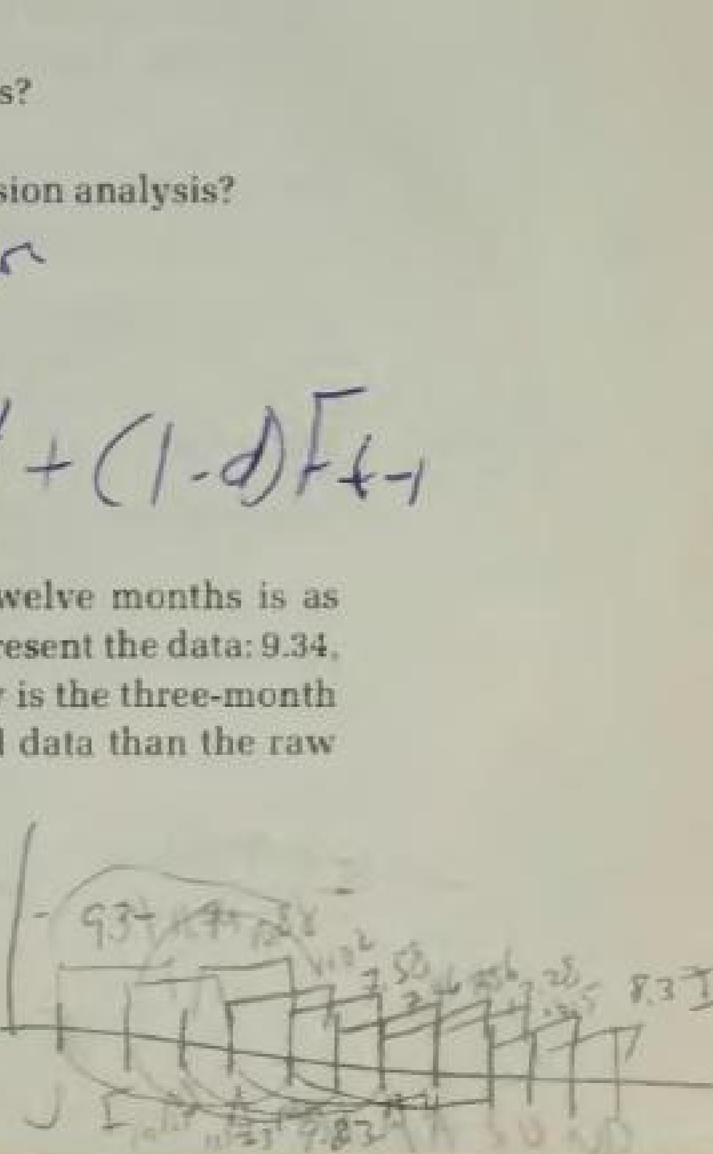
533 Expsm $F_t = \alpha D_t + (1-\alpha)F_{t-1}$

1. If the record of demand for a product for the past twelve months is as follows, compute a three-month moving average to represent the data: 9.34, 11.79, 10.88, 11.02, 7.58, 7.46, 7.56, 13.28, 10.50, 8.32. Why is the three-month moving average a better representation of the demand data than the raw data?

curvefit

10 data elmts

.6



2. The smoothed average for January was 121.2, and the current actual demand in February is 110. Compute the current smoothed average for February if $\alpha = 0.1$. On the basis of these data, what forecast can be made for March?

3. Suppose that, in addition to the data given in Problem 2, we know that the average trend adjustment for January was -1. Compute the expected demand for February. What forecast would be made for March with the additional data? for April?

EXAM

4. Figure 7 is a graph of 5.5 years of monthly average data for the number of in-patients at a large West Coast hospital. Table 6 gives the corresponding data for the first 36 months.
   a. Compute a three- and a five-month moving average for the data.
   b. Compute an exponentially weighted moving average, using Equation 2 and $\alpha = 0.1$. P. 318
   c. Compute an exponentially weighted moving average, using Equation 4 for $\alpha = 0.1$. P. 321

5. Based on moving averages computed in Problem 4a, b, and c, compute the forecasts that would be made for 1969. Compute the mean absolute deviations of errors (MAD) and compare them.

6. Based only on the exponentially weighted forecasts computed in Problem 5 for the months of 1969, what forecasts for the months of 1970 would be

TABLE 6

MONTHLY AVERAGE NUMBER OF IN-PATIENTS AT A LARGE WEST-COAST HOSPITAL

Month	Year		
	1968	1969	1970
Jan.	820	770	795
Feb.	800	795	770
Mar.	810	800	805
Apr.	865	815	810
May	850	850	805
June	745	790	705
July	820	820	800
Aug.	840	880	835
Sept.	820	875	830
Oct.	845	890	810
Nov.	820	775	800
Dec.	795	860	800

Figure 7. Monthly average in-patients at a large West Coast hospital

LNREG

justified, assuming that the actuals for 1970 given in Table 6 did not yet exist. How many months in advance do you feel justified in forecasting, using either Equations 2 or 5a?

7. Given the regression Equation 6 as a means of forecasting, if disposable income for 1971 was 679.4, compute the forecast. Actual sales were 295.79. How do they compare?

CS = A+DPI+NM p. 327 data

24 observations } dependent variable
1 independent "

References

Adam, E. E., Jr.; W. L. Berry; and D. C. Whybark. "Forecasting Demand for Medical Supply Items Using Exponential and Adaptive Smoothing Models." Paper no. 381. Purdue University, Krannert Graduate School of Industrial Administration, November 1972.

Batty, M. "Monitoring an Exponential Smoothing Forecasting System." *Operations Research Quarterly* 20(1969).

Benton, W. K. *Forecasting for Management*. Reading, Mass.: Addison-Wesley Publishing Co., 1972.

Berry, W. L., and F. W. Bliemel. "Selecting Exponential Smoothing Constants: An Application of Pattern Search." *International Journal of Production Research* 12(July 1974): 483–500.

Box, G. E. P., and G. M. Jenkins. *Time Series Analysis, Forecasting, and Control*. San Francisco: Holden-Day, 1970.

Brown, R. G. *Decision Rules for Inventory Management*. New York: Holt, Rinehart, and Winston, 1967.

______. *Smoothing, Forecasting and Prediction*. Englewood Cliffs, N. J.: Prentice-Hall, 1963.

______. *Statistical Forecasting for Inventory Control*. New York: McGraw-Hill, 1959.

Buffa, E. S., and W. H. Taubert. *Production–Inventory Systems: Planning and Control*. Rev. ed. Homewood, Ill.: Richard D. Irwin, 1972.

Chambers, J. C.; S. K. Mullick; and D. D. Smith. *An Executive's Guide to Forecasting*. New York: Wiley-Interscience, 1974.

______. "How to Choose the Right Forecasting Technique." *Harvard Business Review*, July–August 1971, pp. 45–74.

Chow, W. M. "Adaptive Control of the Exponential Smoothing Constant." *Journal of Industrial Engineering*, September–October 1965.

Cogger, K. O. "Extensions of the Fundamental Theorem of Exponential Smoothing." *Management Science* 19(January 1973): 547–554.

Eilon, S., and J. Elmaleh. "Adaptive Limits in Inventory Control." *Management Science*, April 1970.

Farley, J. U., and M. J. Hinich. "Detecting 'Small' Mean Shifts in Time Series." *Management Science*, November 1970.

Geoffrion, A. M. "A Summary of Exponential Smoothing." *Journal of Industrial Engineering* 13(July–August 1962).

Groff, G. K. "Empirical Comparison of Models for Short Range Forecasting." *Management Science* 20(September 1973):22–31.

Hill, A. V., and V. A. Mabert. "A Combined Projection-Causal Approach for Short Range Forecasts." Paper No. 527. Purdue University, Institute for Research in the Behavioral, Economic, and Management Sciences, September 1975.

Huang, D. S. *Regression and Econometric Methods*. New York: John Wiley & Sons, Inc., 1970.

Jantsch, E. "Forecasting and Systems Approach: A Frame of Reference." *Management Science* 19(August 1973): 1355–1367.

Mabert, V. A. "The Box-Jenkins Forecasting Technique." Paper no. 529. Purdue University, Institute for Research in the Behavioral, Economic, and Management Sciences, September 1975.

______. "Statistical Versus Sales Force-Executive Opinion Short Range Forecasts: A Time Series Analysis Case Study." *Decision Sciences* 7(April 1976): 310–318.

Makridakis, S.; A. Hodgsdon; and S. C. Wheelwright. "An Interactive Forecasting System." *American Statistician* 28(November 1974): 153–158.

Miller, J. G.; W. L. Berry; and C. F. Lai. "A Comparison of Alternative Forecasting Strategies for Multi-Stage Production-Inventory Systems." Paper no. 441. Purdue University, Krannert Graudate School of Industrial Administration, April 1974.

Montgomery, D. C. "Adaptive Control of Exponential Smoothing Parameters by Evolutionary Operation." *AIIE Transactions*, September 1970.

Moore, J. R., Jr. "Forecasting and Scheduling for Past-Model Replacement Parts." *Management Science* 18, Part 1(December 1971): 200–213.

Nelson, C. R. *Applied Time Series Analysis for Managerial Forecasting*. San Francisco: Holden-Day, 1973.

Nerlove, M., and S. Wage. "On the Optimality of Adaptive Forecasting." *Management Science* 10(January 1964): 207–224.

Packer, A. H. "Simulation and Adaptive Forecasting as Applied to Inventory Control." *Operations Research* 15(July–August 1967): 600–679.

Parker, G. C., and E. L. Segura. "How to Get a Better Forecast." *Harvard Business Review*, March–April 1971, pp. 99–109.

Raine, J. E. "Self-Adaptive Forecasting Reconsidered." *Decision Sciences* (April 1971).

Rao, A. G., and A. Shapiro. "Adaptive Smoothing Using Evolutionary Spectra." *Management Science* 17(November 1970): 208–218.

Reisman, A., K. Gudapati, R. Chandrasekaran, P. Darukhanavala, and D. Morrison. "Forecasting Short-Term Demand." *Industrial Engineering* 8(May 1976): 38–45.

Roberts, S. D., and R. Reed. "A Development of a Self-Adaptive Forecasting Technique." *AIIE Transactions*, December 1969.

———, and D. C. Whybark. "Adaptive Forecasting Techniques." Paper no. 335. Purdue University, Krannert Graduate School of Industrial Administration, October 1971.

Theil, H., and S. Wage. "Some Observations on Adaptive Forecasting." *Management Science* 10(January 1964).

Torfin, G. P., and T. R. Hoffman. "Simulation Tests of Some Forecasting Techniques." *Production and Inventory Management* 9(1968).

Trigg, D. W., and A. G. Leach. "Exponential Smoothing with an Adaptive Response Rate." *Operational Research Quarterly* 18(March 1967): 53–59.

Wheelwright, S. C., and S. Makridakis. *Forecasting Methods for Management*. New York: John Wiley & Sons, 1973.

Whybark, D. C. "A Comparison of Adaptive Forecasting Techniques." *Logistics and Transportation Review* 8(1972): 13–26.

Winters, P. R. "Forecasting Sales by Exponentially Weighted Moving Averages." *Management Science* 6(April 1960): 324–342.

Chapter 10

Aggregate Plans and Programs

Most managers want to plan and control operations at the broadest level through some kind of aggregate planning that bypasses the details involved in producing individual products or services and scheduling facilities and workers. Management would rather make the basic decisions that relate to programming the use of resources, both by reviewing projected employment and subcontracting levels, and by setting variable activity rates within a given employment level (accomplished by varying hours worked—overtime or undertime). Once these basic decisions have been made for the upcoming period in the planning horizon, detailed scheduling can proceed at a lower level, within the constraints of the broad plan. Finally, last minute changes in activity levels need to be made with the realization that they may affect costs by changing levels, and that they also may affect inventory.

In order to make aggregate plans, some logical overall unit of measuring sales and output first must be developed—for example, gallons in the paint industry; cases (representing an equal quantity, regardless of package style) in the beer industry; or perhaps equivalent machine hours in some mechanical industries. Second, management must be able to forecast for some reasonable planning period, perhaps a year, in these aggregate terms. Finally, management must be able to isolate and measure the relevant costs that we shall discuss in this chapter. Depending on the methods used, these costs may be reconstructed in the form of a model that will permit near optimal decisions for the sequence of planning periods in the planning horizon.

The sequential nature of the decisions should be kept in mind; a decision on employment levels and production rates that is made for an upcoming period cannot be termed either right or wrong, good or bad. Decisions will be made two periods hence, based on the decisions just made; new information about the actual progress of sales; and forecasts for the balance of the planning horizon. As a result, all decisions are right or wrong only in terms of the sequence of decisions over a period of time.

Most of our discussion deals with systems that produce inventoriable items, in which the existence of finished goods inventories can make possible a cost trade-off for the costs of changes in employment level. Near the end of this chapter we shall discuss aggregate planning for noninventoriable items.

Let us place our discussion in context by visualizing the various plans for production in terms of their time horizons. Figure 1 shows five bands of plans. The implementation of the present plans in the shortest time horizon operates under the constraints of the established firm aggregate plans that already have been made. The intermediate- and long-range plans shown in Figure 1 are involved in the broad strategic planning for the enterprise. The plans and decisions related to the upcoming and immediate future periods, however, are called aggregate planning and scheduling.

Planning Time-Horizon

Present period—implementation of firm plans already made	Firm plans for upcoming period	Tentative plans for immediate future periods	Intermediate-range plans for products and capacity	Long-range plans for products, markets, and capacities
Detailed plans and schedules for the use of the existing work force and equipment.	Based on forecasts of demand for this and future periods, make plans for immediate capacity needs. This may involve hiring or laying off personnel, setting production rates, and planning for the use of overtime and inventories.	Size of work force. Production rates. Subcontracting. Overtime. Inventories.	Changes in product mix. Projected capacity needs for equipment and manpower.	Studies of markets and their locations. Studies of the size and location of facilities.

Figure 1. Relationship of plans for various time horizons

NATURE OF AGGREGATE PLANNING

Aggregate planning increases the range of alternatives for ~~the capacity~~ use that management can consider. The term *aggregate planning* includes ~~scheduling~~ in the sense of a program. The economic significance of these ideas is by no means minor. The concepts raise such broad basic questions as: ~~To what extent~~ should inventory be used to absorb the fluctuations in demand that will occur over the next six to twelve months? Why not absorb these fluctuations simply by varying the size of the work force? Why not maintain a fairly stable work force size, and absorb fluctuations by changing production rates through varying work hours? Why not maintain a fairly stable work force and production rate and let subcontractors wrestle with the problem of fluctuating order rates? Should the firm purposely not meet all demands? In most instances it probably is true that any one of these pure strategies would be less effective than a balance among them. Since each strategy has associated costs, we seek a wise combination of the alternatives.

If inventories are used to absorb seasonal changes in demand, then capital and obsolescence costs, as well as the costs associated with storage, insurance, and handling, tend to increase. Beyond the question of seasonal factors, the use of inventories to absorb short-term fluctuations incurs increases in these same costs when they are compared with some ideal or minimum inventory level that is necessary to maintain the productive process. When inventories fall below this ideal or minimum level, stock out costs and all costs associated with short runs increase.

Changes in the size of the work force affect the total costs of labor turnover. When new workers are hired, costs arise from selection, training, and lower production effectiveness. The separation of workers may involve unemployment compensation or other separation costs, as well as an intangible effect on public relations and public image. Large changes in the size of the work force may mean adding or eliminating an entire shift; the incremental costs involved are shift premium as well as supervision and other overhead costs.

If fluctuations are absorbed through changes in the activity rate, overtime premium costs for increases, and probably idle labor costs (higher average labor cost per unit) for decreases, also will be absorbed. Usually, however, managers try to maintain the same average labor costs by reducing hours worked to somewhat below normal levels. When undertime schedules persist, labor turnover and the attendant costs are likely to increase.

Many costs affected by aggregate planning and scheduling decisions are difficult to measure and are not segregated in accounting records. Some, such as

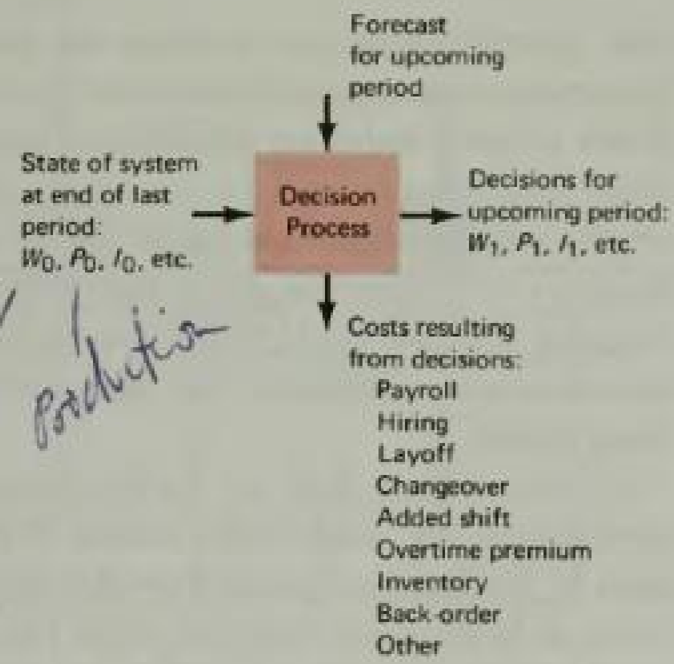


Figure 2. Single stage aggregate planning decision system where planning horizon is only one period. W = size of work force, P is production rate, and I = inventory level

Production

State of work force

interest costs on inventory investment, are alternative costs of opportunity. Other costs, such as those associated with public relations and public image, are not measurable directly. However, all the costs bear on aggregate planning decisions.

Problem Structure

The simplest structure of the aggregate planning problem is represented by the single stage system shown in Figure 2. Here, the planning horizon is only one period ahead; therefore we call Figure 2 a *single-stage* system. The state of the system at the end of the last period (which becomes the initial conditions for the upcoming period) is defined by W_0, P_0, and I_0, the aggregate work force size, production rate, and inventory level, respectively. We have a forecast of the requirements for the upcoming period; through some decision process, decisions are made which set the size of the work force and production rate for the upcoming period. Projected ending inventory I_1 then is determined as the excess of the sum of the inventory carried forward from the previous period, plus production during the period over forecasted sales during the period; that is, $I_1 = I_0 + P_1 - F_1$, where F_1 is forecasted sales.

The decisions made may call for hiring or laying off personnel, thus expanding or contracting the effective capacity of the productive system. The work force size, together with the decision on production rate during the period, then determines: the required amount of overtime; the required amount of inventory levels or back ordering; whether or not a shift must be added or deleted; and other possible changes in operating procedure. The comparative costs that result from alternate decisions on work force size and production rate influence our judgments about the effectiveness of the decisions made and the decision process used. The comparative cost of a sequence of such alternate decisions also influences our judgment about the applicability of the single-stage model.

Let us suppose that we make a sequence of decisions on the basis of the structure of the single stage model of Figure 2. Suppose that the forecasts for each of the first four periods are decreasing progressively. Our decision process leads us to decrease both the work force size and the production rate in some combination, incurring layoff and changeover costs. Then, for the fifth through tenth periods, we find that the period forecasts are increasing progressively, and that, for each period, the decision process calls for hiring personnel and increased production rates, incurring hiring and more changeover costs. The single period planning horizon has made each independent decision seem internally logical, but has caused us to lay off workers only to hire them back again, thus incurring costs for both actions as well as changeover costs due to changing production rates.

Had an appropriate decision process enabled us to look ahead for several periods, we might have decided to stabilize the work force size, at least to some extent, and to absorb the fluctuations in demand in some other way—perhaps solely by changes in production rate through the use of overtime and undertime, or by carrying extra inventories through the trough in the demand curve. It appears that broadening the planning horizon can improve the effectiveness of the aggregate planning system.

Figure 3 shows a multistage aggregate planning system in which the horizon has been expanded by means of forecasts for each period. Our objective is the same as before: to make decisions concerning the work force size and production rate for the upcoming period. In doing so, however, we consider the sequence of projected decisions in relation to forecasts and their cost effects. The decision for the upcoming period will be affected by the future period forecasts, and the decision process must consider the cost effects of the sequence of decisions. The connecting links between the several stages are the W, P, and I values, which occur, in effect at the end of one period and the beginning of the next. The feedback loop from the "decision process" box to the stage boxes indicates that the decision process may involve some iterative or trial and error process in order for us to obtain a solution.

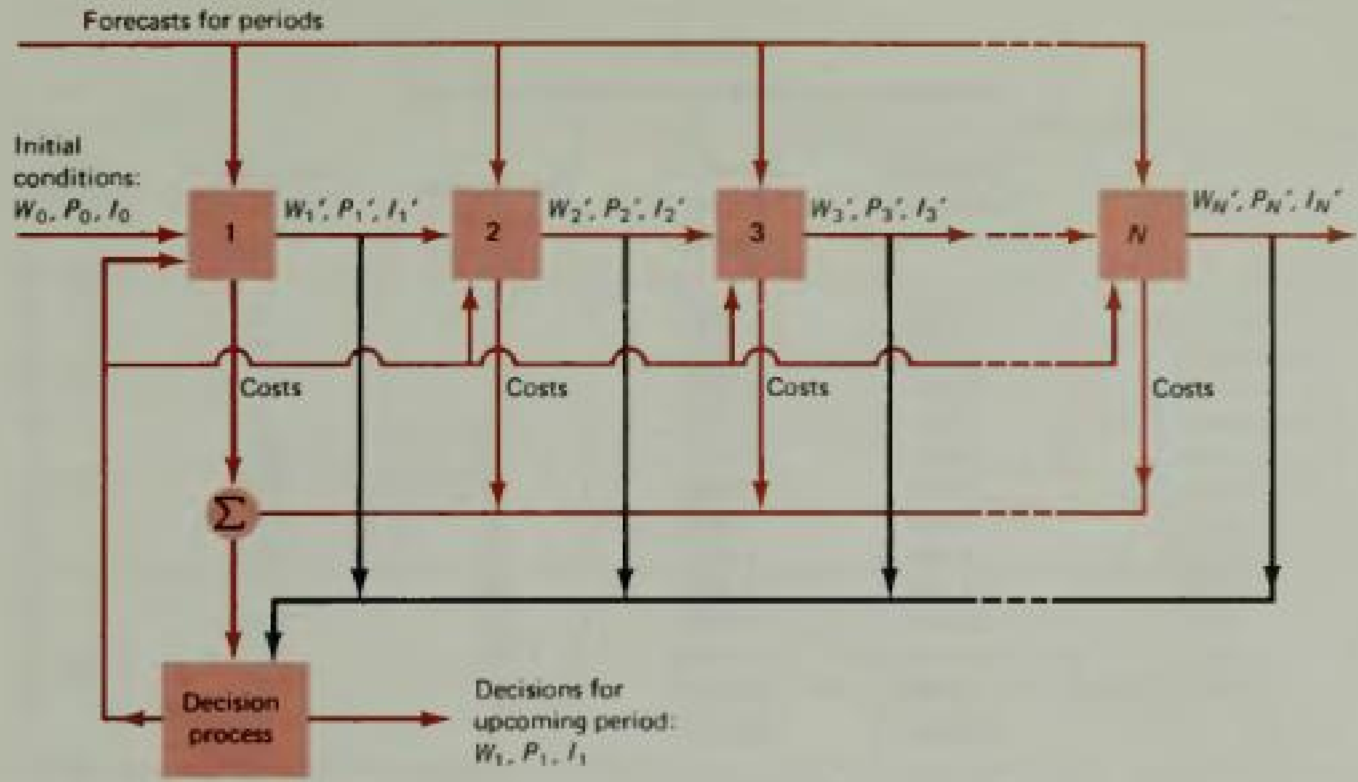


Figure 3. Multistage aggregate planning decision system for planning horizon of N periods

DECISION PROCESSES FOR AGGREGATE PLANNING

Figure 3 generally states the problem, and our interest now is drawn to the decision processes. What are the decision processes used and proposed for use? There are several, and they may be classified as graphic, mathematical, heuristic, and computer search methods. We might classify them further as static versus dynamic, as well as single stage versus multistage, models. We shall discuss the graphic, mathematical, and computer search methods.

Graphic Methods

Table 1 shows a forecast of expected production requirements for a product, together with the required safety stocks for each projected month. The available production days for each month emphasize that, although the monthly seasonal requirement swing is expected to be 13,000/3,000 = 4.33 (peak in September, 13,000; low in March, 3,000), the production rates required to follow the sales pattern are much greater, since the plant shuts down for two weeks in July for vacations. The resulting swing in requirements per production day is from 143 units per day in March to 917 units per day in July, or a ratio 917/143 = 6.41. The

TABLE 1
FORECAST OF SALES AND SAFETY STOCKS

Month	Expected Production Requirements	Required Safety Stocks	Production Days	Cumulative Production Days
Jan.	6,000	3,000	22	22
Feb.	4,000	2,500	19	41
Mar.	3,000	2,100	21	62
Apr.	4,000	2,500	21	83
May	6,000	3,000	22	105
June	9,000	3,500	20	125
July	11,000	4,000	12	137
Aug.	12,000	4,200	22	159
Sept.	13,000	4,400	20	179
Oct.	12,000	4,200	23	202
Nov.	11,000	4,000	19	221
Dec.	9,000	3,500	21	242
	100,000	40,900		

Average safety stock $= \frac{40{,}900}{12} = 3400.$

projected requirements line in Figure 4 is shown in terms of rates per working day, with the peak occurring in July instead of September because of the restriction on production time available in July.

Normal production capacity is 500 units per day, and up to 600 units per day can be produced by resorting to overtime at an extra cost of $10 per unit. To exceed the limit of 600 units per day, subcontracting is required, which costs an extra $15 per unit compared to in-plant normal production.

Figure 4 shows three alternate production programs for meeting the requirements. Plan 1, level production, is in many ways the simplest, since employment levels are stable. No overtime or subcontracting is required because the rate of 413 units per day is well within normal plant capacity. The inventory costs of this plan are high, however. At an inventory carrying cost of $40 per unit, the average seasonal stock of 9,600 units costs $384,000. The seasonal inventory is that which is required to meet projected seasonal sales peaks, over and above the amount being produced—that is, total monthly inventory less safety stock in that month for the particular production plan. See Table 2 for a comparative summary.

Plan 2 attempts to minimize inventory costs by following the requirement curve fairly closely. It accomplishes this very well, since average seasonal inventories are held to about 1,150 units at a cost of only $46,000 per year, a saving of $338,000 per year. However, labor turnover costs are high, since the

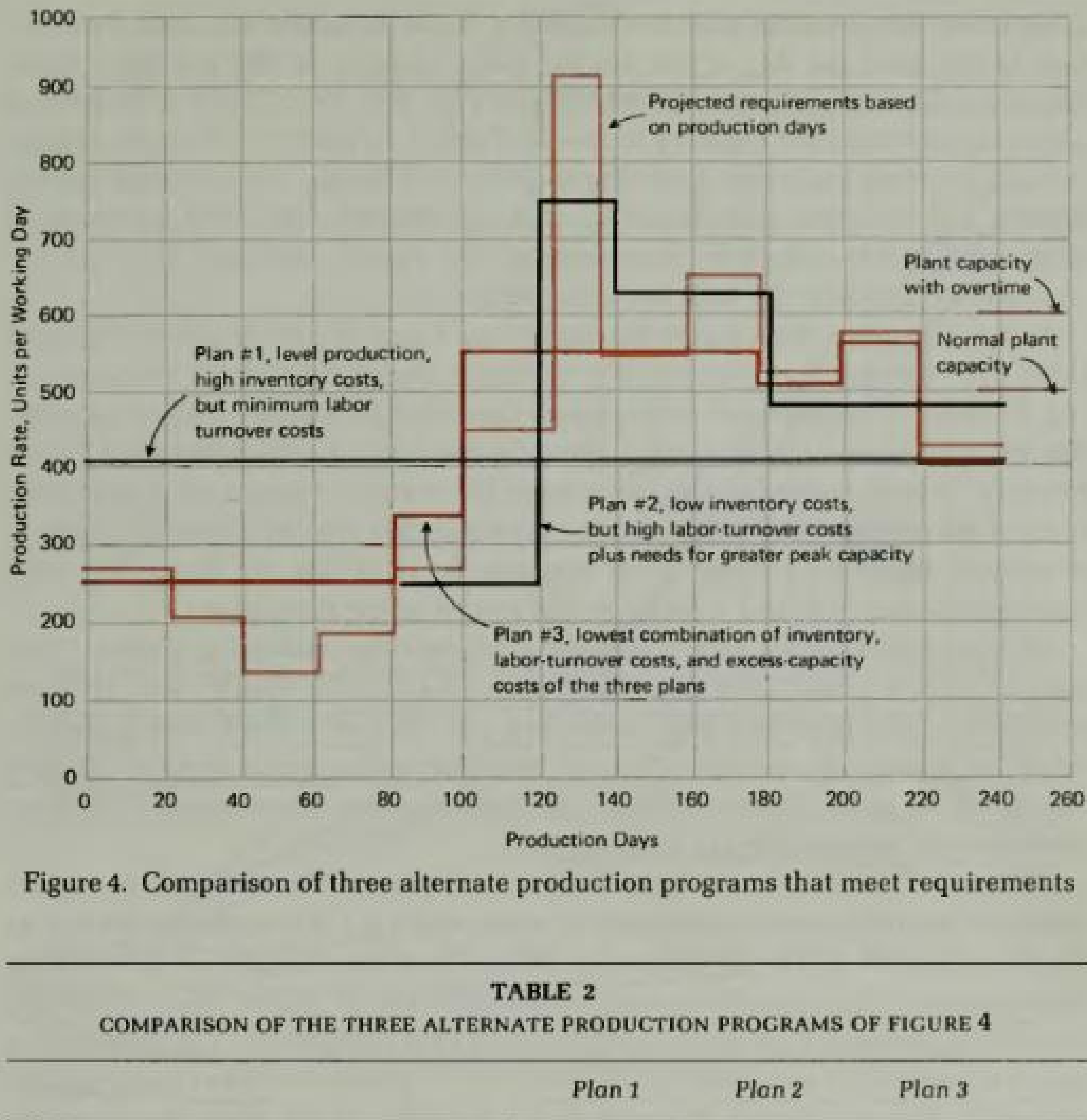


Figure 4. Comparison of three alternate production programs that meet requirements

TABLE 2

COMPARISON OF THE THREE ALTERNATE PRODUCTION PROGRAMS OF FIGURE 4

	Plan 1	*Plan 2*	*Plan 3*
Average seasonal inventory	9,600	1,150	2,275
Average safety stock	3,400	3,400	3,400
Average total inventory	13,000	4,550	5,675
Peak capacity required (No. 1 = 100)	100	181	133
Incremental costs:			
Seasonal inventory cost* (ASI×$40)	$384,000	$ 46,000	$ 91,000
Labor turnover costs†	0	164,300	164,300
Overtime premium‡	0	60,000	52,820
Extra subcontracting**	0	60,000	0
Total	$384,000	$330,300	$308,120

*Inventory carrying cost is $40 per year per unit.

†Assuming that a change in the rate of output of 35 units per day requires the employment or release of 100 workers and the cost of hiring and training an employee is $230.

‡At $10 per unit produced at overtime.

**Extra cost is $15 per unit subcontracted.

plant labor force roughly must be doubled to build the production rate from the low of 250 units per day to the normal plant capacity of 500 per day. These turnover costs are assessed at $164,300 per year. (No extra labor is employed above the normal plant capacity level—see Table 2. In addition, we must engage in both overtime and subcontracting to meet peak loads. The overtime cost is $60,000 and the extra subcontracting costs are $60,000. The total incremental costs are $330,000—a definite improvement over Plan 1, although Plan 2's turnover and excess capacity costs are quite high.

Plan 3 attempts to steer a path between Plans 1 and 2 by holding the employed labor over a longer period in order to eliminate the need for extra subcontracting. This involves a higher inventory cost than does Plan 2, but it eliminates the $60,000 extra subcontracting cost. Labor turnover costs are the same as in Plan 2 because, in both instances, the labor force fluctuation is based on a minimum level of 250 units per day and a maximum level of 500 per day. (No extra labor is employed above the normal plant capacity level of 500 per day.) The total incremental cost of Plan 3 is $308,120, the lowest of the three plans.

Whether either Plan 2 or 3 is acceptable constitutes another question. They both involve a great labor force fluctuation (about 713 employees). If these employees have relatively high skills and are difficult to find, this approach might be impractical and the effect on employee and community relationships would not be good. Other combinations involving less severe labor force fluctuation could be examined in the same way.

Although Figure 4 shows the effects of the production rate changes quite clearly, it actually is somewhat easier to work with a set of cumulative curves, as shown in Figure 5. The procedure is first to plot the cumulative production requirements. The cumulative maximum requirement curve, then, is simply the former curve plus the required buffer stocks for each period. The cumulative graph of maximum requirements can be used to generate alternative program proposals. Any production program that is feasible—in the sense that it meets requirements while providing the desired buffer stock protection—must fall entirely above the cumulative maximum requirement line. The vertical distances between the program proposal curves and the cumulative maximum requirement curve represent seasonal inventory accumulation for the plan in question.

The graphic methods are simple and have the advantage of helping us visualize alternate programs over a broad planning horizon. Graphic models are static in nature, however, and the process is in no sense optimizing. In addition, the process does not generate good programs itself but simply compares the proposals that are made.

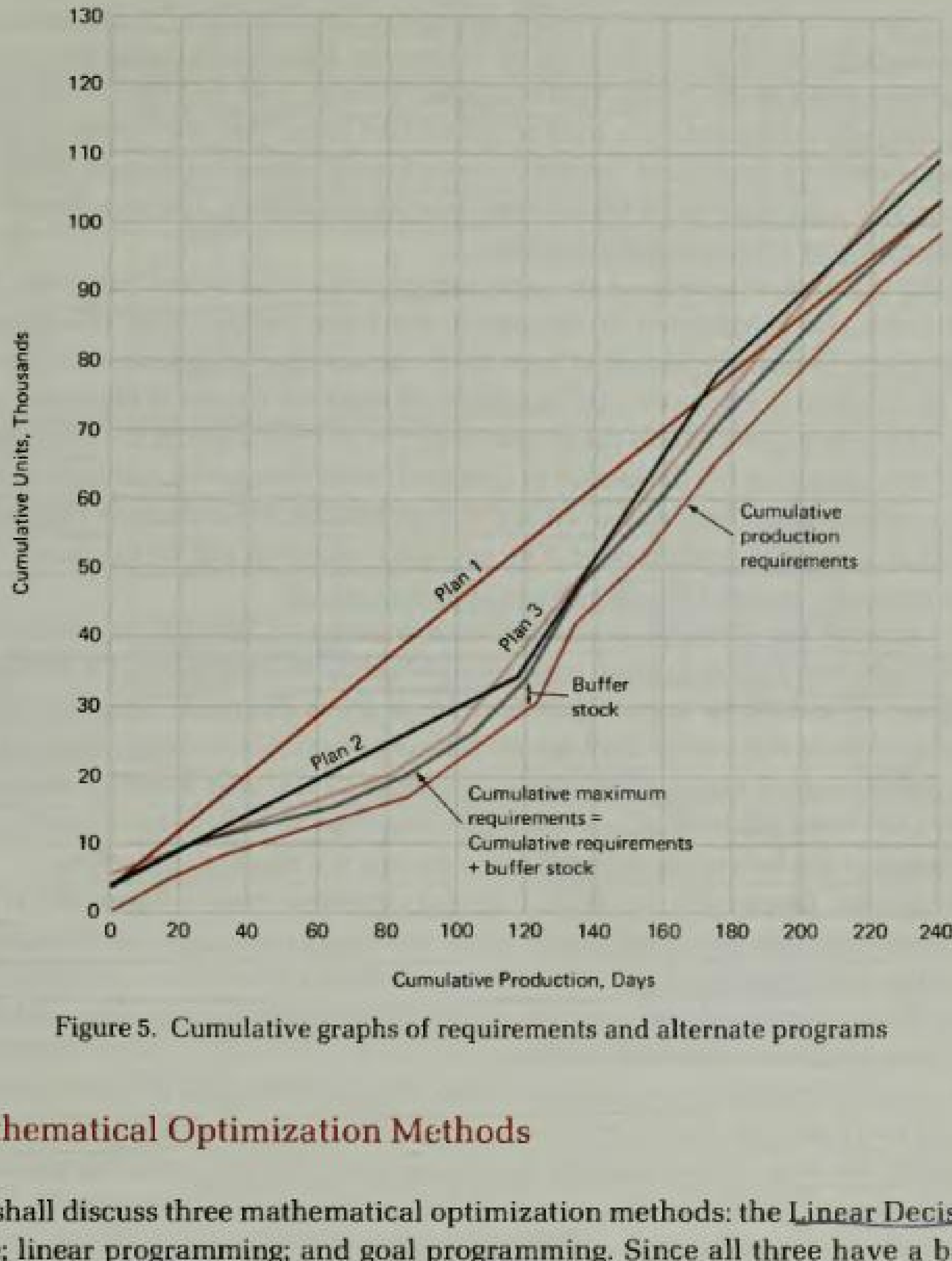


Figure 5. Cumulative graphs of requirements and alternate programs

Mathematical Optimization Methods

We shall discuss three mathematical optimization methods: the Linear Decision Rule; linear programming; and goal programming. Since all three have a basis for optimizing the model developed, our interest is particularly in appraising how closely the three models represent reality.

The Linear Decision Rule. The Linear Decision Rule (LDR) was developed by Holt, Modigliani, Muth, and Simon [1955, 1956] as a quadratic programming approach for making aggregate employment and production rate decisions. The

LDR is based on the development of a quadratic cost function for the company in connection with cost components of: (1) regular payroll; (2) hiring and layoff; (3) overtime; and (4) inventory holding, back ordering, and machine setup costs. The quadratic cost function is used to derive two linear decision rules for computing work force levels and production rate for the upcoming period, based on forecasts of aggregate sales for a preset planning horizon. The two linear decisions rules are optimum for the model.

Figure 6 shows the form of the four components of the cost function. The work force size is adjusted in the model once per period, with the implied commitment to pay employees at least their regular time wages for that period. This is indicated in Figure 6*a*. Hiring and layoff costs are shown in Figure 6*b*, and the LDR model approximates these costs with the curve shown. If the work force size is held constant for the period in question, then changes in production rate can be absorbed by the use of overtime and undertime. Undertime is the cost of idle labor at regular payroll rates. The overtime cost depends on the size of the work force, W, and on the aggregate production rate, P.

The form of the overtime-undertime cost function in relation to production rate is shown in Figure 6*c*, and is approximated by the curve shown. Whether overtime or undertime costs will occur for a given decision depends on the balance of costs defined by the horizon time. For example, in responding to the need for increased output, the costs of hiring and training must be balanced against the overtime costs. Conversely, the response to a decreased production rate requires the balancing of layoff costs against the costs of undertime.

The general shape of the net inventory cost curve is shown in Figure 6*d*. When inventories deviate from ideal levels, extra inventory costs must be absorbed if inventory levels are too high, or else costs of back ordering or lost sales will occur if inventory levels are too low. Again, these costs are approximated by a quadratic function in the LDR model.

The total incremental cost function, then, is simply the sum of the four component cost functions for a particular example. The mathematical problem is to minimize the sum of the monthly combined cost function over the planning horizon time of n periods. The result of this mathematical development is the specification of two linear decision rules that are to be used to compute the aggregate size of the work force and the production rate for the upcoming period. These two rules require as inputs: the forecast for each period of the planning horizon in aggregate terms; the ending size of work force; and inventory level in the last period. Once the two rules have been developed for a specific application, the computations required to produce the decision recommended by the model require only ten to fifteen minutes by manual methods.

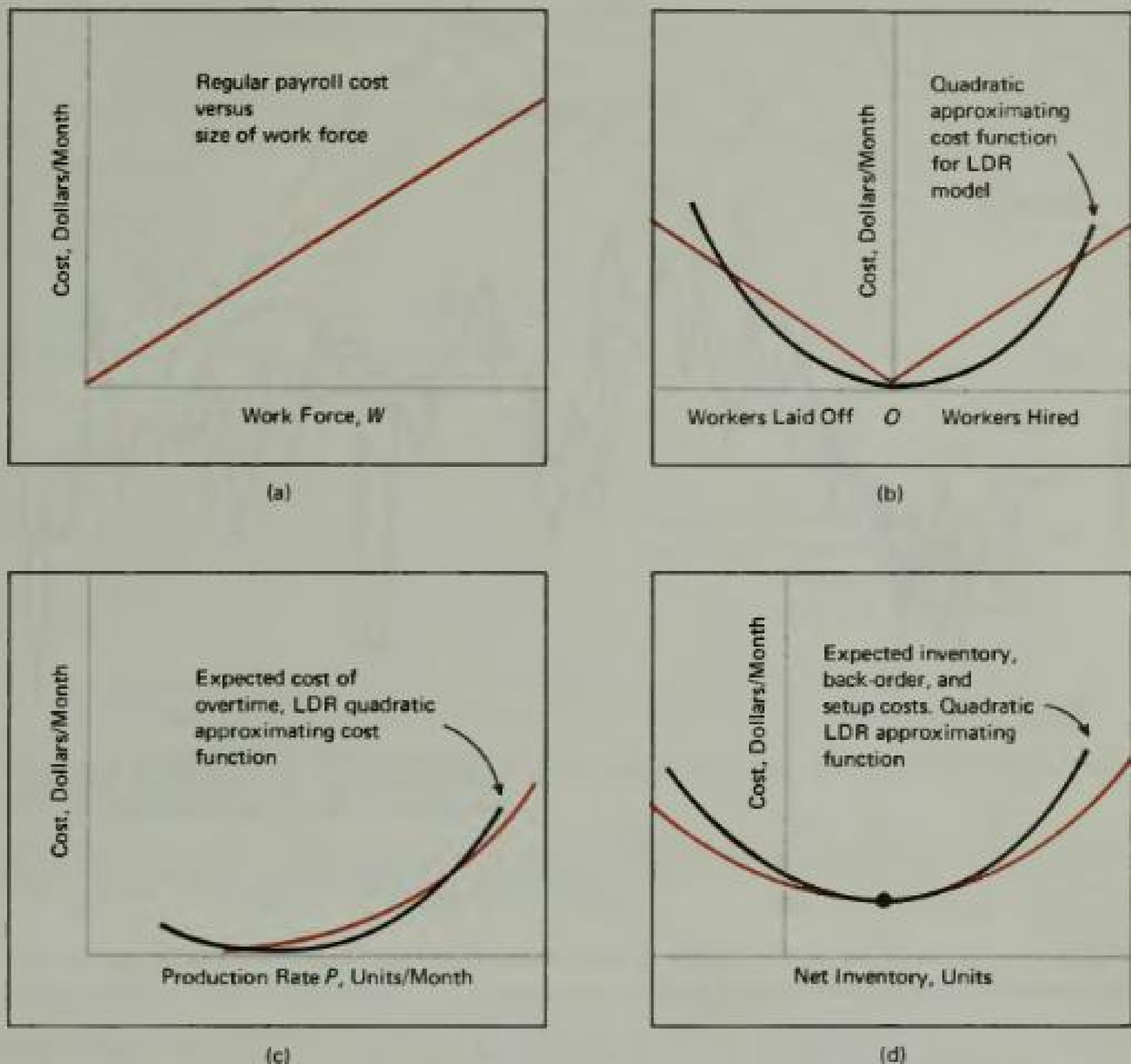


Figure 6. Approximating linear and quadratic cost functions used by the Linear Decision Rule (LDR) model. Colored lines represent presumed actual cost functions and black lines represent the LDR approximating functions

An Example

And LDR model was developed for a paint company and applied to a six-year record of known decisions in the company. Two kinds of forecasts were used as inputs: a perfect forecast, and a moving average forecast. The actual order pattern was extremely variable, involving both the 1949 recession and the Korean War. The graphical record of actual factory performance, compared with the simulated performance of the LDR, is shown in Figures 7 and 8 for production rates and work force levels. Additional graphical results of overtime hours, inventories, and back orders are contained in Holt, et al. [1955, 1960]. Costs were reconstructed for the six-year period of actual operation, and projected for the decision rules based on the nonquadratic cost struc-

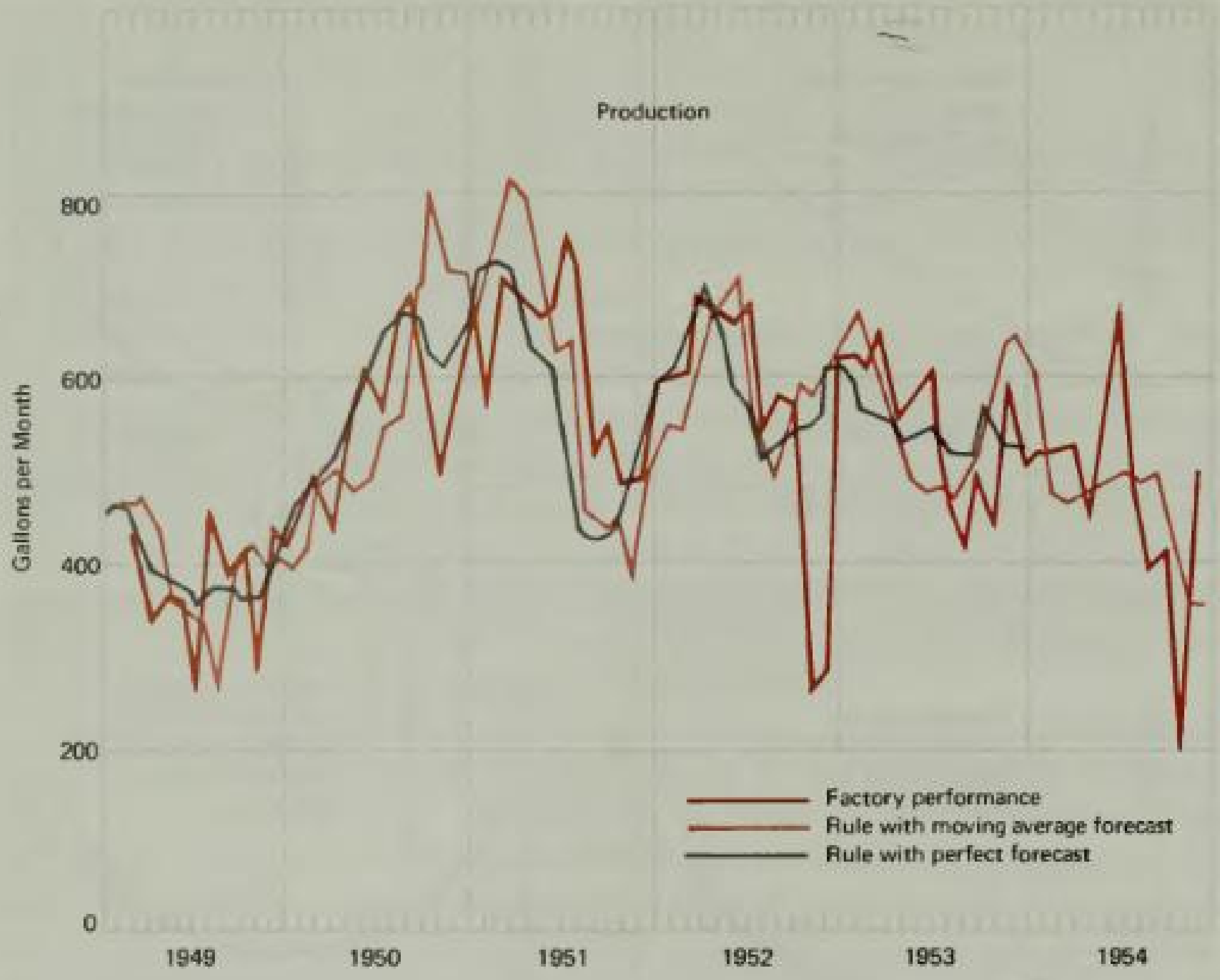


Figure 7. Comparative performance of the Linear Decision Rule (LDR) with actual factory performance for production rates

SOURCE: C. C. Holt, F. Modigliani, and H. A. Simon, "A Linear Decision Rule for Production and Employment Scheduling," *Management Science* 2(October 1955).

ture estimated from paint company data. The cost difference between actual company performance and the LDR with the moving average forecast was $173,000 per year in favor of the LDR.

LDR has many important advantages. First, the model is optimizing, and the two decision rules—once derived—are simple to apply. In addition, the model is dynamic and representative of the multistage kind of system that we discussed in connection with Figure 3. On the other hand, the quadratic cost structure may have severe limitations, and it probably does not adequately represent the cost structure of any organization. Also, since there are no constraints on the size of work force, overtime, inventory, and capital, decisions that are not feasible from some particular points of view may be generated.

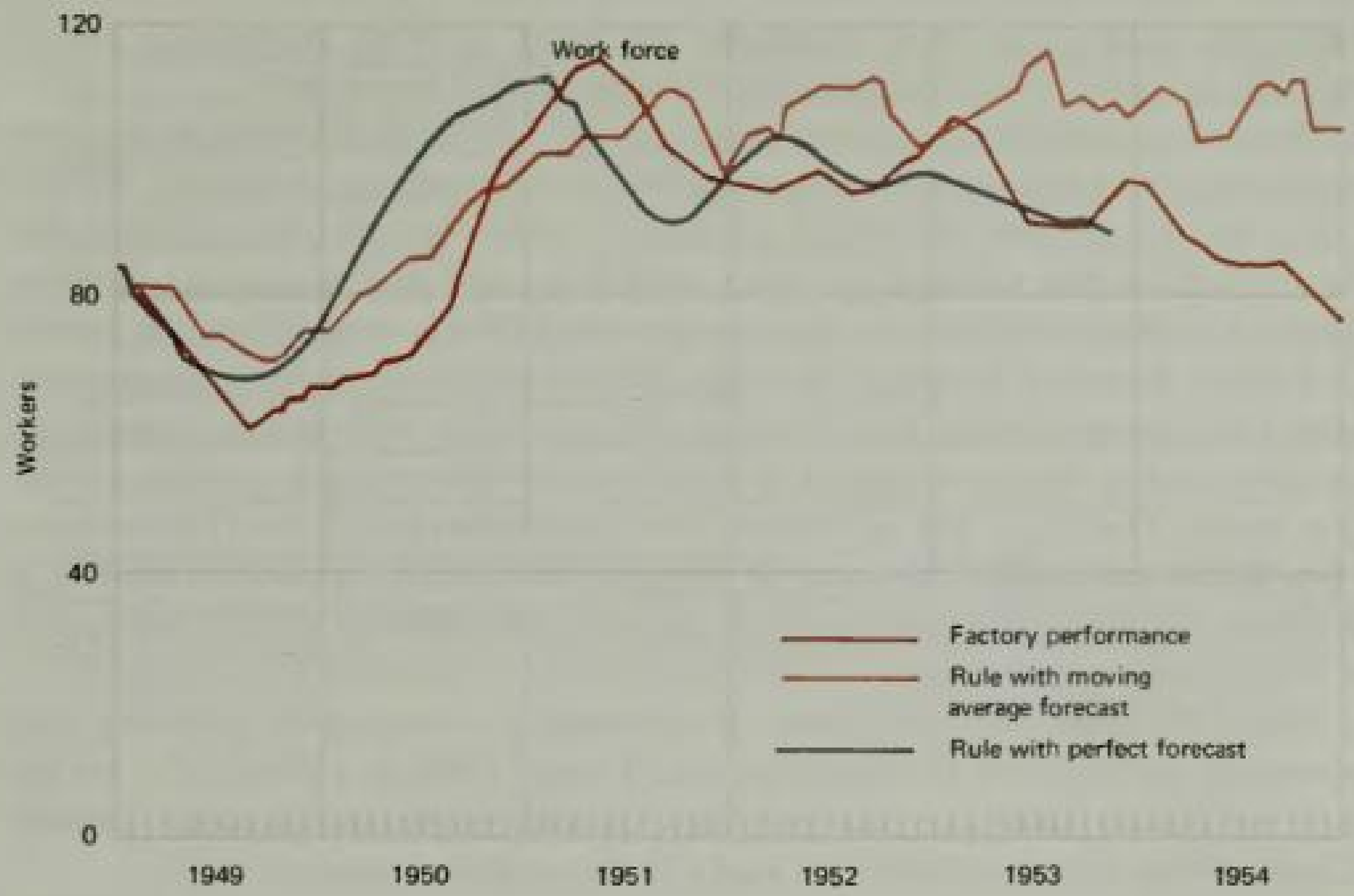


Figure 8. Comparative performance of the Linear Decision Rule (LDR) with actual factory performance for work force size

SOURCE: C. C. Holt, F. Modigliani, and H. A. Simon, "A Linear Decision Rule for Production and Employment," *Management Science* 2(October 1955).

Linear Programming Methods. The aggregate planning problem has been developed in the context of both simplex and distribution models of linear programming. Bowman [1956] proposed the *distribution model* of linear programming as a format for aggregate planning. The objective of this model is to assign units of productive capacity to minimize production plus storage costs, and to meet sales demand within the constraints of available capacity. The rim conditions in the distribution matrix form the constraints that sales requirements and capacity limitations must be met. Both beginning and ending inventories must be specified for the program developed over the *N* periods in the planning horizon. The matrix elements are costs.

Distribution methods of linear programming have serious limitations when applied to the aggregate planning problem. First, the distribution model does not account for production change costs such as hiring and laying off personnel, and there is no cost penalty for back ordering or lost sales. Thus, resulting programs may call for changes in production levels in one period, requiring an expanded work force, only to call for the layoff of these workers in future periods. In addition, the linearity requirement often is too severe.

The *simplex method* of linear programming makes it possible to include production level change costs and inventory shortage costs in the model. Hanssmann and Hess [1960] developed a simplex model that entirely parallels the linear decision rule in terms of: using work force and production rate as independent decision variables; and the components of the cost model and inclusion of a preset planning horizon. The main differences between the Hanssmann-Hess linear programming model and the LDR are that all cost functions must be linear rather than quadratic, and that linear programming is used as a solution mode. Therefore, the advantages and disadvantages of the Hanssmann-Hess model are roughly the same as for the LDR. Which of the two models is preferred depends on a preference for either the linear or quadratic cost model within a given application.

Industrial applications of linear programming to aggregate planning and problems are recorded by Eisemann and Young [1960] in a study of a textile mill; by Fabian [1967] in a study of blast furnace production; and by Greene, Chatto, Hicks, and Cox [1959] in a study of the packing industry.

Goal Programming. Goodman [1973], and Lee and Moore [1974] have applied the general goal programming format to the aggregate planning problem; Lee [1973] has applied it to similar problems in hospitals; and Lee and Clayton [1972] have applied it to academic resource allocation. Goal programming is an extension of linear programming that uses the slack variables so that the user is freed from the usual unidimensional objective function. Managerial goals are identified and placed in priority order. The following is a list of goals that pertain to aggregate planning in a specific example [Lee and Moore, 1974]:

P_1. Operate within the limits of productive capacity.
P_2. Meet the contracted delivery schedule.
P_3. Operate at a minimum level of 80 percent of regular time capacity.
P_4. Keep inventory to a maximum of three units.
P_5. Minimize total production and inventory costs.
P_6. Hold overtime production down to a minimum.

The solution procedure seccessively seeks the achievement of goals in priority order, when higher priority goals are constraints that cannot be violated. Higher priority goals then can be achieved at the expense of lower priority goals.

In an example that used the preceding list of goals, Lee and Moore generated a solution in which, in terms of goal attainment, the results were as follows:

Production capacity (P_1)	Achieved
Product delivery (P_2)	Achieved
Minimum utilization of production capacity (P_3)	Achieved
Inventory capacity (P_4)	Achieved
Minimization of production and inventory cost (P_5)	Not achieved
Minimization of overtime production (P_6)	Not achieved

Therefore, a trade-off was made between (1) production, inventory, and overtime costs, and (2) the higher priority goals of production capacity, product delivery, minimum capacity utilization, and inventory capacity.

Search Decision Rule (SDR)

A computer optimum seeking procedure may be used to evaluate systematically the cost criterion function at trial points. In the process, we hope that an optimum value eventually will be found, but there is no guarantee. In direct search methods of computer optimum seeking, the cost criterion function is evaluated at a point; the result is compared with previous trial results; and a move is determined on the basis of a set of heuristics. The new point then is evaluated and the procedure is repeated either until a better value of the function cannot be found or the predetermined computer time limit is exceeded. Taubert [1968] selected the Hooke–Jeeves [1961] pattern search program as a vehicle for experimenting with the aggregate planning and scheduling problem. To test the feasibility of using such a program, he selected the paint company example that was used to test the LDR.

In general terms, the cost criterion function represents the costs to be minimized over the planning horizon time; it can be expressed as a function of production rates and work force levels in each period of the planning horizon. Therefore, each period included in the planning horizon requires the addition of two dimensions to the criterion function: one for production rate, and one for work force level. Since the particular pattern search program used was written to handle a maximum of twenty independent variables, Taubert's analysis of the paint company limited the planning horizon time to ten months. The search program was set to end whenever the decrease in the objective cost function found by SDR's exploration of the response surface was less than \$0.5 $\times$ 10^{-6}.

TABLE 3

SDR OUTPUT FOR THE FIRST MONTH OF FACTORY OPERATION (PERFECT FORECAST)

A. SDR Decisions and Projections

Month	Sales (Gallons)	Production (Gallons)	Inventory (Gallons)	Work Force (men)
0			263.00	81.00
1	430	471.89	304.89	77.60
2	447	444.85	302.74	74.10
3	440	416.79	279.54	70.60
4	316	380.90	344.44	67.32
5	397	374.64	322.08	64.51
6	375	363.67	310.75	62.07
7	292	348.79	367.54	60.22
8	458	358.63	268.17	58.68
9	400	329.83	198.00	57.05
10	350	270.60	118.60	55.75

B. Cost Analysis of Decisions and Projections (dollars)

Month	Payroll	Hiring and Firing	Overtime	Inventory	Total
1	26,384.04	743.25	2,558.82	18.33	29,704.94
2	25,195.60	785.62	2,074.76	24.57	28,080.54
3	24,004.00	789.79	1,555.68	135.06	26,484.53
4	22,888.86	691.69	585.21	49.27	24,215.03
5	21,932.79	508.43	1,070.48	0.36	23,512.06
6	21,102.86	383.13	1,206.90	7.06	22,699.93
7	20,473.22	220.51	948.13	186.43	21,828.29
8	19,950.99	151.70	2,007.33	221.64	22,331.66
9	19,395.30	171.76	865.74	1,227.99	21,660.79
10	18,954.76	107.95	−1,396.80	3,346.46	21,012.37
					241,530.14

SOURCE: W. H. Taubert, "Search Decision for the Aggregate Scheduling Problem," *Management Science* 14(February 1968).

Table 3 shows a sample of the computer output for the first month of factory operation of the paint company. The computer output gives the first month's decision, as well as an entire program for the planning horizon of ten months. In the lower half of the table, the program prints out for the entire planning horizon the component costs of payroll, hiring and firing, overtime, inventory, and the total of these discretionary costs. Thus, a production manager is provided not only with the immediate decisions for the upcoming month, but also with the projected decisions based on monthly forecasts for the planning horizon time, and the economic consequences of each month's decisions.

Table 4 gives a month-by-month comparison for the first twenty-four months

TABLE 4
A COMPARISON OF LINEAR DECISION RULE AND SEARCH DECISION RULE FOR THE FIRST 24 MONTHS OF OPERATION WITH PAINT COMPANY DATA (PERFECT FORECAST)

Month	Monthly Sales (Gallons)	Production (Gallons)		Work Force (Workers)*		Inventory (Gallons)		Monthly Cost (Dollars)	
		LDR	SDR	LDR	SDR	LDR	SDR	LDR	SDR
0				81	81	263	263		
1	430	468	472	78	78	301	305	29,348	29,705
2	447	442	443	75	74	296	301	27,797	27,930
3	440	416	418	72	71	272	279	26,294	26,460
4	316	382	385	69	68	337	348	24,094	24,415
5	397	377	376	67	66	317	327	23,504	23,436
6	375	368	366	66	64	311	318	22,879	22,672
7	292	360	360	65	63	379	386	22,614	22,539
8	458	382	382	65	63	303	309	23,485	23,322
9	400	377	379	66	64	280	288	23,367	23,331
10	350	366	366	67	64	296	304	22,846	22,569
11	284	365	359	69	67	377	379	23,408	23,004
12	400	404	401	72	70	381	380	25,750	25,654
13	483	447	447	75	74	345	344	28,266	28,367
14	509	477	479	79	78	313	314	30,180	30,408
15	500	495	498	83	81	307	312	31,310	31,479
16	475	511	510	87	86	343	348	32,422	32,481
17	500	543	547	91	90	386	394	34,858	35,074
18	600	595	592	96	94	380	387	38,118	38,216
19	700	641	642	100	98	321	328	40,849	41,110
20	700	661	659	103	101	282	287	41,848	41,898
21	725	659	658	105	103	216	220	41,945	41,981
22	600	627	624	106	105	244	245	39,074	38,940
23	432	605	601	107	106	417	413	38,134	37,928
24	615	653	655	109	108	455	454	41,785	42,003
Totals		11,621	11,619	1,972	1,936	7,859	7,970	734,176	734,982

*Rounded to the next larger number of workers.
SOURCE: W. H. Taubert, "Search Decision for the Aggregate Scheduling Problem," *Management Science* 14(February 1968).

of the results obtained by the SDR program, and those obtained by the two optimum decision rules for the LDR. The month-by-month decisions are not identical, but they are very close to each other; the twenty-four-month production totals differ by only two gallons. The total cost of the SDR program exceeds the LDR total by only $806 or 0.11 percent. This difference may be explained by the fact that the SDR used a planning horizon of only ten months, in comparison with the twelve-month horizon used by the LDR.

Because of the encouraging results in virtually duplicating the performance of LDR in the paint company, it was decided to test SDR in more demanding situations. Thus, models were developed for three rather different situations: (1)

the Search Company [Buffa, 1972; Buffa and Taubert, 1972], a hypothetical organization involving a very complex cost model that included the possibility of using a second shift when needed; (2) the Search Mill [Buffa, 1972; Redwine, 1971], based on disguised data obtained from a major American integrated steel mill; and (3) the Search Laboratory [Buffa, 1972; Buffa and Taubert, 1972; Taubert, 1968a], a fictitious name for a division of a large aerospace research and development laboratory. All three situations represent significant extensions beyond the paint company application, using highly complex cost models and other factors to challenge the SDR methodology.

Comparative Performance of Aggregate Planning Decision Processes

Because the LDR is optimal for the model, it has commonly been used as a standard for comparison by proposers of new decision processes for aggregate planning [Bowman, 1963; Gordon, 1966; Hanssmann and Hess, 1960; Jones, 1967; Taubert, 1968]. The availability of a standard for comparison has been particularly valuable for decision processses that are not mathematically optimal, but optimum seeking in their nature. By substantially duplicating LDR performance on standard problems such as the paint company, the general validity of such methods has been established. Such comparisons, however, fall short of validating the performance of any decision process in real environments. LDR assumes that cost functions are quadratic in form when in fact they may take a variety of mathematical forms. Thus, the best evaluation of the comparative performance of alternate decision processes is in the real world where we are attempting to optimize the costs (or profits) actually found, rather than a restrictive model of cost behavior.

Lee and Khumawala [1974] report a comparative study carried out in the environment of a firm in the capital goods industry having annual sales of approximately $11 million. The plant was a typical closed job shop manufacturing facility in which parts were produced for inventory and then assembled into finished products either for inventory or for customer order. A model of the firm was developed which simulated the operation of the firm from the aggregate point of view. Demand forecasting in the model provided an option to use either a perfect or imperfect forecast, and four alternate aggregate planning decision processes were used to plan production and work force size. The four decision processes were LDR, SDR, Parametric Production Planning, and the Management Coefficients model.

Parametric Production Planning is a decision process proposed by Jones

[1967] which uses a course grid search procedure to evaluate four possible parameters associated with minimum cost in the firm's cost structure. The cost structure is developed for the particular firm and is free of constraints on mathematical form. The four parameters are then inserted into decision rules for work force size and production rate. There is no guarantee of optimality.

The Management Coefficients model was proposed by Bowman [1963] and establishes the form of decision rules through rigorous analysis but determines the coefficients for the decision rules through statistical analysis of management's own past decisions. The theory behind Bowman's rules is that management actually is sensitive to the same behavior used in analytical models, and that management's behavior tends to be highly variable, rather than consistently above or below optimum performance. The Management Coefficients model assumes that management's performance in terms of the decision rules can be improved considerably simply by applying the same rules more consistently.

Results. Table 5 summarizes comparative profit performance for actual company decisions and the four test decision models. When the imperfect forecast that is available to the firm's management is used, all four decision models result in increased profits. The minimum mean profit increase is $187,000 (4 percent) using the Management Coefficients model. The maximum mean increase is $601,000 (14 percent) using SDR. The contrast between the profit figures for the perfect and imperfect forecasts gives a measure of the value of forecast information. Although obviously perfect information has value ($119,000 for SDR), it is less significant than the decision process itself ($601,000 difference between SDR and company decisions, and $414,000 difference between SDR and Management Coefficients).

TABLE 5

COMPARATIVE PROFIT PERFORMANCE

	Imperfect Forecast	Perfect Forecast
Company decisions	$4,420,000	–
Linear Decision Rule	$4,821,000	$5,078,000
Management Coefficients Model	$4,607,000	$5,000,000
Parametric Production Planning	$4,900,000	$4,989,000
Search Decision Rule	$5,021,000	$5,140,000

SOURCE: W. B. Lee and B. M. Khumawala, "Simulation Testing of Aggregate Production Planning Models in an Implementation Methodology," Management Science 20(February 1974): 903–911.

Joint Decisions Among Operations, Marketing, and Finance

The main objective of aggregate planning and scheduling methods is to employ systems concepts in making key decisions for operations. The results of coordinating decisions concerning activity levels, work force size, use of overtime, and inventory levels amply illustrate that these kinds of decisions should be made jointly rather than independently (the latter would create suboptimization). But why stop with the operations function? Would even better results be obtained if some of the key operational decisions were made jointly with other key decisions in marketing and finance?

Tuite [1968] proposed the merger of marketing strategy selection and production scheduling; Holloway [1969] proposed price as an independent variable, coupled with allocations of compensatory promotion budgets; and Bergstrom and Smith [1970] proposed the estimation of revenue-versus-sales curves for each product in each time period, and the consideration of the amount to be sold as a decision variable, dependent on price and possibly on other parameters. Finally, Damon and Schramm [1972] proposed to make decisions jointly in production, marketing, and finance. In their model, marketing sector decisions are made in accordance with price and promotion expenditures, and finance sector decisions are made with respect to investment in marketable securities and short-term debt incurred or retired. The solution technique is a computer search methodology similar to SDR.

AGGREGATE PLANNING FOR NONMANUFACTURING SYSTEMS

The general nature of aggregate planning and scheduling problems in nonmanufacturing settings basically is similar to such problems in manufacturing settings: we are attempting to build a cost or profit model in terms of the key decision variables for short-term capacity. There is likely to be less freedom in adjusting short-term capacity in nonmanufacturing settings, however, because of the absence of inventories and subcontractors as sources of capacity. As a result, the manager is more likely to absorb fluctuations in demand directly, by varying work force size, hours worked, and overtime in a fashion that minimizes cost, maximizes profit, or optimizes some substitute criterion.

We shall not discuss aggregate planning in nonmanufacturing here. Subsequent chapters will deal with operations planning and control in large-scale projects and in service systems, and will include the special problems of both aggregate and detailed schedules in such systems. At that point, we shall discuss applications of aggregate planning in large-scale projects, a research laboratory, hospitals, educational systems, and a public utility.

THE MANAGER'S COST REDUCING RESPONSES

Aggregate scheduling is a formal approach to smoothing which, in its most developed state, extends into the marketing and financial functions. Decisions are made jointly concerning production, employment, prices, compensatory marketing and promotion expenditures, and short-term investments. In addition, smoothing strategies are used by managers that are not included in any formal model. Galbraith [1969] included among these strategies attempts to influence and manipulate the demand function, adaptation of the organization in various ways, and coordination with other organizations.

Influencing Demand

As we noted, some of the formal models include attempts to shift demand from sales peaks to valleys through counterseasonal pricing and promotion of their products and services [Damon and Schramm, 1972; Holloway, 1969; Vergin, 1966]. These models include decision variables regarding price and promotion funding through known relationships regarding price elasticity of demand and response to promotion. Of course, the models are attempts to formalize well-known managerial behavior, such as that illustrated by post-Christmas sales and by airline-offerings of special prices during the off-season and during the night (when equipment otherwise would be utilized poorly). Galbraith also gives examples occurring in the post office and the highly seasonal flower industry.

In addition, managers view their demand in the aggregate, and seek to expand product lines by making products whose demand is counterseasonal to their existing lines but that use the same basic production technology. Vergin [1966] in his analysis of eight manufacturing organizations with seasonal demand patterns, found that the counterseasonal product was the dominant managerial strategy, almost to the exclusion of other very attractive alternatives, such as aggregate scheduling.

Organization Adaptations

Managers make fundamental adaptations of their organizations in an attempt to smooth demand in relation to frequently fixed resources. For example, in a chocolate factory's early application of the LDR, the firm changed its location to a rural area to take advantage of the farm labor available in the fall and winter seasons. Thus, a hiring-and-layoff smoothing strategy became compatible with the seasonal production period [Galbraith, 1969]. The guaranteed annual wage in the meat packing industry makes it possible to vary the length of the work week without causing substantial wage variations. Of course, the use of counterseasonal products is also an adaptive response.

Coordination with Other Organizations

One of the common managerial strategies has been to subcontract needs above certain capacity limits when possible. Such a strategy involves coordination between two (usually different) firms in the marketplace. A larger, integrated firm may make subcontracting arrangements with smaller, more flexible firms operating in the same field. Galbraith quotes examples in the coal and oil refining industries.

In other situations, managerial strategy may involve coordination between producers and customers that may allow the producer to increase lot sizes, split lots, or delay or speed up deliveries in order to smooth work load. In return, the customer receives preferential supply treatment.

monopoly
price-fix

Finally, some organizations may be able to form a coalition that can smooth work loads effectively. For example, electric utilities join together in networks of supply that enable them to meet peak demands for the network system as a whole. If each individual utility had to meet its individual peak demands, higher plant investment would be required in the aggregate, as well as for each organization. The airlines have found that, by sharing equipment when route structures are noncompetitive and counterseasonal, all organizations can achieve somewhat better equipment utilization.

Summary

Aggregate plans and programs are of the greatest importance to operations management, since these plans enable management to deploy the major resources at its command. Management's interest therefore is focused on the most important aspects of this deployment process, such as employment levels, activity or production rates, and inventory levels. If basic plans reflect decisions in these areas, then detailed planning and scheduling of operations can proceed within stated operating constraints.

We have discussed the structure of the aggregate planning problem and a number of decision processes that have been designed to meet the needs of aggregate planning. Currently, the graphic methods probably are used most often. The mathematical and computer search methods have been developed in an effort to improve on traditional methods by making the process dynamic, optimum seeking, and representative of the problem's usual multistage nature. Several models have been valuable mainly as stepping-stones to more useful models that represent reality more accurately. The most important single stepping-stone undoubtedly has been the LDR; however, its original advantage of requiring only simple computations largely has been offset by the computer.

Presently, the computer search methods seem to offer the most promise because of their greater flexibility in representing costs that really occur in organizations. Although some of the mathematical methods do produce optimum solutions, it is the model that is being optimized. The real world counterpart of the model is optimized only if the mathematical model duplicates reality. The computer search methods are only optimum seeking by their nature; however since they do not suffer from the need to adhere to strict mathematical forms in the model, they can duplicate reality more accurately in cost and profit models. After all, it is the real world situation that we wish to optimize, not the model.

The extension of aggregate planning models to make joint decisions among the production, marketing, and finance functions is most encouraging, and demonstrates progress in our ability to employ systems concepts.

Aggregate planning in nonmanufacturing situations is similar conceptually to that of manufacturing situations: we are attempting to make broad level plans and schedules that are sensitive to the relative cost of alternate resources. In the absence of inventories to absorb fluctuations in demand, we must find other variables under managerial control that we can manipulate as a buffer between operations and demand.

Review Questions

1. What is the meaning of the term *aggregate plan*? What are the objectives of aggregate plans? What are the inputs and the nature of outputs?

2. Place aggregate planning in context with the term *planning horizon*. What is the appropriate horizon for aggregate planning?

3. Discuss the relevant cost components involved in aggregate planning decisions.

4. Under what conditions would a single stage aggregate planning decision system be appropriate?

5. In what ways does a multistage aggregate planning decision system consider realities that actually would affect decisions for production rates and work force size?

6. Compare the Linear Decision Rule with the multistage aggregate planning decision system discussed and summarized by Figure 3. What compromises with reality, if any, have been made by the LDR model?

7. What is the nature of the goal programming approach to aggregate planning? What are its advantages? disadvantages?

8. As a decision system, contrast the SDR with the LDR. What are the advantages and disadvantages of each?

9. Referring to Figure 6:
   a. Explain why the cost of overtime should rise at an increasing rate as production rate increases.
   b. If inventory varies from the optimal level (minimum cost level), why should the incremental costs go up at an increasing rate, as represented in Figure 6d?

10. Criticize the usefulness and validity of the strict aggregate planning concept—that is, making decisions solely in aggregate terms of work force size and production rate.

11. What is the meaning of the term *capacity* in aggregate planning models? How does a decision to hire, lay off, or subcontract affect capacity? How does physical or limiting capacity affect these decisions?

12. Cost comparisons between the result of actual managerial decisions and those produced by solving decision rule models typically are made by running both sets of decisions through the cost model and then comparing

the results. Does this methodology seem valid? If not, what other approach might be followed?

13. Account for the difference in performance of the several aggregate planning decision systems in the study summarized by Table 5.

14. What values are gained by expanding aggregate planning decisions systems to produce joint decisions among operations, marketing, and finance? Are there any disadvantages?

Problems

1. Table 6 gives data that show projected requirements for the production of a middle priced camera, together with buffer stock requirements and available production days in each month. Develop a chart of cumulative requirements and cumulative maximum requirements for the year, plotting cumulative production days on the horizontal axis and cumulative requirements in units on the vertical axis.

2. Using the data of Table 6, compare the total incremental costs involved: in level production plan; in a plan that follows maximum requirements quite

P.345

TABLE 6
PROJECTED PRODUCTION AND INVENTORY REQUIREMENTS

Month	Production Requirements	Required Buffer Stocks	Production Days
Jan.	3,000	600	22
Feb.	2,500	500	18
Mar.	4,000	800	22
Apr.	6,000	1,200	21
May	8,000	1,600	22
June	12,000	2,400	21
July	15,000	3,000	21
Aug.	12,000	2,400	13
Sept.	10,000	2,000	20
Oct.	8,000	1,600	23
Nov.	4,000	800	21
Dec.	3,000	600	20
	87,500	17,500	244

closely; and in some intermediate plan. Normal plant capacity is 400 units per working day. An additional 20 percent can be obtained through overtime, but at an additional cost of $10 per unit. Inventory carrying cost is $30 per unit per year. Changes in production level cost $5,000 per 10 units in production rate. Extra capacity may be obtained by subcontracting certain parts at an extra cost of $15 per unit. Beginning inventory is 4,000 units.

3. An organization has forecasted maximum production requirements for the coming year, as shown in Table 7. The present labor force can produce 470 units per month. An employee added or subtracted from the labor force affects the production rate by 20 units per month. The average salary of employees is $660 per month, and overtime can be used at the usual premium of time-and-a-half pay up to 10 percent of time for each employee. Therefore, an employee working the maximum overtime could produce the equivalent of an additional two units per month. Hiring and training costs are $100 per worker, and layoff costs are $200 per worker. Inventory holding costs are $10 per month per unit, and shortages cost $50 per unit short. Changeover costs for any increase or decrease in production rate are $3,000 per changeover, over and above pertinent hiring and layoff costs. These costs include replanning and rebalancing of production lines, and so on. No changeover cost is appropriate when added production is achieved through the use of overtime. What plan do you recommend? What is the incremental cost of your plan?

4. Given the data in Table 2:
   a. What value of inventory carrying cost would make Plans 1 and 2 equally desirable?
   b. What hiring layoff cost makes Plans 1 and 2 equally desirable?
   c. What subcontracting cost makes Plans 1 and 2 euqally desirable?

p. 347

TABLE 7
FORECASTED MAXIMUM PRODUCTION REQUIREMENTS

Jan.	400	July	580
Feb.	510	Aug.	600
Mar.	400	Sept.	300
Apr.	405	Oct.	280
May	460	Nov.	440
June	675	Dec.	500

5. A company manufactures a single product for which Table 8 represents a schedule of forecasted and actual demand in units for one year.

Table 8

Month	Forecasted Demand	Actual Demand
Jan.	23,000	23,000
Feb.	24,000	25,000
Mar.	21,000	20,000
Apr.	23,000	22,000
May	20,000	22,000
June	19,000	24,000
July	17,000	22,000
Aug.	14,000	15,000
Sept.	8,000	6,000
Oct.	10,000	13,000
Nov.	9,000	10,000
Dec.	10,000	14,000
Total	198,000	216,000
Average	16,500	18,000

The initial inventory is 15,000 units. The desired ending inventory is 20,000 units. The cost of storage is $1 per unit per month. It costs $1,000 to change production from 0 to 3,000 units, and $3,000 to change production from 3,001 to 6,000 units. No change larger than 6,000 units is possible in one period. Back orders are permitted at a cost of $5 per unit per period.

a. What is the best production plan for the forecasted demand, if we wish to minimize pertinent costs?
b. Assuming that the year is over, what is the best production plan for the actual demand, utilizing the benefit of hindsight?

References

Baker, K. R. "An Experimental Study of the Effectiveness of Rolling Schedules in Production Planning." *Decision Sciences* 8(January 1977).

Bergstrom, G. L., and B. E. Smith. "Multi-Item Production Planning—An Extension of the HMMS Rules." *Management Science* 16(June 1970): 614–629.

Bowman, E. H. "Consistency and Optimality in Managerial Decision Making." *Management Science* 9(January 1963): 310–321.

______. "Production Scheduling by the Transportation Method of Linear Programming." *Operations Research* 4(February 1956): 100–103.

Buffa, E. S. "Aggregate Planning for Production." *Business Horizons*, Fall 1967.

______. *Operations Management: Problems and Models*. 3d ed. New York: John Wiley & Sons, 1972. Chapter 13.

______, and W. H. Taubert. "Evaluation of Direct Computer Search Methods for the Aggregate Planning Problem." *Industrial Management Review*, Fall 1967.

______, and W. H. Taubert. *Production-Inventory Systems: Planning and Control* Rev. Ed. Homewood, Ill.: Richard D. Irwin, 1972. Chapters 5, 6, and 7.

Damon, W. W., and R. Schramm. "A Simultaneous Decision Model for Production, Marketing, and Finance." *Management Science* 19(October 1972): 161–172.

Eisemann, K., and W. M. Young. "Study of a Textile Mill with the Aid of Linear Programming." *Management Technology* 1(January 1960): 52–63.

Fabian, T. "Blast Furnace Production—A Linear Programming Example." *Management Science* 14(October 1967).

Galbraith, J. R. "Solving Production Smoothing Problems." *Management Science* 15(August 1969): 665–674.

Goodman, D. A. "A New Approach to Scheduling Aggregate Production and Work Force." *AIIE Transactions* 5(June 1973): 135–141.

Gordon, J. R. M. "A Multi-Model Analysis of an Aggregate Scheduling Decision." Ph.D. dissertaion, Massachusetts Institute of Technology, Sloan School of Management, 1966.

Greene, J. H.; K. Chatto; C. R. Hicks; and C. B. Cox. "Linear Programming in the Packing Industry." *Journal of Industrial Engineering* 10(September-October 1959): 364–372.

Hanssmann, F., and S. W. Hess. "A Linear Programming Approach to Production and Employment Scheduling." *Management Technology* 1(January 1960): 46–52.

Hausman, W. H., and R. Peterson. "Multiproduct Production Scheduling for Style Goods with Limited Capacity." *Management Science* 18(March 1972): 370–383.

Holloway, C. A. "A Mathematical Programming Approach to Identification and Optimization of Complex Operational Systems with the Aggregate Planning Problem as an Example." Ph.D. dissertation, University of California, Los Angeles, 1969.

Holt, C. C., F. Modigliani, and J. F. Muth. "Derivation of a Linear Decision Rule for Production and Employment." *Management Science* 2(January 1956): 159–177.

———; ———; ———; and H. A. Simon. *Planning Production, Inventories and Work Force*. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1960.

———; ———; and H. A. Simon. "A Linear Decision Rule for Production and Employment Scheduling." *Management Science* 2(October 1955): 1–30.

Hooke, R., and T. A. Jeeves. "A 'Direct Search' Solution of Numerical Statistical Problems." *Journal of the Association for Computing Machinery*, April 1961.

Jones, C. H. "Parametric Production Planning." *Management Science* 13(July 1967): 843–866.

Lee, S. M., and L. J. Moore. "A Practice Approach to Production Scheduling." *Production and Inventory Management* 1st Quarter (1974): 79–92.

Lee, W. B., and B. M. Khumawala. "Simulation Testing of Aggregate Production Planning Models in an Implementation Methodology." *Management Science* 20(February 1974): 903–911.

———, and C. P. McLaughlin. "Corporate Simulation Models for Aggregate Materials Management." *Production and Inventory Management* 1st Quarter (1974): 55–67.

Moskowitz, H. "The Value of Information in Aggregate Production Planning—A Behavioral Experiment." *AIIE Transactions* 4(December 1972): 290–297.

Redwine, C. N. "A Mathematical Programming Approach to Production Scheduling in a Steel Mill." Ph.D. dissertation, University of California, Los Angeles, 1971.

Silver, E. A. "A Tutorial on Production Smoothing and Work Force Balancing." *Operations Research* 15(November–December 1967).

Taubert, W. H. "Search Decision Rule for the Aggregate Scheduling Problem." *Management Science* 14(February 1968): 343–359.

______. "The Search Decision Rule Approach to Operations Planning." Ph.D. dissertation, University of California, Los Angeles, 1968a.

Tuite, M. F. "Merging Market Strategy Selection and Production Scheduling." *Journal of Industrial Engineering* 19(February 1968): 76–84.

Vergin, R. C. "Production Scheduling Under Seasonal Demand." *Journal of Industrial Engineering* 17(May 1966).

Chapter 11

Inventories in the System

THE FUNCTIONS OF INVENTORIES

Inventories make possible a rational production system. Without them we could not achieve smooth production flow, obtain reasonable utilization of machines and reasonable material handling costs, or expect to give reasonable service to customers on hundreds of items regarded as "stock" items. Inventories serve the vital function of *decoupling* the various operations in the sequence at each stage of both manufacturing and distribution—beginning with raw materials, extending through all the manufacturing operations and into finished goods storage, and continuing to warehouses and retail stores. Inventories make the required operations between each pair of activities in this sequence sufficiently independent of each other that low cost operations can be carried out.

Thus, when raw materials are ordered, a supply is ordered that is large enough to justify the out-of-pocket cost of putting through the order and transporting it to the plant. When production orders to manufacture parts and products are released, we try to make them big enough to justify the cost of writing the orders and setting up machines to perform the required operations. Otherwise, order writing and setup costs easily could become prohibitive. Running parts through the system in lots also tends to reduce handling costs because parts can be handled in groups. Similarly, in distributing finished products to warehouses and other stock points, freight and handling costs per unit go down if we can ship in quantity. Not only are inventories desirable, but they also are vital to low cost manufacture.

These advantages partially are lost when we are dealing with custom manufacture. Because the lot size is dictated by the customer's order, and because the order may never be repeated, we cannot risk making extras. Thus, order writing, material handling, and machine setup costs are just as high for the custom sized lot as for an economically size lot. Similarly, when the custom job is shipped to the customer, we cannot take advantage of carload or truckload freight rates. If the custom order is for only one part, all these changes must be absorbed by the single part. It is easy to see why low volume special orders are exceedingly expensive.

Unfortunately, the inventory question is not a one-sided one, which is precisely why inventories are a problem in the operation of a productive system. If there were no optimal level to shoot for, there would be no problem. Anyone could follow the simple rule: "Make inventories as big as possible." But since inventories require invested capital to be tied up, an appropriate opportunity cost is associated with their value. In addition, they require valuable space and absorb insurance and taxation charges. It is not uncommon for a manufacturing concern to have 25 percent of its total invested capital tied up in inventories; some companies have as little as 10 percent, and others have as much as 55 percent or more.

The significance of the inventory problem is indicated by the inventory changes of Eastman Kodak and U.S. Steel between 1971 and 1972. In 1971, Kodak had \$572.3 million of its assets in inventories. By 1972, this figure had increased to \$649 million, representing a \$15.3 million cost increase, at a carrying cost of 20 percent. In the same period U.S. Steel's inventories decreased from \$840 million to \$791 million, resulting in a decreased inventory carrying cost of almost \$9.9 million (if carrying costs are computed at 20 percent of inventory value).

We have one set of costs that are fixed by the purchase or production order size, and another set of costs that increase with the level of inventory. The first set of costs exerts pressure toward large purchase and production lots to reduce unit order writing and setup costs to a reasonable level. The second set of costs exerts pressure toward small lots in order to maintain inventory costs at reasonable levels. We are dealing with another problem where good management must find policies and decision rules that balance the various opposing costs for the system being considered.

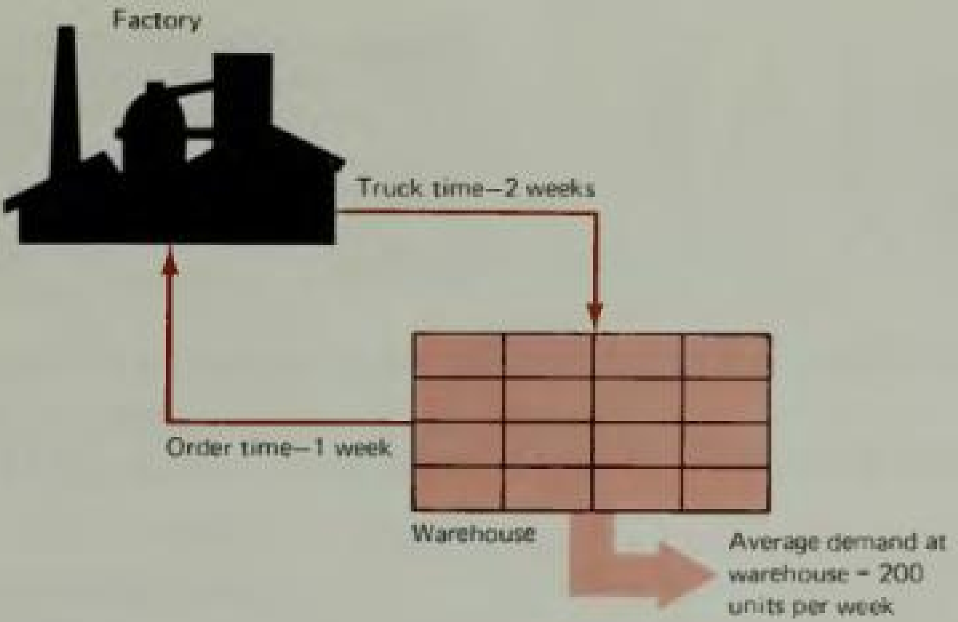


Figure 1. Simple factory-warehouse system

KINDS OF INVENTORIES

To see the various kinds of inventories that might be appropriate for analytical purposes, let us take as an example the simple factory-warehouse system shown in Figure 1. The factory manufactures a number of products, but we shall consider only one. For this particular product, there is an average demand at the warehouse of 200 units per week. The normal warehousing procedure is to prepare a procurement order to the factory when the warehouse inventory falls to a critical level, called the order point. It takes one week to prepare the order, get it approved and mailed, and finally have it be received at the factory. Once the order is received at the factory, it takes two weeks for loading, trucking, and unloading at the warehouse.

Pipeline Inventories

The warehouse must, at a minimum, carry enough stock on hand to meet demand during the transit time. Figure 2 shows an idealized graph of the inventories required at the warehouse just to cover trucking time from the factory. The average transit inventory is the product of the truck time and the demand rate, or $2 \times 200 = 400$ units. At all times, then, 400 units are in motion from the factory to the warehouse; the warehouse must maintain an inventory that takes account of this fact. The order time delay of one week has the effect of a

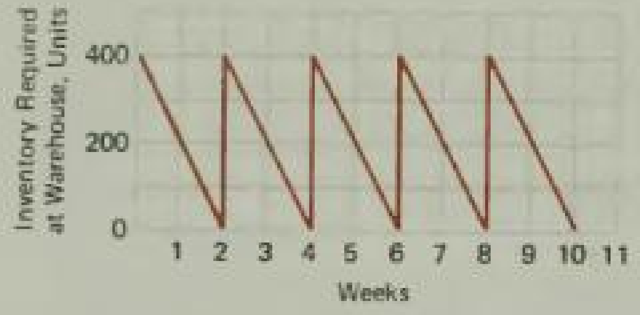


Figure 2. Idealized graph of inventories required at the warehouse to cover trucking time from the factory. The average transit inventory is the product of truck-time and demand rate, or $2 \times 200 = 400$ units. Average inventory at the warehouse is $400/2 = 200$ units

transit time, since the warehouse also must carry inventories to cover this delay. This point is a general one. Every lag in the system generates the need for inventory to fill the pipeline. Within a productive system, we call it *in-process inventory.*

Lot Size or "Cycle" Inventories

Let us return to our example. Since the truck is to make the trip from the factory to the warehouse, a logical question is how many to transport at one time. Most costs of trucking will occur anyway, and this includes the costs of preparing the requisition and the other clerical costs involved. We shall not attempt to answer the question of what the lot size should be in this instance, but the general question will be discussed later. Let us assume that orders are placed for a truckload of 800 units, equivalent to a four-week supply. Figure 3 shows an idealized graph of inventories required at the warehouse when orders are placed for a truckload. We see that the average inventory at the warehouse necessarily must increase.

Buffer Inventories

Of course, Figure 3 is completely unrealistic, because it assumes that demand rate, truck time, and order time all are constant. We know that these factors

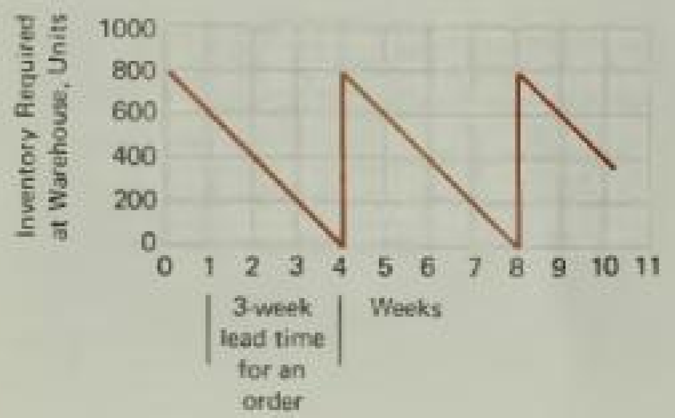


Figure 3. Idealized graph of inventories required at the warehouse when orders are placed for a truckload, 800 units. Average inventory is now $800/2 = 400$ units. Transit inventories are unchanged

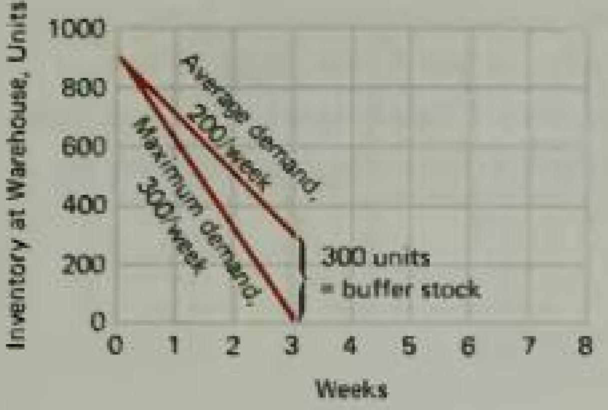


Figure 4. Contrast of idealized inventory levels required for average demand and maximum demand during lead time. The difference, 300 units, represents a minimum inventory or "buffer stock" required to protect against the occurrence of shortages resulting from random fluctuations in demand.

ordinarily are not constant, so a buffer stock is normally required to protect against unpredictable variations in demand and supply time. Figure 4 shows the contrast in inventory levels at the warehouse that might occur if the maximum demand of 300 units per week took place during the supply time of three weeks. In order not to run out of stock, a buffer inventory of 300 units would be required. As we shall see, techniques for taking account of demand variability have great importance in relation to inventory models.

Decoupling Inventories

We referred before to the decoupling function of inventories. Inventories make the required operations sufficiently independent of each other that low cost operations can be carried out. For example, with inventories, the warehousing operation in Figure 1 can proceed relatively independently of manufacturing. Similarly, the existence of inventories at the retail level makes it possible to carry on that function relatively independently; replenishment stocks are ordered only periodically. No extra inventories, over and above those for the other functions we have discussed, are necessary to provide for the decoupling function; but the existence of inventories provides the needed independence between stages.

Seasonal Inventories

Many products have a fairly predictable but seasonal pattern through the year. Where this is true, management has the choice of changing production rates over the year to absorb the fluctuation in demand, or of absorbing some or all of the fluctuation in demand with inventories. If we attempt to follow the demand curve through the seasons by changing production rates, the capital investment for the system must provide for the peak capacity; and we also must absorb costs for hiring, training, and separating labor, as well as overtime costs. The use of inventories often can help strike a better balance of these costs. The concepts and methods for handling this seasonal problem were covered in Chapter 10.

SIZE OF PURCHASE AND PRODUCTION ORDERS

Purchase Order Quantities

The opposing pressures of the various inventory-related costs produce a situation calling for policies and practices that strike a balance among these costs. The question of how many units to buy at one time may be conceptualized by observing the behavior of the costs affected by changes in lot sizes.

First, let us define the system with which we are dealing. Figure 5 is a diagram of what happens to inventory levels for a particular item. If we assume an annual requirement of $R = 2{,}400$ units (demand is assumed constant), or an average of 200 per month, the inventory levels would fluctuate as in Figure 5*a* if we were to order in lots of $Q = 200$. The average inventory level for the idealized situation shown is one-half the number ordered at one time, or $Q/2 = 100$ units. If the item is ordered more often in smaller quantities, as shown in Figure 5*b*, we can see that the inventory level will fall in proportion to the number of units ordered at one time. Of course, inventory level will affect the incremental costs of holding inventory; so in effect, these costs of carrying inventory are proportional to the lot size Q, the number ordered at one time. Figure 5 also shows that the total annual cost of placing orders increases as the number of units ordered at one time decreases. Therefore, we have isolated two types of incremental costs that represent the measures of effectiveness for the system we are developing: the costs associated with inventory level, *holding costs*, and with the number of orders placed, *preparation costs*.

In further defining our system, let us attempt to construct a graph that shows the general relation between Q (lot size) and the incremental costs that we have isolated. We noted in Figure 5 that when Q was doubled, average inventory level was doubled. It costs, perhaps, $c_H = \$0.60$ per year to carry a unit of inventory (interest costs, insurance, taxes, etc.). Since the average inventory level is $Q/2$ and the inventory cost per unit per year is $c_H = \$0.60$, then the annual incremental costs associated with inventory are:

$$\frac{Q}{2} c_H = \frac{Q}{2}(0.60) = 0.30Q \tag{1}$$

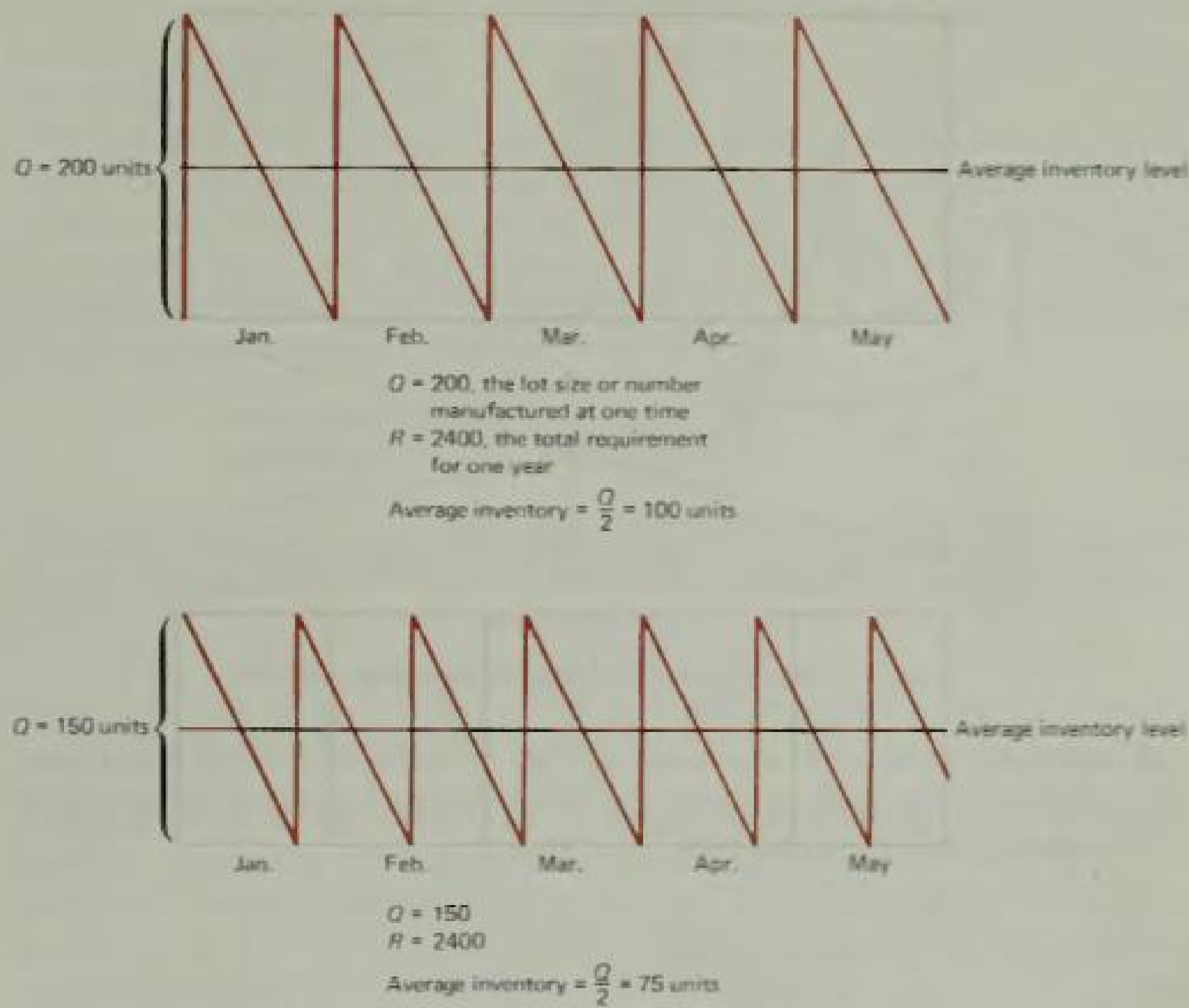


Figure 5. Simplified model of the effect of lot size on inventory levels

Substituting different values for Q, we can plot the result as in Figure 6, curve *a*.

We can plot the costs of ordering in a similar way. Let us use the symbol c_p to represent the costs of preparing and following up on an order. We calculate the number of orders by $R/Q = 2{,}400/Q$. If each order preparation costs $c_p = \$20$, then the total annual incremental cost due to ordering is:

$$\frac{R}{Q}c_p = \frac{2{,}400 \times 20}{Q} = \frac{48{,}000}{Q} \tag{2}$$

Therefore, as Q increases, the annual incremental cost due to ordering decreases. This makes good sense since, as Q increases, the number of order preparations decreases, and therefore the incremental cost decreases. This relationship is plotted in Figure 6, curve *b*, by substituting different values of Q in 48,000/Q.

Figure 6, curve c, shows the resulting total incremental cost curve, which is determined simply by adding the two previous curves. Thus we have a model

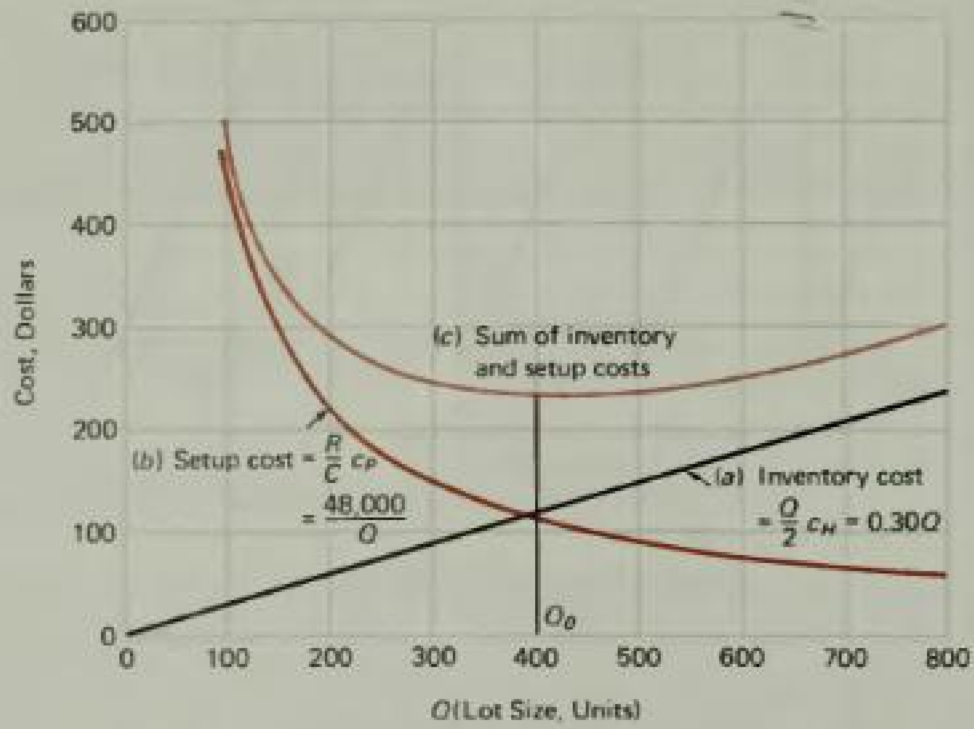


Figure 6. Graphical model of simple inventory problem

that expresses the objective as a function of the variables that define the system. The equation for the total cost curve also is determined by adding the equations for the two separate cost functions. Thus we have:

$$C = \frac{Q}{2} c_H + \frac{R}{Q} c_P \tag{3}$$

$$\begin{matrix}\text{Total incre-} \\ \text{mental cost}\end{matrix} = \begin{pmatrix}\text{average} \\ \text{inventory}\end{pmatrix} \begin{pmatrix}\text{unit inventory} \\ \text{cost per year}\end{pmatrix} + \begin{pmatrix}\text{number of} \\ \text{orders per year}\end{pmatrix} \begin{pmatrix}\text{cost of} \\ \text{an order}\end{pmatrix}$$

Which variables in Equation 3 are controllable and which are uncontrollable? Controllable variables are those which may be manipulated pretty much at the will of management. In our example, the controllable variable is Q (lot size). Uncontrollable variables are those which management cannot control, at least not within the limitations of the problem. For example, consumer demand, taxes, and insurance rates would fall into this category. For our example, uncontrollable variables are c_H (inventory costs per unit), R (demand or requirement), and c_p (order preparation costs). We might argue that management can influence c_H by bringing pressure to bear to have taxes or insurance rates reduced. Nevertheless, management cannot change taxes and insurance rates at will. In general, tax rate changes would be reflected in C. Management would respond by adjusting Q to minimize C again. The new policy would be the best policy, given the new tax rate.

A General Solution. For our simplified inventory model, we can pick off the minimum point from the total incremental cost curve of Figure 6, curve *c*. It is a minimum for $Q_0 = 400$ units. (The symbol Q_0 denotes the optimal value of Q.) This is a solution for the specific problem with the given values of c_H, R, and c_p. The important thing about this model, however, is that we can obtain a general solution that will give us the minimum point directly for any values of c_H, R, and c_p. From the equation for the total incremental cost, we may derive a formula for the minimum point on the curve by means of differential calculus. The formula that represents the general solution for the model is:

$$\boxed{Q_0 = \sqrt{2c_pR/c_H}} \tag{4}$$

This formula gives directly the value of Q_0, which yields the minimum total incremental cost for the model. Substituting the values of our example, we have:

$$Q_0 = \sqrt{\frac{2 \times 20 \times 2{,}400}{0.60}} = \sqrt{160{,}000} = 400 \text{ units}$$

In using the formula, if we desire to express the economic quantity in dollars, then the requirements also must be expressed in dollars. Similarly, if the requirements are expressed in monthly rates, then the inventory costs also must be expressed as a monthly rate. These and other changes in the units lead to apparent modifications of the formula when it is used in practice, although the end-result is the same, in terms of the size of the order. In practice, the formula itself is not used often; rather, charts, graphs, and tables based on the formula are used to minimize computations, or the computations are automated in modern computing systems.

The Effect of Quantity Discounts. The basic economic lot size formula assumes a fixed purchase price. When quantity discounts enter the picture, additional simple calculations determine whether there is a net advantage. As an illustration, assume that a manufacturer's requirement for an item is 2,000 per year. The purchase price is quoted as \$2 per unit in quantities below 1,000, and \$1.90 per unit in quantities above 1,000. Ordering costs are \$20 per order, and inventory costs are 16 percent per year per unit of average inventory value, or \$0.32 per unit per year at the \$2 unit price. Our formula indicates that the economic order quantity is:

discount
R = 2000 unit
< 1000 \$2
> 1000 \$1.9

$$Q_0 = \sqrt{\frac{2 \times 20 \times 2{,}000}{0.32}} = \sqrt{250{,}000} = 500 \text{ units}$$

However, when quantity discounts enter the picture, the total incremental cost equation is no longer a continuous function of order quantity but becomes a step-function, and has components of annual inventory costs, ordering cost, and material cost involving the price-discount schedule. The total incremental-cost equation becomes:

$$\boxed{C = c_H(Q/2) + c_P(R/Q) + p_iR} \tag{5}$$

where p_i is the price per unit for the *i*th price break, and $c_H = p_iF_H$ where F_H is the *fraction* of an inventory value.

Using the preceding data and Equation 5, we can compute C for each of the three price ranges, as shown in Figure 7 (the solid line curves indicate the relationships for valid price ranges):

1. Note that, for the two-dollar price that applies for $Q < 1{,}000$, $EOQ = 500$ units produces the lowest cost of \$4,160.
2. However, between order quantities of 1,000 and 1,999, the price of \$1.90 per unit applies, and when $Q = 1{,}000$, $C = \$3{,}992$–a cost saving of \$168 per year, compared with ordering in lots of 500 units.
3. Finally, at $Q \geq 2000$, the price of \$1.86 per unit applies, and $C = \$4{,}038$ at $Q = 2000$ units.

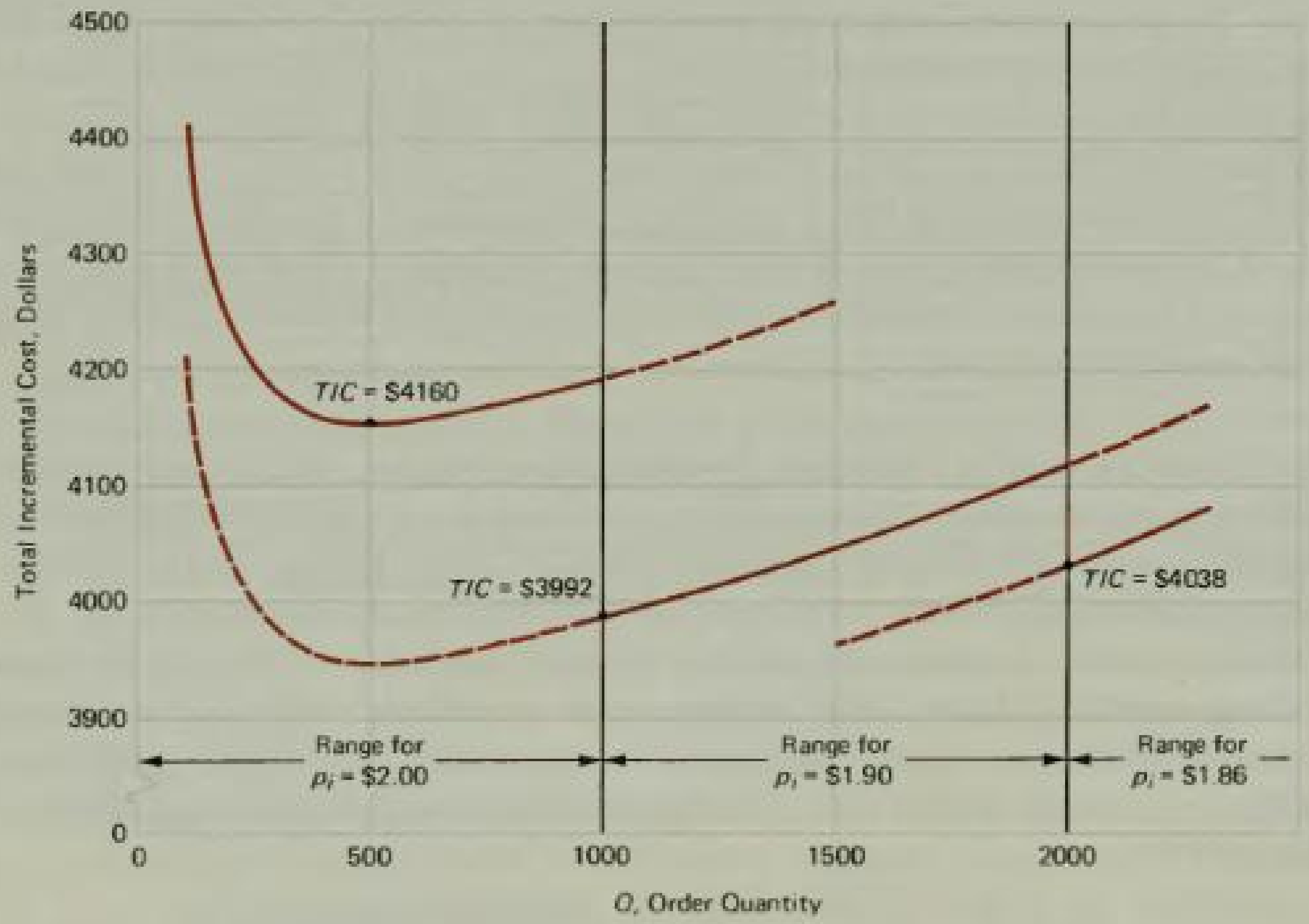


Figure 7. Total incremental cost curves for inventory model with three price breaks. $R = 2000$ units per year, $c_p = \$20$, $F_H = 0.16$

TABLE 1
INCREMENTAL COST ANALYSIS TO DETERMINE NET ADVANTAGE OR DISADVANTAGE WHEN PRICE DISCOUNTS ARE OFFERED

	Lots of 500 Units Price = \$2.00 per Unit	Lots of 1000 Units Price = \$1.90 per Unit	Lots of 2000 Units Price = \$1.86 per Unit
Purchase of a year's supply ($p_i \times 2000$)	\$4000	\$3800	\$3720
Ordering cost ($20 \times 2000/Q$)	80	40	20
Inventory cost (average inventory $\times$ unit price $\times$ 0.16)	80	152	298
Total	\$4160	\$3992	\$4038

The lowest cost ordering policy is to take advantage of the first price break but not the second, and to order in lots of $Q = 1{,}000$ units. Summary calculations are shown in Table 1.

Production Order Quantities

As we have noted, production order quantities are based on the same general concepts as purchase order quantities, and Equation 4 can be applied. (Preparation costs are costs related to machine setups, and incremental paper work costs to write production orders and control the flow of orders through the shop; inventory costs are costs associated with holding in-process inventory.) However, the assumption that the order is received and placed into inventory all at one time often is not true in manufacturing runs. The basic formula assumes the general inventory pattern shown in Figure 8*a*, where the order quantity Q is received into inventory. The inventory then is drawn down at the usage rate, and subsequent orders are placed with sufficient lead time so that their receipt coincides with zero (or minimum stock) inventory.

For many manufacturing situations, the production of the total order quantity Q takes place over a period of time, and the parts go into inventory not in one large batch but in smaller quantities as production continues. This results in an inventory pattern similar to Figure 8*b*. The maximum and average inventory levels are reduced. The minimum cost production quantity formula becomes:

$$Q_p = \sqrt{\frac{2c_pR}{c_H\left(1 - \frac{r}{p}\right)}} \tag{6}$$

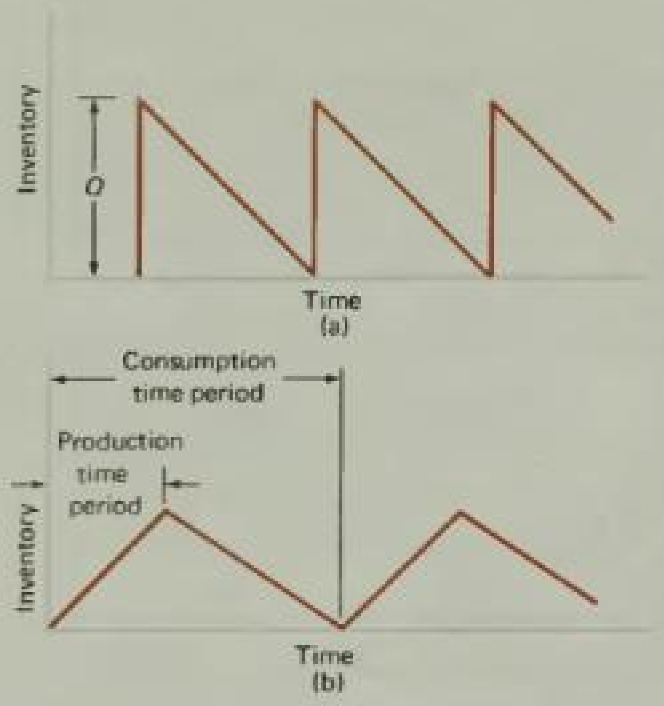


Figure 8. Comparison of inventory balance: (a) when the order quantity, Q, is received all at one time, and (b) when Q is received over a period of time, as would be true in many production situations

where

r = requirements or usage rate (short-term, perhaps daily or weekly)
p = production rate (on the same time base as for r)
Q_p = minimum cost production order quantity

Of course, the Q_p that results is larger than that which would result from the basic formula. Average inventory is smaller for a given Q_p, and so balance between setup costs and inventory costs takes place at a higher value of the order quantity Q_p. By dividing the annual requirement by Q_p, we have the number of production cycles for the part or product which should be produced in a year.

D = f(prod. sched.)

A word of caution is in order concerning the application of EOQ concepts for the size of production runs. The assumption of constant demand often is not a good one because demand for an item more commonly is a function of production schedules, creating what is termed "lumpy" demand. Production order quantities are placed in this general context (in which the assumption of constant demand is questionable), in the following chapter under the heading of "Requirements Planning."

Production Runs for Several Parts or Products. The general problem does not involve the determination of the economical length of a production run for a single part or product, but the joint determination of the runs for an entire group of products that share the use of the same facilities. If each part or product run is set independently, some conflicts of equipment needs are likely to result unless the operating level is somewhat below capacity, and considerable idle equipment time is available. The formal problem is similar to the one-product case, but the objective is to minimize combined set up and inventory costs for the entire set (See Buffa [1972]; Buffa and Taubert [1972]; and Magee and Boodman [1967]).

DEALING WITH VARIABILITY OF DEMAND

In the basic EOQ model and in the simple extensions that we developed, we assumed that both demand and supply lead time were constant. Yet variability of demand and supply lead time are elements of reality which can be very important because they impose risks that commonly are two-sided.

We can cushion the effects of demand and supply lead time variation by absorbing risks in carrying larger inventories, called *buffer* or *safety stocks*. The larger we make these buffer stocks, the greater our risk, in terms of the funds tied up in inventories, the possibility of obsolescence, and so on. However, we tend to minimize the risk of running out of stock. On the other hand, although the inventory risk can be minimized by reducing buffer inventories, the risks associated with poor inventory service increase, including the costs of back ordering, lost sales, disruptions of production, and so on. Our objective, then, will be to find a rational model for balancing these risks.

Service Levels and Buffer Stocks (Constant Lead Time)

Let us begin by dealing with the effects of demand variability and assuming that supply lead time is known and constant. Figure 9 shows the general structure of inventory balance by means of a reordering system similar to the one we developed before. When inventory falls to a preset reorder point P, an order for the quantity Q is placed. The reorder point P is set to take account of the supply lead time L, so that if we experience normal usage rates during L, inventory is reduced to minimum levels when the order for Q units is received. Note, however, that demand may not occur at a constant rate. Inventories may decline to the reorder point P earlier or later than expected. But more importantly, if demand during lead time is greater than expected values, inventory levels may decline below the minimum or planned buffer stock level. In the limiting situation, if we experience maximum demand during lead time (as shown in Figure 9), inventory levels will decline to zero by the time the order for Q units is

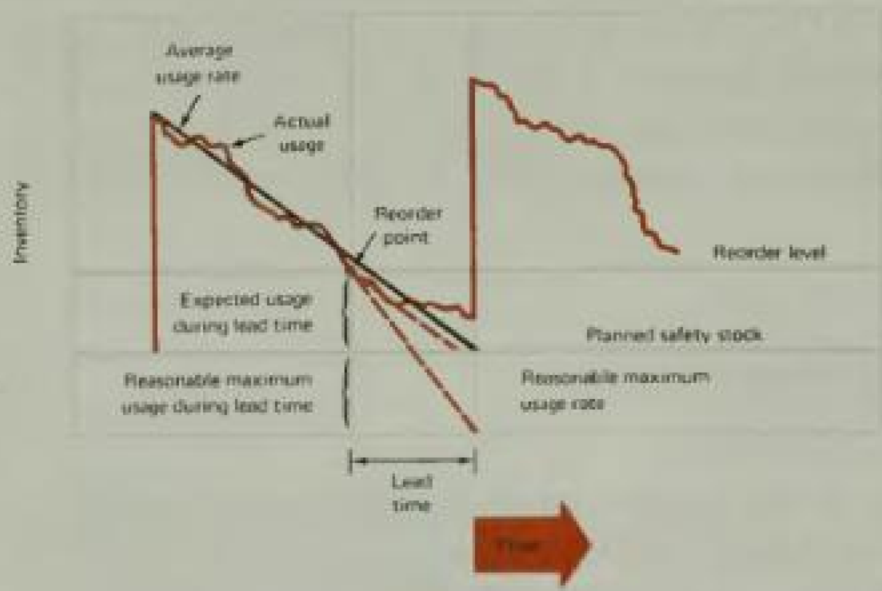


Figure 9. Structure of inventory balance for a fixed order quantity system, with stocks to absorb fluctuations in demand and in supply time. The stock level is set so that a reasonable figure for maximum usage would draw down the inventory to zero during the lead time

received. The size of the needed buffer stock, then, is the difference between the expected or average demand $\bar{D}$ and the maximum demand D_{max} during the supply lead time, or $\boxed{B = D_{max} - \bar{D}}$. The issue concerns how we define D_{max}.

Defining D_{max}. Maximum demand is not a fixed number that we can simply abstract from a distribution of demand, but depends on an analysis of the risks. Let us take as an example the record for the distribution of demand that exceeds a given level for an item shown in Figure 10. This figure represents only the random variations; if there were other effects, such as trend and seasonals, they have been removed by standard statistical techniques. We note that average monthly demand was $\bar{D} = 460$ units.

Since the average monthly usage rate is 460 units, and if we assume a lead time of $L = 1$ month, we could be 90 percent sure of not running out of stock by having 620 units on hand when the replenishment order is placed (see Figure 10 for the demand rate associated with 10 percent). The buffer stock required for this 90 percent service level is $B = 620 - 460 = 160$ units. Similarly, if we wish to be 95 percent sure of not running out of stock, then $B = 670 - 460 = 210$ units. For a 99 percent service level (1 percent risk of stock out), the buffer stock level must be increased to 320 units.

From the shape of the demand curve, it is clear that required buffer stock goes up rapidly as we increase service level, and, therefore, the cost of providing this assurance goes up. These effects are shown by the calculations in Table 2, in which we have assumed the demand curve of Figure 10, assigning a value of $100 to the item and inventory holding costs of 25 percent of inventory value. The average inventory required to cover expected maximum usage rates during the one-month lead time is calculated for the three service levels shown. To offer service at the 95 percent level instead of the 90 percent level requires an incremental $1,250 per year; but to move to the 99 percent level of service from the 95 percent level requires an additional $2,750 in inventory cost.

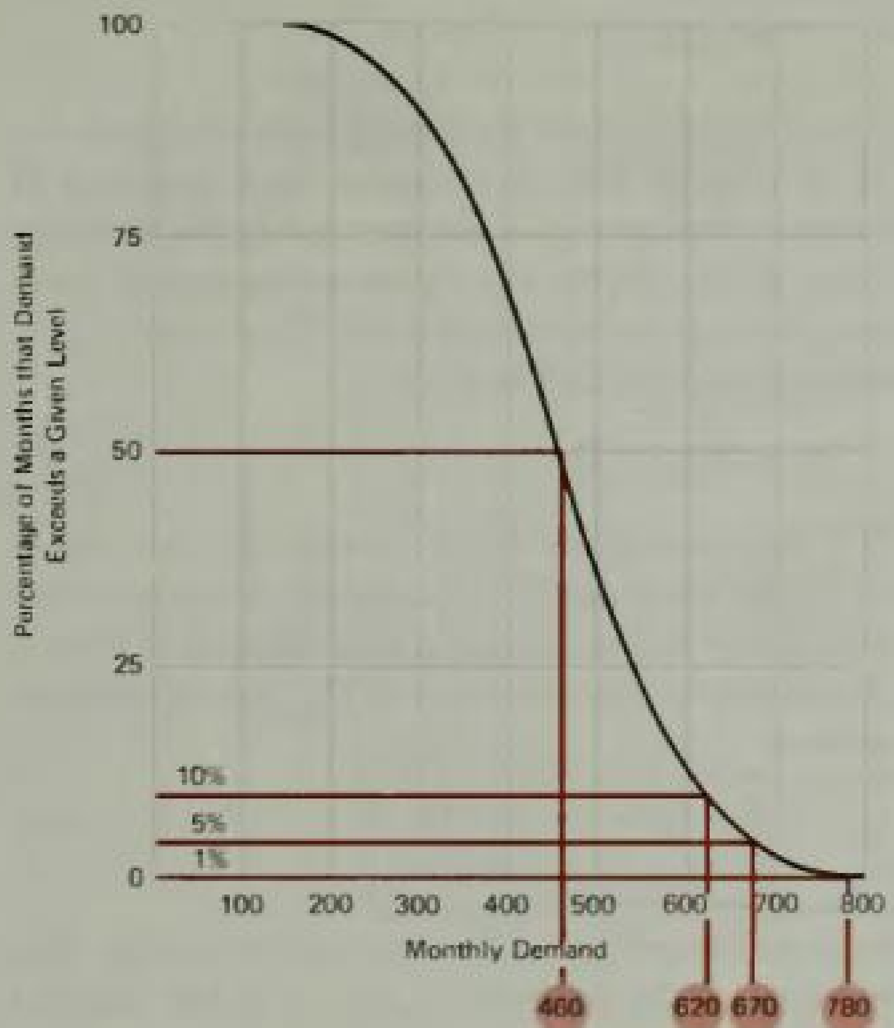


Figure 10. Distribution showing the percentage of months that demand exceeded a given level

Management could define any of the three levels of demand as D_{max} by setting a service level policy. Given the service level policy, the buffer stock required to implement that policy is, simply, $B = D_{max} - \bar{D}$.

TABLE 2

COST OF PROVIDING THREE LEVELS OF SERVICE IN FIGURE 10
(ITEM IS VALUED AT $100 EACH AND INVENTORY HOLDING COSTS ARE 20 PERCENT)

	Service Level		
	90%	95%	98%
Expected maximum usage for one month replenishment time	620	670	780
Buffer stock required, $B = D_{max} - 460$	160	210	320
Value of buffer stock, $100 \times B$	$16,000	$21,000	$32,000
Inventory holding cost at 25 percent	$ 4,000	$ 5,250	$ 8,000

(16000 × .25 = 4000)

Practical Methods for Determining Buffer Stocks

The general methodology that we have discussed for setting buffer stocks is too cumbersome for practical use in systems that may involve large numbers of items. Computations are simplified considerably if we can justify the assumption that the demand distribution follows some particular mathematical function, such as the normal, Poisson, or negative exponential distributions.

First, let us recall the general statement for buffer stocks:

$$B = D_{max} - \bar{D} \tag{7}$$

Note, however, that $D_{max} = \bar{D} + n\sigma_D$; that is, the defined reasonable maximum demand is the average demand $\bar{D}$, plus some number of standard deviation units n that is associated with the probability of occurrence of that demand (n now is defined as the safety factor). Substituting this statement of D_{max} in our general definition of B, Equation 7, we have:

$$\boxed{\begin{aligned} B &= D_{max} - \bar{D} = (\bar{D} + n\sigma_D) - \bar{D} \\ &= n\sigma_D \end{aligned}} \tag{8}$$

This simple statement allows us to determine easily those buffer stocks that meet risk requirements when we know the mathematical form of the demand distribution. The procedure is as follows:

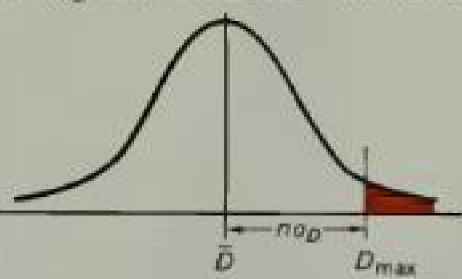
1. Determine whether the normal, Poisson, or negative exponential distribution approximately describes demand during lead time for the case under consideration. This determination is critically important, involving well known statistical methodology.
2. Set a service level based on (a) managerial policy; (b) an assessment of the balance of incremental inventory and stock out costs; or (c) an assessment of the manager's trade-off between service level and inventory cost when stock out costs are not known.
3. Using the service level, define D_{max} during lead time in terms of the appropriate distribution.
4. Compute the required buffer stock from Equation 8, where n is termed the safety factor and σ_D is the standard deviation for the demand distribution.

We will illustrate the methodology in the context of the normal distribution.

Buffer Stocks for the Normal Distribution. The normal distribution has been found to describe many demand functions adequately, particularly at the factory level of the supply–production–distribution system [Buchan and Koenigsberg, 1963]. Given the assumption of normality and a service level of perhaps 95

TABLE 3

AREA UNDER THE RIGHT TAIL OF THE NORMAL DISTRIBUTION (SHOWING THE PROBABILITY THAT DEMAND EXCEEDS $\bar{D} + n_{\sigma_D}$ FOR SELECTED VALUES OF n)



$D_{max} = n\sigma_D$	Probability	
		99.9%
$\bar{D} + 3.090\sigma_D$	0.001	99%
$\bar{D} + 2.576\sigma_D$	.005	
$\bar{D} + 2.326\sigma_D$	.010	97.5%
$\bar{D} + 1.960\sigma_D$	.025	95%
$\bar{D} + 1.645\sigma_D$	.050	
$\bar{D} + 1.282\sigma_D$	0.100	90%
$\bar{D} + 1.036\sigma_D$	.150	
$\bar{D} + 0.842\sigma_D$	.200	80%
$\bar{D} + 0.674\sigma_D$	.250	75%
$\bar{D} + 0.524\sigma_D$	.300	70
$\bar{D} + 0.385\sigma_D$	0.350	65
$\bar{D} + 0.253\sigma_D$	.400	60
$\bar{D} + 0.126\sigma_D$	.450	55
$\bar{D}$	.500	50

percent, we can determine B by referring to the normal distribution tables, a small part of which has been reproduced as Table 3. The normal distribution is a two-parameter distribution, which is described completely by its mean value $\bar{D}$ and the standard deviation σ_D. Implementing a service level of 95 percent means that we are willing to accept a 5 percent risk of running out of stock. Table 3 shows that demand exceeds $D + n\sigma_D$ with a probability of 0.05, or 5 percent of the time, when $n = 1.645$; therefore, this policy is implemented when $B = 1.645\ \sigma_D$. As an example, if the estimate of σ_D is $s = 300$, and $\bar{D} = 1{,}500$ units, assuming a normal distribution, a buffer stock to implement a 95 percent service level would be $B = 1.645 \times 300 = 494$ units. Such a policy would protect against the occurrence of demands up to $D_{max} = 1{,}500 + 494 = 1{,}994$ units during lead time. Obviously, any other service level policy could be implemented in a similar way.

Buffer Stocks When Both Demand and Lead Time Are Variable. The problem of determining buffer stocks when both demand and lead time vary is somewhat more complex. When lead times as well as demand vary, we have an interaction between the fluctuating demand and the fluctuating lead times, similar to the situation shown in Figure 11.

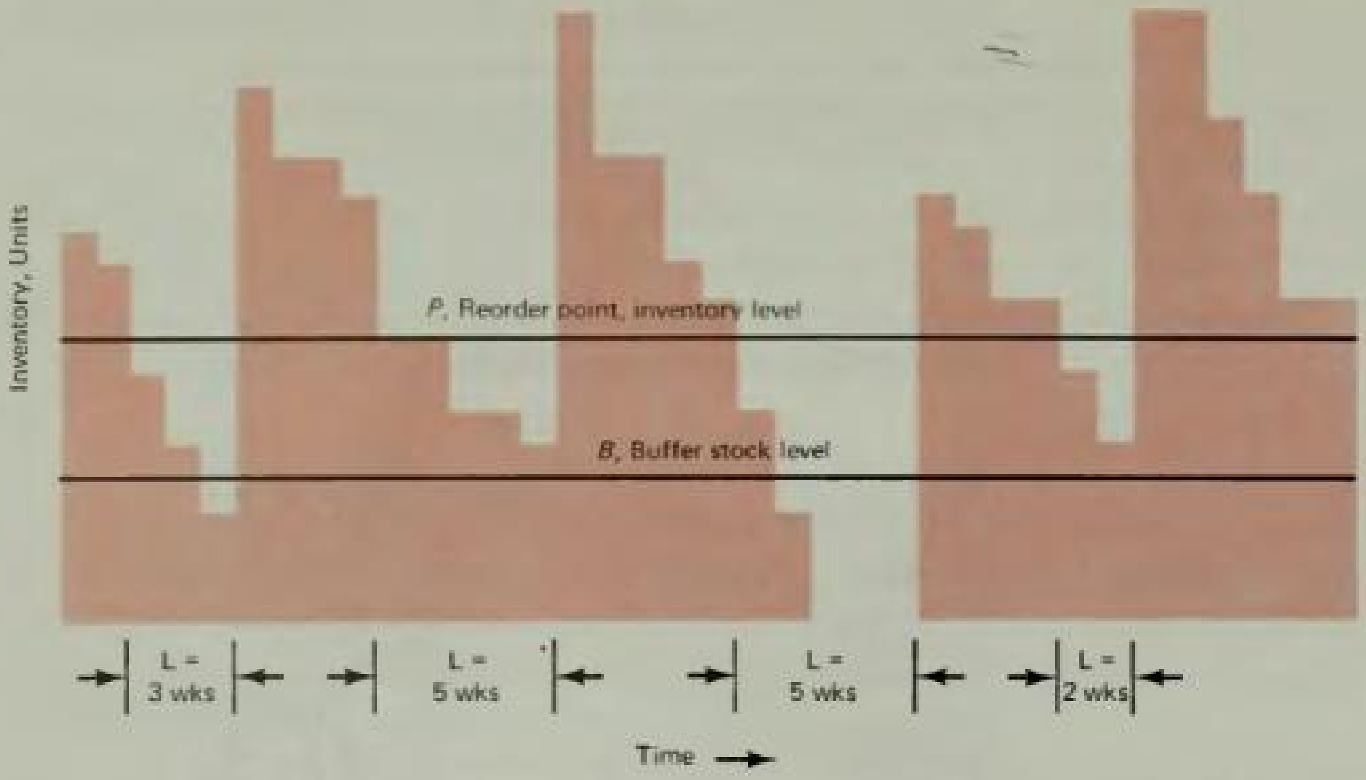


Figure 11. Inventory balance when both demand and lead time vary. When inventory falls to the reorder point P, the quantity Q is ordered. Inventory falls below the buffer stock level B twice, and a stock out occurs during the third cycle

In such situations, a Monte Carlo simulation may determine buffer stocks by means of a straightforward application of the methods presented in Chapter 2. To carry out the simulation, we need data describing both the demand and lead time. Then we can develop buffer stock requirements for the various risk levels of stock out. We can implement whatever risk level we choose by selecting the corresponding buffer stock. The methodology and computed examples are developed in Buffa and Taubert [1972], and McMillan and Gonzalez [1973].

DETERMINING SERVICE LEVELS

The service level states the probability that all orders can be filled directly from inventory during a reorder cycle. As we have stated, the buffer inventory that is designed to provide for the risk of stock out is $B = n\sigma_D$. Assuming a normal distribution and a safety factor of $n = 1.645$, then the chance of a stock out is 0.05 (from Table 3), and the service level is 95 percent.

Now let us examine more closely the meaning of a service level statement of policy. It means that there is one chance in twenty that demand during lead time will exceed the buffer stock *when there is exposure to risk*. It does not mean that

5 percent of the demand is unsatisfied, but that demand during lead time can be expected to exceed buffer stock for 5 percent of the replenishment orders, or that the chance that demand will exceed the buffer stock for any given replenishment order is 5 percent.

Effect of Order Size

This interpretation of service level immediately shows us that the expected quantity short over a period of time is proportional to the number of times we order, since we are exposed to shortages only once for each reordering cycle. For our example, if we ordered Q units twenty times per year, we would expect shortages to occur an average of only once per year. If we ordered in quantities of 2Q only ten times per year, we would expect stock outs to occur only once every other year, on the average. Larger orders provide exposure to risk less often, and will result in lower annual expected quantities short for the same service level. but greater inventory turns

Expected Quantities Short

For a given safety factor and distribution of demand during lead time, we can compute the expected quantity short. Assuming a normal distribution of $\bar{D} = 50$ units during lead time and $\sigma_D = 10$ units, let us determine the expected quantity short for service levels of 80, 90, 95, and 99 percent. Based on the safety factors for the stated service levels (Table 3), the computed buffer stocks are:

80 $B_1 = 0.842 \times 10 = 8.4$, or 9 units
90 $B_2 = 1.282 \times 10 = 12.8$, or 13 units
95 $B_3 = 1.645 \times 10 = 16.45$, or 17 units
99 $B_4 = 2.326 \times 10 = 23.26$, or 24 units

The values of B approximate the stated service levels and give slightly better service because of rounding upwards to integer units.

Brown [1963] has shown that the expected quantity short per order is the product of σ_D and $E(k)$, where $E(k)$ is the partial expectation for a distribution with unit standard deviation. The partial expectation is the expected value of demands beyond some specified level. Brown [1967] developed tables of partial expectations for the normal distribution, and provided a graph of the function which is reproduced in Figure 12. Estimates of the expected quantity short per order can be obtained from Figure 12 for a given safety factor which, in turn, is

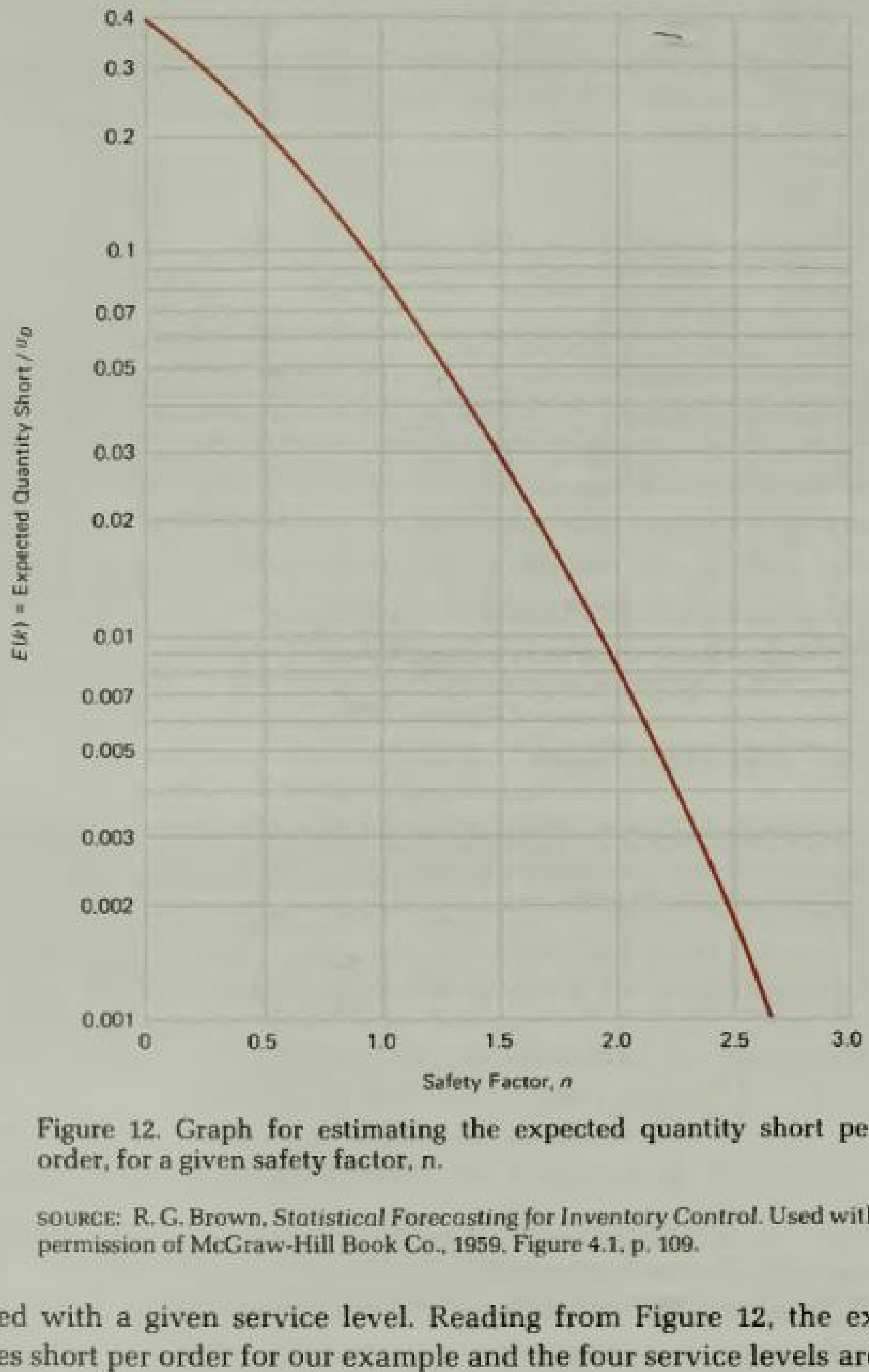


Figure 12. Graph for estimating the expected quantity short per order, for a given safety factor, n.

SOURCE: R. G. Brown, *Statistical Forecasting for Inventory Control*. Used with permission of McGraw-Hill Book Co., 1959. Figure 4.1, p. 109.

associated with a given service level. Reading from Figure 12, the expected quantities short per order for our example and the four service levels are:

Service Level, Percentage	Expected Quantity Short Per Order, Units
80	$0.10 \times 10 = 1.00$
90	$0.04 \times 10 = 0.40$
95	$0.02 \times 10 = 0.20$
99	$0.003 \times 10 = 0.03$

For each of the service levels indicated, the expected quantity short for each order is as given, and these expected shortages are rather startlingly small. Thus, the effect of a given service policy may be misleading unless it is translated into its equivalent expected quantity short per order. While a 90 percent service policy may seem relatively loose, it holds fairly tight control in terms of the expected shortages on each ordering cycle.

Optimal Service Levels, Shortage Costs Known

Now that we have methods for estimating the expected quantity short per order, we can determine the optimal service level if we know the relevant costs. Let us slightly amplify the example that we have been using. Suppose that annual requirements for the example item are $R = 3{,}000$ units per year; inventory holding costs are $c_H = \$20$ per unit per year; ordering costs are $c_P = \$25$ per order; and shortage costs are $c_S = \$100$ per unit short. If the order quantity were $Q = 500$ units, 6 orders per year would be required.

Let us examine the annual buffer inventory and shortage costs for the four different service levels. We already computed the buffer inventory and expected quantities short per order. Since there are six orders per year, the *annual* expected quantity short is $6\sigma_D E(k)$. These values and the relevant costs are summarized in Table 4. The service policy that minimizes relevant costs for these data is the 95 percent policy that involves maintaining a buffer of $B = 17$ units, and that results in an annual expected quantity short of $6 \times 0.20 = 1.20$ units and a minimum total relevant cost of \$460 per year. What would be the optimal service policy if the cost of shortages was only $c_S = \$40$?

Optimal Service Levels, Shortage Costs Unknown

Frequently it is true that we do not know the value of c_S with any degree of confidence. Many factors in a given situation may affect the true cost of shortages. Some of these factors may be reasonably objective but difficult to measure. For example, although part shortages in assembly processes create costly disruptions and delays, measuring these costs is quite another matter. Another example is when shortages occur; then it may be necessary to place the parts on back order, or to expedite them with special handling and extra costs. These incremental costs are real, but they are not segregated in cost records. Thus, making realistic estimates of their value would be costly in itself, and these costs could exceed the value of the information. If a shortage definitely results in a lost sale, we can impute a shortage cost, equal to the lost contribution. But do we

TABLE 4

ANNUAL BUFFER INVENTORY AND SHORTAGE COSTS FOR FOUR SERVICE LEVELS (D = 50 UNITS DURING TIME, σ_D = 10 UNITS, c_H = \$20 PER UNIT PER YEAR, c_S = \$100 PER UNIT SHORT, R = 3000 UNITS PER YEAR, AND Q = 500 UNITS PER ORDER)

	Approximate Service Level, Percentage			
	80	90	95	99
Buffer inventory*, $B = n\sigma_D = 10 \times n$ (p. 359)	9	13	17	24
Expected quantity short per order**, $\sigma_D E(k) = 10 \times E(k)$	1.0	0.4	0.20	0.03
Buffer inventory cost, $c_H B = 20 \times B$	\$180	\$260	\$340	\$480
Shortage cost, $c_S(R/Q) \times$ (expected quantity short per order) $= 100 \times 6 \times$ (expected quantity short per order)	\$600	\$240	\$120	\$ 18
TOTAL INCREMENTAL COSTS	\$780	\$500	\$460	\$498

*Values of n from Table 3 for given service level. Values of B rounded to next highest integer.
**$E(K)$ estimated from Figure 12.

know whether the sale is lost for certain, or must we merely estimate the probability of a lost sale? Finally, the loss may be intangible, such as the loss of goodwill of a valued customer who receives poor service.

For all the preceding reasons, we may not be able to estimate values of c_S with sufficient precision to justify an analysis similar to that given in Table 4 as a basis for selecting an optimum service level policy.

Nevertheless, in the absence of known shortage costs, we still have valuable data from Table 4. We have objective annual buffer inventory costs for various service level policies, and we have the expected annual quantities short that would result from each service level policy. The graphical relationship between buffer inventory cost and quantities short is shown in Figure 13. These data provide the manager with the basis for determining his own utility function so he can make a trade-off analysis.

MANAGERIAL CONTROL SYSTEMS

Let us now focus on the results of inventory models for managerial use. What variables are under managerial control, and how can they be incorporated into

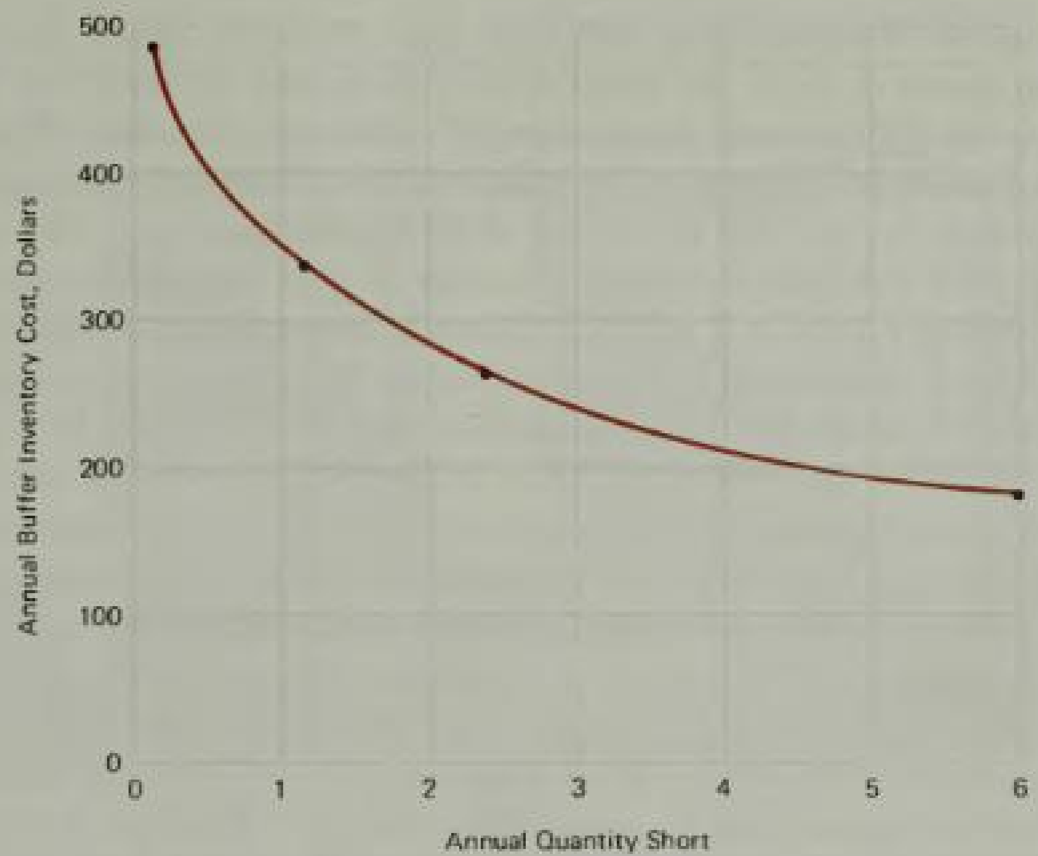


Figure 13. Annual buffer inventory costs versus annual quantity short for a system where: $D = 50$ units during lead time, $\sigma_D = 10$ units, $c_H = \$20$ per unit per year, $R = 3000$ units per year, and $Q = 500$ units per order

useful control systems? The basic control variables are the quantity ordered at one time, the service level, and the particular way in which these elements are combined in a control system.

The elements of inventory models that normally lie outside of managerial control are the costs c_H, c_p, and c_s, and the supply lead time. These are parameters that management cannot change at will, though it may try to reduce or control them in the longer term. The annual requirement, R, is dependent on external market factors; and again, although the manger can try to influence R through marketing techniques, he/she cannot purposely decide what it will be.

Although we recognize that there are many variations in practice, we shall summarize two basic kinds of control systems that are widely used.

The Fixed Reorder Quantity System

We have used the fixed reorder quantity system for illustrative purposes in developing inventory models. Its structure is best illustrated by Figure 9, in which a reorder level has been set by the point P, which allows the inventory level to be drawn down to the buffer stock level within the lead time if average usage rates are experienced. Replenishment orders are placed in a fixed, predetermined amount (in practice, not necessarily the minimum cost quantity, Q_0) that is timed to be received at the end of the supply lead time. The maximum inventory level becomes the order quantity Q plus the buffer stock B. The average inventory, then, is $B + Q/2$. Usage rates are reviewed periodically in an attempt to react to seasonal or long-term trends in requirements. At the time of the periodic reviews, the order quantities and buffer stock levels may be changed to reflect the new conditions. Buffer stock levels are set based on determinations of the appropriate service level policy. This policy reflects the balancing of buffer inventory costs and shortage costs, or the manger's tradeoff between buffer inventory cost and the expected quantity short.

Fixed reorder quantity systems are common where a perpetual inventory record is kept, or where the inventory level is under sufficiently continuous surveillance that notice can be given when the reorder point has been reached. One of the simplest methods for maintaining this close watch on inventory level is the use of the "two-bin" system. In this system, the inventory is physically (or conceptually) separated into two bins, one of which contains an amount equal to the reorder inventory level. The balance of the stock is placed in the other bin, and day-to-day needs are drawn from it until it is empty. At this point, it is obvious that the reorder level has been reached, and a stock requisition is issued. Then stock is drawn from the second bin, which contains an amount equal to the average use over the lead time plus a buffer stock. The stock is replenished when the order is received, and the physical segregation into two bins is made again; the cycle then is repeated. Fixed reorder quantity systems are common with low-valued items, such as nuts and bolts.

Fixed Reorder Cycle Systems

In fixed cycle systems, control is maintained by ordering regularly on some fixed cycle, perhaps each week or each month. In its simplest form, the amount ordered is the quantity needed to replenish inventory to I_{max}, a preset level. This reorder quantity Q is variable, and is equal to the amount used during the current period plus the expected usage during the supply lead time, as shown in Figure 14. The service level policies and resulting buffer stock requirements are based on the same general concepts and methods as before.

The managerial control variable is now the length of the reorder cycle,

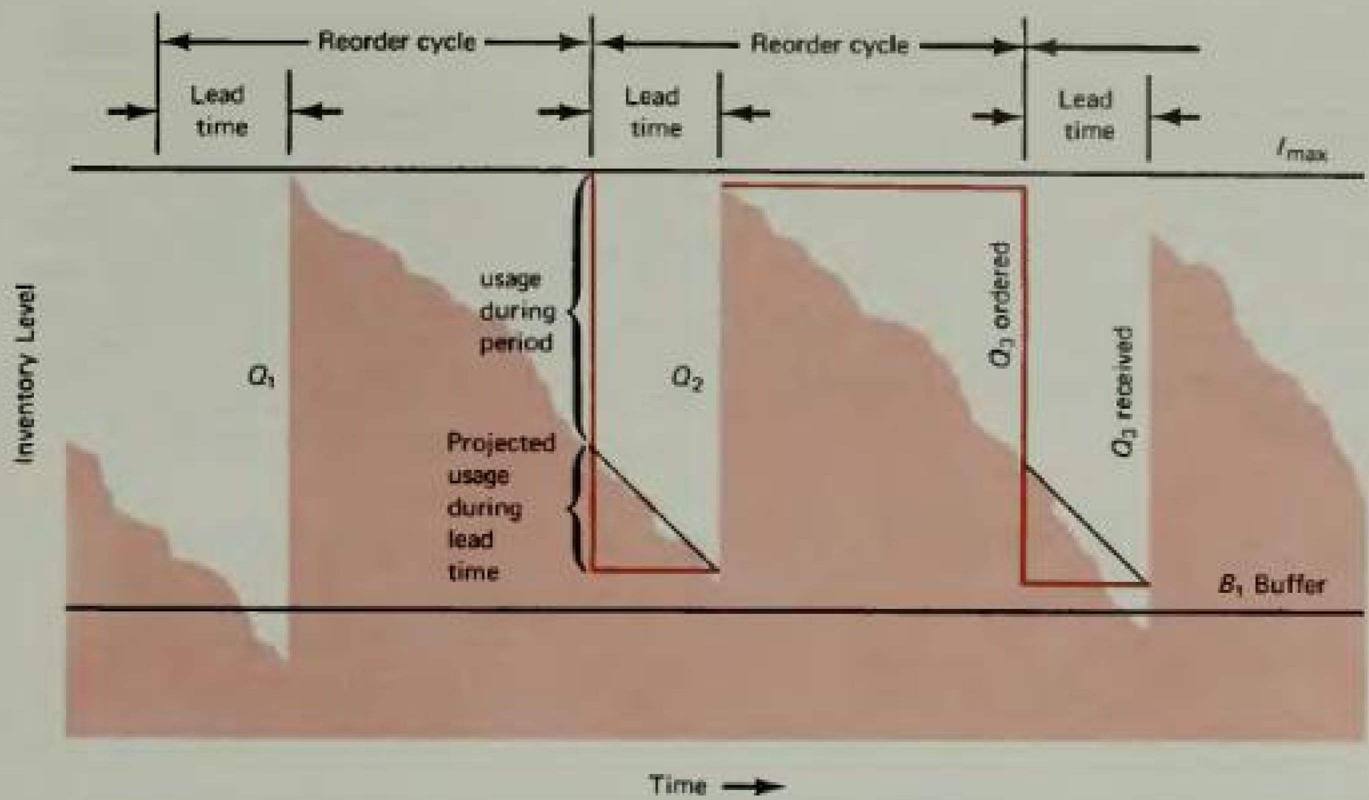


Figure 14. Fixed reorder cycle system of control. An order is placed at regular intervals which replenishes stock by an order of variable size, Q, which is the sum of usage during the immediate past period plus projected usage during the supply lead time

commonly called the *review period*, since inventory status is reviewed for replenishment orders during the period. By means of this system, optimal review periods can be derived that are equivalent to the optimum order quantity Q_0 in the fixed reorder quantity system. The optimal review period can be approximated by Q_0/R, where Q_0 is computed from Equation 4. As we shall note, however, the optimal review period is seldom used because other criteria tend to dominate in reorder cycle selection.

Fixed reorder cycle systems are prominent with higher valued items, and in situations where a large number of items are ordered regularly from the same vendor. Thus, one of the significant advantages of the fixed cycle system is that freight cost advantages often can be gained by grouping these orders together for shipment. Also, the regular review of all items for replenishment ordering on the same basic cycle has procedural advantages, and provides close surveillance over inventory levels for all items. Thus, while optimal review periods that depend on the cost parameters of the model can be determined for each item, review periods actually are set on the basis of all the considerations that we have discussed, including individual optimal review periods.

Combination systems exist in which stock levels are reviewed on a periodic basis but orders are placed only when inventories have fallen to a predetermined reorder point. When an order is placed, its size is sufficient to replenish inventories, as in the fixed cycle system. Periodic reviews of current usage rates may result in changes in I_{max}, P, and B. Combination systems have the advantage of the close control that is associated with the fixed cycle system; but since replenishment orders are placed only when the reorder point has been reached, fewer orders are placed on the average. Therefore, annual ordering costs are comparable to those that are associated with the fixed quantity systems.

Summary

Inventories serve a crucial function in productive systems and, while they may be costly, it usually is more costly to try to minimize them. There are several kinds of inventories, and they are classified in terms of the functions they serve: pipeline, cycle, buffer, and seasonal. In addition, inventories make it possible to decouple successive stages or operations in the productive process, so that the effects of disruptions are minimized.

The concept of an economical order quantity (EOQ) is based on determining an order size that is a compromise between ordering and inventory holding costs. By balancing these two kinds of costs, we can derive the mathematically minimum cost quantities that we should order at one time. When price-volume discounts enter the picture, additional incremental savings may result.

The economical size of production runs can be formulated in a way that is parallel to minimum cost purchase quantity concepts; however, the assumption of constant demand often is not a good one.

Because there usually are random variations in demand, and because shortages involve (often very large) costs, it is economical to carry extra inventories as a buffer. The function of the buffer stock is to absorb demand variation, and its rational determination depends on a knowledge of the distribution of demand.

Based on the general concepts of inventory management, three inventory control systems have gained prominence: the fixed reorder quantity system; the fixed reorder cycle system; and systems that combine the reordering policies of the first two.

Review Questions

1. Define each of the following terms:
   a. Decoupling function of inventories.
   b. Pipeline inventory.
   c. Cycle inventory.
   d. Transit inventory.
   e. Buffer inventory.

2. If system volume increases by a factor of two, what would you expect to happen to:
   a. Pipeline inventories.
   b. Cycle inventories.
   c. Buffer inventories.

3. What costs are relevant to the managerial decision regarding the size of purchase orders? the size of production runs?

4. Write the total incremental cost equation appropriate for:
   a. A simple purchase reordering decision.
   b. A reordering decision in which price discounts are to be considered.
   c. A reordering decision in which shortage costs are known.

5. Explain the rationale for the derivation of the simple EOQ formula (Equation 4).

6. In the quantity discount example used in the text, the value of c_H for a price of \$1.86 per unit would be $c_H = 0.16 \times 1.86 = \$0.30$ per unit per year, given $F_H = 0.16$. Since $R = 2{,}000$ units per year, and $c_P = \$20$ per order, Equation 4 yields EOQ = 516 units. Why is $Q_0 = 516$ not a valid answer?

7. Explain the assumptions made in the derivation of Equation 4.

8. Explain the reasons for the difference between Equations 4 and 6. Would Equation 4 be appropriate if the entire lot were produced before it was placed in inventory?

9. If several parts or products share the use of the same facilities, and we

wish to determine the size of production runs for each, how can we avoid conflicts in equipment usage when the operating level is below capacity? above capacity?

10. What risks result from the fact that demand or usage rates vary, instead of being constant (as assumed in Equation 4)?

11. Define the following terms:
    a. Order point.
    b. Lead time.
    c. Reorder level.
    d. Maximum demand.
    e. Safety factor.
    f. Service level.

12. If the average demand during the supply lead time is 100 units, and the buffer stock is 75 units, what is the implied maximum demand?

13. Explain the concept of service level in terms of maximum demand.

14. State the equation for the buffer stock B in terms of the safety factor and the standard deviation of demand for the normal distribution.

15. Explain the general rationale for determining service levels when shortage costs are known, and when they are not known.

16. Why are large order sizes desirable, from the point of view of minimizing the risk of stock out?

17. Explain the concept of *expected quantity short*; of *annual expected quantity short*.

18. What happens to the annual expected quantity short if:
    a. The buffer stock is increased.
    b. The order size is increased.
    c. Annual requirements double, same order size.
    d. Through a policy change, service level is increased.
    e. Annual requirements double, EOQ policy.
    f. Variability of demand increases.

19. What "triggers" a replenishment order in each of the following managerial control systems?
    a. Fixed reorder quantity system.
    b. Fixed reorder cycle system.

20. Explain the concept of the *two-bin* system.

21. Under what conditions would we use the fixed reorder quantity system in preference to the fixed reorder cycle system, and vice versa?

22. As time passes, any inventory control system may become dated as demand, costs, and competitive pressures change. Thus, periodic review of parameters is important. What are the parameters that should be reviewed for the fixed reorder quantity, and fixed reorder cycle systems?

aging process
↓
in-process inv

Problems

1. The manager of a beer company is always astounded by the size of aggregate finished goods inventories in his system. He is perfectly aware of the in-process inventory requirements that are determined largely by the aging process, but once the beer is bottled or canned, he feels that it should be moved out, unless he has gauged the market incorrectly. He decides to compute the minimum finished goods inventory for an average system-volume of 7,000 cases per day (seven-day operation), using established policies for service levels, and inventory replenishment at the plant distributor and retail stock points. The company owned distributors have the responsibility for controlling retail inventory to ensure that supplys do not become outdated. (Contrary to widespread opinion, beer degenerates with time, once bottled.)

   volume out
   once
   S+H or canned

   The manager decides that three components of inventory must be considered: pipeline, cycle stock, and buffer stock. He feels that if he computes inventory requirements for these three components under present conditions and policies, he should have a good idea of the minimum system inventory requirement. He then would regard inventories beyond this minimum figure as excessive.

   Looking at the system from the time that the beer comes off the bottling-and-packaging line, the manager follows the process to see what happens and how long it takes. The average delays and transit times that result are as follows:

Plant to plant warehouse	1 day
Delay at plant warehouse	4 days
Warehouse to distributors	2 days
Delays at distributor warehouse	2 days
Distributors to retailers	1 days
Delays at retailers before sale	7 days

Next, the manager examines the inventory replenishment policies at the plant warehouse, distributors, and retailers. The plant warehouse has no independent policy for replenishment, since it merely receives the entire daily output of the plant. The distributors, however, receive shipments every 2 weeks based on replenishment orders that they place once per 2 weeks. There are 14 distributors, and they are on regular shipment cycles; thus, the plant–warehouse load is nicely balanced. On the average, each distributor places an order for 7,000 cases every 2 weeks. Each distributor services an average of 700 retailers who are on a weekly ordering cycle. On the average, each retailer places an order for 5 cases per week, his/her average usage.

Finally, the manager looks at the buffer stock policies. After examining the demand distributions and the basic company policy for off-the-shelf service, he establishes that a reasonable maximum demand during the distributors' 2 week replenishment cycle is 8,500 cases per 2 weeks. Similarly, looking at retailers' requirements, he establishes that a reasonable maximum demand is 10 cases per week. He wants to provide service at both the distributor and retail levels that takes account of these maximum demand levels.

a. Compute the pipeline inventory requirements.
b. Compute the cycle inventory requirements.
c. Compute the buffer inventory requirements.
d. What are the system finished goods inventory requirements?

2. A hospital maintains inventories of a large number of items in order to provide medical service to patients. Some of the items are of relatively low value and are not considered critical, such as soap, floor wax, cleaning supplies, etc. These items are being considered for routine inventory control, using an EOQ policy for inventory replenishment. Studies indicate that the cost of placing orders is $10, and that the average inventory cost,

in terms of the percentage of inventory value, is 25 percent. A typical item costs $20 per case, and the hospital uses 10 cases per month.

a. Compute the inventory holding cost, c_H.
b. Compute EOQ.
c. How many orders are required per year?

3. What change in EOQ results if each of the following variables change by 25 percent?
   a. c_P
   b. R
   c. c_H

4. A manufacturer fabricates a part in lots, and stores the parts as a manufactured item for use in a continuous assembly process. Since the assembly process is continuous, only the excess is stored temporarily.

   The cost of writing production orders and setting up machines to run the order is only $10. The annual requirements for the part are 5,000, and the inventory holding cost is 5 cents per unit per year. When the fabrication process is operating, parts are produced at the average rate of 141.4 parts per week.
   a. Compute the manufacturing lot size that would minimize incremental costs.
   b. How many production orders are required per year?

5. The hospital discussed in problem 2 is offered a price discount for the item, as follows:

   $20 per case in quantities of less than 50 cases.
   $18 per case in quantities from 50 to 99 cases.
   $17 per case in quantities of 100 cases or more.

   a. Are the price discounts attractive?
   b. Why not simply buy annual requirements, once per year?

6. Weekly demand for a product is represented by the empirical distribution shown in Figure 15, exclusive of trend and seasonal variations. What buffer stock is required to ensure that we would not run out of stock more than 10 percent of the time, 5 percent of the time, and 2 percent of the time? (Normal lead time is one week.)

7. If the product in Problem 6 has a unit value of $50, and inventory holding

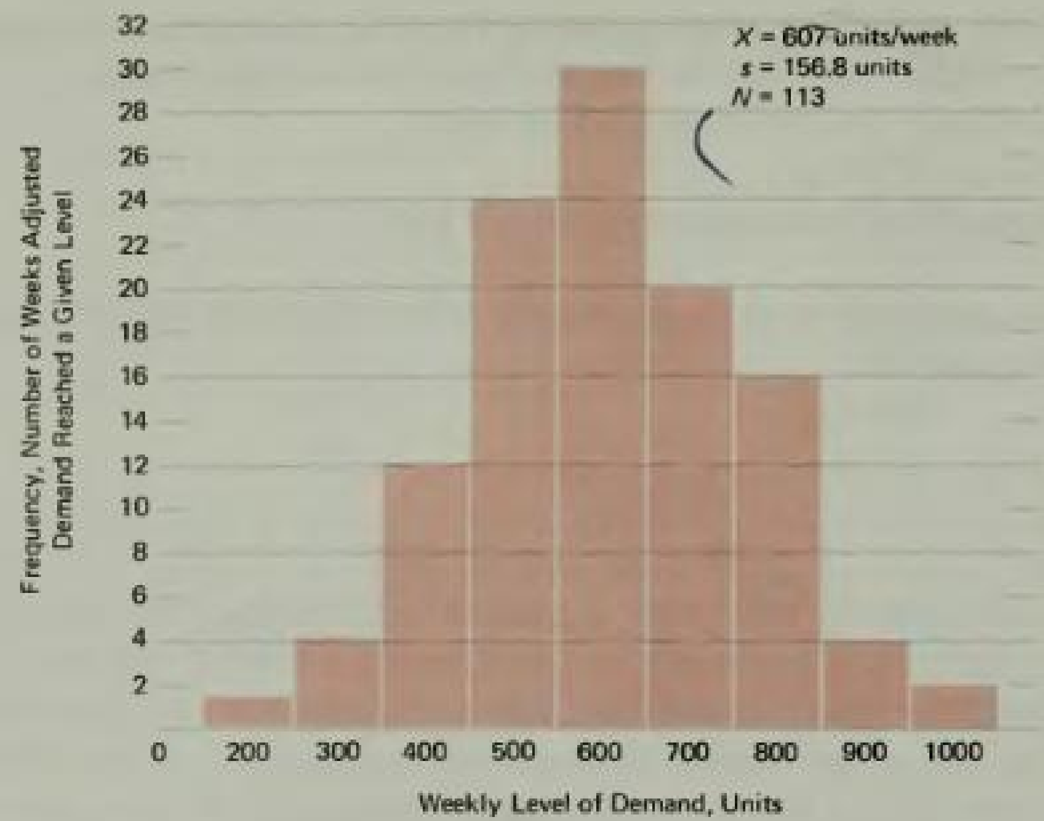


Figure 15. Distribution representing expected random variation in weekly sales, exclusive of seasonal and trend variations

costs are 25 percent of inventory value, what is the cost of providing service at the 90, 95, and 98 percent levels?

8. If we can assume that Figure 15 is reasonably represented by a normal distribution, what buffer stocks would be required for the 90, 95, and 98 percent service levels? How do these values compare with those obtained directly from the empirical distribution?

9. Suppose that a distributor for home appliances experiences demand of $\bar{D}$ = 18 units per week, and that the distribution of demand is found to be approximated by the *normal* distribution. What buffer stock should be maintained in order to offer a 90 percent service level? What is the implied safety factor?

10. If the demand distribution is normal with $\bar{D} = 1{,}000$ units per week and $s = 200$ units, what is the:
    a. Buffer stock required for the 90 , 95 , and 99 percent service levels?
    b. What is the expected quantity short per order for each service level?
    c. If inventory holding costs are $c_H = \$10$ per unit per year, $c_P = \$25$ per order, and $c_S = \$15$ per unit short, what is the most economical service

(4) p.387 (4) 390 (C) p.342

level if 32 orders per year are placed? (EOQ = 1,612, or approximately 32 orders per year to meet requirements.)

d. If order size is reduced to Q = 1,000 units, how is service policy affected?

e. If order size is increased to Q = 2,600 units, how is service policy affected?

f. What appears to be the best combination of order size and service policy?

References

Brown, R. G. *Decision Rules for Inventory Management*. New York: Holt, Rinehart, & Winston, 1967.

———. *Smoothing, Forecasting and Prediction*. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

———. *Statistical Forecasting for Inventory Control*. New York: McGraw-Hill, 1959.

Buchan, J., and E. Koenigsberg. *Scientific Inventory Control*. Englewood Cliffs, N. J.: Prentice-Hall, 1963.

Buffa, E. S. *Operations Management: Problems and Models*. 3d ed. New York: John Wiley & Sons, 1972. Chapters 7, 8, and 9.

———, and W. H. Taubert. *Production-Inventory Systems: Planning and Control*. Rev. ed. Homewood, Ill.: Richard D. Irwin, 1972. Chapters 3 and 4.

Fetter, R. B., and W. C. Dalleck. *Decision Models for Inventory Management*. Homewood, Ill.: Richard D. Irwin, 1961.

Hadley, G., and T. M. Whitin. *Analysis of Inventory Systems*. Englewood Cliffs, N. J.: Prentice-Hall, 1963.

Lewis, C. D. *Demand Analysis and Inventory Control*. London: Saxon House, Lexington Books, 1975.

Magee, J. F., and D. M. Boodman. *Production Planning and Inventory Control.* 2d ed. New York: McGraw-Hill, 1967.

McMillan, C., and R. F. Gonzalez. *Systems Analysis: A Computer Approach to Decision Models.* 3d ed. Homewood, Ill.: Richard D. Irwin, 1973.

Starr, M. K., and D. W. Miller. *Inventory Control: Theory and Practice.* Englewood Cliffs, N. J.: Prentice-Hall, 1962.

Wagner, H. M. *Statistical Management of Inventory Systems.* New York: John Wiley & Sons, 1962.

Chapter 12

Industrial Planning, Scheduling, and Control

While the particular methods of aggregate planning that are used for industrial situations depend on whether or not the product is inventoriable, the detailed planning methods depend more on the flow characteristics within the productive system. When we deal with high-volume standardized products, the likelihood is high that flow is continuous and direct, and that operation sequences are designed into the system. Under these circumstances, scheduling the line as a whole schedules each operation in the sequence, since the line is, in fact, an integrated machine. Although detailed scheduling is not necessarily simple (as we shall see), we do not need to schedule and control each operation individually.

On the other hand, when we are dealing with systems in which equipment is time-shared among many different parts and products so that we can obtain reasonable equipment utilization, flow is likely to be intermittent. Flow paths vary from product to product and from time to time. Parts and products are processed on a cyclical basis, and they flow through the plant in lots or production orders, either because they are custom designed or because the annual volume does not justify continuous output. Planning, scheduling and control are focused on the individual lot or order.

We will discuss these two quite different approaches to industrial planning, scheduling, and control.

HIGH-VOLUME CONTINUOUS SYSTEMS

As we noted, high-volume continuous systems operate as a giant machine. If the production rate is set, then the dovetailing and sequencing of operations already is taken care of by the system design. Isn't the detailed scheduling problem already accomplished by the system design? In a sense, yes. However, given aggregate planning decisions that set employment levels and production rates, the remaining problems still are significant.

First, how can we determine the amounts of each product to be produced? The aggregate plan specifies an overall production rate, but this overall total must be allocated to the various sizes and types of items produced. Second, the aggregate plan may call for a change in the size of the work force. A change in manpower level must be translated into a rebalancing of facilities, along with modified work assignments for the crew. Therefore, while the aggregate plan specifies the quantitative change in work force size, there still remains the matter of how to implement this change in a way that makes the most sense. In other words, which workers and what kinds of skills are involved? If the aggregate plan calls for a decrease in employment, union aggrements may specify who will be laid off. Third, the combined work force size and production rate decisions determine the aggregate amount of overtime or undertime to be worked. How will the overtime or undertime be allocated to products and manpower? And fourth, do the projected inventory levels that result from a proposed detailed schedule of items agree with the aggregate planning level, and with the policies and procedures developed for individual items?

NATURE OF THE PRODUCTION-DISTRIBUTION SYSTEM

To understand the nature of the scheduling and control problems for high-volume continuous systems, we need to understand the overall flow, the importance of system inventories, and the system dynamics. Two crucially important factors to keep in mind are that the scheduling of the factory output is very dependent on the behavior of distributors and retailers downstream in the flow system, and that these elements of the system normally lie outside of management's control.

Let us begin by examining the production-distribution system diagrammed in Figure 1. Assume that the system represents the manufacture and distribution of a small appliance. Figure 1 shows the major functions performed in the

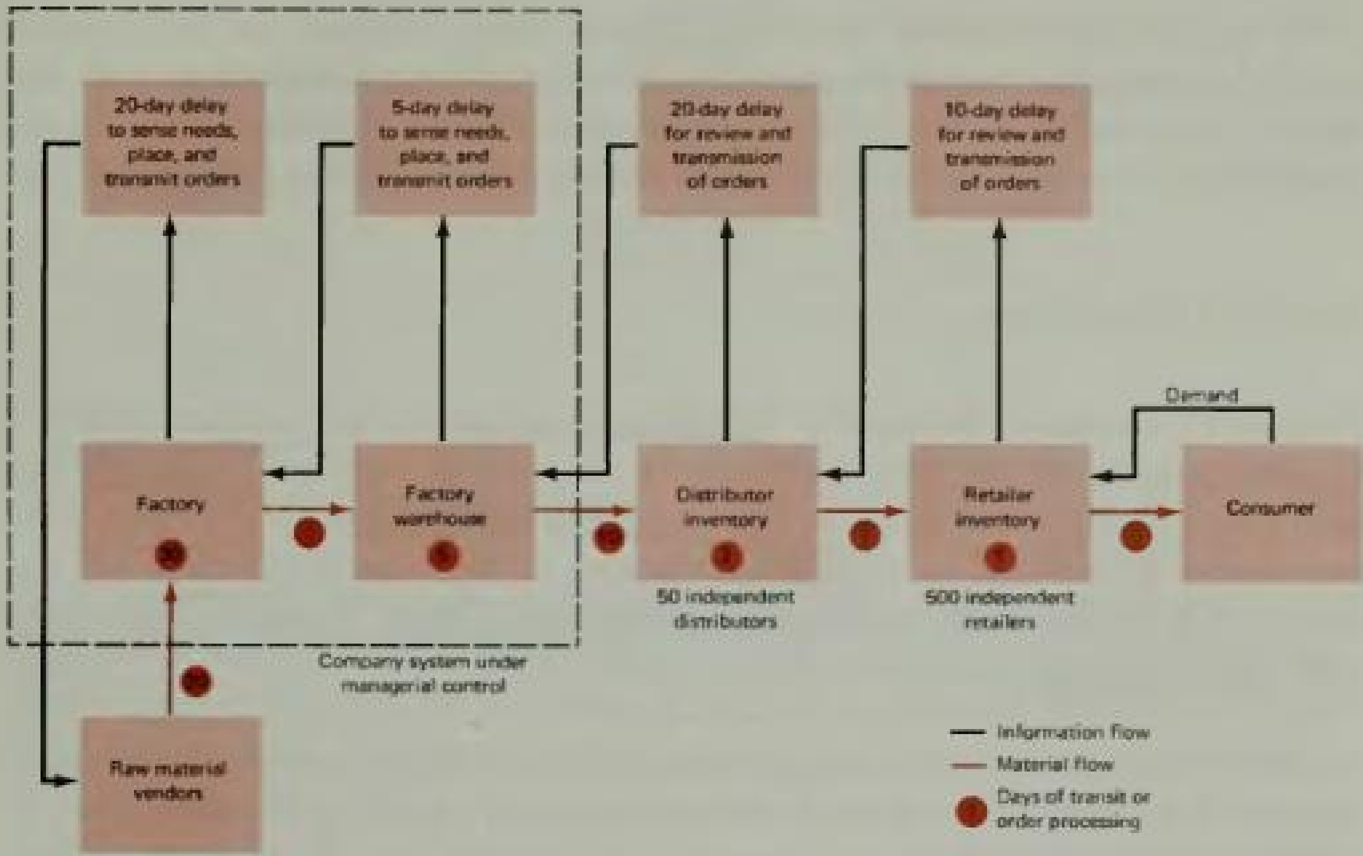


Figure 1. Production–distribution system for high-volume continuous system showing broad flow of materials and information, transit times, and order processing delays (System volume averages 2000 units per week)

production and distribution of the appliance, beginning with raw material procurement, extending through manufacture, and ultimately—through the sequence of distribution steps—to the consumer. There are 500 independent retailers, 50 independent distributors, and a single factory in the system. Each of the distribution steps involves a stock point for the finished product. Therefore, one way of looking at the system downstream from the factory is to envision it as a *multistage inventory* system.

Note also that Figure 1 indicates the basic elements of the ordering procedure for replenishing the finished goods inventory at each stage. Thus, each retailer has a replenishment cycle, which involves the review of retailing demand and inventory status, the transmission of orders to the supplying distributor, and the filling and shipping of the order by the distributor. Similarly, each distributor has an equivalent replenishment cycle, based on the assessment of demand from retailers, the transmission of orders to the factory warehouse, and the filling and shipping of the orders. Similar cycles are required for the factory warehouse in ordering from the factory, and for the factory itself in ordering raw materials from vendors.

We are interested in some aspects of all these steps, although our key interest is in scheduling the factory to produce. Before concerning ourselves with the factory's scheduling, however, we want to make some rough calculations to get some idea of how much inventory will be required to make the system function.

System Inventories

Chapter 11 discussed the several functions of inventories. Some inventory is required just to fill the flow pipelines of the system; another component is required to take account of the periodic nature of the ordering cycles; and a third component is required to absorb fluctuations in demand. Depending on the nature of the policies for factory scheduling, additional seasonal inventories also might be required.

Pipeline Inventories. If the average system volume is 2,000 units per week and it takes 1 day to transport from the factory to the factory warehouse, then there are 2,000 (1/7) units in motion at all times. If the order processing delay at the factory warehouse is 5 days, then there are 2000 (5/7) units tied up at all times because of the delay. Table 1 summarizes the pipeline inventory requirements for finished goods for the entire system, indicating a minimum of 7,144 units required just to fill the pipelines. These inventories cannot be reduced unless transit times, delays, and handling times can be reduced. The inventories are proportional to the system volume and the physical flow times. If system volume increases, then this component of inventories must increase to keep the pipelines full and the system functioning.

TABLE 1

SUMMARY OF PIPELINE INVENTORY REQUIREMENTS FOR FINISHED GOODS (AVERAGE SYSTEM VOLUME IS 2,000 UNITS PER WEEK)

	Average Delay Time, Days	*Average Pipeline Inventory, Units (Days/7) × 2000*
Factory to factory warehouse	1	286
Delays at factory warehouse	5	1,429
Warehouse to distributors	10	2,857
Delays at distributors	2	571
Distributors to retailers	5	1,429
Delays at retailers	1	286
Retailers to customers	1	286
Totals	25	7,144

TABLE 2
SUMMARY OF CYCLE INVENTORY REQUIREMENTS FOR FINISHED GOODS
(AVERAGE SYSTEM VOLUME IS 2.000 UNITS PER WEEK)

	Reorder Cycle Time, Weeks	Average Cycle Inventory Units
500 Retailers	2	2.000
50 Distributors	4	4.000
Factory warehouse	6	6.000
Total		12.000

Cycle Inventories. We will assume that the fixed reorder cycle system of inventory replenishment is being used at the factory warehouse, distributor, and retailer levels. The average order size is set by the ordering frequency at each stage. For example, the average retailer orders once each 2 weeks, following a review of sales (the information and physical flow cycle is shown in Figure 1). Therefore, a retailer's order must be for a 2 week supply just to meet average demand. The average retailer sells 2.000/500 = 4 units per week, or 8 units during the 2 week ordering cycle. Therefore the retailer must have no less than 8 units on hand to service sales during the replenishment period; and the average inventory for this purpose is half this amount, or 4 units. The cycle inventory for the entire system of 500 retailers is 4 × 500 = 2000 units. Table 2 summarizes the cycle stock requirements of the system, and shows that 12.000 units are required in average inventory because of the periodic nature of ordering.

Buffer Inventories. In Chapter 11, we stated that buffer inventories are designed to absorb random variation in demand, and that the size of buffer inventories depends on the nature of the distribution of demand and the service levels that we wish to maintain. Table 3 shows the computation of buffer stocks for the system. For example, the buffer stock required for each retailer is the difference between the estimated maximum demand over the 17 day review plus supply lead time, and the average demand, or 18 − 9.7 = 8.3 units. The average system buffer stock for 500 retailers, then, is 8.3 × 500 = 4.150 units. The buffer stock requirements for the entire system are 13.864 units.

In summary, the inventories in the system that are needed to accommodate all three functions are as follows:

Pipeline inventories	7.144
Cycle inventories	12.000
Buffer inventories	13.864
Total	33.008

TABLE 3
SUMMARY OF BUFFER INVENTORY REQUIREMENTS FOR FINISHED GOODS
(AVERAGE SYSTEM VOLUME IS 2,000 UNITS PER WEEK)

	Average Demand per Week	Lead Time, Days	Average Demand over Lead Time	Maximum Demand over Lead Time	Average System Buffer Stock
500 Retailers	4	17	9.7	18	4,150
50 Distributors	40	35	200.0	300	5,000
Factory Warehouse	2,000	36	10,285.7	15,000	4,714
Total					13,864

These are inventories required by the structure of the system and the ordering rules and service levels used. They represent the minimum possible amounts necessary to operate the system. Overall inventories might be larger than this minimum if controls were not effective, or if seasonal inventories also were accumulated in the system.

The impact of inventories on the problem of factory scheduling is underlined by the fact that management has control over only approximately one-third of these inventories; the balance is held by distributors and retailers. Also, management does not have control over the ordering policies and procedures of the retailers and distributors.

System Dynamics

Now let us attempt to understand the impact of activities downstream in the production-distribution system on the scheduling of factory operations. Let us look again at the multistage system diagrammed in Figure 1. Though highly idealized and simplified in many respects, this model is sufficient to show us some of the system's important material and information flow.

We already are familiar with the significance of the time lags in relation to inventories needed for replenishment and for filling the distribution pipelines. Now, however, we will find it valuable to look more closely at the behavior of the system and to examine its significance for factory scheduling. The general behavior of the system depends on the periodic review of inventory needs, and on the preparation and transmission of orders for replenishment to the next stage upstream.

Suppose that consumer demand falls by 10 percent from its previous rate.

During the next review of inventory needs, the retailer reflects this decrease in orders for replenishment that are sent to the distributor; however, ten days have elapsed. Similarly, the distributor reflects the decrease in the next orders for replenishment to the factory warehouse; however, an additional twenty days will have elapsed before the factory warehouse is aware of the fall in sales. Thus, is we add up all the time delays in the information system, we find that the factory will not learn of the 10 percent fall in demand until thirty-five days have passed. Meanwhile, the factory has been producing 1.00/0.90, or 111 percent, of the new consumer requirement. An excess of 11 percent will have accumulated each day in the inventory at the various stock points. This system inventory will have increased to $11 \times 35 = 385$ percent of the usual normal day's supply.

To react to the change, retailers, distributors, and the factory warehouse decrease the quantities ordered. To take account of the excess inventory, the factory will have to cut back by substantially more than the 10 percent. We now can see that the time lags in the system amplify the original 10 percent change at the consumer level to a much greater change in production levels than would have seemed justified by the simple 10 percent decrease in consumer demand. We also see that inventories have increased instead of decreased. Obviously, a more direct communication of changes in demand can reduce the magnitude of this amplification.

Suppose we insert in Figure 1 a more direct information feedback loop in the form of a system for assessing actual demand and forecasting demand in the immediate period ahead. Assume a ten day delay in assembling the actual demand and forecasting information. This reduces the total delay by twenty-five days. Under this system, a 10 percent decrease in sales would mean that an excess inventory of only $11 \times 10 = 110$ percent of the normal levels would accumulate before the factory was aware of the change. Obviously, the forecast, combined with the aggregate scheduling system shown in Figure 2, would stabilize the effects even more.

SCHEDULING AND AGGREGATE PLANNING

The broad outlines of the scheduling process for the multistage production-distribution system in relation to the aggregate planning process is shown in Figure 2. Within the "Company System under Managerial Control" and above the dashed line in Figure 2 we see the activities that are related directly to scheduling.

Forecasts of demand that are based on information concerning the progress of

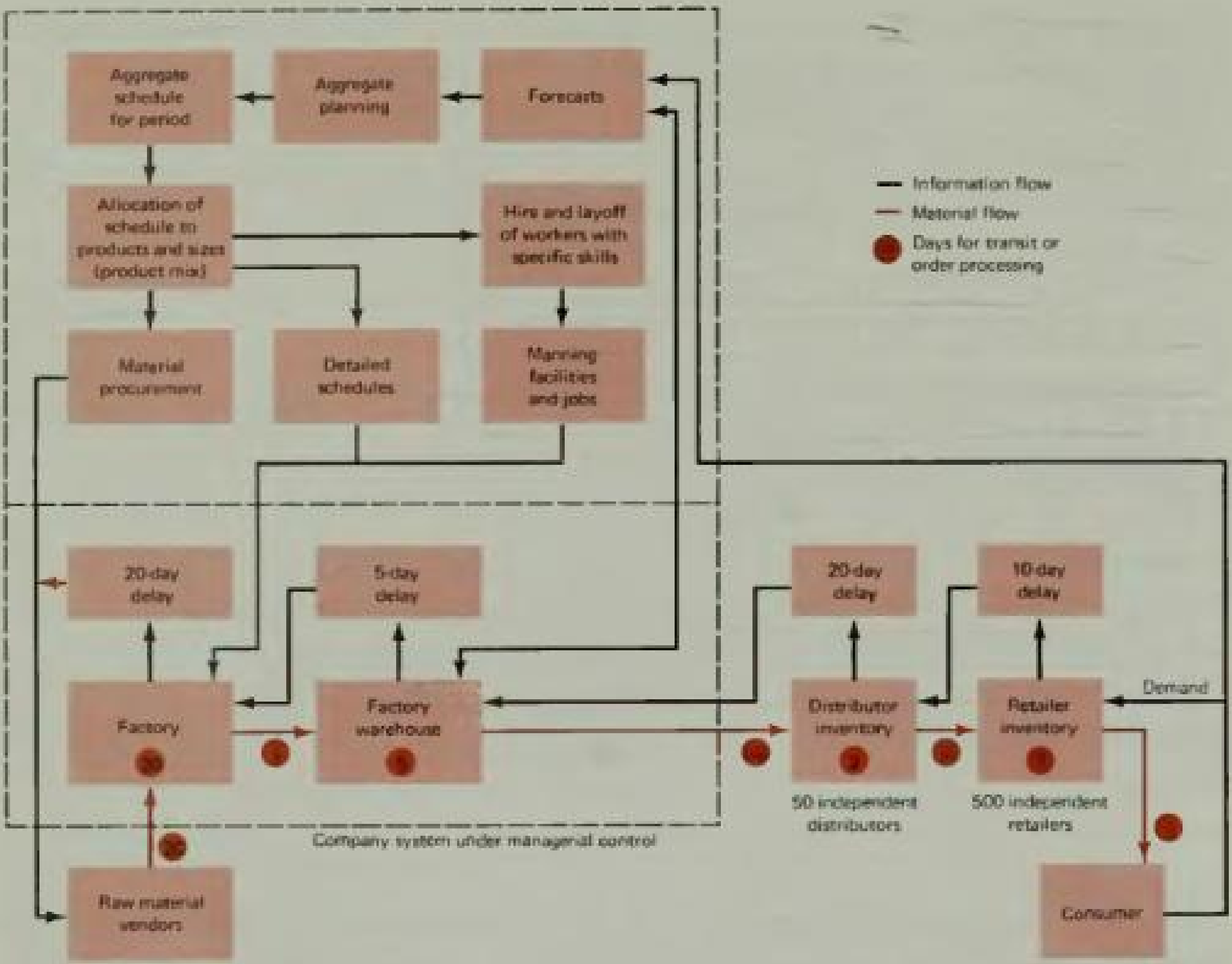


Figure 2. Relationship of forecasts, aggregate scheduling, and detailed scheduling to production–distribution system

sales at the retail–consumer level, as well as on information concerning order rates at the factory warehouse, are fed into the forecasting model. The forecasting model produces period forecasts over the planning horizon for the aggregate planning process, which in turn produces basic decisions on production rate and work force levels for the upcoming period. From the basic decisions produced by the aggregate planning process, we either can compute the overtime and subcontracting required, as well as the projected end-of-period inventory, or produce these data as part of the output of the aggregate planning procedure, depending on the aggregate planning model used. The next task is translating these aggregate plans into detailed working schedules.

The Product-Mix Problem

The effect of an aggregate planning decision on employment levels and production rates is the adjustment of the system's effective capacity for the upcoming period. In other words, setting capacity limits creates capacity constraints that must be observed in the detailed schedule that is generated. Thus, the allocation of the limited capacity to product types and sizes becomes an important economic problem.

The product-mix problem may be most complex in chemical industries such as oil refining, in which interdependencies exist in the quantites of different products that can be produced. If more of one product, (e.g., aviation gasoline) is to be produced, then less of some other product must be produced by the very nature of the refining process. The profitability of various products may be different, and the markets for each product are limited. The result is a complex programming problem whose objective is to determine the best product-mix. Interdependencies always exist in oil refining, for example, because the basic raw material, crude oil, can be processed into many different products. An increase or decrease in one petroleum product always means a change in the quantities of some other products.

In the mechanical industries, the interdependencies among products are more likely to stem only from the capacity limits imposed by (1) the aggregate planning decisions, and, in some instances, (2) the capacity of time-shared facilities that are used to produce a variety of types and sizes. In all instances, however, we face an allocation problem of how best to use available limited capacity. These problems of product-mix have been approached through linear programming. In present day practice, it is common to ignore the allocation problem, and to determine the product-mix by applying percentage sales figures, along with modifications based on personal judgment.

The allocation to types and sizes gives basic information for detailed scheduling, detailed hiring and layoff instructions, and material procurement schedules, as shown in Figure 2. Capacity limitations either for labor or facilities may create conflicts in detailed schedules at this point, since detailed scheduling involves setting individual type and size production rates and manpower schedules.

In setting the individual type and size production rates and manpower schedules, the scheduler may be faced either with great rigidity or reasonable flexibility, depending on the nature of the processes and the design of the production system. Production lines are quite rigid in their nature: once designed and set up, the line produces assembled units at a fairly fixed hourly rate, since all operations have been balanced to coordinate with the preset hourly rate.

What flexibility does the scheduler have to obtain a certain target weekly or monthly rate of output? The scheduler has basically two alternatives: scheduling the work force on the line to work shorter or longer hours (including overtime); or rebalancing the entire line to achieve a somewhat higher or lower hourly rate or output. Obviously, the latter alternative would be used to achieve more drastic changes in output rate, since it involves hiring or laying off workers. On the other hand, simply changing hours worked can be done cheaply, unless overtime must be scheduled. In any case, the scheduler will be following the basic instructions given by the aggregate plan, which, presumably, has considered the relative costliness of changing production rates by means of changing hours, using overtime, and rebalancing (hiring and laying off).

The aggregate plan establishes the constraints under which the scheduler must perform. Thus, assembly line balance should concern us not only in terms of original design of productive systems, but also in terms of continued operation, since it is through rebalancing that the scheduler can change the basic hourly production rate of the system. For example, rebalancing commonly is used in automotive assembly lines to achieve different hourly rates of output.

If the line is completely rigid in design (which easily can happen, if it is paced mechanically), then the scheduler can only change total output for the period either by changing the number of hours per period that the man-machine system is operated or by changing the line speed. The rigid system, of course, is used often.

The Detailed Scheduling Process

Since the broadly based optimization already has taken place in the development of the aggregate plan, the main problems of detailed scheduling involve devising ways of following out the aggregate plan as far as possible. After aggregate production has been allocated to product types and sizes, an iterative process begins, which develops a tentative plan and checks back to see whether the details of the plan fit the stated constraints of the aggregate plan and other possible company policies.

The first question raised in the schematic diagram of Figure 3 involves the possible changes in employment levels that are called for by the aggregate plan. These changes will require a rebalancing of manpower to facilities. In the case of complex assembly lines, the methods and procedures of Chapter 8 may be used. In many instances, alternate manning plans that are based on previous careful studies probably exist for various output rates of facilities. In any case, rebalancing will result in hiring or laying off personnel in specific skill categories, checking to see whether the result is within the aggregate plan. If the aggregate planning model has been constructed carefully to reflect the relationship between productivity and manpower, it should be possible to adjust manning assignments within the constraints of the plan. However, our study of

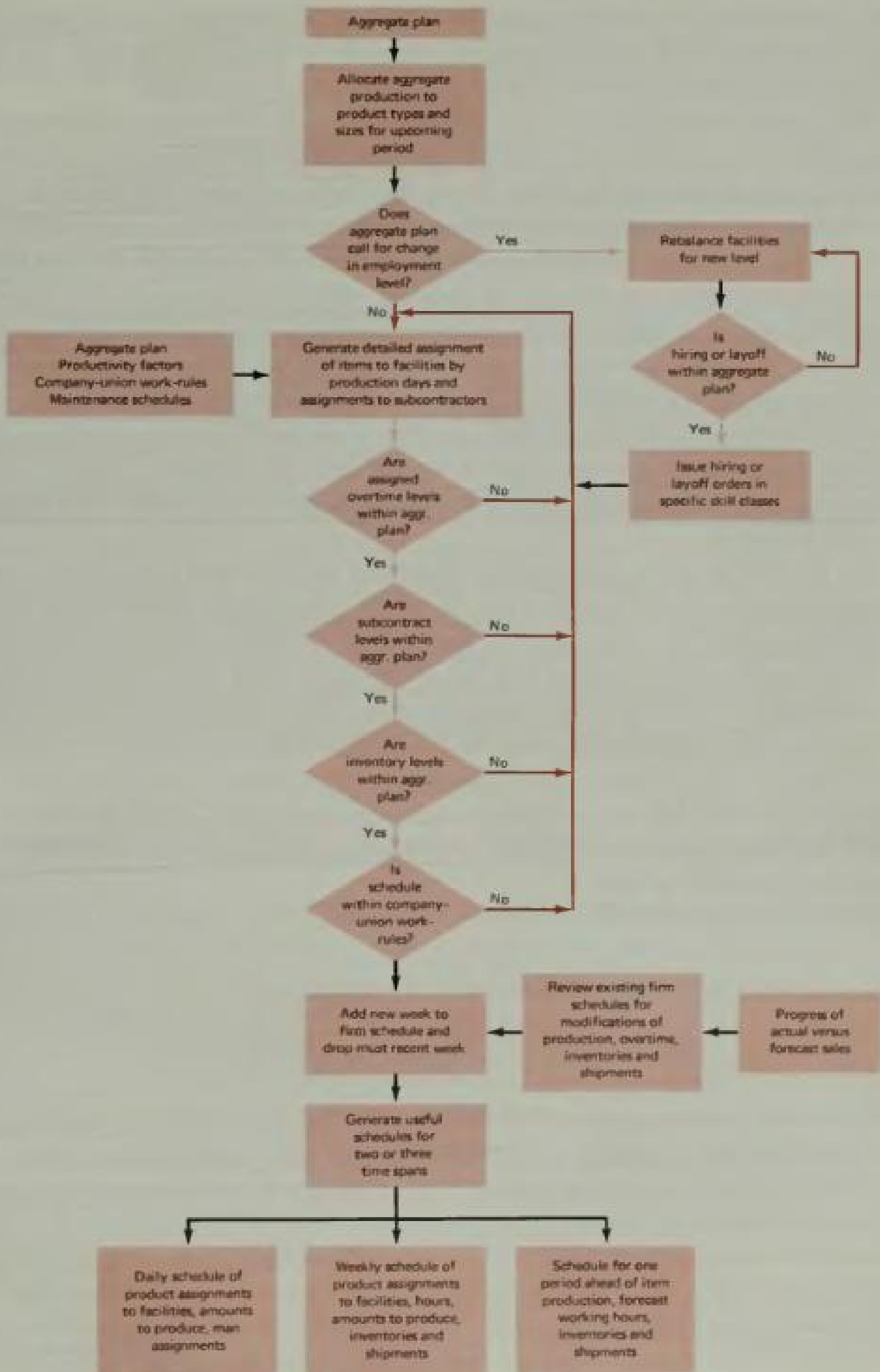


Figure 3. Flow diagram of operations scheduling for continuous systems

line balancing showed that balance solutions must deal with the assignment of whole manned-units so that cost and capacity increase or decrease in step-fashion, rather than in a continuous relationship to manpower.

When the hiring-layoff question has been resolved by means of rebalancing, this information becomes an input to the initial attempt to generate a detailed schedule of products to facilities. Using the aggregate plan and the new basic production rate—based on rebalancing, productivity factors, company-union work rules, and a knowledge of existing maintenance schedules—the scheduler generates a detailed assignment of product items to facilities by production days, as well as assignments to subcontractors (if that is appropriate). The scheduler then checks the result in an iterative fashion to see whether assigned overtime levels, subcontracting levels, and inventories meet the requirements of the aggregate plan as well as other constraints (such as maintenance schedules and agreed company-union work rules). The result is a feasible daily schedule that will add an additional period, perhaps a week, to the firm's schedule (see Figure 3).

In a parallel way, the existing firm schedules are reviewed for possible modification based on new information about the progress of actual sales versus forecasted sales. The schedules of production commitments are updated by adding the newest information and dropping off the most recent week or period.

Based on the resultant information, the scheduler may develop full schedules for two or three time-spans, such as those shown at the bottom of Figure 3. For some specific company situations, most of the process could be computerized. We assume that, at some point, our technical capability will improve so much that aggregate and detailed scheduling largely can be combined into one "optimum-seeking" model.

FORMAL METHODS FOR COUPLING AGGREGATE AND DETAILED SCHEDULING PLANS

The preceding material describes the complex problems involved in coupling and coordinating aggregate and detailed plans and schedules. A number of formal methods for achieving optimal coupling have been proposed, most of which involve some degree of disaggregation; that is, developing plans and schedules by product line or by major unit or department, instead of considering all products, sizes, and types in the aggregate. This work has pointed out the inadequacy of a totally aggregate plan for organizations that produce a variety of products. The decisions must be more specific concerning the products for which W should be increased or decreased, etc. Though the decision process may call for a smaller total work force in the upcoming period, if we have n products, W_{1t} may have to be increased even though W_{2t} and W_{3t} should be decreased.

> As an extreme example of this, consider the situation where a facility's two products are lawn mowers and snow blowers. In this case, specification of an aggregate production plan neglects the most interesting question; namely, the correct production plan for each individual item. [Bergstrom and Smith, 1970]

Several methods for achieving the disaggregation and coupling have been proposed. First, Bergstrom and Smith [1970] and Chang and Jones [1970] have achieved disaggregation by product line within the framework of the Linear Decision Rule methodology. Taubert [1968] has applied the general SDR methodology to a research laboratory, disaggregated to reflect the individual demands and costs for six different departments. Zoller [1971] has developed a general mathematical framework for the optimal disaggregation of aggregate plans. Green [1971] and Shwimer [1972] have proposed heuristic methods for coupling aggregate and detailed scheduling models. Finally, Hax and Meal [1975] have proposed and applied hierarchical planning as a methodology by which aggregate level decisions provide constraints that more detailed levels must meet. Hierarchical planning is an elaboration of the general conceptual framework centered around Figures 2 and 3 and we shall discuss it in some detail.

Hierarchical Planning for Coupling Aggregate and Detailed Plans

Hierarchical planning is a methodology that enables aggregate level decisions to provide constraints that more detailed levels must meet. The execution of plans at the detailed levels provides useful feedback that may generate change in higher-level, more aggregated decisions. Hierarchical planning systems are responsive to the organizational structure of the firm and define a framework for partitioning and linking the planning activities. Hax and Meal [1975] report an application of such a plan in a process manufacturing firm similar to a chemical plant or steel mill, and state that it is difficult to construct general purpose hierarchical planning systems that could be applied to any kind of industry. Rather, the planning effort must be tailored to a specific organization and its idiosyncracies.

The example firm is a multiplant, multiproduct operation with three distinct seasonal demand patterns. The firm has a strong incentive for maintaining a nearly level manufacturing rate, for the following reasons:

1. The capital cost of equipment is very high compared with the cost of shift premium for labor, and the plants normally operate three shifts five days a week, with occasional weekend work.
2. The labor union is very strong and exerts pressure to maintain constant production levels throughout the year for employment stabilization.

At the beginning of the study, the problem symptoms were poor customer service, excessive inventory, and high production costs. High inventory in the face of poor customer service was caused by excessive seasonal stock accumulation for some items and by shortages for others. High production costs primarily were due to runs that were uneconomically short, with consequent high setup costs and low productivity. The setup costs resulted from the high stock out rate and the consequent need to produce a small amount of each of many items in order to satisfy back orders. The problem to be solved involves

> . . . planning aggregate production levels, particularly in allocating available production capacity among several product types with differing seasonal demand patterns, and in the subsequent detailed scheduling of each item belonging to a product type. [Hax and Meal, 1975.]

System Structure. The first step toward solving the problem of developing the planning and control system was to define levels of aggregation for the various items produced by examining the extent to which sets of decisions regarding production were interdependent. If two sets of decisions were found to be independent, they were separated totally in the hierarchy of decisions.

Beginning at the most detailed level, items sharing a major setup cost were grouped into "families." Thus, scheduling decisions for items within a family were very dependent, while the opposite was true for items within different families. Also, decisions for a family in one time period were tied strongly to decisions for the same family in other time periods. This time dependence resulted from the need to accumulate seasonal inventories in both product families.

Product families were aggregated into "types" if they shared a common seasonal pattern and production rate. This categorization facilitated seasonal planning, since only the aggregate for all families in the type had to be considered in developing the plan.

The next step in the process was to develop a hierarchy of decisions based on the relationships developed in the aggregation process. The following steps were developed:

1. Assignment of families to plants.
2. Seasonal planning.

3. Scheduling of families.
4. Scheduling of items.

In addition to the preceding steps, basic inventory methods were used to establish minimum run lengths and overstock limits.

Results. Hax and Meal reported a total development cost in the range of \$150,000–\$200,000. While exact benefits were not reported, cost reductions from smoother production, fewer emergency interruptions, and reduced inventory carrying costs were expected to be more than \$200,000 per year in each plant.

INTERMITTENT SYSTEMS

The differences in planning, scheduling, and control between high-volume and intermittent systems are very substantial. While both *may* produce finished products for inventory, the periodic or intermittent nature of the latter system's production of parts, components, and products creates a more complex detailed scheduling problem. The nature of forecasting and planning production lot sizes is unique because of the dependent nature of the demand for parts and components—that is, their demand depends on the production schedule for the primary product, which in turn depends on market demand. In the case of complex assembled products with important subassemblies that, in themselves, may be produced for inventory, this demand dependence may be second- or third-order. The demand may depend on the production schedules for the subassembly, which in turn depends on the production schedule for the final product.

Efforts to couple aggregate and detailed plans and schedules for intermittent systems have been made by heuristic methods [Green, 1971]; [Shwimer, 1972], and by coupling a linear program with a detailed heuristic procedure [Hong, 1974]. These approaches have theoretical interest and may have future practical value. However, the procedure that is used most commonly in industry is requirements planning. Requirements planning's great value is that its logic comes from the details of the structure of orders flow, and that it has been computerized for efficient day-to-day use.

We shall discuss the unique aspects of intermittent productive systems, as well as the concepts of requirements planning, production lot sizes that consider these unique characteristics, scheduling concepts, and integrated systems of planning and control now in use in industry.

REQUIREMENTS PLANNING CONCEPTS

Let us start with a simple product used as an example by New [1973], which will illustrate many of the concepts that concern us. Figure 4*a* is a drawing of a coffee table that shows the parts or components required. The table consists of a plywood top covered with veneer, four legs, and short and long stays or stringers between the legs for strength and rigidity.

Figure 4*b* is a simplified operation process chart showing the sequence of major operations required on each component to fabricate and assemble the table. Note that the frame, consisting of the four legs and stays, is assembled separately as a subassembly, finished, and then entered into the final assembly process in which the plywood top and veneer are added and finished. The veneer, glue, and finished materials are purchased outside: that is, unlike the other raw materials (plywood and block wood), they are not processed before use.

The operation process chart of Figure 4*b* specifies and generates directly what must be done, and the sequences required, if the table is to be a "one-of-a-kind" custom table. However, if a substantial number of tables are to be made, we have alternatives that complicate the problem but offer planning and scheduling opportunities for efficient manufacture.

Effects of Demand Dependence

Suppose that we have translated the current demand for tables into a master schedule, and that the tables are to be produced in lots of 100 every 2 weeks. They are not produced continuously, because the production rates for each operation in Figure 4*b* are relatively high; this would result in low utilization of workers and machines. We can presume that the enterprise uses the workers and the machines for other products (perhaps other table sizes and designs). Our table will be produced periodically in lots to satisfy demand.

There are many alternatives. We can consider the table as a unit and produce enough legs, stays, and tops to assemble 100 tables every two weeks. But since the cost of setting up the equipment for the various operations is different, and since the variable costs of production are different, we may be able to produce more efficiently by considering the manufacture of each component—perhaps manufacturing legs every 4 weeks in lots of 800 to key in with the master schedule.

Let us take this combination as an example: tables assembled every 2 weeks in lots of 100, and legs assembled every 4 weeks in lots of 800. Because the demand

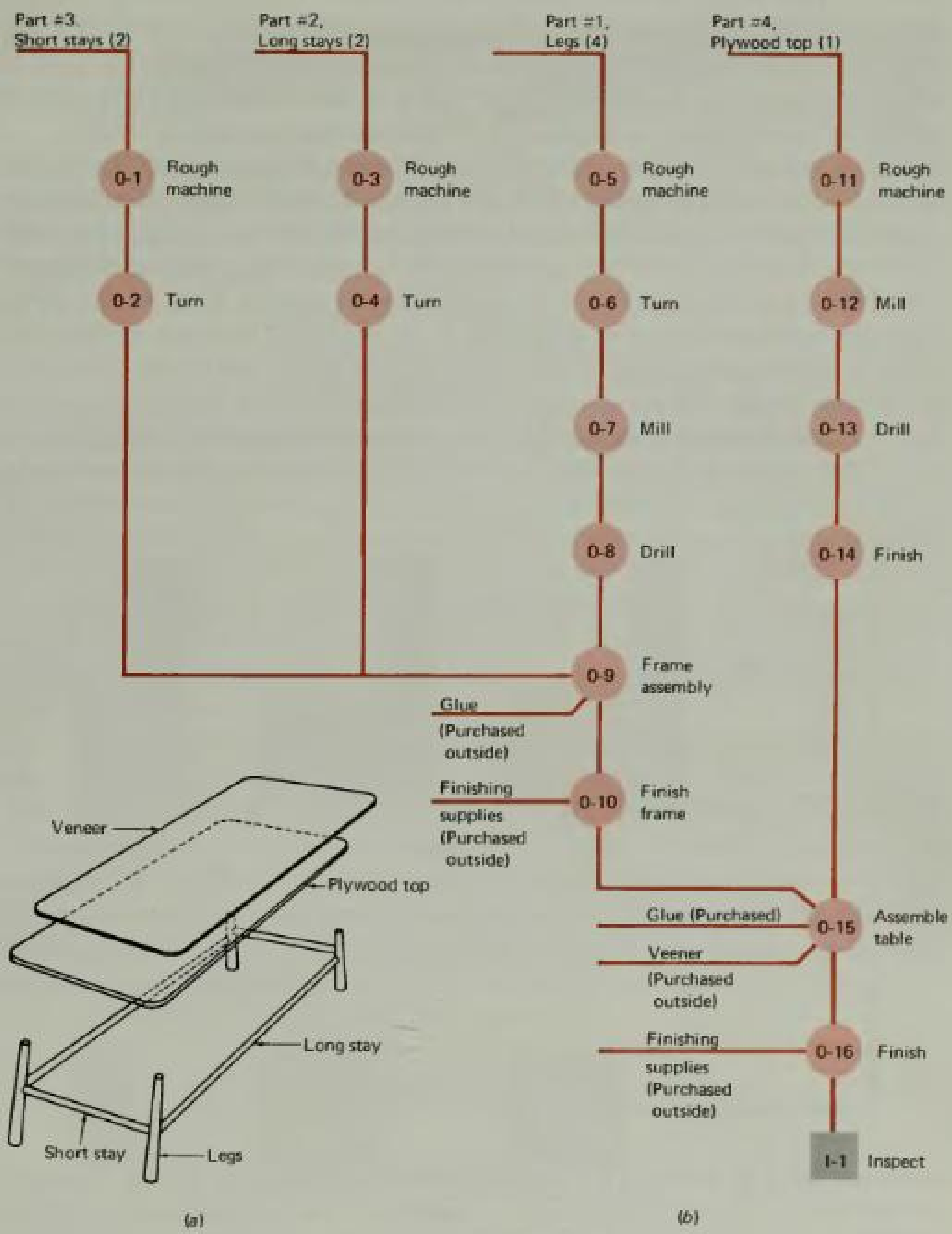


Figure 4. (a) Part components for a coffee table
(b) Operation process chart for the coffee table

ADAPTED FROM: C. New, *Requirements Planning* (Great Britain: Gower Press Ltd., and U.S.A.: Halsted Press, 1973).

for the table legs depends entirely on the *production schedule* for tables, the time phasing of the production of lots of legs with respect to table lots greatly affects the in-process inventory of legs. This is shown in Figure 5, in which the effects of two different schedules of producing legs are shown. Tables are produced in lots of 100 in weeks 3, 5, 7, etc. In Figure 5*a*, legs are produced in lots of 800 every 4 weeks in weeks 1, 5, 9, etc. Legs go into inventory when the lot is completed and are available for use the week following; the average in-process inventory of legs is $\bar{I} = 400$ units. Note that the time phasing of leg production in lots of 800 is terribly important; if we produce legs in weeks 2, 6, 10, etc., as in Figure 5*b* the in-process inventory is reduced to = 200 units. Therefore, the problem is not simply to produce legs in lots of 800 every 4 weeks, but to time-phase the production of legs with respect to the production of the primary item—the table. The demand for tables presumably is independent; however, the production of

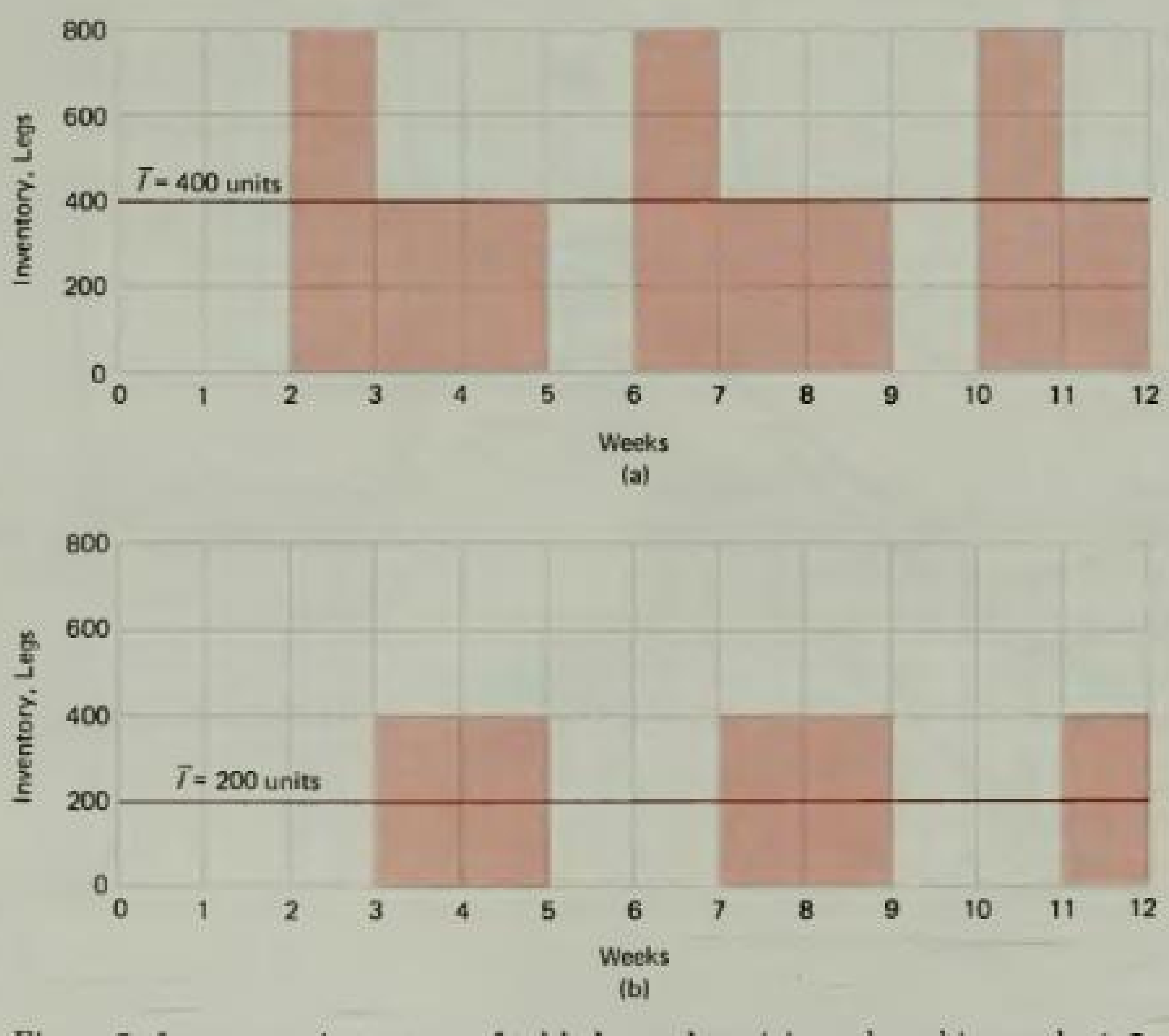


Figure 5. In-process inventory of table legs when, (a) produced in weeks 1, 5, 9, etc., and (b) produced in weeks 2, 6, 10. Tables go into inventory at the end of the week in which they are produced. Tables are assembled in weeks 3, 5, 7, etc.

legs must become a *requirement* as soon as the table production schedule is set. If the proper time-phasing is ignored, the price paid will be in higher in-process inventory of components.

Multiple Products with Some Common Components. Now let us consider a slightly more complex situation involving three different tables, presented by New [1973]. There are two different table tops—round-edge and squared-edge—and two leg types, as follows:

Table A—round-edge, leg-type 1
Table B—round-edge, leg-type 2
Table C—square-edge

We will follow through the structure of requirements determination that culminates in total requirements for plywood sheet for all 3 tables (Figure 6). Primary demand for the 3 tables is 20, 30, and 65 units per week, respectively. Production schedules to meet demand for each of the 3 tables are in lots of 40, 80, and 220, respectively (we will discuss lot size later). Since round-edge tops are used both in Tables A and B, requirements for round-edge tops represent the period-by-period summing from the production schedules for Tables A and B. For example, the requirement for round-edge tops in Period 1 is the sum of the production schedules of 40 + 80 for the 2 tables.

Plywood requirements are, in turn, the period-by-period summation of the production schedules for round-edge tops plus square-edge tops used in Table C. Even though plywood requirements represent aggregated usage in 3 different tables, they have a very uneven usuage pattern—470 in Period 1, 0 in Periods 2 and 3, 220 in Period 4, etc. Professionals term such a pattern "lumpy" demand. Figure 6 is a requirements plan for the 3 tables, the table tops, and the plywood raw material. A complete plan would provide similar information for all components and raw materials.

Forecasting versus Requirements

Figure 6 provides information concerning the use of plywood as a raw material. Let us now suppose that we wish to set up an inventory control system for the plywood so that it can be reordered as needed and so that a stock can be maintained. If the requirements for plywood (as shown in Figure 6) represent demand, can we apply standard forecasting techniques?

Figure 7 shows the application of an exponential forecasting system with

Table A.

Period	1	2	3	4	5	6	7	8
Forecast demand	20	20	20	20	20	20	20	20
Production	40	–	40	–	40	–	40	–

Table B.

Period	1	2	3	4	5	6	7	8
Forecast demand	30	30	30	30	30	30	30	30
Production	80	–	80	–	–	80	–	–

$\sum_i \text{Prod}_i$ =

Table C.

Period	1	2	3	4	5	6	7	8
Forecast demand	65	65	65	65	65	65	65	65
Production	220	–	–	220	–	–	220	–
Production Squegie edge tops	220	–	–	220	–	–	220	–

Round-edge tops

Period	1	2	3	4	5	6	7	8
Require-ments	120	–	120	–	40	80	40	–
Production	250	–	–	–	250	–	–	–

$\sum_i \text{Prod}_i$ =

Plywood sheets

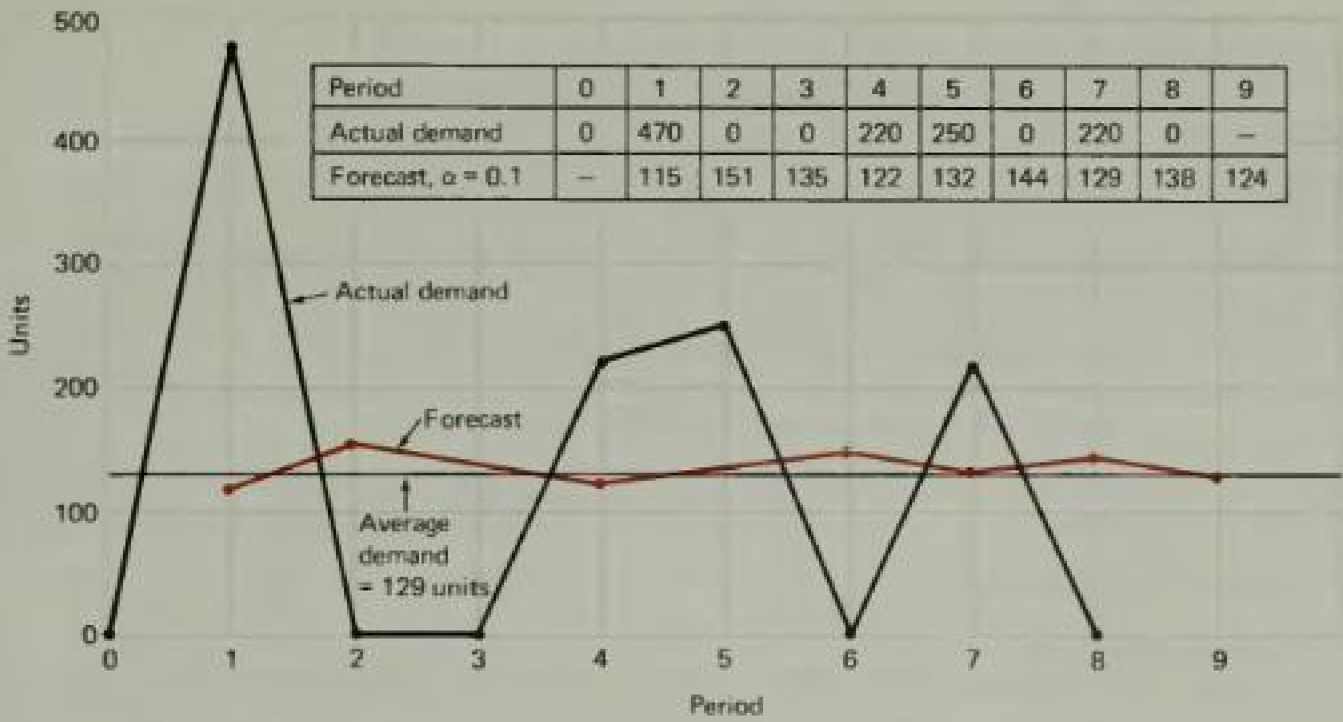
Period	1	2	3	4	5	6	7	8
Require-ments, Plywood sheets	470	–	–	220	250	–	220	–

Figure 6. Requirements determination for round edge tops, square edge tops, and finally plywood sheets. Requirements for round edge tops is *dependent* on production schedules for tables A and B. Requirements for plywood sheets is in turn dependent on production schedules for round and square edge tops.

ADAPTED FROM: C. New, *Requirements Planning* (Great Britain: Gower Press Ltd., and U.S.A.: Halsted Press, 1973).

Exp smt

$\alpha = 0.1$ to the actual demand data. Average actual demand for the 8 periods is 129 units, and the forecast has the usual smoothing effect (the initial forecast figure of 115 was taken as the sum of the average demand for the 3 tables.) However, we do not want to smooth demand, because the requirements schedule for plywood is the best forecast that we can obtain. Using exponential smoothing (or any other forecasting technique) is a misapplication of forecast-



Period	0	1	2	3	4	5	6	7	8	9
Actual demand	0	470	0	0	220	250	0	220	0	–
Forecast, $\alpha = 0.1$	–	115	151	135	122	132	144	129	138	124

Figure 7. Effect of using an exponential forecast ($\alpha = 0.1$) for dependent demand of plywood sheets

ADAPTED FROM: C New. *Requirements Planning* (Great Britain: Gower Press Ltd., and U.S.A.: Halsted Press, 1973).

ing here, because the demand for plywood is dependent. Compare the forecast line with actual demand in Figure 7. The forecast errors are very large: $MAD = 150.4$. Where, then, can we apply statistical forecasting? It can be applied to the primary demand for the 3 tables but not to the dependent components and raw materials. The requirements of the latter elements are derived directly from the production schedules, which have considered primary demand and other factors. Thus, inventory control systems for dependent items will take the requirements schedule directly as an input, rather than interposing a forecasting system.

The basic ideas of requirements generation developed in connection with the simple example of the three tables, is carried forward into larger-scale, more complex systems. As Figure 8 shows, the form of the plan normally includes: (1) the master schedule for the independent item at the head; followed by (2) requirements plans for dependent components; which are keyed in with (3) the master schedule as shown; indicating (4) production lead time for the component, expected usage, planned deliveries from production, a planned stock record, and the timing of production order releases phased to consider lead time.

Master Assembly Schedule - TABLE 13-76

Week	9	10	11	12	13	14	15	16	17
Planned requirements	-	600	-	-	700	-	400	-	200

Requirements, Part 13-76 L (Legs)(4 per assembly)

Production lead time, 2 weeks

Week	9	10	11	12	13	14	15	16	17
Expected Usage	-	2400	-	-	2800	-	1600	-	800
Planned deliveries	-	-	-	-	2400	-	600	-	600
Planned stock on hand, end of week	4000	1600	1600	1600	1200	1200	200	200	200
Planned production order release	-	-	2400	-	600	-	600	-	-

Requirements, Part 13-76 SL(Long stay)(2 per assembly)

Production lead time, 1 week

Week	9	10	11	12	13	14	15	16	17
Expected Usage	-	1200	-	-	1400	-	800	-	400
Planned deliveries	-	1200	-	-	1500	-	800	-	300
Planned stock on hand, end of week	200	200	200	200	300	300	300	300	200
Planned production order release	1200	-	-	1500	-	800	-	300	-

Figure 8. Requirements plan for Master Assembly Schedule and two dependent components.

LOT SIZE DECISION POLICIES

In the examples discussed so far, we assumed the size of the production run or lot size. Now we shall consider some of the alternate decision policies that might be used, and examine how they fit in with requirements planning.

First, however, let us review what we have already learned about requirements planning that may help us choose appropriate policies. First, we know that demand for components is dependent, and that it must be thought of as requirements that are generated to key in with the production schedule of the master assembly. The nature of the resultant demand distributions is not uniform and continuous, as it is for primary items whose demand is the aggregation of independent orders from multiple sources. Demand is lumpy because it is dependent, and demand variations do not generally result from random fluctuations. Thus, some of the important assumptions of traditional inventory control theory may not apply to dependent items. We must keep these assumptions in mind as we discuss alternate lot size policies, for they explain why some policies perform better than others. A comparative study by Berry [1972] is reported in the following sections.

Economic Order Quantity (EOQ)

The EOQ policy, or the nonoptimal fixed reorder quantity alternatives, probably have been used most in the past. In Chapter 11 we said that a reorder point is set that is designed to cover expected usage during lead time, and that an order for a fixed quantity (EOQ) is placed at that time. The EOQ is computed by the equation, $EOQ = \sqrt{2c_pR/c_H}$, which we developed in the last chapter. The three important numbers that enter the computation are: R = requirements per time period; c_p = preparation costs per order; and c_H = inventory carrying costs per unit per time period.

Table 4 shows the performance of an EOQ model for the requirements schedule given. In order for the system to function over the 12 weeks, an order must be placed whenever beginning inventory less requirements is less than 280 units. Table 4 shows that 9 orders of $Q_0 = 166$ units were placed in order to meet requirements. The ordering cost, then, is $9 \times 300 = \$2{,}700$. Inventory costs are based on the average for beginning and ending inventory in each period; for example, average inventory for Period 1 is $(166 + 156)/2 = 161$ units. The inventory cost for the 12 periods is the simple sum of the period average inventories multiplied by the unit inventory cost, or $2 \times 3{,}402.48 = \$6{,}805$. The total incremental cost for the EOQ model for 12 weeks is \$9,505. As Berry points out, this example illustrates several problems with the EOQ procedure.

> When the demand is not equal from period to period, as is often the case in requirements planning forecasts, one of the assumptions underlying the EOQ formula is violated. Since demand does not occur at a constant rate, as is assumed by the EOQ formula, the restriction of fixed lot sizes results in larger inventory carrying costs. This occurs because of the mismatch between the order quantities and the demand values, causing excess inventory to be carried forward from week to week. [Berry, 1972.]

weakness of EOQ

TABLE 4

PERFORMANCE OF AN *EOQ* MODEL FOR A GIVEN REQUIREMENTS SCHEDULE
($\bar{R}$ = 92.1 UNITS PER WEEK, c_p = \$300 PER ORDER,
c_H = \$2 PER UNIT PER WEEK, Q_0 = 166 UNITS)

Week number	1	2	3	4	5	6	7	8	9	10	11	12
Requirements	10	10	15	20	70	180	250	270	230	40	0	10
Quantity ordered	166	166		166		166	166	166	166	166	166	
Beginning inventory	166	156	312	297	443	373	359	275	171	107	233	399
Ending inventory	156	312	297	443	373	359	275	171	107	233	399	389

Ordering cost:	\$2,700
Inventory carrying cost:	6,805
Total incremental cost:	\$9,505

$\sum \frac{166+156}{2} + \cdots$ $\frac{399+389}{2}$

Adapted from: W. L. Berry, "Lot Sizing Procedures for Requirements Planning Systems: A Framework for Analysis," *Production & Inventory Management*, 2d Quarter, 1972.

Table 4 results from the rigid application of the EOQ procedure. If we allow the order quantity to be increased in those periods in which the demand exceeds the economic lot size plus the inventory carried over into the period, Berry shows that we can reduce the average inventories considerably, resulting in a 12 period inventory cost of \$3,065 and a total incremental cost of \$4,865.

Fixed Reorder Cycle System

$\frac{(10+10+15+20+70 \ldots +0+10)}{12}$

Recall that we can approximate the economic time interval between replenishment orders by dividing the EOQ by the mean demand rate. For our example, $EOQ/\bar{R} = 166/92.1 = 1.8$, or approximately 2. When we apply this procedure to the same requirements schedule that we used before, we get the reorder and inventory pattern of Table 5, in which the order quantity is the sum of requirements for the next two weeks. The system requires 6 orders, but the lot sizes range from 20 to 520 units. The result is that ordering costs are reduced by 33 percent, the inventory carrying cost is reduced by 68 percent, and the total incremental cost is reduced by 58 percent.

The fixed reorder cycle system improves inventory cost performance by allowing lot sizes to vary. However,

reread

> Like the EOQ procedure it, too, ignores much of the information contained in the requirements schedule. That is, the replenishment orders are constrained ~~to occur~~ at fixed time intervals, thereby ruling out the possibility of combining orders during

TABLE 5

PERFORMANCE OF A FIXED REORDER-CYCLE MODEL FOR A GIVEN REQUIREMENTS SCHEDULE ($\hat{R}$ = 92.1 UNITS PER WEEK, c_P = \$300 PER ORDER, c_H = \$2 PER UNIT PER WEEK, T_0 = 2 WEEKS)

Week Number	1	2	3	4	5	6	7	8	9	10	11	12
Requirements	10	10	15	20	70	180	250	270	230	40	0	10
Quantity Ordered	20		35		250		520		270		10	
Beginning Inventory	20	10	35	20	250	180	520	270	270	40	10	10
Ending Inventory	10	0	20	0	180	0	270	0	40	0	10	0
	Ordering Cost:							1,800				
	Inventory Carrying Cost:							2,165				
	Total Incremental Cost:							\$3,965				

Adapted from: W. L. Berry, "Lot Sizing Procedures for Requirements Planning Systems: A Framework for Analysis," *Production and Inventory Management*, 2d Quarter, 1972.

periods of light product demand, e.g., during weeks 1 through 4 in the example. If, for example, the orders placed in weeks 1 and 3 were combined and a single order was placed in week 1 for 55 units, the combined costs can be further reduced by \$160. [Berry, 1972]

Part-Period Total Cost Balancing

Another procedure involves the attempt to balance the total incremental costs in each ordering decision. This procedure also is called the *part-period algorithm*, and it uses all the information provided by the requirements schedule and attempts to equate the total costs of placing orders and carrying inventories. The procedure considers the alternative lot size choices available at the beginning of Week 1; that is, it places an order to cover the requirements for:

Week 1 only
Weeks 1 and 2
Weeks 1, 2, and 3
Etc.

The inventory carrying costs for these alternatives are as follows:

Week 1: $c_H(\bar{I}_1) = 2 \times 0.5 \times 10 = \$10$

Weeks 1 and 2: $c_H(\bar{I}_1 + \bar{I}_2) = 2(0.5 \times 10 + 1.5 \times 10) = \$40$

Weeks 1, 2, and 3: $c_H(\bar{I}_1 + \bar{I}_2 + \bar{I}_3) = 2(0.5 \times 10 + 1.5 \times 10 + 2.5 \times 15) = \$115$

Weeks 1, 2, 3, and 4: $c_H(\bar{I}_1 + \bar{I}_2 + \bar{I}_3 + \bar{I}_4) = 2(0.5 \times 10 + 1.5 \times 10 + 2.5 \times 15 + 3.5 \times 20) = \$255$

Weeks 1, 2, 3, 4, and 5: $c_H(\bar{I}_1 + \bar{I}_2 + \bar{I}_3 + \bar{I}_4 + \bar{I}_5) = 2(0.5 \times 10 + 1.5 \times 10 + 2.5 \times 15 + 3.5 \times 20 + 4.5 \times 70) = \$885$

By scanning the preceding set of calculations, we see that the fourth alternative—ordering 55 units to cover the demand for the first 4 weeks—approximates the ordering cost of $300. Thus, we conclude that an order should be placed at the beginning of the first week to cover the first 4 weeks' requirements.

Applying this procedure to the same requirements schedule results in the data shown in Table 6. The total incremental cost is reduced by an additional $480 or 12 percent, compared with the results for the fixed reorder cycle system in Table 5. The part-period total cost balancing procedure allows both the lot size and the time between orders to vary; thus, in periods of light demand, there are smaller lot sizes and longer time intervals between orders, compared with periods of high demand.

Obviously, the part-period balancing procedure (total incremental cost balancing) gives the best performance of the three reordering policies reported. First, we should understand that it performs so well because of its flexibility in considering replenishments involving both variable reorder quantity and variable reorder frequency. It considers several different possible horizons and selects the one in which order and inventory costs approximately are equated. In periods of low requirements for several periods, the part-period balancing procedure will group requirements, thus saving possible high ordering costs. If demand increases, the policy reacts by closing down its horizon, thus saving

TABLE 6

PERFORMANCE OF A TOTAL COST BALANCING MODEL (PART-PERIOD BALANCING) FOR A GIVEN REQUIREMENTS SCHEDULE ($\bar{R}$ = 92.1 UNITS PER WEEK, c_P = \$300 PER ORDER, c_H = \$2 PER UNIT PER WEEK)

Week Number	1	2	3	4	5	6	7	8	9	10	11	12
Requirements	10	10	15	20	70	180	250	270	230	40	0	10
Quantity Ordered	55				70	180	250	270	270			10
Beginning Inventory	55	45	35	20	70	180	250	270	270	40	0	10
Ending Inventory	45	35	20	0	0	0	0	0	40	0	0	0

Ordering Cost:	\$2,100
Inventory Carrying Cost:	1,385
Total Incremental Cost:	\$3,485

Adapted from: W. L. Berry, "Lot Sizing Procedures for Requirements Planning Systems: A Framework for Analysis," *Production and Inventory Management*, 2d Quarter, 1972.

incremental inventory costs. In short, the policy uses the information provided in the requirements plan fully to its advantage.

Second, however, the part-period balancing procedure is not necessarily optimal because it does not evaluate *all* possible alternatives for lot sizes. The Wagner-Whitin [1958] dynamic programming algorithm produces an optimal solution; however, since it is much more complex and requires greater computer time, it is not often used in practice. Berry included the Wagner-Whitin algorithm in his study, and it produced a solution that was $240 or 7 percent lower in cost than the part-period balancing policy. The part-period procedure did not consider the possibility of combining orders placed in Weeks 9 and 12. By carrying an extra 10 units in inventory for 3 weeks at a cost of $60, the Wagner-Whitin policy saved placing an order in the twelfth week; the net cost reduction was 300 − 60 = $240.

From a practical point of view, we need a lot size algorithm that performs well in a reasonable amount of computing time. At this point, the part-period procedure seems to be an outstanding candidate, based on comparative studies and experience [Berry, 1972; New, 1973; Orlicky, 1975]. The same lot size techniques can be applied to the determination of lot sizes for the primary products.

Buffer Stocks in Requirements Systems

We have already noted that dependent items are not subject to the kind of random variations in demand that characterize primary product demand. The demand variability largely is planned and is lumpy in nature. We do not need buffer stocks to absorb these kinds of fluctuations, since the requirements plan is designed to counter the variations with production orders, as shown in Figure 9. These kinds of variation are under managerial control.

There are sources of variation for which buffer stocks are a logical counter

Period	1	2	3	4	5	6	7	8
Expected usage	–	2400	–	–	2000	1500	–	2000
Planned deliveries	–	–	–	–	2000			
Planned stock on hand, end of week	5000	2600	2600	2600	2600 ~~600~~			
Planned production order release	–	–	2000					

Buffer stock = 1000 units, Lead time = 2 periods, EOQ = 2000 units

Figure 9. Projected inventory falls below the buffer stock level of 1000 units in period 5, triggering a prduction order release in the third period

measure, and they largely are connected with variations in supply schedule. The time required for processing orders through an intermittent system is variable because of delays, breakdowns, plan changes, etc. In addition, the actual quantity delivered from production is variable because of scrap. As a result, we need a cushion to absorb variations in the supply time and the quantity actually delivered.

Buffer Stock Levels

The nature of requirements systems changes the concept of reorder levels in both the EOQ and part-period methods of control. In Chapter 11, we noted that the reorder level was set to cover normal usage during the supply lead time *plus* the buffer stock (see Chapter 11, Figure 9). In requirements systems, however, the buffer stock level becomes the trigger or reorder level because we work with future inventory levels in a coordinated plan. Therefore, a production order release—phased by the lead time to be delivered fast enough to eliminate the impending shortage—can anticipate a projected fall in inventory level to or below the buffer level. Figure 9 shows how a production order release, triggered by a projected decrease in inventories below the buffer level of 1,000 units, is scheduled in the third period.

Buffer stock levels can be determined on the basis of general experience with supply lead times and an estimate of the maximum usage that is likely to occur per period. Other variables have the equivalent effect of buffer stocks. First, "safety factors" may be involved in the lead time estimates. If the lead time estimate is the time that includes 90 percent of the cases (and most production orders are received before then), then we have a buffer, in effect. Inflated requirements schedules also have an equivalent buffering effect. All these techniques are used; when they are used in combination, the equivalent buffer can be large and (unfortunately) partially hidden.

Computerized Requirements Planning

There are computer packages for requirements planning that make the entire system internally consistent and capable of rapid response for planning, as well as for required replanning. These programs generate detailed requirements and planned production orders. Usually, options can be called for determining the timing and the size of production orders, including the part-period reordering policy.

SHOP LOADING AND SCHEDULING

The requirements plan is also a general schedule of production orders that are to be processed through the shop. Each order is for a lot of specific part or component, and each may require a sequence of processing through a number of different functional work centers. Orders go to each work center according to the sequence of operations required, and they enter a queue of orders to be processed by that work center according to some priority sequence (the simplest is first come–first served).

Research on Priority-Dispatch Decision Rules

Looking at the system as a whole, we find a set of service centers with orders that require various sequences of processing and variable process or service times. In effect, this is a network of queues. A great deal of research has viewed the scheduling problem for intermittent systems in this context, using large-scale simulation models to test alternative policies. The queue discipline (the sequence of processing orders) has been the focus of many of these simulation studies [Carroll, 1965; Jones, 1973; LeGrande, 1963, Rowe, 1958]. In shop parlance, the queue discipline is the "priority system," commonly called priority dispatching decision rules.

In one such study at the Hughes Aircraft Company, LeGrande [1963] studied the effects of six different decision rules, using a variety of criteria reflecting flow time, on-time delivery of orders, and utilization factors. When the criteria were weighted equally, the SOT rule gave the best results (the SOT rule assigns the next job to be worked on by choosing the one with the shortest operation-time). However, when the criteria were weighted in terms of getting orders out on time, a due-date-oriented rule gave superior results. Carroll [1965] experimented with a family of rules that sequenced orders according to the largest ratio of downstream waiting time costs to operation process time, or c/t (called c over t, or COVERT). In comparison with five other rules tested, the COVERT rule was extremely effective in meeting due dates.

Tracking the Progress of Orders

Unfortunately, we cannot predict the flaws in our original plans. Machines may break down, work may pile up behind some critical machine, and dozens of other unexpected production troubles may occur that interfere with original schedules. With hundreds or even thousands of current orders in a shop, the only way to be sure that orders ultimately will meet schedules is to provide information feedback and a system of corrective action that can compensate for delays. Returned job tickets, move orders, and inspection reports can provide formal information that can be compared with schedules. But in order to be valuable, this information flow must be rapid so that we can take action on up-to-date reports. Usual company mail systems are much too slow to serve the needs of production control systems. Therefore, special communication systems are common, such as special mail services, intercommunication systems, teletype writers interconnected to central offices, pneumatic tube systems, and remote data collection centers tied directly in with automatic computing systems. These systems, in combination with rapid data processing, can grind out current reports that can be used to expedite and reschedule orders.

Computerized Planning, Scheduling, and Control Systems

As we mentioned earlier, the complexities of job shop planning, scheduling, and control, together with the difficulties of maintaining graphic methods have led to the use of computer based systems [Bulkin, Colley, and Steinhoff, 1966; LeGrande, 1963; Moodie and Novotny, 1967; Reiter, 1966]. One such system, installed at the Hughes Aircraft Company, is particularly interesting because it combines shop-load forecasting, daily simulation scheduling, remote data collection, and a series of important reports in a computer based system [Bulkin, Colley, and Steinhoff, 1966; LeGrand, 1963].

The Hughes Job Shop Control System. The Hughes plant in which the system is installed has approximately 1,000 machines and/or work centers that are grouped into 120 functional machine or work centers. The work centers are manned by 400 direct workers. Approximately 2,000 to 3,000 orders are in process at any single time, with an average of 7 operations per order. The average total processing time for an order is 2.5 hours, and the average order cycle time is 3 to 4 weeks.

The broad outlines of the entire system are shown in Figure 10. Operational process planning is developed from engineering blueprints to yield the critical data of processes to be performed, standard process times, and standard flow allowances through the shop. These data are placed in the disk file of the computing system. When a fabrication requirement notice is received from one of the assembly product line departments, a fabrication order is transmitted to data processing, which initiates the preparation of a fabrication shop order and establishes an open order in the fabrication open order master file. A deck of

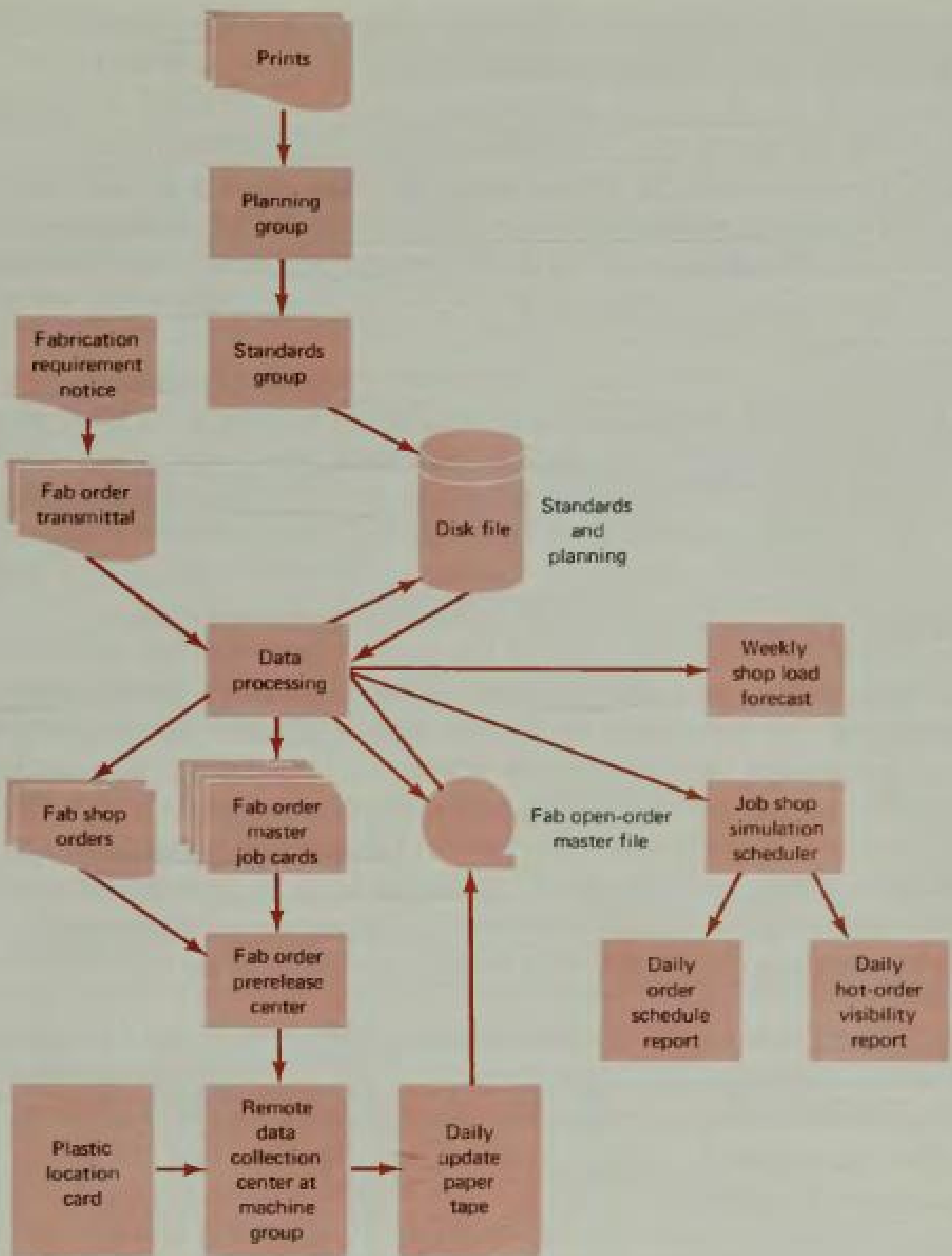


Figure 10. Hughes Aircraft Company's production, planning, and control system

ADAPTED FROM: E. Le Grande, "The Development of a Factory Simulation System Using Actual Operating Data," *Management Technology* 3(May–June 1963).

master job cards also is produced, which accompanies the order through the shop. These papers then are sent to a fabrication order control prerelease center, from which orders are dispatched to the shop.

After release to the shop, the master job cards are used in conjunction with plastic location cards at machine groups to actuate the remote data collection devices. At the end of the day, a paper tape containing all order moves transacted during the day is used to update current order locations and to code completed operations in the fabrication open order file. On a daily, updated basis, the fabrication open order master file contains the information on every open order. It is that heart of the system, which produces the three documents: the *weekly shop-load forecast*, the *daily order-schedule report*, and the *daily hot-order visibility report*.

The *daily order-schedule report* and the *hot-order visibility report* are developed by the job shop simulation scheduler. These documents are used by shop foremen and expediters in the daily sequencing of orders at the machine groups in the shop. The simulation scheduler is a separate computing routine that is used to simulate one shift of shop activity and to generate the two daily reports. Involved in the simulation scheduling process is the use of a priority dispatch decision rule that is related to slack time in the schedule.

The reports of the performance of the Hughes system given by LeGrande [1963] indicate that the computer based integrated system yields subsantial improvements. For example, in the first six months of operation, the number of orders completed on time increased by 10 percent. The average order cycle time was reduced by one week, thus reducing in-process inventory and increasing worker and machine utilization significantly. The new system seemed to relieve organization stress and pressure, as indicated by a 60 percent reduction in expediting effort. Such effort occurs when special personnel are assigned to hot orders to make sure that these orders receive special attention at each stage.

Western Electric Interactive System. An interesting development that uses the capabilities of time-shared computing has promise for the job shop scheduling problem. Godin and Jones [1969] developed interactive scheduling systems that use either typewriter or display terminals and thereby place the production supervisor, or a scheduler, in a loop with a computer program. By interacting with the computer program, the scheduler develops and/or alters a schedule. Schedules are generated by making choices from among sets of decision rules: for example, rules for the acceptance or rejection of orders, rules for sequencing orders, and rules for allocating the use of overtime. A schedule is developed by testing the effects of various combinations of possibilities and by simulating various alternative assignments and their effects.

Godin and Jones developed a small-scale application of the interactive scheduling concept in one of the coil winding shops of the Western Electric Company. Of the 200 types of coils manufactured, approximately 65 are active at any point in time, and each type usually is produced for 3 or 4 different orders.

There are 20 of a particular type of winding machine and 2 of a second type; the work force consists of from 20 to 35 machine operators working on 2 or 3 shifts, depending on shop load.

Because of differences in skills, not all operators can wind all types of coils, and performance against the standard differs considerably for various combinations of operators and coil types. Similarly, some machines cannot be used to wind some types of coils. The production supervisor must assign workers and machines to each other in a way that balances pressures for on time deliveries against the other factors that make for production effectiveness.

The system functions through a main program and twenty-three subroutines. To enable humans to communicate with the program through the typewriter terminal, a simple conversational language was developed. A single key stroke indicates a whole series of commands. For example, when the key marked "Load" is depressed, the computer understands it as a request to print the winding shop load on the printer.

The system allows the production supervisor to consider seven variables in constructing schedules:

1. Skill levels of workers.
2. Skill needs of products.
3. Capability of machines.
4. Availability of workers, machines, and materials.
5. Quantities required.
6. Completion dates.
7. Existing machine setups.

The computer program can carry out simulations of shop schedules so that the supervisor can test various alternative assignments and try to anticipate future problems. Various reports—such as shop status, a history of work by operator or machine, load summarized by standard hours, etc.—can be called for at any time.

Summary

High-volume continuous systems are production-distribution systems; they begin with raw material suppliers and end with consumption. Normally, only a portion of this system is under control of one manager. However, if the manager is also the manufacturer, he/she must view the system as a whole because of the inventories at various stock points. While seasonal inventories can help solve problems in aggregate scheduling at the factory level, they also cause problems. If demand decreases by 10 percent, for example, the effect at the factory level can be a substantial curtailment of output while excess system inventories are being used up. Recall the effects of the 1974 decrease in automobile demand on employment in automobile factories.

The magnitude of system inventories required for high-volume systems to function is rather surprising. These inventories are needed to fill the production-distribution pipelines, to accomodate economical systems for periodic replenishment, to cushion the effects of demand variability, and to smooth the effects of seasonal variation. Or, if we view inventories in terms of different classifications, they are needed in several phases: raw materials, in-process, and finished goods.

Proposals for relating aggregate and detailed plans have included disaggregation, heuristic coupling of the two levels of plans, and hierarchical planning. Currently, the latter appears to be both a practical and effective methodology for advanced concepts linking modular programs, while also considering the idiosyncracies of individual organizations.

Requirements planning coordinates aggregate and detailed plans and schedules for intermittent systems. In the conceptual framework of requirements planning, primary demand for products is reduced to economical production lots (or may have a continuous schedule); however, the production of parts and components depends on the production schedule of the primary item. Often, parts and components may be used in more than one primary product; thus, the requirements schedule for such an item is derived by a complex summing of the needs of all primary products. The demand dependence of parts and components profoundly affects the policies used to determine the relative timing of production orders, as well as the determination of production lot sizes.

The demand for parts and components depends not only on the quantities needed, but also on the timing of delivery. Since we are dealing with an interlocked structure, the components must be ready for use at a precise time. If they are too late, production of primary items will be disrupted and delayed. If they are too early, in-process inventory costs will increase.

Since demand for components is dependent and usually lumpy, some of the usual assumptions in traditional inventory models are not valid. For example, demand does not occur at a constant rate, since it is the aggregation of independent demands from multiple sources; and the demand variations are not due to random fluctuations. As a result, the so-called Economic Order Quantity (EOQ) policy belies its name, and substantial improvements can be obtained by periodic and part-period reordering policies.

Computerized requirements planning systems now commonly are available. These systems generate requirements for complex product structures and create lot size policy options.

We can visualize the flow of production orders through the shop as a network of queues, and a great deal of simulation research work has focused on the queue discipline as a variable that is under managerial control. There are many different criteria by which to judge the effectiveness of priority dispatch decision rules (queue disciplines); when they are weighted equally, the SOT rule gives the best results. When the criteria that emphasize getting orders out on time are given heavier weight, then a due-date oriented rule is favored. Rules that anticipate downstream waiting time cost are particularly effective in meeting due dates.

With the availability of present day computers, integrated systems of planning and control are possible. One example is the Hughes Job Shop Control System. Such systems load work centers based on some priority rule and produce updated daily status reports on department schedules and weekly reports on shop load. Interactive systems also are available now, such as the one at the Western Electric Company.

Review Questions

1. Under what circumstances is rebalancing of production lines necessary when adjusting to a new aggregate plan?

2. Given the aggregate plan as a constraint on operations scheduling in continuous systems, what flexibility is left to planners in developing detailed schedules?

3. What flexibility is left to management in continuous systems for adjusting deviations in actual output from planned levels within a given period?

4. In the model of the production-distribution system of Figure 2, note that information concerning demand at the retail level is fed back directly to the factory, and that, as a result, the magnitude of oscillations in orders and production activity are reduced considerably. What are the implications of this result for organizing and establishing lines of authority and responsibility for a multistage production-distribution system?

5. Using Figure 2 as a background, discuss the aggregate plan as a constraint on operations scheduling.

6. What are the basic reasons why we prefer to deal with disaggregate plans for products rather than planning entirely on an aggregate basis?

7. Contrast optimal disaggregation and hierarchical planning as methods for coupling aggregate and disaggregate plans.

8. In terms of the coffee table that is shown in Figure 4 and used as an example for requirements planning, why not produce enough components every 2 weeks to key in with a schedule of 100 completed tables per 2 weeks? What are the disadvantages?

9. Still using the coffee table as an example, let us suppose that the tables are produced in lots of 400 in weeks 3, 6, 9, etc. Legs are produced in lots of 1,600 every 6 weeks. How should the leg production be phased with respect to the table assemblies? Why?

10. What are the definitions of *dependent* and *independent* demand items? What kinds of forecasting methods are appropriate for each as a basis for production and inventory control?

11. What kinds of variation in demand for dependent items are not taken into account by the straightforward computation of requirements? How can these kinds of variation be absorbed?

12. What is a *dispatching decision rule*? How does it relate to the structure of a waiting line model?

13. If management's objectives place a high value on low in-process inventories, what kind of priority dispatching rule seems best?

14. If management's objectives place a high value on on-time delivery of orders, what kind of priority-dispatching decision rule seems best?

15. Examine the Hughes Aircraft Company's computerized planning, scheduling, and control system summarized by Figure 10. Does the Hughes system incorporate the functions of requirements planning? If so, how? If not, how does the Hughes system deal with similar problems?

Problems

1. The manufacturer of a small appliance is trying to understand what happens in his system as a result of a change in demand at the retail level. There is a time lag of one week for a retailer's order to be received from the wholesaler's warehouse. When the wholesaler orders from the factory, it takes two weeks for the factory to adjust its production level to meet the wholesaler's order. Inventory levels are not changed to adjust to increases or decreases in sales.

   The manufacturer writes down the following model to express the relationships of interest to him:

   Retail receipts$_n$ = retail orders$_{n-1}$
   Retail inv.$_n$ = retail inv.$_{n-1}$ + retail orders$_{n-1}$ − retail sales$_n$
   Retail orders$_n$ = X - retail inv.$_n$
   X = 100, for weeks 1–3 (old sales figure)
   X = 110, for weeks 4–11 (new sales figure)
   Warehouse receipts$_n$ = warehouse orders$_{n-2}$
   Warehouse Inv.$_n$ = warehouse Inv.$_{n-1}$ + warehouse orders$_{n-2}$ − retail receipts$_n$
   Warehouse Orders$_n$ = Y − warehouse Inv.$_n$
   Y = 200 per two weeks for weeks 1–6
   Y = 220 per two weeks for weeks 7–11
   Factory production rate$_n$ = warehouse orders$_{n-2}$ next page

   Using the form shown in Figure 11, determine the effect of a 10 percent increase in retail sales. Note that after three periods of increased sales, the retailer will accept the new sales figure as constant. Also, after retail orders have stabilized, the wholesaler will accept the new order level as the new permanent level.

2. Consider a supply-production-distribution system for a high-volume standardized product, perhaps a small appliance, which has the physical flow, transit, and handling times indicated in Figure 12. The total system volume is 2,000 units per day. Compute the equivalent pipeline inventory required for the system. p.443?

3. Referring to the system shown in Figure 12 and discussed in Problem 2, consider the effects of the ordering behavior of retailers, distributors, and the factory warehouse.

   a. The system volume of 2,000 units per day is not equally distributed

Period (Weeks)	Retail Sales	Retail Repts.	Retail Invty.	Retail Orders	W-hse Repts.	W-hse Invty.	W-hse Orders	Factory Rate
1	100	100	100	100	100	200	100	100
2	110	99						
3	110							
4	110							
5	110							
6	110							
7	110							
8	110							
9	110							
10	110							
11	110							
12	110							

Figure 11. Form for computing the effect of a 10-percent sales increase

10 Dist
200 units per day

among the 1,000 retailers, nor is the retailers' ordering behavior identical. Three-fourths of them carry no inventory but depend on the distributor to supply their needs when sales occur. These retailers account for 25 percent of the total sales volume. The other 250 retailers account for 1,500 sales units per day and carry inventories so that they can give customers immediate service. A study of their average sales volumes and ordering behavior shows that: the largest 10 retailers sell 50 per day; the next largest 40 sell 10 per day; and the remaining group of 200 sell 2 per day. The 50 largest retailers place orders every 10 days to cover sales (fixed reorder-cycle system). The remaining retailers who carry inventories order every 20 days to cover sales. The total supply lead time is 5 days.

5 company owned DIS differ from indep. DIST.

b. The 10 distributors each have approximately equal volume of 200 units per day. The ordering behavior of the 5 company owned distributors differs somewhat from that of the independent distributors. Since the company owned distributors are not charged for their inventories, they seem to order less often and in larger quantities to make sure that the retailers always get excellent service. The distributors review needs monthly and order one month's (30 days') supply. It takes: approximately 16 days for them to review sales records and prepare orders; 2 days for them to transmit these orders to the factory warehouse, and 10 days for shipment plus transit to the distributor. The distributor then takes 2 days to receive, unload, and store merchandise in order to make it available for future shipment to retailers. The company owned distrib-

not charged for inventory, order less often & larger qtys

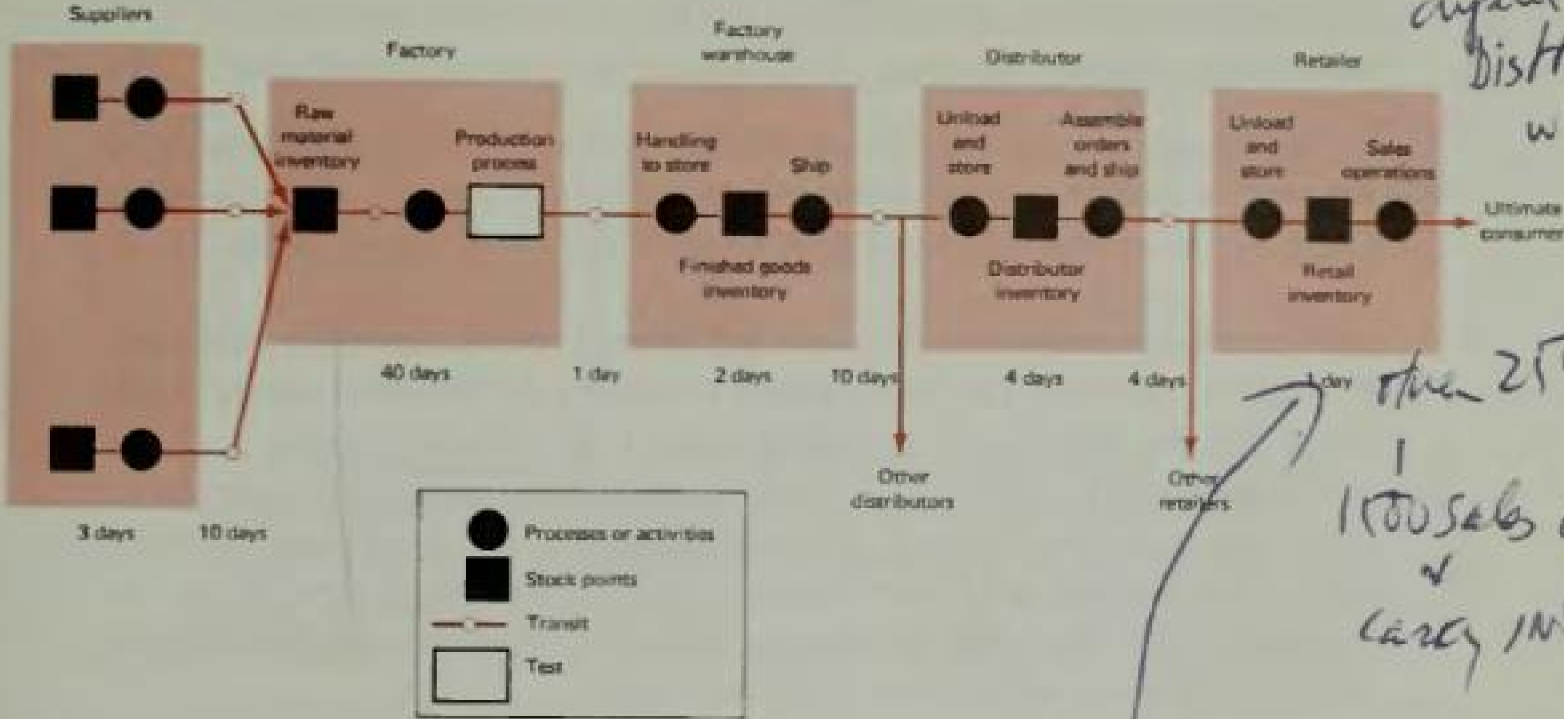
Dist checks needs monthly & orders one month's supply

16 Days - review sale record, prepare orders
2 " - transmit orders to fact. W.H.
10 " - shipment & transit to DIST

2000 units

10 no Retailers

2x% of total sales
3 Retaile carry no INV
dependent on Dist to supply when demand occurs.



then 250 retailers
100 sales unit per day
carry INV.

Figure 12. Supply-production-distribution system for a small appliance showing physical flow, assumed handling and transit times, and major stock points (five company-owned and five independent distributors, one thousand retailers)

p. 408

Indep. Dist. inv Bimonthly

(1) largest retailers sell 50 per day

next largest to sell 15 per day

rest of group 200 sells 2 per day

utors therefore have a replenishment cycle of 16 + 2 + 10 + 2 = 30 days. The independent distributors review needs twice each month (15 days) and order 15 days' supply. It takes them 6 days to review needs and prepare orders, and the remaining replenishment times are the same as for the company owned distributors, resulting in a replenishment cycle of 6 + 2 + 10 + 2 = 20 days.

Question

Compute the cycle inventories required at the retail and distributor stages of the system. What effects do the replenishment cycles themselves have on the size of the cycle inventories?

4. Referring again to the system shown in Figure 12 and discussed in Problems 2 and 3, consider the problems faced by the factory warehouse, distributors, and retailers in trying to give good delivery service. All the usage levels that we mentioned before were average rates. The various units have experienced extreme demand levels that would cause them to run out of stock (stock out) if they carried no buffer stock. The reasonable rates expected are as follows:

Largest 10 retailers	350 per 5 days per retailer
Next largest 40 retailers	70 per 5 days per retailer
Next largest 200 retailers	15 per 5 days per retailer

Dist - 2 days to receive unload, store MLSE

50 largest retailer orders every 10 days
rest of retailers - every 20 days
total supply lead time is 5 days

Five company owned distributors	6,800 per 30 days per distributor
Five independent distributors	4,400 per 20 days per distributor

The factory warehouse carries a buffer inventory of 10 days' supply to ensure that extreme demand from the distributors can be accommodated. Compute the system's finished goods buffer inventory requirement.

5. Referring again to the system shown in Figure 12 and discussed in Problems 2, 3, and 4, compute the total system inventory for finished goods. How could this finished goods system inventory be reduced? Since 5 of the distributors and all of the 1,000 retailers are outside managerial control and own their inventory, should the management of the enterprise (factory, factory warehouse, and 5 distributors) be concerned about how the rest of the system operates? That is, does the behavior of the independent distributors and retailers affect them, and if so, how?

6. Referring again to the system diagram of Figure 12, suppose that consumer demand increases by 10 percent from its previous rate of 2,000 units per day. What impact will this have on system inventories and the ability of the system to provide service in the short term?

7. The requirements for a motor-drive unit to be assembled into a dictating machine follow the assembly schedule for the completed unit. The assembly schedule requires motor-drive units timed for use as shown in Table 7. Other data for the motor-drive unit are: average requirements are $\bar{R} = 116$ units per week, $c_P = \$400$ per lot, and $c_H = \$4$ per unit per week. What is the inventory record and total incremental cost under each of the following lot size policies: 1780
   a. Economic lot size.
   b. Economic fixed reorder cycle model.
   c. Part-period total cost balancing.

   Account for the differences in performance of the three lot-size policies.

TABLE 7

REQUIREMENTS SCHEDULE FOR A MOTOR-DRIVE UNIT

Week number	1	2	3	4	5	6	7	8	9	10	11	12
Requirements, units	25	30	75	125	200	325	400	100	0	100	0	10

Total requirements for 12 weeks, 1,400 units.

1.3149961 115.83

TABLE 8

STABILIZED REQUIREMENTS SCHEDULE FOR A MOTOR-DRIVE UNIT

Week number	1	2	3	4	5	6	7	8	9	10	11	12
Requirements, units	300	300	300	300	350	350	400	400	350	350	350	325

Total requirements for 12 weeks, 4,375 units.

8. The requirements for the motor-drive unit described in Problem 7 have been stabilized considerably by compensatory promotion of the dictating machine and by the design of a line of portable tape recorders that have general use and a counterseasonal cycle. Table 8 shows the new requirements schedule. What is the inventory record and total incremental cost of the same three lot size policies discussed in Problem 7? Account for the differences in performance of the three policies.

9. Job orders are received at a work center whose characteristics are indicated by the data in Table 9. In what sequence should the orders be processed at the work center if the priority dispatch decision rule is:
   a. FCFS (first come–first served).
   b. SOT (shortest operation time).
   c. SS (static slack; that is, due-date less time of arrival at work center).
   d. FISFS (due-date system; first in system–first served).
   e. SS/RO (static slack divided by remaining number of operations).

   Compute priorities for each rule and list the sequence in which orders would be processed. Which decision rule do you prefer? Why?

TABLE 9

ORDER AND PROCESSING DATA FOR SIX JOBS

Order Number	*Due Date*	*Date and Time Received at Center*	*Operation Time, Hours*	*Remaining Operations*
1	May 1	Apr. 18, 9 A.M.	6	3
2	Apr. 20	Apr. 21, 10 A.M.	3	1
3	June 1	Apr. 19, 5 P.M.	7	2
4	June 15	Apr. 21, 3 P.M.	9	4
5	May 15	Apr. 20, 5 P.M.	4	5
6	May 20	Apr. 21, 5 P.M.	8	7

References

Bergstrom, G. L., and B. E. Smith. "MultiItem Production Planning—An Extension of the HMMS Rules." *Management Science* 16(June 1970): 614–629.

Berry, W. L. "Lot Sizing Procedures for Requirements Planning Systems: A Framework for Analysis." *Production and Inventory Management* (2d Quarter 1972): 19–34.

Bitran, R., and A. Hax. "On the Design of Hierarchical Production Planning Systems." *Decision Sciences* 8(January 1977).

Brown, R. G. *Decision Rules for Inventory Management.* New York: Holt, Rinehart, & Winston, 1967.

Buffa, E. S., and W. H. Taubert. *Production-Inventory Systems: Planning and Control.* Rev. ed. Homewood, Ill.: Richard D. Irwin, 1972.

Bulkin, M. H.; J. L. Colley; and W. H. Steinhoff. "Load Forecasting, Priority Sequencing, and Simulation in a Job-Shop Control System." *Management Science* 13 (October 1966): 29–51.

Carroll, D. C. "*Heuristic Sequencing of Single and Multiple Component Jobs.*" Ph.D. dissertation, Massachusetts Institute of Technology, 1965.

Chang, R. H., and C. M. Jones. "Production and Work Force Scheduling Extensions." *AIIE Transactions* 2(December 1970): 326–333.

Crowston, W. B.; W. H. Hausman; and W. R. Kampe II. "Multistage Production for Stochastic Seasonal Demand." *Management Science* 19(April 1973): 924–935.

Forrester, J. *Industrial Dynamics.* Cambridge, Mass.: MIT Press, 1961.

Fryer, J. S. "Labor Flexibility in Mulhechelon Dual-Constraint Job Shops." *Management Science* 20(March 1974): 1073–1080.

———. "Organizational Segmentation and Labor Transfer Policies in Labor and Machine Limited Systems." *Decision Sciences* 7(October 1976).

Gelders, L., and P. R. Kleindorfer. "Coordinating Aggregate and Detailed Scheduling Decisions in the One-Machine Job Shop." *Operations Research* 22(January-February 1974): 46-60.

GH20-0751-1, OS/360 "Requirements Planning Application Description." *IBM* 2d ed., June 1972.

Godin, V., and C. H. Jones. "The Interactive Shop Supervisor." *Industrial Engineering* (November 1969): 16-22.

Green, R. S. "Heuristic Coupling of Aggregate and Detail Models in Factory Scheduling." Ph.D. Dissertation, Massachusetts Institute of Technology, 1971.

Greene, J. H. *Production and Inventory Control: Systems and Decisions*. Rev. ed. Homewood, Ill.: Richard D. Irwin, 1974.

Hausman, W. H., and R. Peterson. "Multiproduct Production Scheduling for Style Goods with Limited Capacity." *Management Science* 18(March 1972): 370-383.

Hax, A. C. *Integration of Strategic and Tactical Planning in the Aluminum Industry*. Technical Report No. 86. Massachusetts Institute of Technology, Operations Research Center, 1973.

______, and H. C. Meal. "Hierarchical Integration of Production Planning and Scheduling". *Studies in the Management Sciences: Vol. 1, Logistics*, edited by M. A. Geisler, Holland: North Holland-American Elsevier, 1975.

Holloway, C. A., and R. T. Nelson. "Job Shop Scheduling with Due Dates and Variable Processing Times." *Management Science* 20(May 1974): 1264-1275.

Holstein, W. K. "Product Planning and Control Integrated." *Harvard Business Review*, May-June 1968.

______, and W. L. Berry. "The Labor Assignment Decision: An Application of Work Flow Structure." *Management Science* 16(February 1970): 324-336.

Hong, M. L. "Experiments with a Job Shop Production Model Coupling a Linear Program and a Heuristic Scheduling Procedure." Master's thesis, University of California, Los Angeles, 1974.

Jones, C. H. "An Economic Evaluation of Job Shop Dispatching Rules." *Management Science* 20(November 1973): 293–307.

Krajewski, L. J., and L. P. Ritzman. "Dissaggregation in Manufacturing and Service Organizations: Survey of Problems and Research. *Decision Sciences* 8(January 1977).

Lee, W. B., and C. P. McLaughlin. "Corporate Simulation Models for Aggregate Materials Management." *Production and Inventory Management* (1st Quarter 1974): 55–67.

LeGrande, E. "The Development of a Factory Simulation System Using Actual Operating Data." *Management Technology* 3(May 1963).

Magee, J. F., and D. M. Boodman. *Production Planning and Inventory Control.* 2d ed. New York: McGraw-Hill, 1967.

"Material Requirements Planning by Computer." *American Production and Inventory Control Society*, 1971.

Miller, J. G., and W. L. Berry. "The Assignment of Men to Machines: An Application of Branch and Bound." *Decision Sciences* 8(January 1977).

Moodie, C. L., and D. J. Novotny. "Computer Scheduling and Control Systems for Discrete Part Production." *Journal of Industrial Engineering* 19(July 1967): 648–671.

Moore, F. G., and R. Joblanski. *Production Control.* 3d ed. New York: McGraw-Hill, 1969.

New, C. *Requirements Planning.* Epping, Essex, Great Britain: Gower Press, Ltd., 1973.

Oral, M., and J. L. Malouin. "Evaluation of the Shortest Processing Time Scheduling Rule with Truncation Process." *AIIE Transactions* 5(December 1973): 357–365.

Orlicky, J. *Material Requirements Planning.* New York: McGraw-Hill, 1975.

Plossl, G., and O. Wight. *Production and Inventory Control: Principles and Techniques.* Englewood Cliffs, N.J.: Prentice-Hall, 1967.

Reiter, S. "A System for Managing Job Shop Production." *Journal of Business* 39(July 1966): 371–393.

Reitman, J. *Computer Simulation Applications.* New York: John Wiley & Sons, 1971. Chapter 9 is a computer simulation of a job shop manufacturing system.

Ritzman, L. P., L. J. Krajewski, and M. T. Showalter. "The Dissaggregation of Aggregate Manpower Plans." *Management Science* 22(July 1976) 1204–1214.

Rowe, A. J. "Sequential Decision Rules in Production Scheduling." Ph.D. dissertation, University of California, Los Angeles, August 1958.

Shwimer, J. *Interaction Between Aggregate and Detailed Scheduling in a Job Shop.* Technical Report No. 71. Massachusetts Institute of Technology, Operations Research Center, 1972.

Taubert, W. H. "The Search Decision Rule Approach to Operations Planning." Unpublished Ph.D. dissertaion, University of California, Los Angeles, 1968.

Thurston, P. H. "Requirements Planning for Inventory Control." *Harvard Business Review* (May–June 1972).

Wagner, H. M., and T. M. Whitin. "Dynamic Version of the Economic Lot Size Model." *Management Science* 5(October 1958).

Whybark, D. C., and J. G. Williams. "Material Requirements Planning Under Uncertainty." *Decision Sciences* 7(October 1976).

Zimmerman, H. J., and M. G. Sovereign. *Quantitative Models for Production Management.* Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Zoller, K. "Optimal Disaggregation of Aggregate Production Plans." *Management Science* 17(April 1971): 533–549.

Chapter 13

Planning, Scheduling, and Control of Large Projects

In light of the relatively recent emphasis on research and development and on very large-scale, one-time projects in the aerospace and construction industries, the planning methods described in Chapters 5, 7, and 12 are not adequate. While assembly and operation process charts still are valid mechanisms for making more detailed studies, they imply that we are dealing with substantial numbers of items that would justify careful analysis of minute details. But for the large-scale project, an entirely different level of planning becomes necessary. For example, the entire manufacturing process represented by the operation process chart for the capacitor in Chapter 7, Figure 6 is condensed into the question: "When must these capacitors be available to dovetail with the assembly of the guidance computer component?" Or the production of the capacitor may completely depend on the answer to another question: "When must the guidance computer component be available for test and assembly into the guidance system as a whole?" This question does not consider the fabrication of the capacitor component, since it probably would be subcontracted, the subcontractor would have to consider the detailed plans for its manufacture. The project as a whole must focus on planning the important activities in larger blocks and giving close attention to how these activities must dovetail to achieve the total end result. Which activities depend for their execution on the completion of other activities? Which activities can proceed relatively independently or in parallel with other activities? As we shall see, these kinds of plans are bound up intimately in the basic sequencing required and the schedule.

The very nature of the one-time large project demands that *what* has to be done and the *schedule* for performance must be planned together. These factors are interdependent, and the planning of large projects involves planning for deployment of resources to the total project. To accomplish this, we must determine the activities required, the timing and interdependencies, the requirements of various possible schedules for manpower and other resources, and the relationship of all the foregoing to a project completion date. The project completion date most often is part of a contract that carries penalties for nonperformance. Therefore, the complexities and the one-time nature of the project require a coordinated plan that involves activities required, scheduling, and deployment of resources. The great complexity of such projects calls for special methods; network planning techniques have been developed to meet this need.

Network planning techniques go under a confusion of acronyms with variations. The two original names, PERT (Performance Evaluation and Review Technique) and CPM (Critical Path Methods), have been differentiated into a variety of brand names that essentially have been applied to the same basic methodology. Some of the alternative names used are: CPS (Critical Path Scheduling), LES (Least-cost Estimating and Scheduling), Micro-PERT, 1-time-PERT, PERT/COST, and PEP. The variety of names for PERT/CPM techniques at least are a measure of the degree of interest that has developed.

ORIGIN OF NETWORK PLANNING

Network planning methods seem to have been developed by two different groups independently. As an internal project of the DuPont Company, critical path methods were developed to plan and control the maintenance of chemical plants; subsequently they were used widely by DuPont for many engineering functions. Parallel efforts were undertaken by the U.S. Navy at about the same time to develop methods for planning and controlling the Polaris missile project. We can glimpse the magnitude of the task when we realize that approximately 3,000 separate contracting organizations were involved. The result was the developxent of the PERT methodology. The immediate success of both the CPM and PERT methodologies may be gauged by the following facts. DuPont's application of its technique to a Louisville maintenance project resulted in reducing down time for maintenance from 125 to 78 hours; the PERT technique was credited widely with helping to shorten by 2 years the time originally estimated for the completion of the Polaris missile engineering and development program.

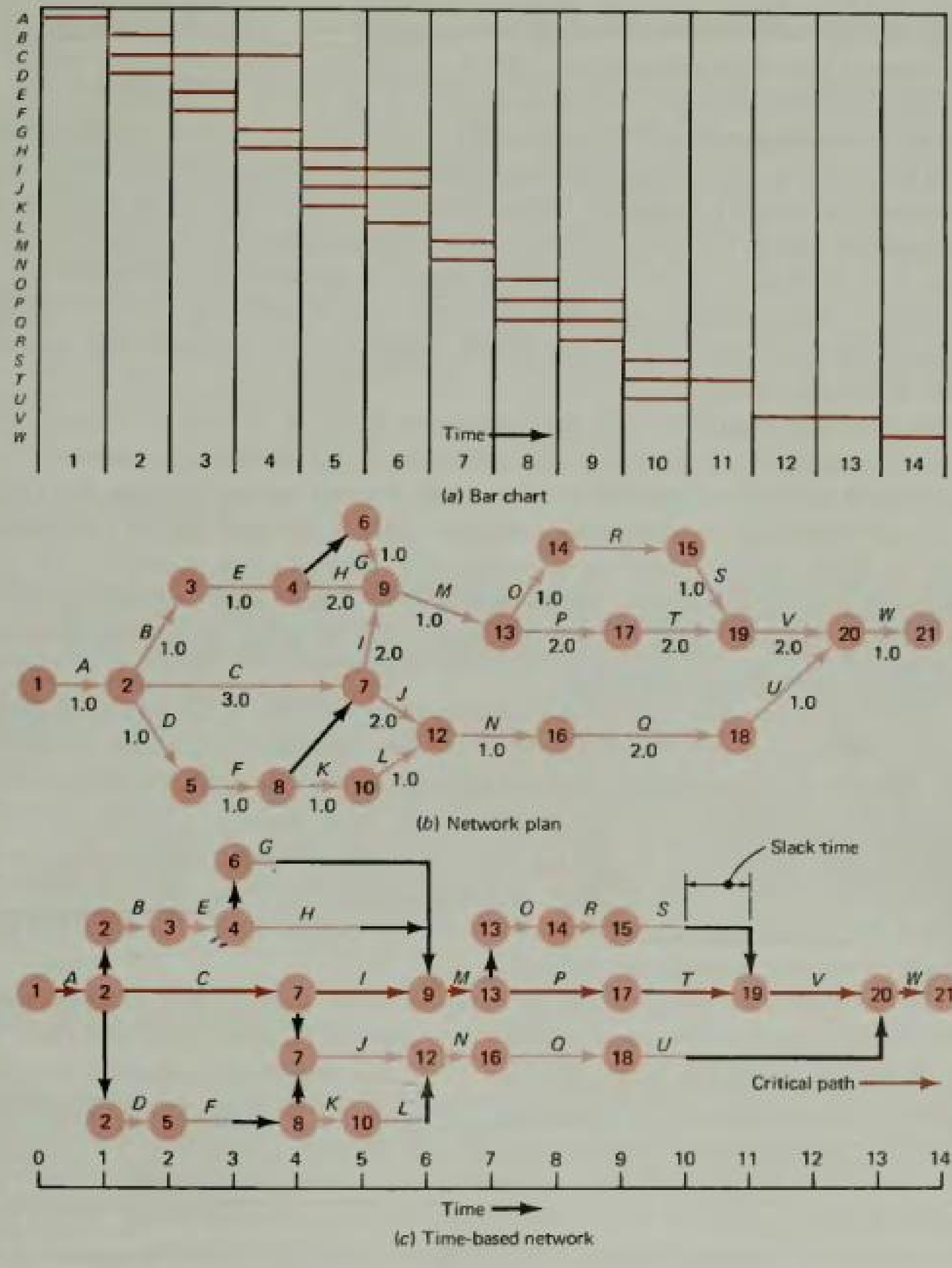
PERT and CPM are based on substantially the same concepts, although there are some differences in details. First, as originally developed, the PERT methods were based on probabilistic estimates of activity times that resulted in a probabilistic path through a network of activities (such as that shown in Figure 1) and a probabilistic project completion time. The CPM methods, however, assumed constant or deterministic activity times. Actually, either the probabilistic or deterministic model is equally applicable to and usable by either technique. As a matter of fact, most present day applications of PERT methods have dropped the use of the probabilistic activity times, and they now use the slightly simpler deterministic time estimates. The second difference between the two techniques is in the details of how the arrow diagram is prepared. In the discussion that follows, we shall note more fully both the probabilistic-deterministic and the arrow diagram differences.

In the following sections, we shall develop PERT planning methods and then show the differences between PERT and CPM methods.

PERT PLANNING METHODS

The essence of PERT planning is based on the development of a network representation of the required activities, as indicated by Figure 1*b*. Here, the arrows represent the required activities coded by the letters, with estimated performance times shown near the arrows. In network planning, the length of the arrows ordinarily has no significance (for reasons that we shall comment on later). The numbered circles define the beginning and end points of activities and are called *events* or *nodes*. The direction of arrows indicate flow in the sense that node 2 marks the end of activity *A* and the beginning of activities *B*, *C*, and *D*; node 3 marks the end of activity *B*. The network also represents the required precedence relationships of all activities. For example, activities *B*, *C*, and *D* cannot start until activity *A* has been completed, but activities *B*, *C*, and *D* all can proceed simultaneously.

Figures 1*a*, *b*, and *c* all represent the same plan. Although it undoubtedly is easier to visualize the fixed time relations in the bar chart of Figure 1*a*, the bar chart does not contain all the crucial information about precedence requirements. It does show the earliest time that each activity can begin and end, but it



(a) Bar chart

(b) Network plan

(c) Time-based network

Figure 1. Comparison of (a) bar chart, (b) network, and (c) time-based network plans. The heavy colored line indicates the critical path; the heavy black lines indicate slack time

SOURCE: R. D. Archibald and R. L. Villoria, *Network-Based Management Systems*, New York: John Wiley & Sons, 1967).

does not show the latest time that each activity can begin in order that its completion time does not interfere with the beginning of following activities.

Note, however, that the "time-based network" in Figure 1c shows this information. Here, the length of the horizontal colored lines represents the estimated activity times, and the black lines represent slack in the schedule. Note, for example, that activity S requires 1 time unit and can begin as early as time $t = 9$, but need not begin before $t = 10$ in order to be completed by its latest finish-time, $t = 11$. If activity S is completed any time between $t = 10$ and $t = 11$, activity V, which follows, can begin on time. Note also that activity V, which depends on the completion of both S and T by $t = 11$ for its start time, has no slack in its schedule.

We see from Figure 1c that only activities G, F, H, S, and U have slack. However, some other activities can use some or all of the slack shown. For example, if activity J is delayed by 1 time unit, it is not serious because N, Q and U all can be set back by 1 time unit, using up 1 unit of the slack shown associated with U. If any of them uses up as much as 3 units of slack, however, all the following activities become *critical*; that is, they must begin and end by their latest event times or else the entire project will be held up and the project completion date will not be met.

Now let us focus attention on the sequence of activities A–C–I–M–P–T–V–W in the time-based network of Figure 1c. They are connected by the heavy colored line, labeled the *critical path*. There is no slack in the schedules of any of these activities, nor can any of them borrow slack from another activity (as was possible with activities J, N, and Q). This set of activities defines a critical path through the network. Each of these critical path activities must begin and end on time if the project completion time of $t = 14$ is to be met. The critical path is the longest time path through the network.

The network plan summarizes, in compact form, a great deal of important information: the activities required, their precedence relationships, and slack in the schedule. From the basic network plan, we easily can calculate crucial information regarding: the earliest and latest start- and finish-times; available slack in the schedules of activities; and the critical path. Ordinarily, the network plan similar to Figure 1b is used rather than the time-based network plan. The reason is simply that the entire system normally is computer based. Given the activities and their precedence relationships, standard computer programs will provide all the schedule information for each activity, indicating which activities are on the critical path. Thus, as we noted before, the length of arrows in the network plan need not be significant since the network itself is only an input to the computation of the other important schedule data. Of course, the time-based network can be constructed from the data given by the computer output if we feel that it would be valuable to visualize the schedule through graphic means.

Let us note again that the interdependent nature of the set of activities, the importance of a project completion date, and the one-time nature of projects require that the planning for what must be done and the schedule considerations be intimately bound up together. The equally important related plan for the use of manpower and other resources will be discussed later in this chapter.

Keeping the generalities of PERT methods in mind, let us utilize a relatively simple example, house construction, to develop the methods used in generating the network representation of a project. The phases of development may be divided into (1) activity analysis; (2) arrow diagramming; and (3) node numbering.

Activity Analysis

This is comparable to what a production engineer or production planner does when he/she is specifying operations, work methods, and tooling for fabricated parts and products. For large projects, however, a degree of such complexity is imposed by the extremely large number of components and required activities that it is possible to overlook the need for some activities. Therefore, while professional planning personnel commonly are used, the activity list is often partially generated in meetings and round table discussions that include managerial and operating personnel. The result of the entire process is a list of required activities, such as the one shown in Table 1, showing the basic activities that are required to construct a house.

Arrow Diagramming

This requires a consideration of the precedence relationships among activities and must be based on a complete, verified, and approved activity list. The important information required for the arrow diagram is generated by the following three questions:

1. Which activities must be completed *before* each given activity can be started?
2. Which activities can be carried out in *parallel*?
3. Which activities immediately *follow* other given activities?

The common practice is simply to work backwards through the activity list.

TABLE 1
PRECEDENCE CHART SHOWING SEQUENCE OF ACTIVITIES AND REQUIRED TIMES TO FINISH A HOUSE

Job No.	*Description*	*Immediate Predecessors*	*Time (days)*
a	Start		0
b	Excavate and pour footings	*a*	4
c	Pour concrete foundation	b	2
d	Erect wooden frame including rough roof	c	4
e	Lay brickwork	d	6
f	Install basement drains and plumbing	c	1
g	Pour basement floor	*f*	2
h	Install rough plumbing	*f*	3
i	Install rough wiring	d	2
j	Install heating and ventilating	d,g	4
k	Fasten plaster board and plaster (including drying)	*i,j,h*	10
l	Lay finishing flooring	*k*	3
m	Install kitchen fixtures	*l*	1
n	Install finish plumbing	*l*	2
o	Finish carpentry	*l*	3
p	Finish roofing and flashing	e	2
q	Fasten gutters and downspouts	p	1
r	Lay storm drains for rain water	c	1
s	Sand and varnish flooring	o,t	2
t	Paint	m,n	3
u	Finish electrical work	*t*	1
v	Finish grading	q,r	2
w	Pour walks and complete landscaping	v	5
x	Finish	s,u,w	0

SOURCE: F. K. Levy, G. L. Thompson, and J. D. Wiest, "The ABC's of the Critical Path Method," *Harvard Business Review*, September–October 1963.

generating the immediate predecessors for each activity listed, as shown in Table 1 for the house construction project. The estimated normal time for each activity also is shown in the table (although this information is not necessary at this point). The arrow diagram then may be constructed to represent the logical precedence requirements shown in Table 1.

Dummy Activities. Care must be taken to represent correctly the actual precedence requirements in the arrow diagram. For example, look at the immediate predecessor activities for activity s, sand and varnish flooring, and activity u, finish electrical work. Activity s has as immediate predecessors o and t, finish carpentry and painting, respectively, while u has a predecessor of only activity t. The relationship shown in Figure 2*a* does not correctly represent this situation because it specifies that the beginning of u is dependent on both o and t, and this

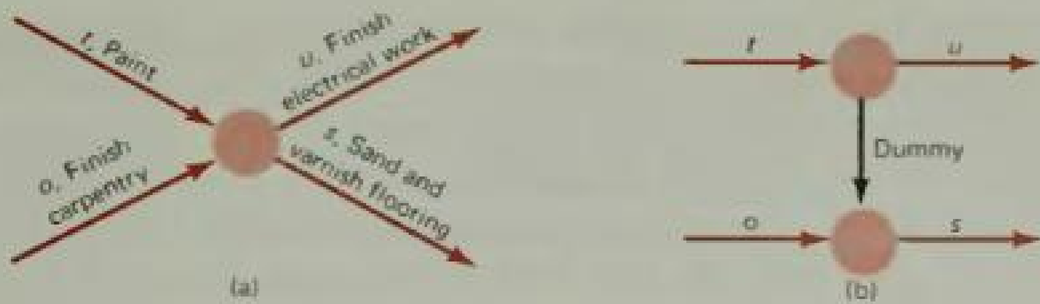


Figure 2. (a) Diagram does not properly reflect precedence requirements since u seems to be dependent on the completion of both o and t but actually depends only on t. (b) Creating two nodes with dummy activity between provides the proper predecessors for both activities s and u

is not true. To represent the situation correctly, we must resort to the use of a dummy activity that requires zero performance time. Figure 2b now represents the stated requirement. Finish electrical work, u, now depends only on the completion of painting, t. Through the dummy activity, however, both finish carpentry and painting must be completed before activity s, sand and varnish flooring, can be started. The dummy activity provides the logical sequencing relationship; but since it is assigned zero performance time, it does not alter any scheduling relationships that will be developed later.

Another use of the dummy activity is to provide a specific and separate beginning and ending event or node for each activity that cannot be confused. In Table 1, for example, note the relationship between activities *l*, *m* and *n*, and *t*. The activity *l*, lay finish flooring, must precede both *m* and *n*, install kitchen fixtures and install finish plumbing, respectively. But *m* and *n* are the predecessors of *t*, paint. Therefore, the functionally correct relationship is represented by Figure 3*a*. But if this figure were used, we could not identify each activity by its predecessor and successor events because both activities *m* and *n* would begin

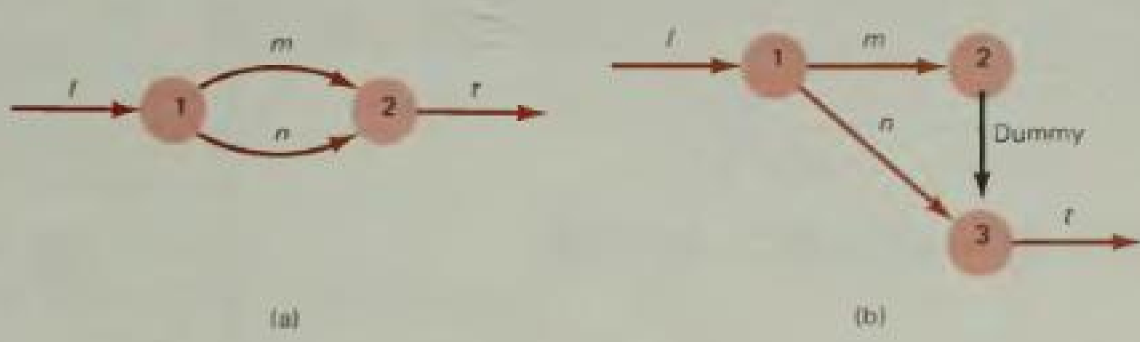


Figure 3. (a) Activities *m* and *n* may be carried out in parallel but result in identical beginning and end events. (b) Use of dummy activity makes possible separate ending event numbers, thus making activity identification by pairs of event or node numbers unambiguous

and end with the same node numbers. This is particularly important in larger networks employing computer programs for network diagram generation. The computer always is programmed to identify each activity by a pair of event numbers. The problem is solved through the insertion of a dummy activity, as shown in Figure 3b. The functional relationship is identical since the dummy activity requires zero time, but now both *m* and *n* are identified by different combinations of node numbers.

Figure 4 shows the completed arrow diagram for the house construction project. Activities are identified with their required times in days, all the nodes are numbered, and the critical path is shown. Of course, the activity times have not been used up to this point, nor were they necessary for the construction of the diagram; we have not yet shown how to determine the critical path. However, the activity times will have great significance in the generation of schedule data, the determination of the critical path, and the generation of alternatives for the deployment of resources.

Node Numbering

The *node numbering* shown in Figure 4 has been done in a particular way. If we identify each activity by its tail (*i*) and head (*j*) numbers, we have numbered the nodes so that, for each activity, *i* always is less than *j*, $i < j$. The numbers for every arrow are progressive, and no backtracking through the network is allowed. This convention in node numbering is effective in computing programs to develop the logical network relationships and to prevent the occurrence of cycling or closed loops.

A closed loop would occur if an activity were represented as going backwards

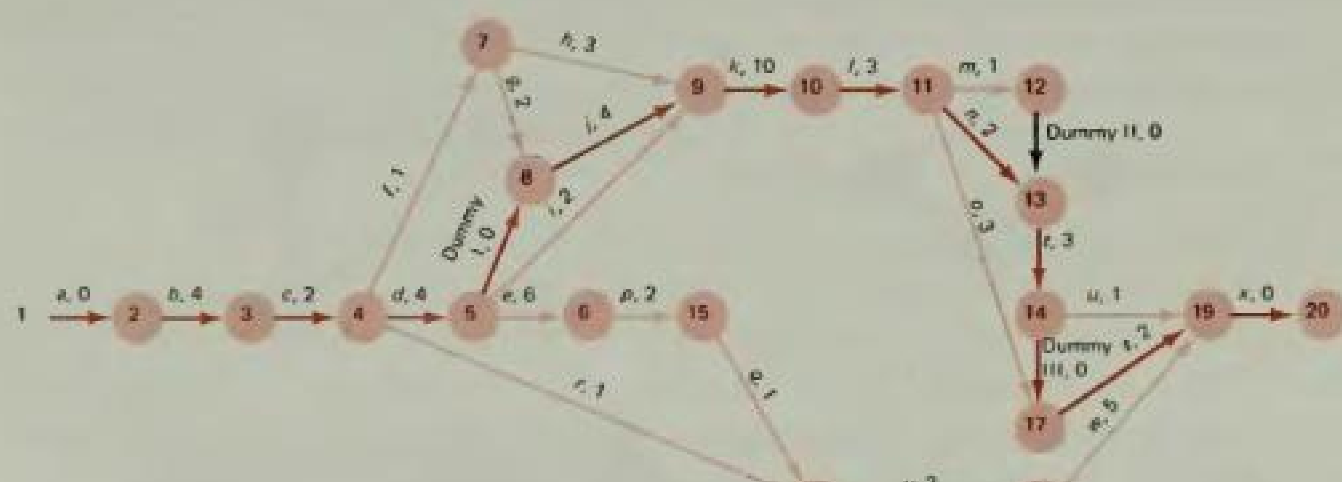


Figure 4. PERT network diagram for the house construction project

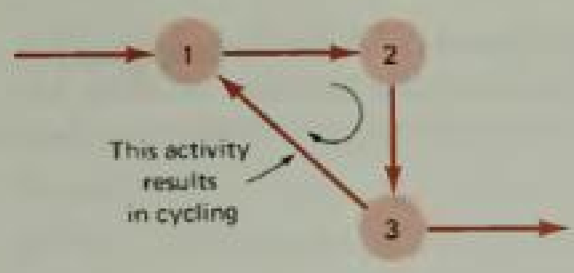


Figure 5. Example of a closed loop or cycling in a network diagram

in time. This is shown in Figure 5, which is simply the structure of Figure 3b with the direction of activity n reversed. Cycling in a network can be caused by a simple error or, in developing the activity plans, by trying to show the repetition of an activity before beginning the next activity. A repetition of an activity must be represented by additional separate activities that are defined by their own unique node numbers. A closed loop would produce an endless cycle in a computer program if it did not have a built-in routine for detection and identification of the cycle. Thus, one property of a correctly constructed network diagram is that it is *noncyclical*.

no loop

CRITICAL PATH SCHEDULING

With a properly constructed arrow diagram, it is a simple matter to develop the important schedule data for each activity, and for the project as a whole. The data that interest us are the earliest and latest start- and finish-times, available slack for all activities, and the critical path through the network.

Earliest Start- and Finish-Times

If we take zero as the starting time for the project, then for each activity there is an earliest starting-time (*ES*) relative to the project starting-time, which is the earliest possible time that the activity can begin, assuming that all the predecessors also are started at their *ES*. Then, for that activity, its earliest finish (*EF*) is simply *ES* + activity time.

Latest Start- and Finish-Times

Now let us assume that we have a target time for completing the project, which, for the ~~house construction example~~, is three days after the *EF* time, or thirty-seven days. This is called the latest finish-time (*LF*) of the project and of the final activity x. The latest start-time (*LS*) is the latest time at which an activity can start if the target or schedule is to be maintained. Thus, *LS* for the final activity x is *LF* − activity-time. Since the finish activity requires zero time units, *LS* = *LF*.

Critical Path and Slack

Existing computer programs may be used to compute these schedule data automatically, requiring as inputs the activities, their performance time requirements, and the precedence relationships established by the *ij* numbers of the tails and heads of arrows. The computer output might be similar to Figure 6, which shows the schedule statistics for all activities when three days of slack have been allowed in the overall project completion time. Note that all critical activities, marked with an asterisk (*), have slack in their schedules of three days. All other activities have greater schedule slack. Dummy activities I and III are on the critical path, but they only establish correct precedence relationships, since their activity times are zero. The schedule slack is simply the difference between computed early and late start-times (*LS* − *ES*), or between early and late finish-times (*LF* − *EF*).

Actually, there are twenty-two different paths from start to finish through the network. The shortest path requires fourteen days by the sequence

Critical Path	Sequence I	Sequence J	Activity Description	Activity Time (days)	Start Early	Start Late	Finish Early	Finish Late	Slack
*	1	2	a, Start	0	0	3	0	3	3
*	2	3	b, Excavate and pour footings	4	0	3	4	7	3
*	3	4	c, Pour concrete foundation	2	4	7	6	9	3
*	4	5	d, Erect frame including rough roof	4	6	9	10	13	3
	5	6	e, Lay brickwork	6	10	21	16	27	11
	4	7	f, Install basement drains and plumbing	1	6	10	7	11	4
	7	8	g, Pour basement floor	2	7	11	9	13	4
	7	9	h, Install rough plumbing	3	7	14	10	17	7
	5	9	i, Install rough wiring	2	10	15	12	17	5
*	5	8	Dummy I	0	10	13	10	13	3
*	8	9	j, Install heating and ventilating	4	10	13	14	17	3
*	9	10	k, Fasten plaster board and plaster, including drying	10	14	17	24	27	3
*	10	11	l, Lay finish flooring	3	24	27	27	30	3
	11	12	m, Install kitchen fixtures	1	27	31	28	32	4
	12	13	Dummy II	0	28	32	28	32	4
*	11	13	n, Install finish plumbing	2	27	30	29	32	3
	11	17	o, Finish carpentry	3	27	32	30	35	5
	6	15	p, Finish roofing and flashing	2	16	27	18	29	11
	15	16	q, Fasten gutters and downspouts	1	18	29	19	30	11
	4	16	r, Lay storm drains for rain water	1	6	29	7	30	23
*	14	17	Dummy III	0	32	35	32	35	3
*	17	19	s, Sand and varnish flooring	2	32	35	34	37	3
*	13	14	t, Paint	3	29	32	32	35	3
	14	19	u, Finish electrical work	1	32	36	33	37	4
	16	18	v, Finish grading	2	19	30	21	32	11
	18	19	w, Pour walks and complete landscaping	5	21	32	26	37	11
*	19	20	x, Finish	0	34	37	34	37	3

Figure 6. Sample computer output of schedule statistics and critical path for the house construction project. Slack in project completion and for all critical activities is three days

a-b-c-r-v-w-x, and the longest, or limiting, path requires thirty-four days by the critical sequence a-b-c-d-j-k-l-n-t-s-x. In a small problem such as this one, we could enumerate all the alternative paths to find the longest path; however, there is no advantage in doing so, since the critical path is determined easily from the schedule statistics, which themselves are useful.

Manual Computation of Schedule Statistics

Manual computation is appropriate for smaller networks and helps to convey the significance of the schedule statistics. To compute ES and EF manually from the network, we proceed as follows (see Figure 7):

1. Place the value of the project start-time in both the ES and EF positions near the start activity arrow (see the legend for Figure 7). We shall assume relative values, as we did in the computer output of Figure 6, so the number 0 is placed in the ES and EF positions for the start activity. (PERT does not require that we include the start activity with zero activity duration. We have included it to make the activity list of this example parallel with the comparable CPM example of Figure 9. The start and finish activities often are necessary in CPM.)

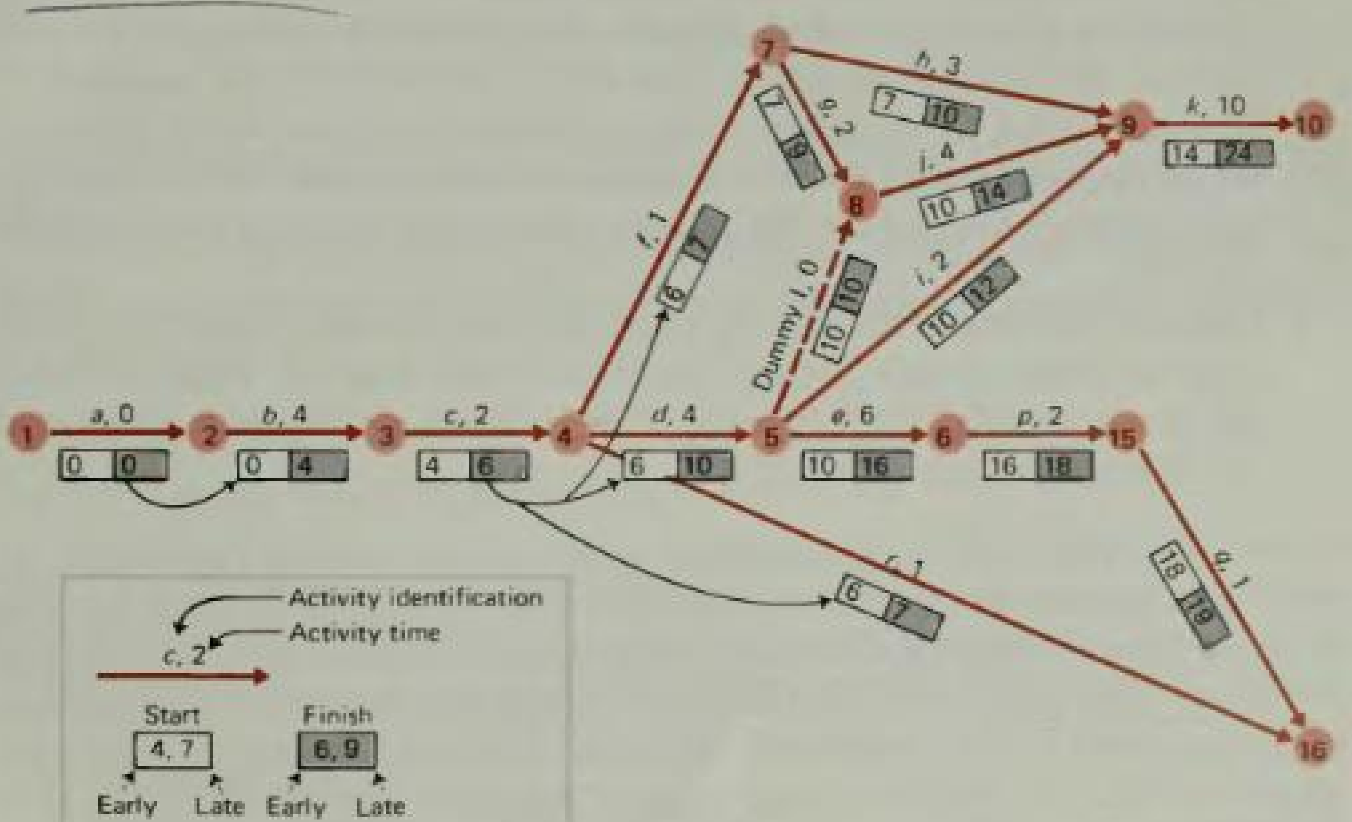


Figure 7. Flow of calculations for early start (ES) and early finish (EF) times

2. Consider any new unmarked activity, *all of whose predecessors have been marked* in their *ES* and *EF* positions, and mark in the *ES* position of the new activity the *largest* number marked in the *EF* poistion of any of its immediate predecessors. This number is the *ES* time of the new activity. For activity *b* in Figure 7, the *ES* time is 0, since that is the *EF* time of the preceding activity.
3. Add to this number the activity time and mark the resulting *EF* time in its proper position. For activity *b*, $ES + 4 = 4$.
4. Continue through the entire network until the "finish" activity has been reached. As we showed in Figure 6, the critical path time is thirty-four days, so $ES = EF = 34$ for the finish activity.

To compute the *LS* and *LF*, we work backwards through the network, beginning with the finish activity. We already have stated that the target time for completing the project is 3 days after the *EF* time, or 37 days. Therefore, $LF = 37$ for the finish activity, without delaying the total project beyond its target date. Similarly, the *LS* time for the finish activity is *LF* minus activity time. Since the finish activity requires 0 time units, $LS = LF$. To compute *LS* and *LF* for each activity, we proceed as follows (see Figure 8):

1. Mark the value of *LS* and *LF* in their respective positions near the finish activity, according to the information in Figure 7.
2. Consider any new unmarked activity, all of whose successors have been marked, and mark in the *LF* position for the new activity the smallest *LS* time marked for any of its immediate successors. In other words, *LF* for an activity equals the earliest *LS* of the immediate successors for that activity.
3. Subtract from this number the activity time, which becomes the *LS* for the activity.
4. Continue backwards through the chart until all *LS* and *LF* times have been entered in their proper positions on the network diagram. Figure 8 shows the flow of calculations, beginning with the finish activity and going backwards through several activities.

As we discussed before, the schedule slack for an activity represents the maximum amount of time that it can be delayed beyond its *ES* without delaying the project completion time. Since critical activities are those in the sequence of the longest time path, it follows that the activities will have the minimum possible slack. If the project target date coincides with the *LF* for the finish activity, all critical activities will have zero slack. If, however, the project date is later than the *EF* of the finish activity, as it is in the house project example (three days), all critical activities will have slack equal to this time-phasing differ-

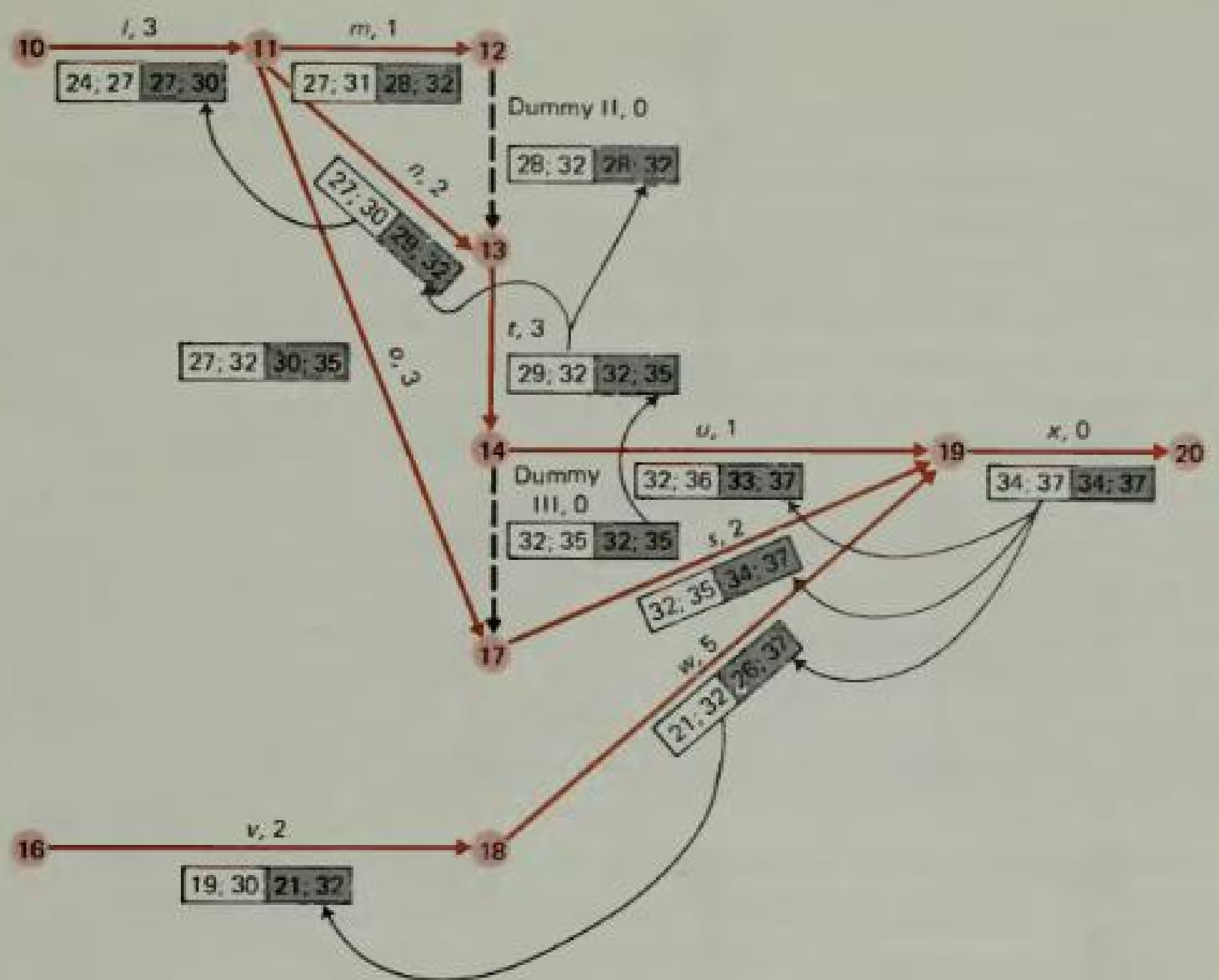


Figure 8. Flow of calculations for late start (LS) and late finish (LS) times

ence. The manual computation of slack is simply $LS - ES$, or, alternatively, $LF - EF$.

The preceding discussion of the computation of schedule statistics is an application of the longest path algorithm in network optimization. See Buffa and Dyer [1977].

Pert/CPM—Arrow Diagram Differences

Up until now, we have used the PERT arrow diagramming procedure to illustrate network methods. The CPM procedure creates a slightly simpler network system by representing activities as occurring at the nodes, with the arrows showing the sequences of activities required for the project. The advantage of the CPM methodology is that we do not have to invoke the use of dummy activities in order to represent the proper sequencing.

Figure 9 shows the CPM network for the house construction project, which

Figure 9. CPM project graph of activities for a house-building project

SOURCE: F. K. Levy, G. L. Thompson, and J. D. Wiest, "The ABCs of the Critical Path Method," *Harvard Business Review*, September-October 1963.

may be compared with the PERT network shown in Figure 4. The analysis by which the early and late start- and finish-times and slack times are developed is identical with the procedure that we outlined before. The net results of both systems are the schedule statistics, which are computed. Since these are the relevant data, and since the entire procedure normally is computerized for both methodologies, the choice between the two may depend on other criteria (such as the availability and adaptability of existing computer routines).

PROBABILISTIC NETWORK METHODS

The network methods that we have discussed so far may be termed *deterministic*, since estimated activity times are assumed to be the expected values. But deterministic methods do not recognize the fact that the mean or expected activity time is the mean of a distribution of possible values that could occur. Rather, these methods assume that the expected time is actually the time taken.

Probabilistic network methods assume the reverse, more realistic, situation, in which activity times are presented by a probability distribution. With such a basic model of the network of activities it is possible to develop additional data that are important to managerial decisions. Such data help managers assess planning decisions that might revolve around such questions as: What is the probability that the completion of activity *A* will be later than January 10? What is the probability that the activity will become critical and affect the project completion date? What is the probability of meeting a given target completion date for the project? What is the risk of incurring cost penalties for not meeting the contract date?

The nature of the planning decisions based on such questions might involve the allocation or reallocation of manpower or other resources to the various activities in order to derive a more satisfactory plan. Thus, a "crash" schedule with extra resources might be justified to ensure the on-time completion of certain activities. The extra resources needed are drawn from noncritical activities or activities in which the critical probability is small.

The discussion that follows can be applied equally to either the PERT or CPM basic format, although the probabilistic methods originally were developed as part of PERT. The probability distribution of activity times is based on three time estimates for each activity.

Optimistic Time

Optimistic time, a, is the shortest possible time in which to complete the activity if all goes well. It is based on the assumption that there is no more than one chance in a hundred of completing the activity in less than the optimistic time.

Pessimistic Time

Pessimistic time, b, is the longest time in which to complete an activity under adverse conditions but barring acts of nature. It is based on the assumption that there is no more than one chance in a hundred of completing the activity in a time greater than b.

Most Likely Time

Most likely time, m, is the modal value of the activity-time distribution.

The three time estimates are shown in relation to an activity completion time distribution in Figure 10. The computational algorithm reduces these three time estimates to a single average or expected value, t_e, which actually is used in the computing procedure.* The expected value is also the one used in computing schedule statistics for the deterministic model. The example distribution in Figure 10 represents only one possibility. Actually, the time distributions could be symmetrical or skewed either to the right or the left.

With a probabilistic model, we can see the probability that seemingly noncritical activities may become critical. This could happen either if a long performance time occurs for the activity in question, or if short performance times occur for activities that already are on the critical path. This is a signal that the schedule plans that have been developed are likely to change. As actual data on the progress of operations come in, we may have to make changes in the allocation of resources in order to cope with the latest set of critical activities.

DEPLOYMENT OF RESOURCES

Given the activity network, the critical path, and the computed schedule statistics, we have a plan for the project. But is it a good plan? We can abstract from our data some additional data on the demand for resources of the early-start schedule. By using the schedule flexibility available through slack in certain activities and/or slack in the project completion date, we can generate alternative schedules and compare the use of important resources with the objective of *load leveling*.

---

*The usual model assumes that t_e is the mean of a beta distribution. The estimates of the mean and variance of the distribution may be computed as follows:

$$\bar{x} = \tfrac{1}{6}[A + 4M + B]$$
$$s^2 = [\tfrac{1}{6}(B - A)]^2$$

where A, B, and M are estimates of the values of a, b, and m, respectively, and $\bar{x}$ and s^2 are estimates of the mean and variance, t_e and σ_t^2.

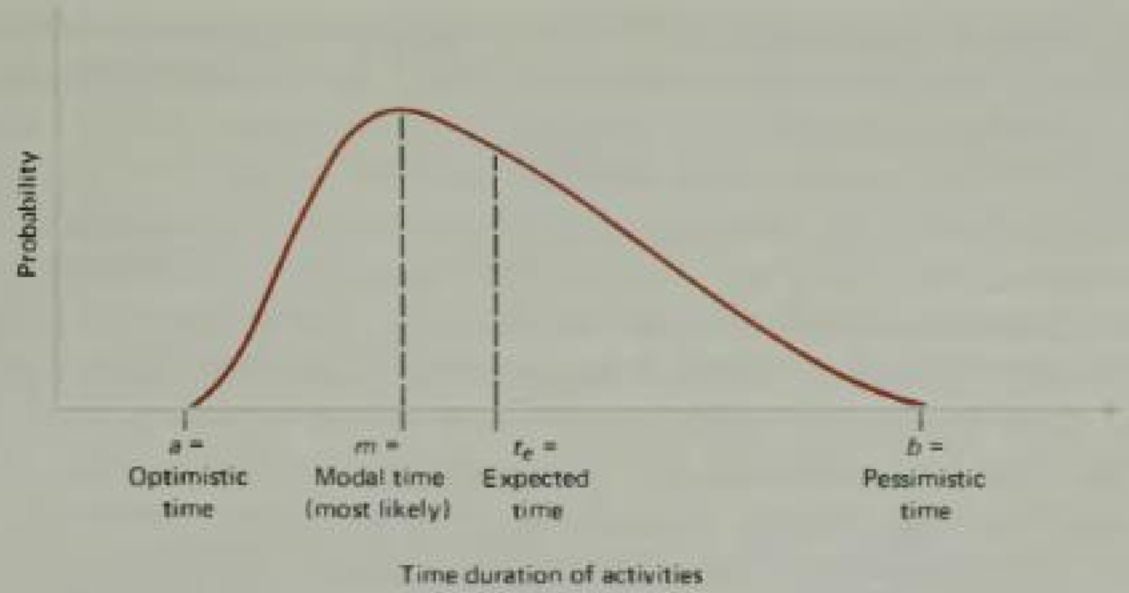


Figure 10. Time values in relation to a distribution of activity time

Another way to look at the initial or raw plan is in terms of activity costs. The initial activity duration estimates are based on an assumed level of resource allocation. Is it possible to alter activity times by pouring in more or fewer resources? Activity times for some activities can be affected directly in this way. For example, adding carpenters usually will shorten the time it takes to frame a house (this was a critical activity in our house construction example). Would it be worthwhile to pour in more manpower on the critical framing and allocate less to the noncritical brickwork, which has eleven days of slack? Would the alternative plan be more or less expensive? Would shortening the critical path be advantageous? *Least costing* considerations are worth examining.

Finally, in some situations we may be faced with a demand for some critical resource that is limited in supply. The raw plan may not be feasible if it implies the use of the only available power shovel in two places at the same time. The raw plan must be examined with the objective of the feasible scheduling of *limited resources*, again using available slack time where possible, or even lengthening the project in order to generate a feasible plan.

Load Leveling

Why level loading? What are the costs of not attempting to level loads in an already feasible schedule? Some factors that enter the problem occur in the following example of a major oil refinery repair and overhaul project [Archibald and Villoria, 1967, pp. 269–280]. After a raw plan was developed, a series of computer runs was made to examine manpower requirements for the refinery project. In the first run it was found that the schedule required 50 boilermakers for the first 4 hours, 20 for the next 6 hours, and 35 for the period immediately following. Similar fluctuations in requirements were found for other crafts. The possible cost of idle labor is associated with this fluctuation.

For example, in the first 10 hours of the refinery project, the peak requirement of 50 boilermakers probably will mean productive work of 50×4 hours $+ 20 \times 6$ hours $= 320$ work-hours. But the likelihood is that it will be difficult to assign the extra 30 workers for the balance of the 8 hour day; thus, in the first 10 hours of the project, the payroll may reflect $50 \times 8 + 20 \times 2 = 440$ work-hours, 120 of which are idle labor. Figure 11 shows the deployment of manpower *after* leveling; some of these same kinds of problems remain. Other costs that may be implicit in manpower fluctuation are hiring and separation costs in projects that extend over long periods. The objective of load leveling is to reduce idle labor costs, hiring and separation costs, or the cost of any resource, such as an equipment rental, which may be affected by fluctuations in the demand for its use.

For very large and complex projects, a computer based leveling model may be required. Simulation methodology commonly is used to generate alternative solutions. The starting solution might be the early start schedule, and a first

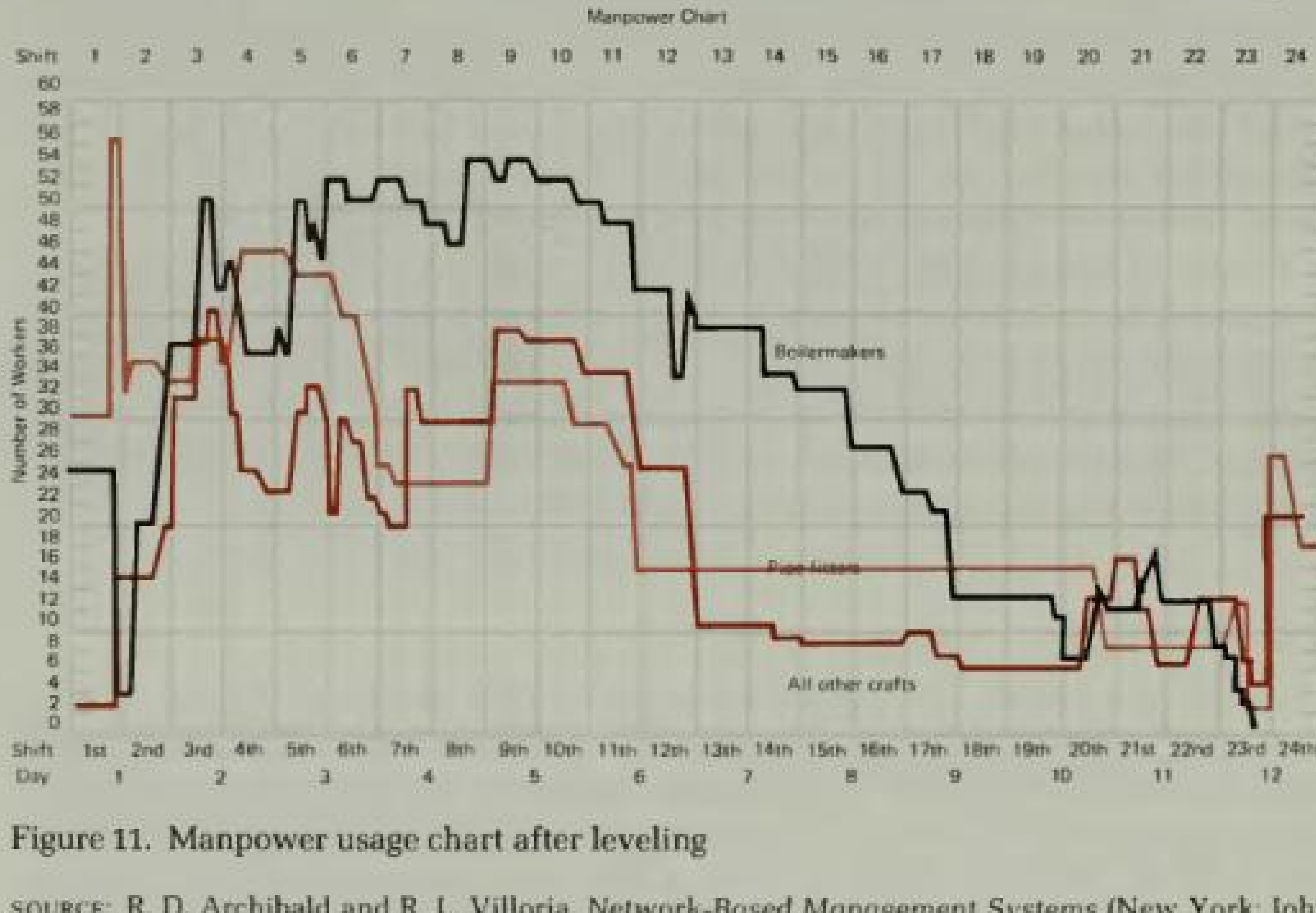


Figure 11. Manpower usage chart after leveling

SOURCE: R. D. Archibald and R. L. Villoria, *Network-Based Management Systems* (New York: John Wiley & Sons, 1967), p. 274.

attempt at leveling then could set a maximum of the resource in question just below the highest peak level that was recorded in the raw plan. The simulation program would proceed as indicated by the arrow diagram, beginning with all activities leaving node 1, and keeping track of the amount of resources used and available. As the calendar is advanced and as activities are completed, resources would be returned to the "available" pool; as new activities are started, resources would be drawn from the pool. Simulation would proceed until an activity required resources from a temporarily exhausted pool. Depending on the decision criteria used by the simulator, the activity might be delayed, even past its latest starting-time, until resources were available. Other decision criteria would "bump" noncritical jobs and reassign resources to the delayed job when the latest starting-time was reached. By a progressive lowering of the resource limits in such a simulation program, the leveling effect would take place until a satisfactory deployment of resources was achieved. A survey of rigorous methods of resource allocation in project network models has been developed by Davis [1974]; specific resource leveling models have been developed by Dewitte [1964], by Levy, Thompson, and Wiest [1963], and by Woodworth and Willie]1975].

Least Costing

Least costing concepts are based on cost versus activity time curves such as those in Figure 12. Different activities respond differently to changes in the application of resources, and some of the activities may not be responsive to changes in resources. Figure 12*a* may be typical of an activity like house framing, in which crash, normal, and slow schedules are progressively less costly. A

Figure 12. Typical activity time–cost curves

curve similar to Figure 12*b*, in which the slow schedule costs more than the normal schedule, could be typical in a situation where the meager resources associated with a slow schedule enforced the use of inefficient methods. The cost trade-offs are possible partially because of the differential cost–time characteristics of different activities.

Formal methodologies for least costing have been developed by Fulkerson [1961] and Kelly [1961] in the form of linear programming models that idealize the cost–time functions by connecting the crash and normal points with a straight line. The total project cost is taken as the sum of the linear activity cost functions. This total project cost function is minimized by the linear programming algorithm.

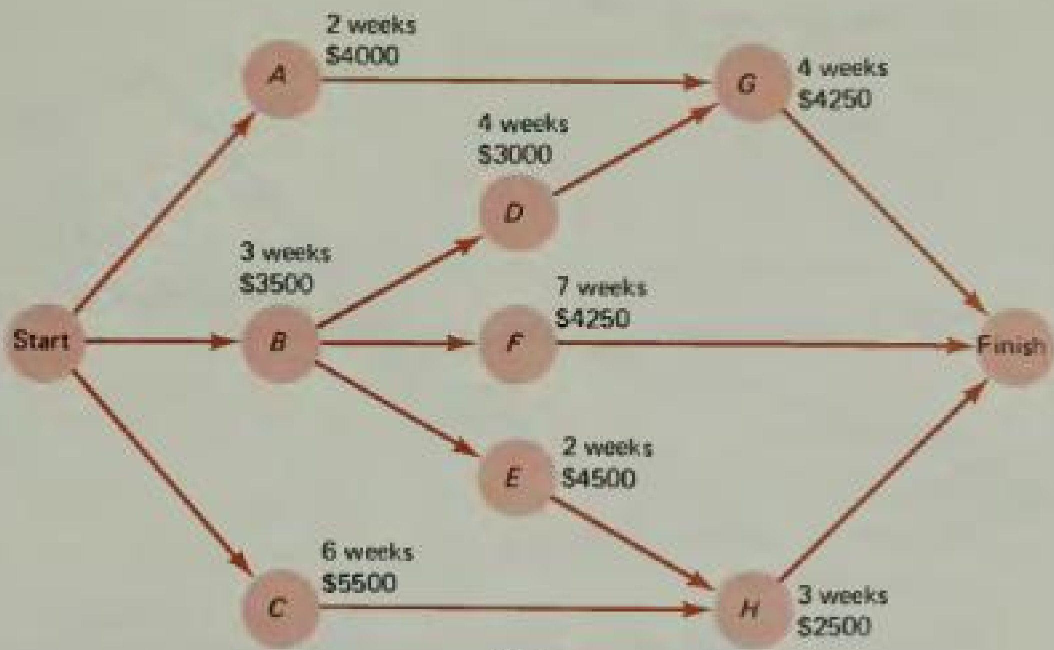
Computing programs are available that take as input activity cost and the incremental cost for time reductions.

An Example. Figure 13*a* is a CPM network diagram for a construction contract that shows the time in weeks to complete each activity and the normal schedule estimated cost of each activity. Figure 13*b* shows the computer output and indicates an overall project time of 11 weeks, the critical path, and the scheduling statistics for each activity. The total contractor's cost is $32,500, but the contract price is $45,000.

The contractor's problem is that his costs are based on a normal completion time of 11 weeks, but the customer insists on a 10 week time and a penalty for late performance of $2,000 per week. The contractor likes the agreed price, since it provides him with a $12,500 profit, and so it is worth his time to discover the circumstances under which he might be able to meet the 10 week schedule.

The contractor examines the various activities to see which ones could be speeded up, and at what cost. Using a LEAST COSTING program [Buckley et at., 1974], he enters data concerning the cost distributions of the activities for which he has obtained "crash" time and cost estimates, as shown in Figure 14*a*. For example, activity *A* costs $4,000 by the normal schedule, but 1 week can be saved at an incremental cost of $1,500. Activity *G* gives estimates for 3 points on its time-cost curve: the normal cost is $4,250, but 1 week can be saved at an incremental cost of $1,500, and a second week can be saved at a cost of $2,000. Activities for which no crash schedule costs have been estimated have an entry of 0. Finally, the program requests the target completion time.

Given the time–cost distribution data input, and the 10 week target, the program produces a new critical path schedule, at minimum cost, which meets the 10 week completion time, as shown in Figure 14*b*. This schedule indicates the new schedule statistics, as well as a cost comparison at the bottom. The optimized cost is $1,500 greater than the normal schedule cost, and activity *G* is



(a)

THE CRITICAL PATH IS

START → B → D → G → FINISH

THE LENGTH OF THE CRITICAL PATH IS 11

NODE	*DURATION*	*EARLY START*	*EARLY FINISH*	*LATE START*	*LATE FINISH*	*TOTAL SLACK*	*FREE SLACK*
START	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A	2.00	0.00	2.00	5.00	7.00	5.00	5.00
B	3.00	0.00	3.00	0.00	3.00	0.00	0.00
C	6.00	0.00	6.00	2.00	8.00	2.00	0.00
D	4.00	3.00	7.00	3.00	7.00	0.00	0.00
E	2.00	3.00	5.00	6.00	8.00	3.00	1.00
F	7.00	3.00	10.00	4.00	11.00	1.00	1.00
G	4.00	7.00	11.00	7.00	11.00	0.00	0.00
H	3.00	6.00	9.00	8.00	11.00	2.00	2.00
FINISH	0.00	11.00	11.00	11.00	11.00	0.00	0.00

(b)

Figure 13. Construction contract project: (a) CPM network diagram showing time in weeks for the completion of each activity, and normal schedule activity costs, and (b) the computer output indicating the critical path, the overall project-time of 11 weeks, and the schedule statistics

shortened by one week. Activity *G* was on the original critical path, but it is not on the crash program critical path. The critical path ~~has changed from~~ *B–D–G* to *B–E–H*.

The contractor can avoid a \$2,000 penalty by spending \$1,500 to speed up activity *G*. If he feels that the new profit of \$12,500 − 1,500 = \$11,000 still is

```
THE PROPER ENTRY FORM IS:
ORIGINAL COST, INCREMENT, INCREASE IN COST, ETC.

START  :0
A      :4000 1 1500
B      :3500 1 6000
C      :0
D      :3000 1 2000
E      :0
F      :0
G      :4250 1 1500 1 2000
H      :0
FINISH :0
ENTER THE AMOUNT OF TIME AVAILABLE FOR THIS PROJECT 10 WEEKS
```

(a)

```
THE CRITICAL PATH IS

START   B → E → H → FINISH

THE LENGTH OF THE CRITICAL PATH IS 10

THE TOTAL FREE SLACK IS 0
```

CODE	DURATION	EARLY START	EARLY FINISH	LATE START	LATE FINISH	TOTAL SLACK	FREE SLACK	TOTAL FREE SLACK
START	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A	2.00	0.00	2.00	5.00	7.00	5.00	5.00	5.00
B	3.00	0.00	3.00	0.00	3.00	0.00	0.00	0.00
C	6.00	0.00	6.00	1.00	7.00	1.00	0.00	1.00
D	4.00	3.00	7.00	3.00	7.00	0.00	0.00	0.00
E	2.00	3.00	5.00	5.00	7.00	2.00	1.00	2.00
F	7.00	3.00	10.00	3.00	10.00	0.00	0.00	0.00
G	3.00	7.00	10,00	7.00	10.00	0.00	0.00	0.00
H	3.00	6.00	9.00	7.00	10.00	1.00	1.00	1.00
FINISH	0.00	10.00	10.00	10.00	10.00	0.00	0.00	0.00

ACTIVITY	ORIGINAL COST	OPTIMIZED COST
START	0.00	0.00
A	4000.00	4000.00
B	3500.00	3500.00
C	0.00	0.00
D	3000.00	3000.00
E	0.00	0.00
F	0.00	0.00
G	4250.00	5750.00
H	0.00	0.00
FINISH	0.00	0.00

TOTAL INCREASE IN COST 1500.00

(b)

Figure 14. Input and output of LEAST COSTING program: (a) cost distribution data for activities selected for possible crash schedules, and (b) new schedule statistics and cost comparisons which meet the target completion time of 10 weeks. Critical path has changed to "START→B→E→H→FINISH"

satisfactory, his best decision would be to spend the money to finish the project in 10 weeks. Or he could experiment with additional crash schedule alternatives that involve other activities.

Limited Resources

A limited resource model called SPAR (Scheduling Program for Allocation of Resources) was developed by Wiest [1967]. SPAR is a heuristic scheduling model for limited resources that is designed to handle a project with 1,200 single

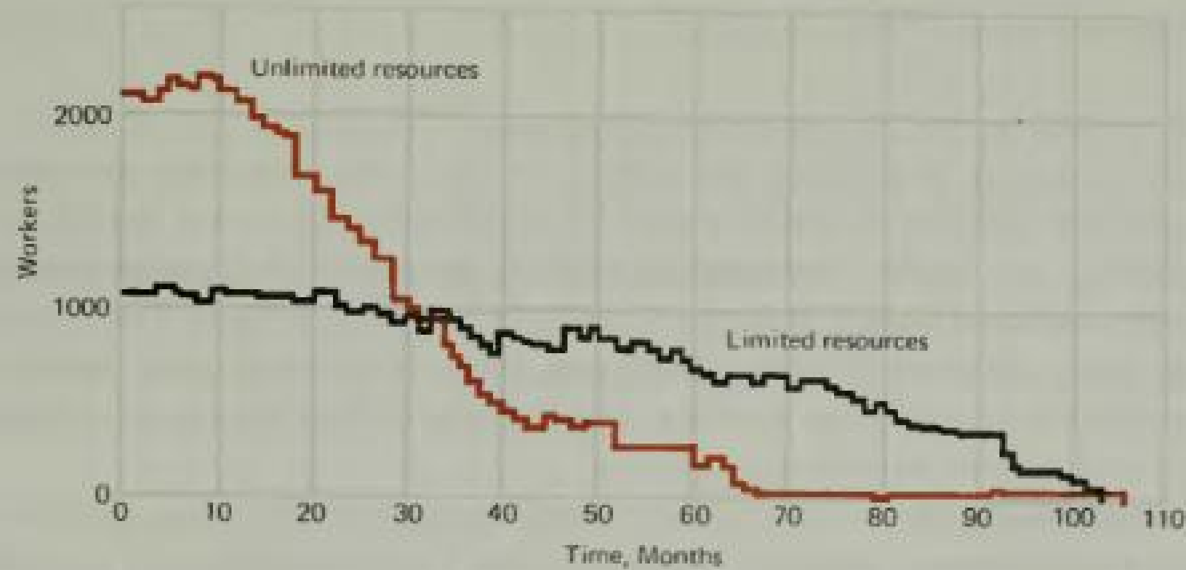


Figure 15. Manpower loading schedule for the space vehicle project

SOURCE: J. D. Weist. "A Heuristic Model for Scheduling Large Projects With Limited Resources," *Management Science* 13(February 1967).

resource activities, 500 nodes, and 12 shops over a span of 300 days. The model focuses on available resources that it allocates period-by-period to activities listed in order of their early start-times. The most critical jobs have the highest probability of being scheduled first, and as many jobs are initially scheduled as available resources permit. If an available activity fails to be scheduled in a particular period, an attempt is made to schedule it in the next period. Finally, all jobs that have been postponed become critical and move to the top of the priority list of available activities.

Wiest applied the SPAR program to a space vehicle project that required large block engineering activities with up to 5 different types of engineers and involving 300 activities. Figure 15 shows an overall manpower loading chart for the program. The unlimited resources line resulted from a conventional PERT schedule, in which all activities were at their early start-times. The limited resources line results from the SPAR schedule, in which peak manpower requirements were reduced considerably. The total length of the project was shortened by 5 months, and the number of gross hirings of personnel was reduced by 30 percent as a result of the SPAR schedule.

Woodworth and Willie [1975] have developed a new heuristic algorithm, in which the completion dates are the most crucial constraint. The algorithm was programmed to handle up to 10 projects, each having up to 500 activities, 20 different resource types, and 200 unique paths. More recently, Suraphongschai [1976] has experimented with a heuristic scheduling procedure for resource constrained projects that seems to hold considerable promise.

Summary

Network planning techniques are unique in the form that they have taken, especially the concepts of critical path. The associated concepts of load leveling, least costing, and limited resource scheduling have provided project management with a rational base that rests on carefully laid, broad plans. The plans are: derived from analyses of some prominent alternatives; computer based and thus applicable to very large systems; and flexible, so that they can be changed as actual experience with them develops.

It is interesting that the independent development of PERT and CPM in two different environments produced methodologies that essentially are equivalent but that carry the stamps of their spawning grounds. CPM developed in engineering maintenance operations, where a great deal of experience existed and activity times were relatively well known. Thus, CPM developed as a deterministic model. PERT, however, developed in an environment of research and development, where great uncertainty of activity times exists. The result was the probabilistic model.

Review Questions

1. What are the unique factors in the nature of large-scale, one-time projects that require special planning methods? Why are the general methods of Chapters 5, 7, and 12 inadequate?

2. Why is it necessary to perform the planning of what must be done and schedule planning together for large projects?

3. Discuss the origin of network planning methods.

4. In the context of network planning methods, define the following terms: *activity*, *event*, *node*, and *critical path*.

5. For PERT planning methods, discuss and interrelate the three phases:
   a. Activity analysis.
   b. Arrow diagramming.
   c. Node numbering.

6. What are the functions of dummy activities in PERT network diagrams?

7. What is the convention for numbering nodes in a PERT network? Why is this convention used?

8. Why must activity networks be noncyclical?

9. Define the following terms: *early start* (*ES*), *early finish* (*EF*), *latest start* (*LS*), *latest finish* (*LF*), and *slack*.

10. Outline the procedure for manual computation of schedule statistics.

11. What are the differences in the construction of the arrow diagram between the PERT and CPM methodologies? How can the probabilistic network model provide additional data that will help managers make decisions?

12. Define the terms *optimistic time*, *pessimistic time*, *most likely time*, and *expected time* in probabilistic PERT networks.

13. What is meant by *load leveling*? How may it be accomplished?

14. Discuss the concepts of *least costing* in relation to crash, normal, and slow schedules.

15. What is the nature of the SPAR limited resource model?

16. Account for the differences between PERT and CPM as they originally were developed.

Problems

1. Table 2 lists a set of activities, sequence requirements; and estimated activity times required for the renewal of a pipeline. Prepare both a PERT and CPM project diagram.

2. For the data of Problem 1 and the arrow diagram generated there:

TABLE 2

ACTIVITIES, SEQUENCE REQUIREMENTS, AND TIMES FOR THE RENEWAL OF A PIPELINE.

Activity	*Letter Code*	*Code of Immediate Predecessor*	*Activity Time Requirement (Days)*	*Crew Requirements per Day*
Assemble crew for job	*A*	—	10	—
Use old line to build inventory	B	—	28	—
Measure and sketch old line	C	*A*	2	—
Develop materials list	*D*	C	1	—
Erect scaffold	E	D	2	10
Procure pipe	F	D	30	—
Procure valves	G	D	45	—
Deactivate old line	*H*	*B, D*	1	6
Remove old line	*I*	*E, H*	6	3
Prefabricate new pipe	*J*	F	5	20
Place valves	K	*E, G, H*	1	6
Place new pipe	L	*I, J*	6	25
Weld pipe	M	L	2	1
Connect valves	N	K, M	1	6
Insulate	O	K, M	4	5
Pressure test	P	N	1	3
Remove scaffold	Q	N, O	1	6
Clean up and turn over to operating crew	*R*	*P, Q*	1	6

a. Compute *ES*, *EF*, *LS*, and *LF* for each of the activities.

b. For the data generated in (a), compute slack for the system. Which activities can be delayed beyond their respective *ES* times without delaying the project completion time of 65 days? Which activities can be delayed, and by how many days, without delaying the *ES* of any other activity?

c. Determine the critical path for the pipeline renewal project.

3. Table 3 shows additional information in the form of optimistic, most likely, and pessimistic time estimates for the pipeline renewal project. Compute variances for the activities. Which activities have the greatest uncertainty in their completion schedules?

4. Suppose that, due to penalties in the contract, each day that the pipeline renewal project can be shortened is worth $100. Which of the following possibilities would you follow, and why?

a. Shorten t_e of activity *B* by 4 days at a cost of $100.

b. Shorten *b* of activity *G* by 5 days at a cost of $50.

Most likely time

465 → 466 Pessimistic Time

250

5×100 - 50 ≈ 450

TABLE 3
TIME ESTIMATES FOR THE PIPELINE RENEWAL PROJECT.

Activity Code	Optimistic Time Estimate of, a	Most Likely Time Estimate of, m	Pessimistic Time Estimate of, b	Expected Time Estimate of, t_e
A	8	10	12	10
B	26	26.5	36	28
C	1	2	3	2
D	0.5	1	1.5	1
E	1.5	1.63	4	2
F	28	28	40	30
G	40	42.5	60	45
H	1	1	1	1
I	4	6	8	6
J	4	4.5	8	5
K	0.5	0.9	2	1
L	5	5.25	10	6
M	1	2	3	2
N	0.5	1	1.5	1
O	3	3.75	6	4
P	1	1	1	1
Q	1	1	1	1
R	1	1	1	1

2x100=200-$150=50

c. Shorten t_e of activity O by 2 days at a cost of $150.

d. Shorten t_e of activity O by 2 days by drawing resources from activity N, thereby lengthening its t_e by 2 days.

5. Table 2 indicates the crew requirements per *day* for each activity in the pipeline renewal project.

7

a. Prepare a crew size versus time chart representing the deployment of manpower for the ES schedule generated in Problem 2. Assume that workers on all crews are completely interchangeable; that is, any person can do any task.

b. Now assume that the work days allocated for each activity can be deployed in any way you wish. For example, activity L has a crew requirement of 25 workers per day for 6 days from Table 2, or 150 work days. These 150 days may be allocated over any chosen activity time; for example, 10 workers per day for 15 days, or vice versa. To achieve load leveling of the total crew, reallocate the work days required for each activity so that no activity has a labor rate greater than a crew size of 10 workers per day. This will require extending the activity times of some activities.

(1) Compute the schedule statistics and new critical path.

(2) Prepare a new crew size versus time chart representing the deployment of manpower, but with the restriction that maximum total crew size per day is 10, using available slack as necessary. Also, use available slack to make this labor schedule as compact as possible in terms of the fewest number of fluctuations in overall crew size and condensed in time span.

REFERENCES

Abernathy, W. J., and J. S. Demski. "Simplification Activities in a Network Scheduling Context." *Management Science* 19(May 1973): 1052–1062.

Archibald, R. D., and R. L. Villoria. *Network-Based Management Systems*. New York: John Wiley & Sons, 1967.

Buckley, J. W.; M. R. Nagaraj; D. L. Sharp; and J. W. Schenck. *Management Problem Solving with APL*. Los Angeles: Melville Publishing Co., 1974.

Buffa, E. S., and J. S. Dyer. *Management Science/Operations Research: Model Formulation and Solution Methods*. Santa Barbara, Ca.: Wiley/Hamilton, 1977.

———; and W. H. Taubert. *Production-Inventory Systems: Planning and Control*. Rev. ed. Homewood, Ill.: Richard D. Irwin, 1972. Chapters 13 and 14.

Burgess, A. R., and J. B. Killebrew. "Variations in Activity Level on a Cyclical Arrow Diagram." *Journal of Industrial Engineering* 13(March–April 1962): 76–83.

Davis, E. W. "CPM Use in Top 400 Construction Firms." *Journal of the Construction Division* 100(March 1974): 39–49.

———. "Project Network Summary Measures and Constrained Resource Scheduling." *AIIE Transactions* 7(June 1975): 132–142.

———. "Project Scheduling Under Resource Constraints—Historical Review and Categorization of Procedures." *AIIE Transactions* 5(December 1973): 297–313.

Dewitte, L. "Manpower Leveling in PERT Networks." *Data Processing Science/Engineering* (March–April 1964).

Elmaghraby, S. E. "On Generalized Activity Networks." *Journal of Industrial Engineering* 17(November 1966): 621–631.

Fulkerson, D. R. "A Network Flow Computation for Project Cost Curves." *Management Science* 7(1961).

Kelly, J. E. "Critical Path Planning and Scheduling: Mathematical Basis." *Operations Research* 9(1961).

Levin, R. I., and C. A. Kirkpatrick. *Planning and Control with PERT/CPM*. New York: McGraw-Hill, 1966.

Levy, F. K.; G. L. Thompson; and J. D. Wiest. "Multi-Shop Work Load Smoothing Program." *Naval Research Logistics Quarterly* (March 1963).

_____; _____; and _____. "The ABCs of the Critical Path Method." *Harvard Business Review*. (September–October 1963): 98–108.

MacCrimmon, K. R., and C. A. Ryavec. "An Analytical Study of the PERT Assumptions." *Operations Research*. 12(January–February 1964).

Malcolm, D. G.; J. H. Roseboom; C. E. Clark; and W. Fazar. "Application of a Technique for Research and Development Program Evaluation." *Operations Research* 7(September–October 1959).

Moder, J. J., and C. R. Phillips. *Project Management with CPM and PERT*. New York: Reinhold, 1964.

Shaffer, L. R.; J. B. Ritter; and W. L. Meyer. *The Critical Path Method*. New York: McGraw-Hill, 1965.

Suraphongschai, V. "A Heuristic Procedure for Resource-Constrained Project Scheduling." Ph.D. dissertation, University of California, Los Angeles, 1976.

Wiest, J. D. "A Heuristic Model for Scheduling Large Projects with Limited Resources." *Management Science* 13(February 1967): 359–377.

______. "Heuristic Programs for Decision-Making." *Harvard Business Review* (September–October 1966): 129–143.

______. "Some Properties of Schedules for Large Projects with Limited Resources." *Operations Research* 12(May–June 1964).

______, and F. K. Levy. *A Management Guide to PERT/CPM*. Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Woodworth, B. M., and C. J. Willie. "A Heuristic Algorithm for Resource Leveling in Multi-Project, Multi-Resource Scheduling." *Decision Sciences* 6(1975): 525–540.

Chapter 14

Service and Nonmanufacturing Systems

Although service and nonmanufacturing systems exhibit great diversity in our society, they have common problems of planning and control. In every case, we can abstract from the unique situation problems of aggregate planning within which detailed plans and schedules must be coordinated. Manpower planning and scheduling is a common focus, although in some instances the scheduling of equipment and equipment maintenance takes precedence. The methods used for planning and scheduling vary widely, but generally they are sensitive to the same kinds of costs and pressures that can be found in industry.

In the Part IV introduction, service nonmanufacturing systems were classified as "systems for noninventoriable items." Since none of these systems produce an inventoriable output, we must eliminate this variable. However, eliminating inventories of the output from the system complicates scheduling problems. We were able to use inventories effectively in manufacturing systems to absorb variations in demand throughout the system. The decoupling function of inventories allowed us to carry on each set of activities relatively independently. We used buffer inventories to decouple demand variations from disrupting activities and to provide good service. We used seasonal inventories to buffer the effects of seasonal variation in demand on factory employment and to get by with a smaller peak physical capacity.

In service and nonmanufacturing systems there is no inventory buffering.

Often, the full impact of demand variations is transmitted to the producing system. When we discussed systems stability in Chapter 3 we stated that fixing a variable may render the remainder unstable. We also stated the simple principle that if the environment imposes some shock to the system, something must be allowed to vary in response. Therefore, in service and nonmanufacturing systems, one of the important system characteristics is the lack of inventories; and one of the crucial characteristics of managerial planning and control systems is the mode by which buffering of demand variations is achieved.

The impact of demand variation is important for both service and nonmanufacturing systems, but the problem is greater in service systems because of the importance of short-term variations. In some systems the demand is immediate, such as in emergency calls for ambulance, fire protection, and police services. Other systems, such as nonemergency medical care, postal services, and trash collection, have short-term demand variations as do some of the nonmanufacturing systems but their demand for service is not so immediate that management cannot buffer at least some of the effects in some way.

NATURE OF SERVICE SYSTEMS

To understand some of the important characteristics of service systems, we shall use as an example a report [Rising, Baron, and Averill, 1973] of an outpatient clinic at the University of Massachusetts. During the period of the study (1970–1971), the outpatient clinic treated an average of 400–500 patients per day with a staff of 12 full-time physicians. Because the physicians had a variety of other duties, only 260 physician-hours per week were available during regular clinic hours, or nearly 22 hours per physician per week. Only about half the patients were seen by a physician on an appointment or walk-in basis. The others were treated by nurses under a physician's supervision, or in specialized subclinics involving tests and immunizations, etc.

In aggregate terms, for the fall of 1969, approximately 178 patients per day needed to use an average of 52 available physician-hours. Thus, the average time spent with a physician was about 17.5 minutes per patient visit.

Demand For Service

Part of the difficulty in rendering service is seen in Figure 1. Demand is not uniform through the week: it rises approximately 20 percent above the average

Figure 1. Percentage of patients arriving at a university health service to see either a physician or a nurse

SOURCE: E. J. Rising, R. Baron, and B. Averill, "A Systems Analysis of a University-Health-Service Outpatient Clinic," *Operations Research* 21(September–October 1973), p. 1034.

on Mondays, falls off during the week to the low of 84 to 88 percent on Thursday, and increases slightly on Friday.

Furthermore, the daily variation as shown in Figure 2 also is significant. This figure shows arrival data for Monday and Thursday (the days with the heaviest and lightest loads, respectively) and highlights great demand variation during the day (peaks are at 8 A.M., 10:00 A.M., and 2:00 P.M.). When the arrival data are placed on an interarrival-time basis (time between arrivals), they exhibit a negative exponential distribution, as shown in Figure 3.

Time For Service

The time that physicians spent with patients was measured in three separate categories: walk-in, appointment, and second-service times. Figure 4 shows histograms of service times reported for the three categories. The second service category represents a return of the patient to the physician, following diagnostic tests or other intervening procedures. Although the three distributions are different, they share the common general property of being skewed to

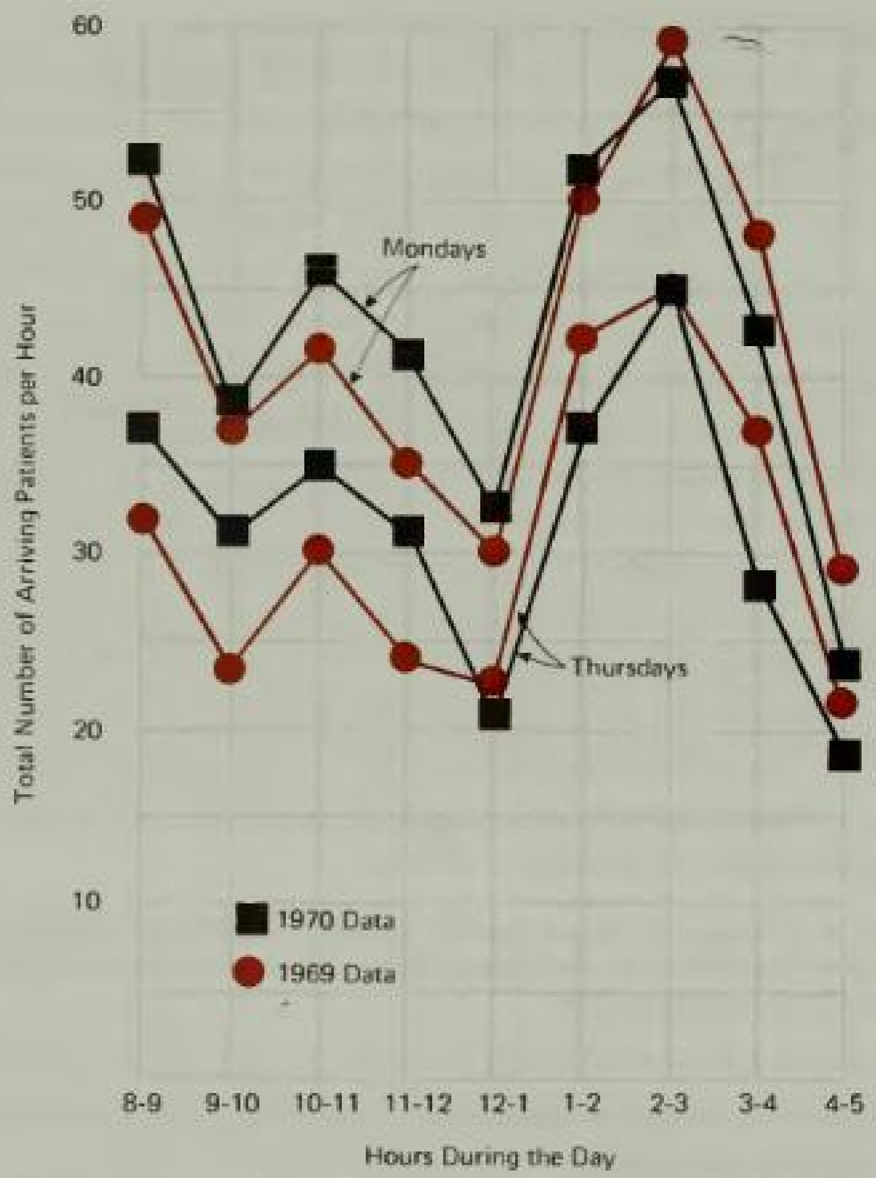


Figure 2. Hourly arrivals at the Student Health Service (Monday and Thursday averages for the fall semesters in 1969 and 1970)

SOURCE: E. J. Rising, R. Baron, and B. Averill, "A Systems Analysis of a University-Health-Service Outpatient Clinic," *Operations Research* 21(September-October 1973): 1030–1047.

the right and having relatively large variances. Thus, the average appointment service time is only 12.74 minutes, but the variance is nearly 10 minutes and the maximum recorded time is 40 minutes. These typical service time distributions reflect the variety of tasks involved in a consultation, depending on the nature of the complaint.

System Design Problems

Because of variable arrival patterns on both a day-of-the-week and hour-of-the-day basis and highly variable service times (depending on the type of pa-

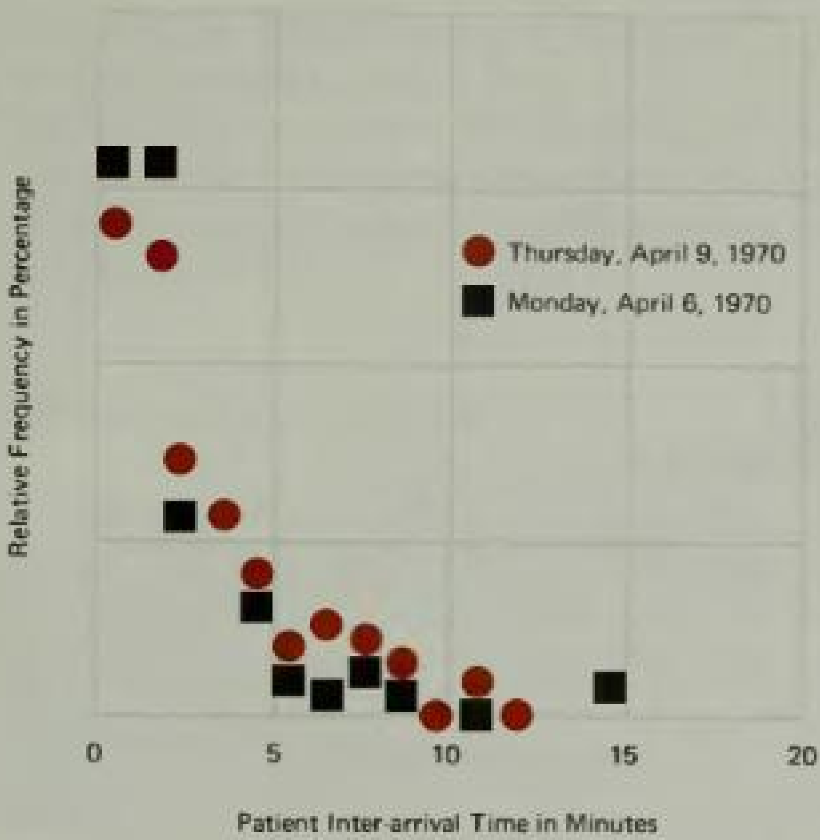


Figure 3. The frequency distribution of patient interarrival times. Monday, April 6, 1970: $\bar{x} = 2.167$, $s = 2.402$, $n = 237$. Thursday, April 9, 1970: $\bar{x} = 2.626$, $s = 2.838$, $n = 202$

SOURCE: E. J. Rising, R. Baron, and B. Averill, "A Systems Analysis of a University-Health- Service Outpatient Clinic," *Operations Research* 21(September-October 1973): 1030–1047.

tient), a number of important problems result. First, how can we schedule appointments in order to smooth the patient load on physicians? How should we arrange appointment schedules through the week and the day, in light of demand variation? What overall capacity for service really is needed? How long can patients reasonably be expected to wait for service? Is physician idle time justified? Would a system of priorities help to level loading? These system design questions are relevant to the outpatient clinic example, but many of them are common to service systems in general.

In analyzing the outpatient clinic as a system, Rising, Baron, and Averill [1973] identified the following conditions as problems:

1. There was a long waiting time for patients.
2. The professional staff felt overworked and harassed.
3. There was much confusion and crowding in the waiting room at predictable times (Monday, Tuesday, and Friday afternoon).
4. The physicians still were seeing patients as long as an hour past closing time.
5. During the day, physicians sometimes were idle because patients did not always keep appointments that often were scheduled several weeks in advance.
6. The current building was overcrowded, since it was designed for nearly half the size of the current student body at the time of the study.

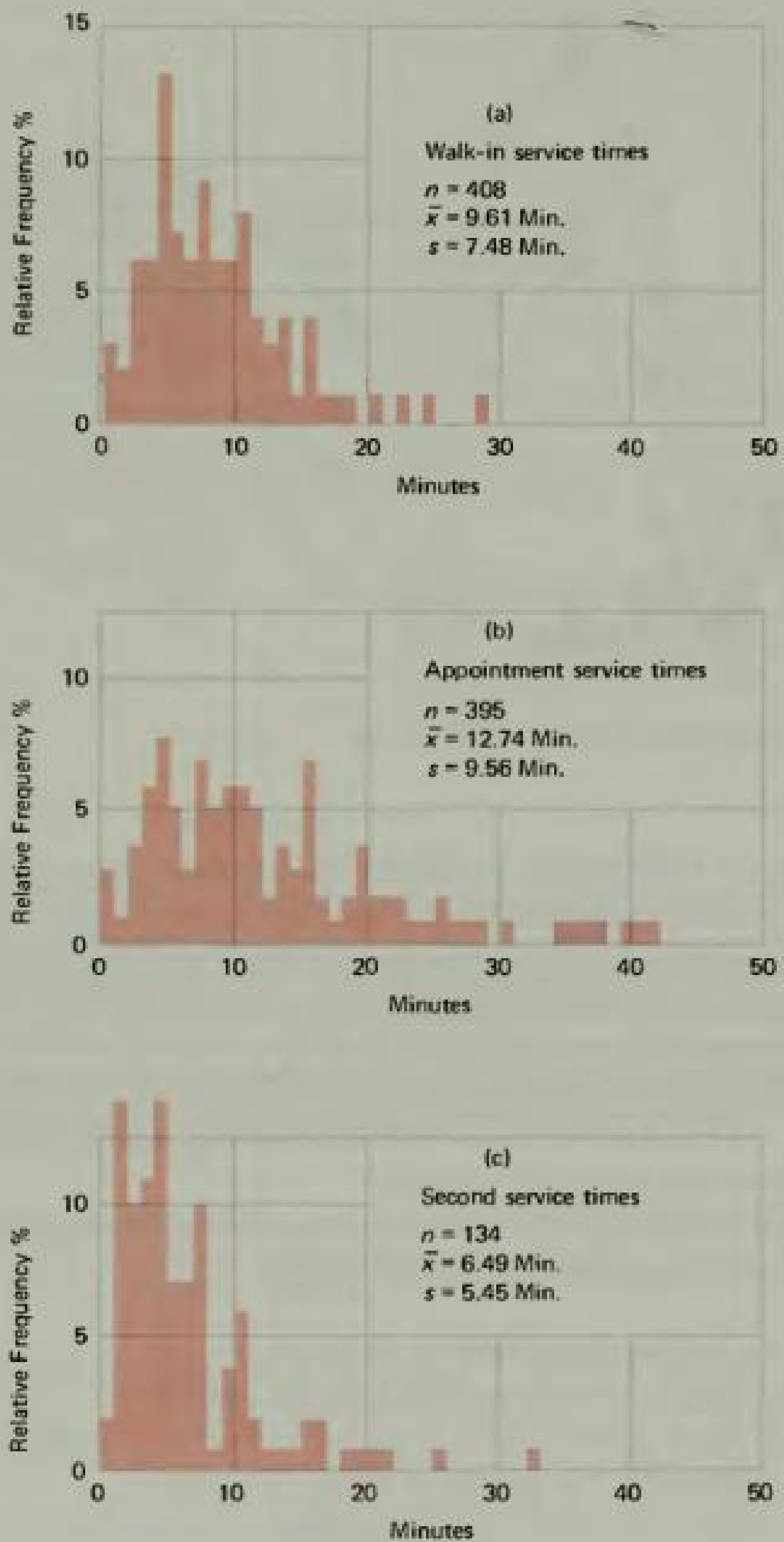


Figure 4. Histograms of service time for (a) walk-in, (b) appointment, and (c) second patients.

SOURCE: E. J. Rising, R. Baron, and B. Averill, "A Systems Analysis of a University-Health-Service Outpatient Clinic," *Operations Research*, 21(September-October 1973): 1030–1047.

In general, the problem was diagnosed as resulting from congestion in a complex queuing system. "The procedure developed to solve this problem was based on the assumption that improved management of demand, through an expanded appointment system, better resource management, and physician scheduling, would make the system function more efficiently." [Rising, Baron, and Averill, 1973]

Simulation and Results

Following general system design methodology, a Monte Carlo simulation of the system was developed to evaluate the effects of alternate decision rules for scheduling both patients and physicians. The proposals included: an improved appointment system that recognized the weekly demand variability but tended to smooth the demand for physicians' appointments; the scheduling of physicians' time to take account of weekly and daily variations; and the establishment of a priority system in which emergencies were given first priority, advance appointments second priority (including second-service), and walk-in patients last priority. Simulation of both the existing and the proposed system provided a measure of improvement in these proposals. Comparisons of measured and simulated waiting times are given in Rising, Baron, and Averill [1973].

Following installation, it was found that there was a 13.4 percent increase in the number of patients seen by physicians, and a 5.1 percent decrease in the number of physician-hours allocated to walk-in and appointment patients. The new operating policies involved less physician overtime, yet there was a 5 percent increase in the overall average time that physicians spent with patients. While the overall waiting time did not change, the mean waiting time for walk-in patients decreased by approximately 10 minutes, and the mean waiting time for appointment patients increased by approximately 14 minutes. Simulation provided a mechanism for studying alternative proposals for managing the demand function. While waiting line models assume a given demand function, management assumes that demand may be partially controllabel and/or that scheduling can compensate. Thus, by managing the demand or arrival process, management absorbs at least some of the demand variation.

Methodology for Service Systems

If we think in terms of the service facility module shown in Figure 5, we have the basic unit on which to build service systems. In each instance, work comes to a service facility or center, some service or processing is performed, and com-

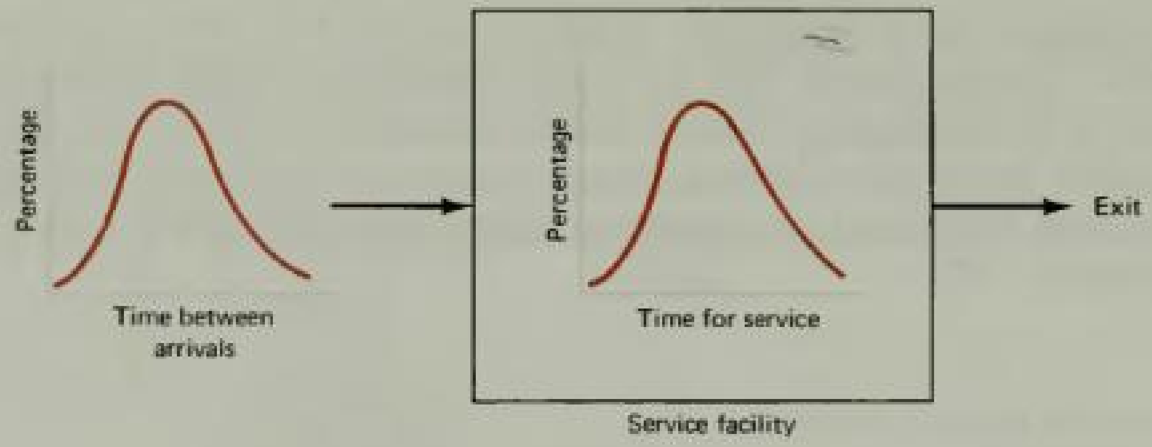


Figure 5. Service facility module

pleted work flows out. The modular unit can be thought of in very general terms. The work that "arrives" for processing or service may be: individuals coming to a medical facility, a bank, or a post office; the breakdown of machines requiring repair by a crew; a product on an assembly line, arriving at the next operation; or a worker in a factory, arriving at the service window of a storeroom to obtain materials. These and other comparable examples represent the types of situations that occur in various kinds of productive systems.

The module is simply the input–transformation–output module that we discussed previously. However, in viewing it as a "service system," we now focus on the probabilistic nature of such systems. In most systems like those in the previous examples, the work does not arrive at a constant rate, nor is the service given at a constant rate. The interplay between variations in arrival rates and variable service times creates unique problems in designing service systems. The concepts provided by waiting line theory (see Appendix D) help us understand how such systems function; and Monte Carlo simulation provides a practical technique for analyzing more complex service systems that are made up of networks of the service facility module shown in Figure 5.

Finally, mathematical programming has been used effectively in resource-allocation problems at the aggregate and detailed manpower scheduling levels in hospital administration, refuse collection, and work shift scheduling.

PLANNING, SCHEDULING, AND CONTROL FOR SERVICE SYSTEMS

HOSPITAL SYSTEMS

Hospital systems as productive systems have been the subject of considerable interest since the early 1960s. Many systems problems have been studied; however, the impact of varying demand on the system has been felt to be highly significant [Fetter and Thompson, 1965; Flagle, 1962; Horvath, 1967; Milsum, Turban, and Vertinsky, 1973]. Monthly averages vary as much as 10 to 15 percent, and there appears to be a seasonal pattern with a low spike in December, followed by relatively high demand from January through April and relatively low demand during the summer months [Horvath, 1967]. The daily fluctuations are somewhat larger and also are somewhat more difficult to deal with. Figure 6 shows a six-and-a-half-year record of an in-patient census at a large West Coast hospital.

The level of bed occupancy reflects the admission rate and the length of hospital stay, both of which are random variables. Statistical studies have indicated that the admission rate is described adequately by a Poisson distribution [Flagle, 1960]. Studies at the Johns Hopkins Hospital have indicated that the distribution of length of hospital stay often can be described adequately by a Gamma distribution with a modal value of 6–8 days. The distribution is skewed strongly to the right, with maximum values of about 50 days.

While the arrival process can be used to aid in the design of bed capacity that is adequate for some percentage of demand, such a methodology merely would accept the random nature of events without attempting to control them. Queu-

Figure 6. Monthly average in-patients at a large West Coast hospital

ing is a commonly acceptable method of dealing with overload in noncritical situations, but it is felt to be an unsatisfactory solution for medical systems in which excessive waiting times add to the anxiety of the patient and may have medical consequences.

A constructive way of looking at the problem is to separate admissions into two groups: emergency cases that obviously are unscheduled, and elective admission cases that have been scheduled ahead of time and therefore can be deferred, if necessary. The resulting proposal is a "call list" of scheduled admissions, that can be used as a "buffer" to minimize fluctuations in bed occupancy level. In other words, the physical queue would become a paper queue (represented by the call list). Figure 7 represents a decision structure for hospital admissions, that uses a call list to smooth occupancy level. During the day, physicians discharge some patients and decide to continue the hospital stay of other patients. Meanwhile, unscheduled admissions are taking place. At a given time in the day, the occupancy state is observed and compared with a decision level. If the state is below the decision level, admissions are called for from the call list to make up the difference. For example, if the decision level were $B = 25$

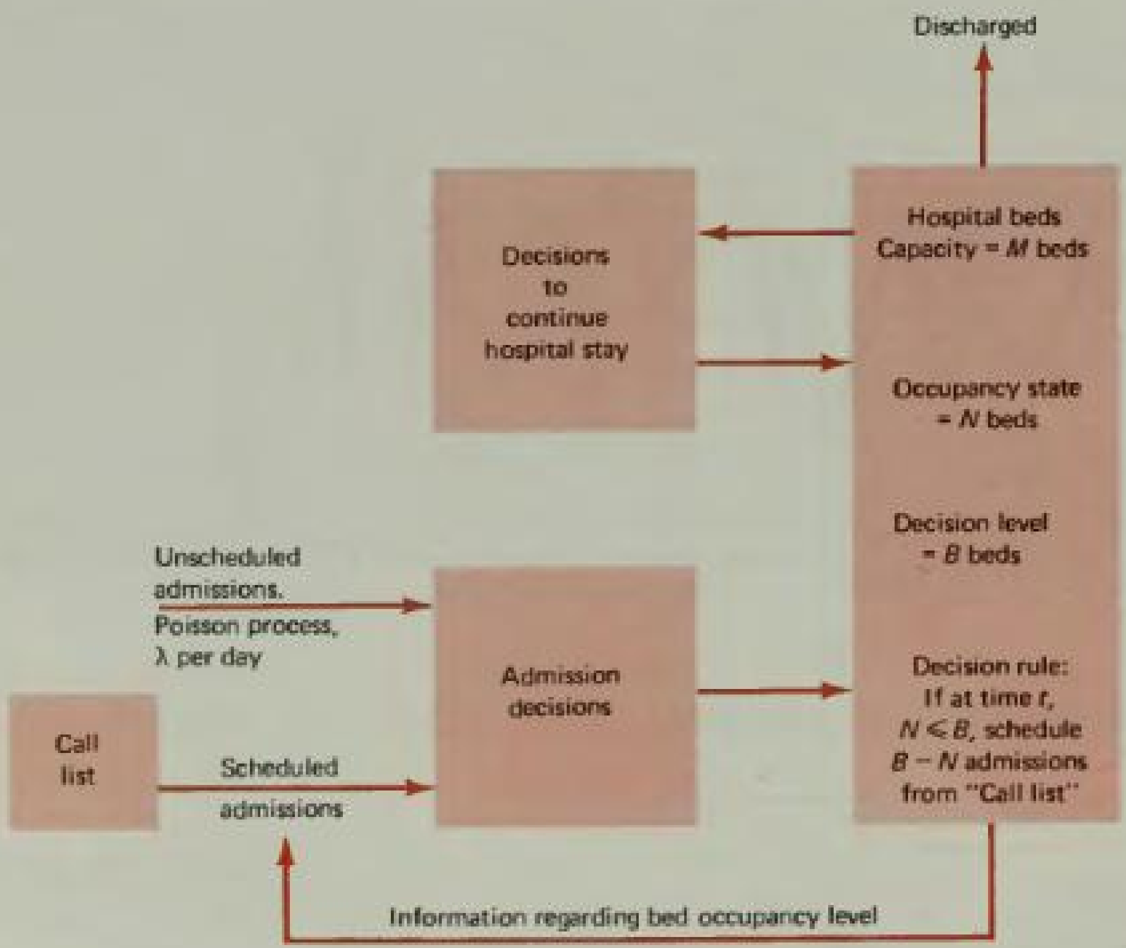


Figure 7. Decision structure for hospital admissions with a "call list" to smooth occupancy level

beds and the occupancy state were $N = 20$ beds at decision-time, then 5 admissions from the call list would be made. This type of feedback control on scheduled admissions reduces fluctuations in occupancy level.

Staff Nursing Requirements. Staff nursing requirements also are affected by varying demand. While cost limitations do not permit staffing for peak levels, it is necessary to maintain a high level of service, which mitigates against staffing to some average level of occupancy. One method of staffing has been to maintain a basic staff in each ward to handle the average load. A flexible core of nurses then is assigned to overload wards on a daily basis.

Since nursing requirements depend on individual patient needs as well as on the number of patients, a method of classifying patients according to their relative nursing needs was developed [Flagle, 1960]. Patients were classified into 3 groups in terms of nursing care needs: self care, partial care, and total care. A sample survey indicated that the distributions of minutes of care were very different for the three categories. The distribution for self care patients had a mean of 0.5 hour and a very small standard deviation. The distribution for partial care patients had a mean of 1 hour and a somewhat larger standard deviation; and the distribution for total care patients had a mean of 2.5 hours and a very large standard deviation. Studies of the number of patients in the total care category in 4 different wards varied significantly. However, there seemed to be no correlation in work load between the wards, thus supporting the idea of using a flexible core of nurses who could be assigned where needed.

A similar study of load fluctuation in England indicated similar results between wards—no correlation in weekly overload between wards. Figure 8 indicates these load variations for the two wards separately (on the left) and when combined (on the right. The study indicated that the variance was reduced by about 20 percent when the two wards were combined. This, of course, follows the queuing principle (discussed in Appendix D) that one large facility can give better service than two equivalent but smaller facilities. Reorganizing the nursing staff into one larger group and regarding the two wards as a single waiting line can produce higher levels of service, lower costs, or possibly both.

Simulation of Hospital Operations. Fetter and Thompson [1965] developed a generalized hospital simulation program with the objective of studying bed utilization in a maternity hospital, using the SIMSCRIPT simulation language. The study was based on observed distributions of labor, delivery, and postpartum recovery time, and on a Poisson distribution of admissions for maternity patients. The results were particularly useful in studying the effect of load variation on facility utilization, as well as the effect on utilization by scheduling some deliveries through drug induced labor.

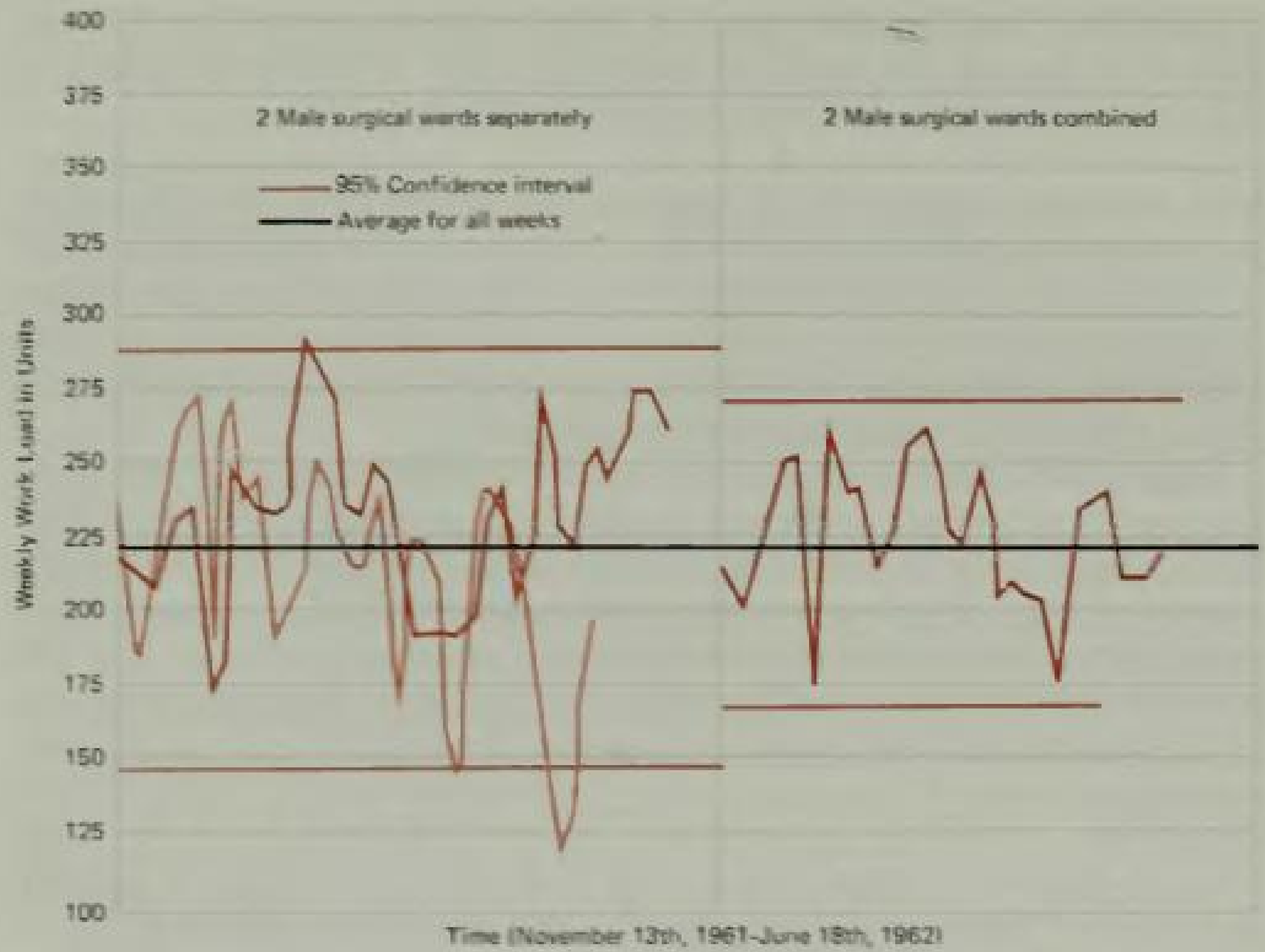


Figure 8. Effect of combining two wards on fluctuation of work load

SOURCE: A. Barr, "Measuring Nursing Care—Operational Research in Nursing," in *Problems and Progress in Medical Care*, (London: Oxford University Press, 1964), pp. 77–90.

Figure 9 shows the simulated effect of increasing load on facility utilization. The average load was 1,320 patients per year, and the effect on utilization was observed when load was decreased by 15 percent to 1,130 patients per year and when it was increased by 25 percent to 1,660 patients per year. The load effects on utilization generally are linear and very substantial for the hospital beds, of some significance for labor rooms, and relatively unimportant for the delivery and postpartum facilities.

To test the value of induction as a means of leveling facility requirements, an experiment was run that allowed admissions twice per day, with a maximum of seven patients per admission, under an elective induction policy. The interesting result of the experiment was the insensitivity of facility requirements to increasing percentages of patients who elected induction. The same number of rooms and other facilities was required for simulated levels of elective inductions of 0 percent, 10 percent, and 20 percent.

This material has provided a background on the general operations environ-

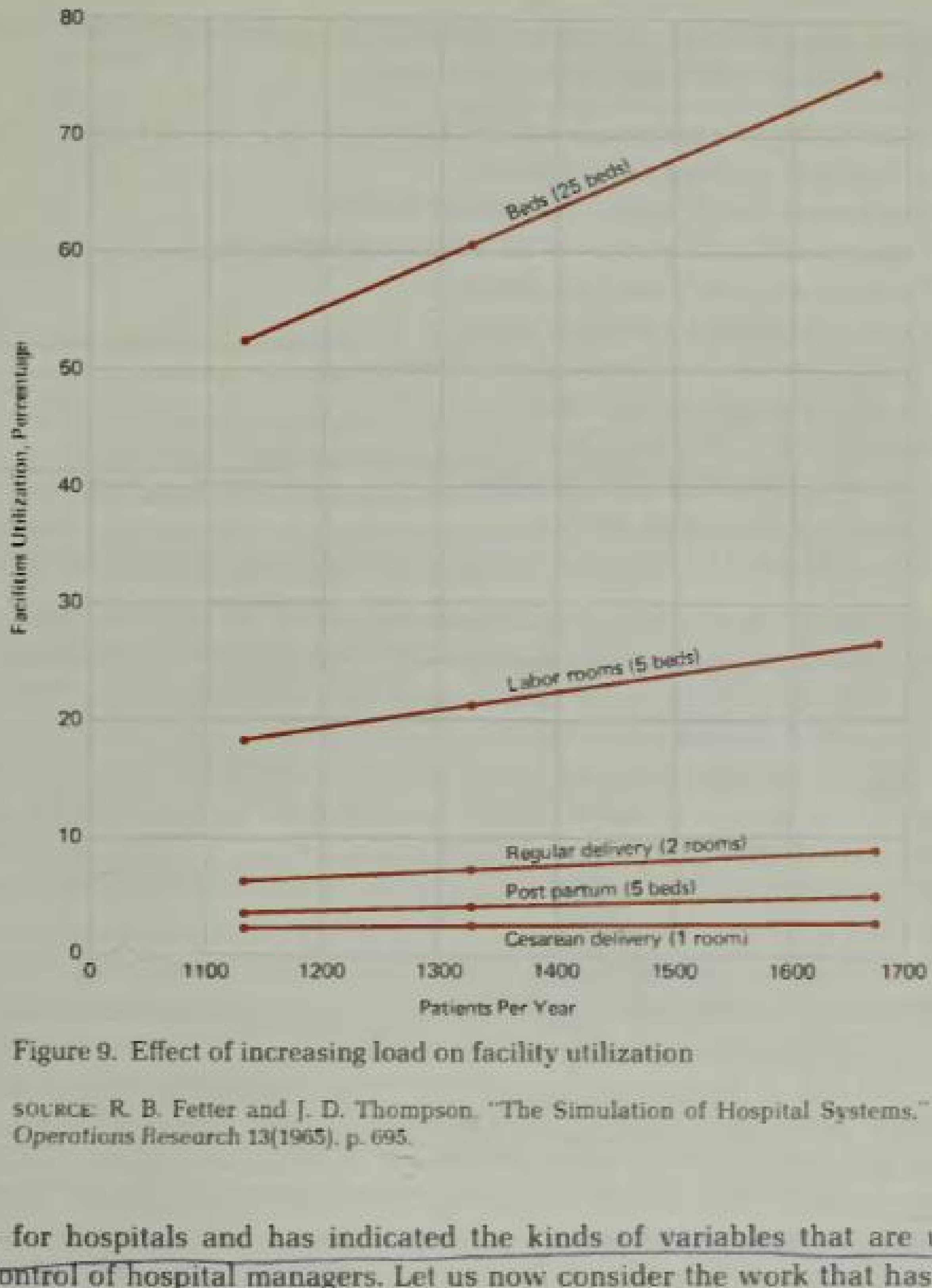


Figure 9. Effect of increasing load on facility utilization

SOURCE: R. B. Fetter and J. D. Thompson. "The Simulation of Hospital Systems," *Operations Research* 13(1965), p. 695.

ment for hospitals and has indicated the kinds of variables that are under the control of hospital managers. Let us now consider the work that has been directed toward operations planning and control.

Aggregate Planning for Hospitals

Lee [1973], has developed an aggregate resource allocation model for hospitals by means of a goal programming approach. This model is very broad, and allocates resources to meet the stated multiple goals in priority sequence for a community hospital in southwestern Virginia. The goals listed by the hospital administrator in priority order, were as follows:

1. Secure adequate personnel and funds, including projected cost increases to maintain existing service levels.
2. Break even, that is, equate revenue and costs.
3. Employ additional personnel for expanded services.
4. Purchase new radiology equipment.
5. Reserve $200,000 for contingencies.
6. Set up research facilities for birth defects.
7. Grant desired salary increases.
8. Employ two staff physicians.
9. Improve emergency room facilities.
10. Acquire mobile clinic facilities.
11. Set up research facilities for children's diseases and maladjustments.

Using the goal programming aggregate planning model, the first nine goals were achieved, the tenth goal was only partially achieved, and the eleventh goal was not achieved. Broad allocations like these provide a framework in which to develop day-to-day, week-to-week, and month-to-month plans and schedules.

The hospital admissions system is used as an overall mechanism for planning and scheduling the use of hospital facilities. The basic unit of capacity is the hospital bed, and bed-occupancy level is the variable that is under management control. Bed-occupancy level, in turn, opens up revenue flow from the broadly based health care services of the hospital. The utilization and resulting revenue from these hospital services then depends on patient-mix selection, which in turn affects length of patients' stay. The demand on other hospital services, such as X-ray and laboratories, then flows from patient-mix.

The problem of operations management is not simply to maximize bed-occupancy level, but also to accommodate emergency demands and the increased costs of the services that are incurred when bed-occupancy levels are too high [Milsum, Turban, and Vertinsky, 1973]. Thus, an aggregate planning and scheduling model must consider scheduling the flow of elective arrivals and the resultant patient-mix. Since the actual admission date for such cases can be scheduled, this flow can be regulated to a substantial degree, and this regulation represents the hospital managers' buffer against the seeming capriciousness of demand. The flow of emergency cases, however, occurs on a random arrival basis, and compounds the problem greatly. The cost effectiveness of the entire hospital admissions system depends on variables that follow probabilistic distributions. Most costs are fixed (rooms, beds, laboratory facilities, basic staff, etc.), comprising up to 75 percent of total hospital costs. Yet the demand on these services is highly variable. Because of the system's complexity, computer simulation has been viewed as the most useful analytical tool. An excellent summary of the work that was done in hospital admissions systems is contained in [Milsum, Turban, and Vertinsky, 1973].

The Nurse Scheduling Problem

Broad level aggregate planning in hospitals must be translated into useful operating plans. While there obviously are many resources to be concerned with in a hospital, the most crucial is nursing. We shall focus our attention on planning and scheduling for this resource. Abernathy, Baloff, and Hershey [1971] have analyzed the nurse staffing problem from the viewpoint of operations planning and control and have pointed out the reasons for underutilization of nurses. We shall review these problems with existing practices to see what fruitful directions are possible. The mounting problems of providing nursing care in part are due to the increase in the intensity of care—the number of nurse-hours per patient-day has increased because of more complex care, shorter average hospital stays, and greater demands by physicians. The difficulty of servicing this increase is compounded by the loss of flexibility in using nurses, which is caused by trends toward specialization.

Abernathy, Baloff, and Hershey examined the general hospital practices in terms of planning, short-term scheduling, and evaluation and control.

Planning Practices. As a basis for planning, it is common to predict demand by forecasting in-patient census six to eighteen months in advance. Based on these forecasts, nursing personnel allocations are made in the form of budgets of FTEs (Full-Time Equivalent nursing personnel).

The forecasts are for total in-patient census, similar to the data in Figure 6. The level of aggregation of such data may be useful for overall aggregate planning, but it will not help us forecast the load on Ward *A* unless a high correlation between census and patient load in specific wards can be shown. In fact, studies have revealed that these factors are not well correlated and that patient load can be measured more accurately by a patient's condition than by the mere fact of bed occupancy.

The second major problem with forecasting methods is that their horizon is too long in the absence of mechanisms for updating. The result is that short-term fluctuations are not detected, and actual occupancy may differ significantly from forecast levels for several months without being adjusted. In other words, the control loop on forecasts is not closed. The forecasting system is static and unresponsive to changing conditions. Abernathy, Baloff, and Hershey reported that, in some instances, forecasts were made simply by restating historical patterns.

"Budgeting to meet demand" is how forecasts for patient demand are translated into authorized FTEs in each shift, for each ward, for each type of nursing personnel: RN = registered nurse, LVN = licensed vocational nurse, and AID = nurse's aide. The defects in the budgeting system arise in the way in which the allocations are made. In some hospitals, the projected average annual census is translated directly into nurse staffing needs at a constant level. Even in situations in which census is forecast on a monthly basis, FTEs are budgeted at a constant level. From a cost point of view, it is particularly significant that the FTE levels are set to meet *peak* demand. Thus, the hallmarks of budgeting practice are the infrequency and static nature of forecasting, the focus on census rather than load, and the setting of constant nurse staffing levels at projected peak demand levels.

The budgeting behavior is understandable: the nursing supervisors distrust the forecasts, both because these forecasts are unreliable about what they purport to measure (census) and because they forecast nothing that was relevant to nurse staffing (load). Thus, without meaningful information, the nursing supervisors logically are more concerned about providing adequate care than about the problems of overstaffing.

? ?

is this good?

Short-Term Scheduling Practices. Staffing supervisors can allocate available labor to wards to meet fluctuating demand on a day-to-day basis. However, very little flexibility is available. Overtime, part-time and "floating" nurses provide some flexibility. Unless current demand is extremely high or low, nursing supervisors must meet daily load with the assigned staff. Thus, while short-term load varies, the system is designed as if load were stable.

Evaluation and Control Practice. Abernathy, Baloff, and Hershey found that the best measure of utilization used by hospital administrators was FTE/PD (Full-Time Equivalent employees per Patient Day). Yet planning and short-term scheduling practitioners say that if nursing supervisors are held accountable for utilization, they cannot take action to stabilize and improve utilization, since the basic policies for forecasting, planning, budgeting and scheduling ~~put~~ appropriate action beyond their control. In system theory terms, nursing supervisors lack the requisite variety* for control. For each action of the system that affects

*The term *variety* defines the ratio of the number of alternatives that are available to the system, divided by the number of alternatives that are available to the controller. When the controller has the *requisite variety*, the ratio is one or less. In other words, the controller must have at least a countermeasure for every action taken by the system, if he/she is to remain in control.

them, a counteraction is not available to them. Abernathy, Baloff, and Hershey conclude that piecemeal patch-up work cannot achieve the needed results, and that an integrated system is required.

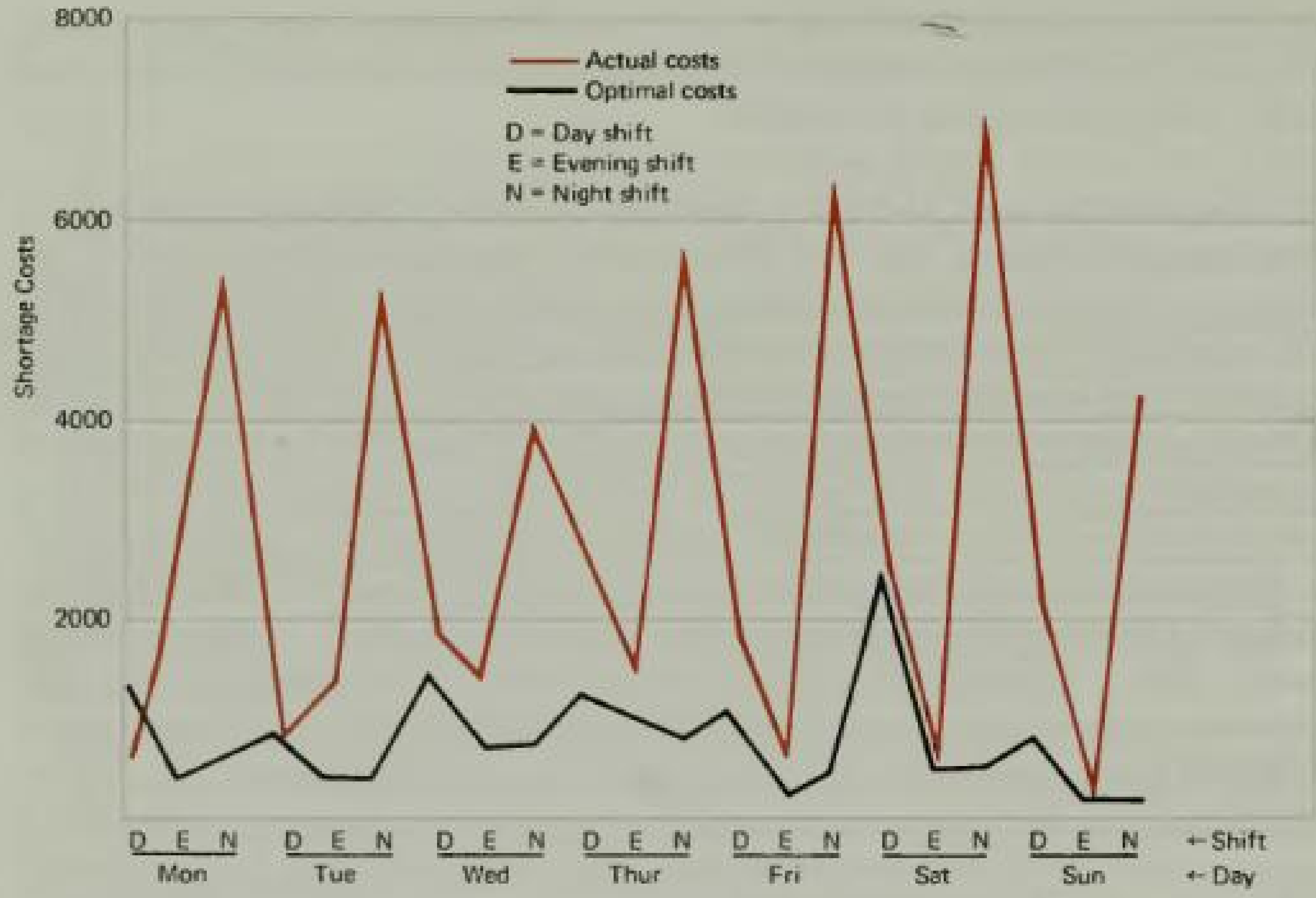
An Integrated System. Abernathy, Baloff, and Hershey suggest a structure for a nursing staff planning and scheduling system. The heart of the system is the aggregate budgeting subsystem, which must achieve balance between the size of the nursing staff and expected demand. Forecast load is an input that must be updated regularly. As the actual scheduling period becomes imminent, forecasts about local public health disease-rate statistics, the impact of catastrophes, and other local conditions can be refined. In the very short term, the effects of current occupancy rates, scheduled admissions, and discharges can be assessed.

Flexible short-term scheduling can use "controlled variable staffing" to maximize the amount of freedom that people are allowed in order to meet changing loads. This is achieved by pooling nursing resources for wards that require approximately the same types of skills. Records indicate little correlation in load between similar wards; by combining wards for staffing purposes (as in Figure 8), the net load fluctuation is reduced. The result is a smaller aggregate staff, a more uniform level of care, and equalized day-to-day staff work loads. Hershey, Abernathy, and Baloff [1972] compared the performance of constant and variable staffing policies in a simulation model.

Detailed Nurse Scheduling Models. Warner and Prawda [1972] developed a mathematical programming model for scheduling nursing personnel. The model specifies the number of nursing personnel that are to be scheduled among wards and shifts in each skill category. In so doing, it satisfies constraints of total personnel capacity and integral assignment, and minimizes a "shortage cost" of nursing care services. Six wards of a 600-bed general hospital were used to test the model's effectiveness.

The model provides a daily staffing report, organized by shifts, that gives the nursing supervisor for each ward the required and integerized optimal numbers of staff in each skill class, the optimal cost, and a sensitivity analysis. The sensitivity analysis shows the cost effect of adding or subtracting one nurse in each skill category, and indicates when a change would not be feasible. Figure 10 shows the actual versus optimal staffing costs for a one-week test, and suggests that using the model can reduce costs by 70 percent. The largest differences between actual and optimal staffing occur on the night shift. In addition to the general cost advantage, shortage costs were reduced by 68 percent, indicating that service had improved.

Liebman [1971] developed a mathematical programming model of personnel allocation and applied it to an extended care nursing unit. The load in ex-



	Actual	Optimal
Mean Staffing Cost of All Wards, All Shifts:	4587	1461
Variance:	33.6×10^6	4.5×10^6

Figure 10. Actual vs. optimal staffing costs. Graphed are the total (sum of the six wards) staffing costs per shift for the actual staffing pattern and the pattern suggested by the optimal solution of the scheduling model

SOURCE: D. M. Warner and J. Prawda, "A Mathematical Programming Model for Scheduling Nurses," *Management Science* 19(December 1972): 411–422.

tended care units is somewhat more stable because of the longer-term hospital stay patterns, and because emergency admissions do not constitute a large factor.

Abernathy, Baloff, Hershey and Wandel [1973] propose a three-stage planning and scheduling model, following the general results of Abernathy, Baloff, and Hershey [1971]. They divide the nurse staffing process into three decision levels: (1) policy and operating procedures; (2) staff planning, including hiring, discharge, training, and reallocation decisions; and (3) short-term scheduling within the constraints of (1) and (2). They formulate the planning and scheduling stages as a stochastic programming problem, and they give an example.

Finally, Maier-Rothe and Wolfe [1973] developed a cyclical scheduling and allocation model for nurse staffing. The objectives of the scheduling and allocation procedure were to assign working days and days off for three shifts to individuals so that:

1. Adequate patient care was assured.
2. A desirable distribution of days off was achieved.
3. Individual members of the nursing staff were treated fairly.
4. Individuals would know their schedule well in advance.

An integer programming model was used and applied in a hospital in which previous scheduling had been done manually. The computer based model was simpler to work with, and provided a more even coverage and a more just distribution of days off. In addition, the scheduling and assignment procedure reduced the floor-to-floor variations of work load index from about 50 percent to 15 percent, and reduced the day-to-day work load index variations for a given floor by 60 percent.

EMERGENCY SYSTEMS

Recently, the number of studies dealing with emergency systems from a planning, scheduling, and control point of view have increased tremendously. The heavy pressure on public budgets, together with the growing complexity of maintaining service levels in larger urban environments, has motivated studies to discover the demand distributions, evaluate alternative location and deployment policies, and develop systems for scheduling manpower.

Emergency Medical Systems

In Chapter 6 we discussed one aspect of emergency medical systems—the location and deployment of ambulances. Fitzsimmons [1970, 1973] used a simulation model to evaluate a wide variety of alternative systems involving different numbers of ambulances and hospitals, and applied the model to determining ambulance locations for a portion of the city of Los Angeles.

From the point of view of operating such systems, the daily load variation shown in Figure 11 immediately suggests the nature of the problem of providing service. The *mean* call rate is plotted in Figure 11, and it varies from a low of about 0.5 calls per hour at 6:00 A.M. to a peak of 3.5 calls per hour at 5:30 P.M.—a ratio of 7 to 1.

However, the 7-to-1 ratio represents only the variation in the mean rate.

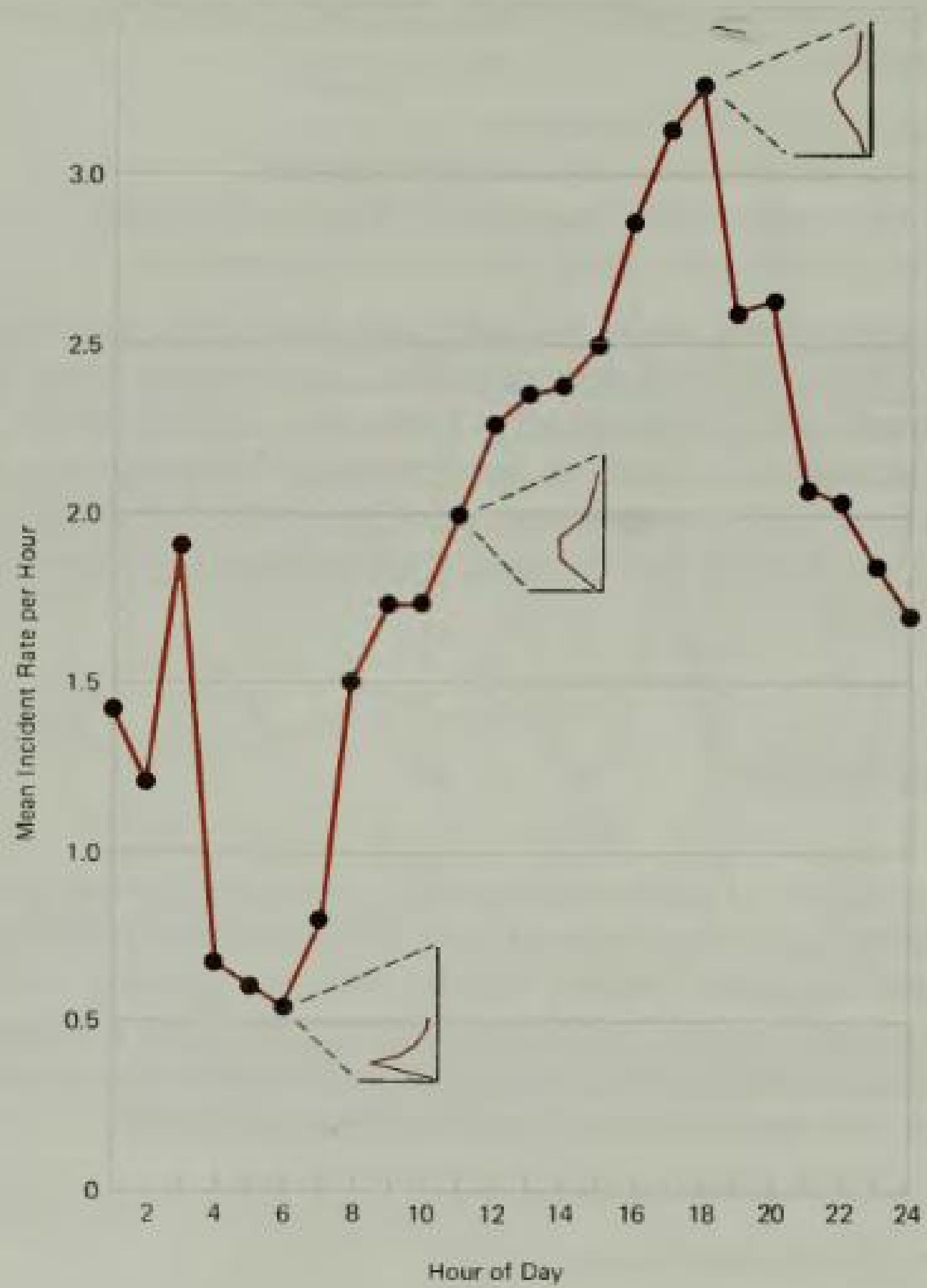


Figure 11. Mean "call-for-service" rate (incident rate) for the Los Angeles emergency medical system

SOURCE: J. A. Fitzsimmons, "Emergency Medical Systems: A Simulation Study and Computerized Method for Deployment of Ambulances," Ph.D. diss., University of California, Los Angeles, 1970).

Studies have shown that the Poisson distribution is a good description of the call rates for the mean values of Figure 11. The shape of the Poisson distribution depends on the mean, and we have sketched in three sample distributions that represent different load levels (shown in Figure 11). At low loads, the Poisson distribution has a fairly sharp peak, with a relatively high probability of call rates (arrival rates) near the mean. At high loads, however, the variability is rather great, and there is a reasonably high probability of call rates that are substantially lower and higher than the mean rate. There is an almost 30 percent probability that the 5:30 P.M. peak could be as high as 5 calls per hour, and a 6 percent probability that it could be 7 per hour.

These are not unrealistic estimates of peak load; the system must be able to accommodate them, or else it will provide poor service. The trade-off between providing capacity versus service is a cruel one; if we allow substantial queuing of calls as a buffer to absorb demand variations, we are trading people's lives for the cost of capacity.

Fitzsimmons [1973] also developed a computer based system—"Computer Ambulance Location Logic" (CALL)—to deploy and dispatch ambulances. Ambulances could be dispatched en route on the basis of evaluations of existing ambulance locations and a computer search routine that directs changes in ambulance location to progressively decrease the system mean response time.

Stevenson [1972] presented cost data to provide service at different probability levels, as well as general analyses of the ambulance problem.

Police Protection

New York City has installed an emergency telephone number that citizens can use [Larson, 1972]. The system has large "trunk" capacity, so that the probability of a busy signal response is very small. A call to the system automatically either is assigned to an idle operator or, by means of a first come-first served queue, to the first available operator. This automatic system provides data about the frequency and timing of calls. Figure 12 shows the daily variation in the number of calls in July and August. The high for the 2 months is approximately 22,000 calls per day, and the low is 13,000—a load of only 41 percent of the peak. Such variations have an enormous influence on the manpower planning and scheduling problem and the resultant costs.

Figure 13 shows the hourly variations for Saturday, August 10 (note that most of the peaks in Figure 12 occur on Saturdays). Calls are highest for approximately 8:00 P.M. until midnight (about 1,400 per hour) and lowest at 5:00 A.M. (about 100 calls per hour—only 7 percent of the peak load). The manning levels needed to receive the emergency calls vary by only approximately 60 percent. Similar distributions are given [Larson, 1972] for other days of the week that have lower peaks; each day the load was met by a somewhat different manning schedule. The data were used to develop empirical probability distributions of operators needed to man the system versus calls received per service time

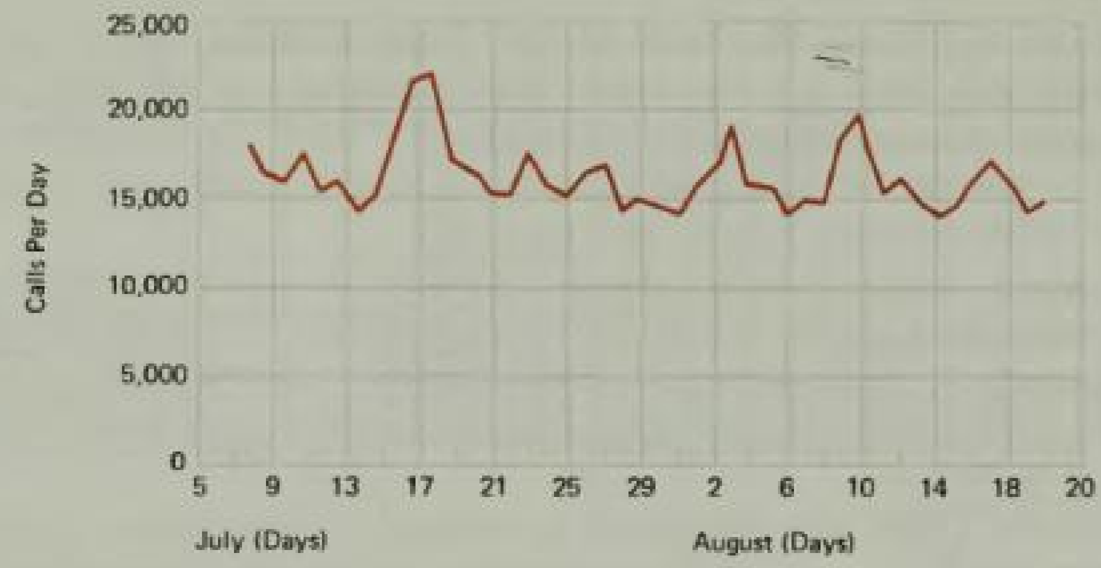


Figure 12. Number of calls received per day

ADAPTED FROM: R. C. Larson, "Improving the Effectiveness of New York City's 911," in *Analysis of Public Systems*, A. W. Drake, R. L. Keeney, and P. M. Morse, eds. (Cambridge, Mass.: MIT Press, 1972).

unit. These distributions, in turn, were components of a manpower scheduling system.

The portion of the manpower scheduling system that deals with emergency operators is only the tip of the iceberg. Scheduling the manpower levels for the police units themselves according to time of day and day of week involves various factors, including the design of response areas [Carter, Chaiken, and Ignall, 1972] and the allocation of emergency units [Chaiken and Larson, 1972]. Larson [1972a] addresses the comprehensive set of problems.

Fire Protection

Figure 14 gives another example of both hourly and daily variation in the demand for emergency service: the total fire alarms that are received in New York City, each hour of the day, for representative low- average-, and high-demand days. On an average day, the 8:00–9:00 P.M. peak is 7.7 times the low point that occurs at 6:00 A.M. The peak of July 4–5 is 3.7 times the peak in the low-day distribution. Traditional deployment policy in fire departments has been to keep the same number of workers and units on duty around the clock [Blum, 1971, 1972] However, if we know the demand pattern, we can devise a deployment plan that takes account of high-risk days and times of the day, and that should be somewhat less costly.

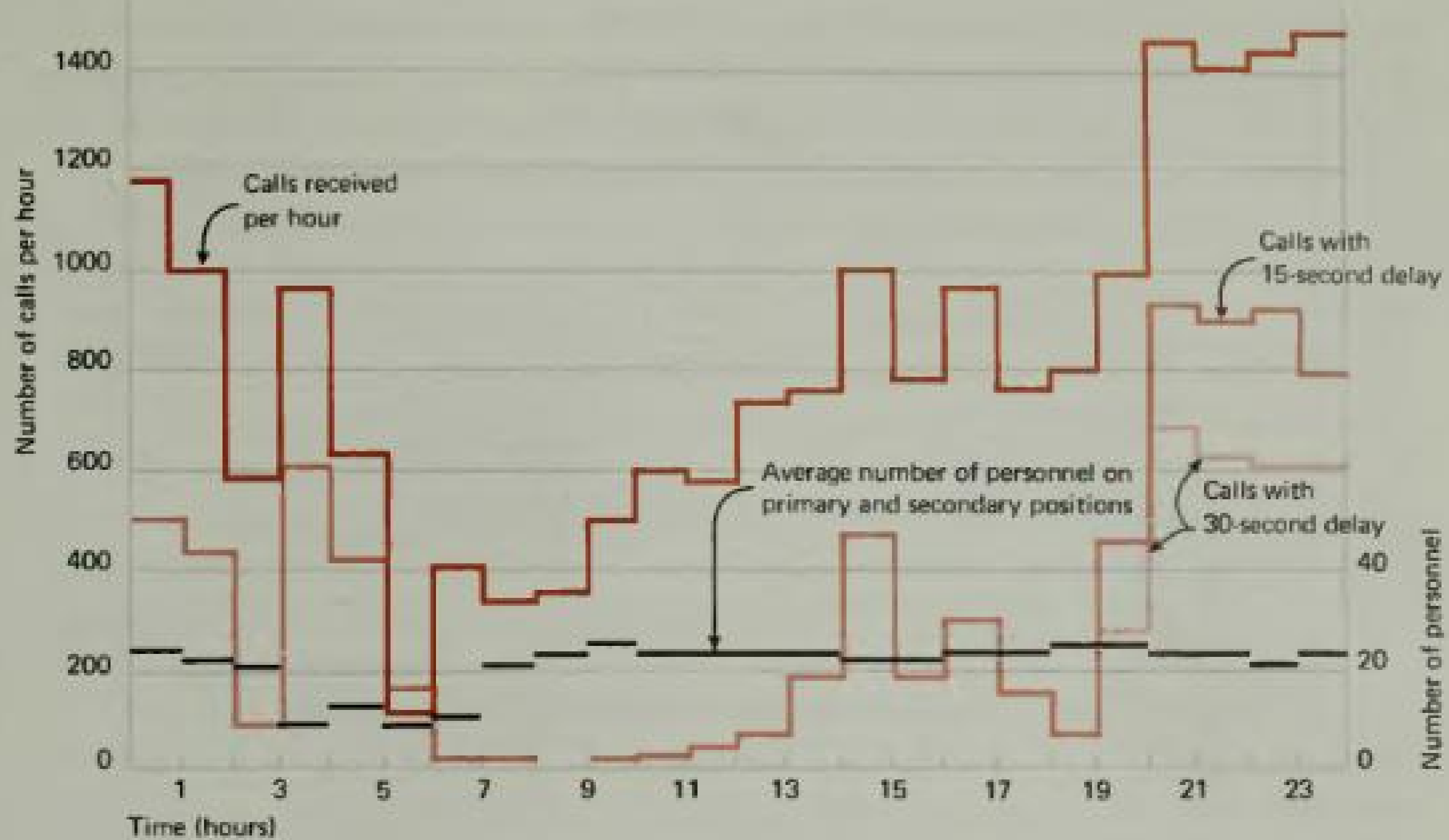


Figure 13. Distribution of calls, delays, and manning levels (Sat., Aug. 10)

SOURCE: R. C. Larson, "Improving the Effectiveness of New York City's 911," in *Analysis of Public Systems*, A. W. Drake, R. L. Keeney, and P. M. Morse, eds. (Cambridge, Mass.: MIT Press, 1972).

Traditionally, fire departments tried to maintain a "standard response" of workers and equipment to alarms in most areas at all times. This policy was pursued despite (1) the considerable variation among areas and times of day in fire hazards, and (2) the likelihood that an initially ordinary alarm fire will turn out to be a serious one [Blum, 1971, 1972].

Obviously, the demand distributions for all the emergency services that we have discussed require severe peak capacities. As a strategy, queuing calls cannot reasonably be expected to absorb demand variability. The common approach for absorbing demand variations is differential manning of the system. This enables us to adjust short-term capacity to meet demand.

GOVERNMENT SERVICES

Manpower Replanning in the Food and Drug Administration

The FDA (see Rosenthal and Murphy [1974]) operates through a network of forty-two district and regional offices and through national laboratories for

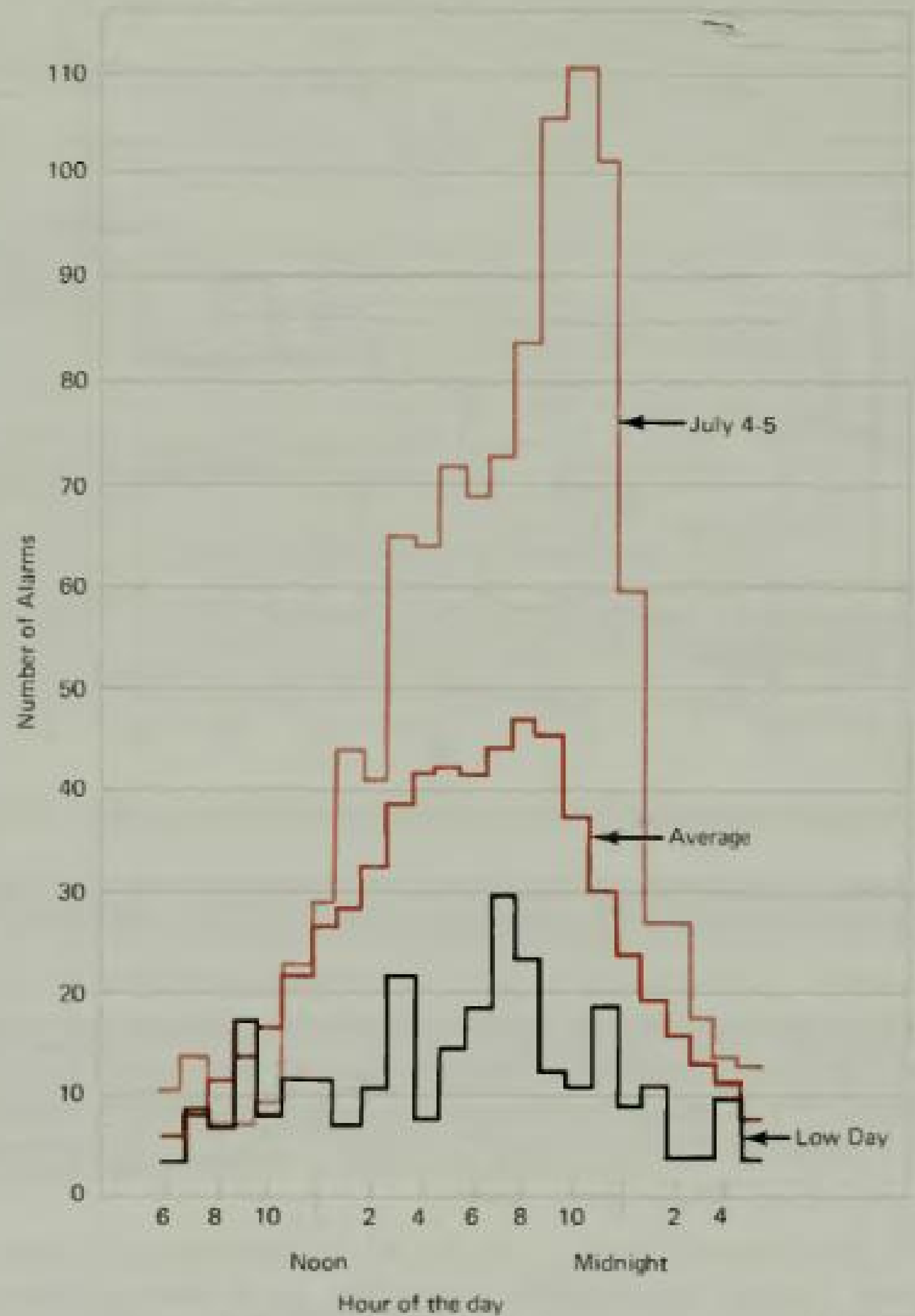


Figure 14. Total fire alarms received in New York City by hour (1963 data)

SOURCE: E. H. Blum. *Deployment Research of the New York City Fire Project*. R-968. The New York City Rand Institute. May 1962.

sample analysis. The FDA's primary activities relate to the inspection of facilities, the collection of samples, and laboratory analysis at the various production and distribution points for consumer goods moving through the economy.

The goals and objectives of the regional operations inspectors are established by the various bureaus of the FDA and sometimes by congressional and executive action. These goals and objectives then are implemented in the field. A central planning function is responsible for the development and coordination of the specific compliance programs that are to be executed by the field staff. Invariably, these programs exceed available manpower resources and budgets.

The planning process is similar to that found in any large organization: it is iterative in nature. This means that it involves the sequential testing of various combinations of program plans in order to balance the requirements with the available resources. This is accomplished by running through innumerable calculations to convert inspection and sampling levels into manpower requirements, and by assessing the feasibility of these levels. The modification of program plans, and recycling through the computation of manpower requirements, may result. This process is designed to achieve a balanced, feasible plan, and is affected critically by the computational cycle time, since several iterations may be necessary.

The cycle time needed to convert requirements into manpower and to test availability is approximately three months. Calculating these results manually requires so much time that, frequently, the final approved plan may not balance with available staff, and distortions in setting goals for regional offices may result. Crisis situations, such as the discovery of bacterial contamination in the processed-mushroom industry, add extra complications to the planning process. Such crises may mean a full-scale, regional investigation to remove the products from distribution channels. Immediate decisions must be made about how many investigators should be diverted, and in which areas. In addition, such investigations significantly affect the remaining programs. The trade-offs associated with the reduced availability of manpower effectively require the entire compliance program for all activities to be revamped. In the past, the long cycle time has impeded a thorough trade-off analysis.

Recently, however, the FDA developed a computer based planning model, called the "Automated Planning Tool," which reduces the planning cycle time to a few minutes. There are 5 levels of consolidation of activities that are built into the system. These include 300 compliance programs, 65 project management system projects, 15 categories, 5 budget groups, and the consolidated total of operations. At each level of consolidation, forecasts can be displayed at a computer terminal according to: district; region; operations; manpower class; and, optionally, work-hours required, work-years required, number of personnel, costs (in terms of average salary grade rates), and work space required.

The planning system is composed of separate modules, which can be interrogated independently and at the option of the analyst. The forecasts can be injected throughout the system. The system has two modes of operation; these might be characterized as top-down and bottom-up. The latter mode can work from the officially established requirements and move up through the various consolidations. In the top-down mode, broad questions can be asked and actions can be simulated (such as changing the emphasis between inspections and sampling at a total bureau level). The results are displayed at a terminal, in the format of reports that are consistent with those that were used before the automated planning tool was developed. The computer based planning concept may have wide application to many levels of government that must deploy a limited number of skilled specialists over a wide area.

Municipal Refuse Collection

Ignall, Kolesar, and Walker [1971] developed a linear programming model for crew assignments in refuse collections systems. The problem was to assign crews to shifts on a district basis in order to achieve a balance between payroll costs and missed collections, while simultaneously to satisfy the constraints due both to available equipment and to policies for granting three-day weekends. A cost of $1 per ton was imputed for uncollected refuse.

Summary results are shown in Table 1 for 15, 16, and 19 trucks, and numbers of crews between 18 and 22. Following are the authors' observations about the results:

TABLE 1
VALUES OF MEASURES OF EFFECTIVENESS FOR CASES RUN

Number of Trucks	Number of Crews				
	18	19	20	21	22
15—					
Manpower cost	$14.774	$13.200	$13.217	$13.225	$13.574
uncollected refuse (tons)	321	245	136	80	62
16—					
Manpower cost	$14.668	$13.129	$13.144	$13.152	$13.541
uncollected refuse (tons)	315	237	132	88	61
19—					
Manpower cost		$13.005		$13.048	
uncollected refuse (tons)		230		68	

SOURCE: E. Ignall, P. Kolesar, and W. Walker, "Linear Programming Models of Crew Assignments for Refuse Collection," P-4717, New York City Rand Institute, November 1971.

1. The system appears to be crew-constrained rather than truck-constrained. The addition of an extra truck leads to an extra first shift crew in most instances, with slightly higher productivity and slightly lower cost (less than a 1 percent reduction). The amount of uncollected refuse changes very little for solutions near the optimum.
2. Manpower costs are minimized with 19 crews (see Figure 15 for 15 trucks). With too few crews, there is no way to collect all refuse without incurring costly overtime pay premiums. With 18 crews, there are over 20 crew-shifts of overtime, yielding an average of more than a 6-day week for each shift, plus 9 or 10 crew- shifts of Sunday work. With too many crews, slack appears, and tons of refuse collected per crew declines. Notice how flat the manpower cost curve of Figure 15 is between assigned crews of 19 and 21.
3. Past the optimum cost point associated with 19 crews, increased manpower cost must be traded off against the social cost of having uncollected refuse. Total uncollected refuse, as a function of the number of crews for 15 trucks, is shown in Figure 16. Uncollected refuse begins to level out at 21 crews. Since the manpower cost curve is relatively flat between 19 and 21 crews, it probably would be worthwhile to hire 21 crews in this district. For only $25 more per week, uncollected refuse could be reduced from 245 to 80 tons. However, to reduce uncollected refuse to 62 tons would cost an additional $349 per week.
4. If the objective function were changed from a weighted cost to the minimization of uncollected refuse, uncollected refuse could be reduced to 70 tons with 21 crews and 15 trucks; however, manpower costs would increase by $241 per week.
5. Uncollected refuse also could be decreased by modifying the three-day-weekend rule. However, since little or no refuse is picked up on Sunday, Monday is usually the peak day of the week in terms of refuse to be collected. By changing to a Friday-Saturday-Sunday weekend, the uncollected refuse could be reduced considerably.

Miller and Burgess [1975] used a simulation model to evaluate alternative schedules, equipment-crew configurations, and policies for uncontained refuse collection in the city of Phoenix. Uncontained refuse consists of items that are too large for containment, such as tree trimmings and discarded hot-water heaters. In general, two methods were used to collect uncontained refuse: one used a hand crew consisting of two men, a truck, and a trailer; the other used three men, two trucks, two trailers, a mechanical loader called a "termite," and a truck and trailer for transporting the termite.

The simulation model reflected: the probability that a given house would generate trash in a given weight-range; probability distributions on interhouse

Figure 15. Manpower cost as a function of number of crews assigned to a district (M = 15 trucks)

SOURCE: E. Ignall, P. Kolesar, and W. Walker, "Linear Programming Models of Crew Assignments for Refuse Collection," P-4717, New York City Rand Institute, November 1971.

travel-time; pickup-time trips to one of four land-fills; and a variety of other aspects of the operating system.

The simulation model was used to evaluate beweekly versus monthly collection cycles, decision rules on crew operation in relation to equipment use, and the effects of seasonality.

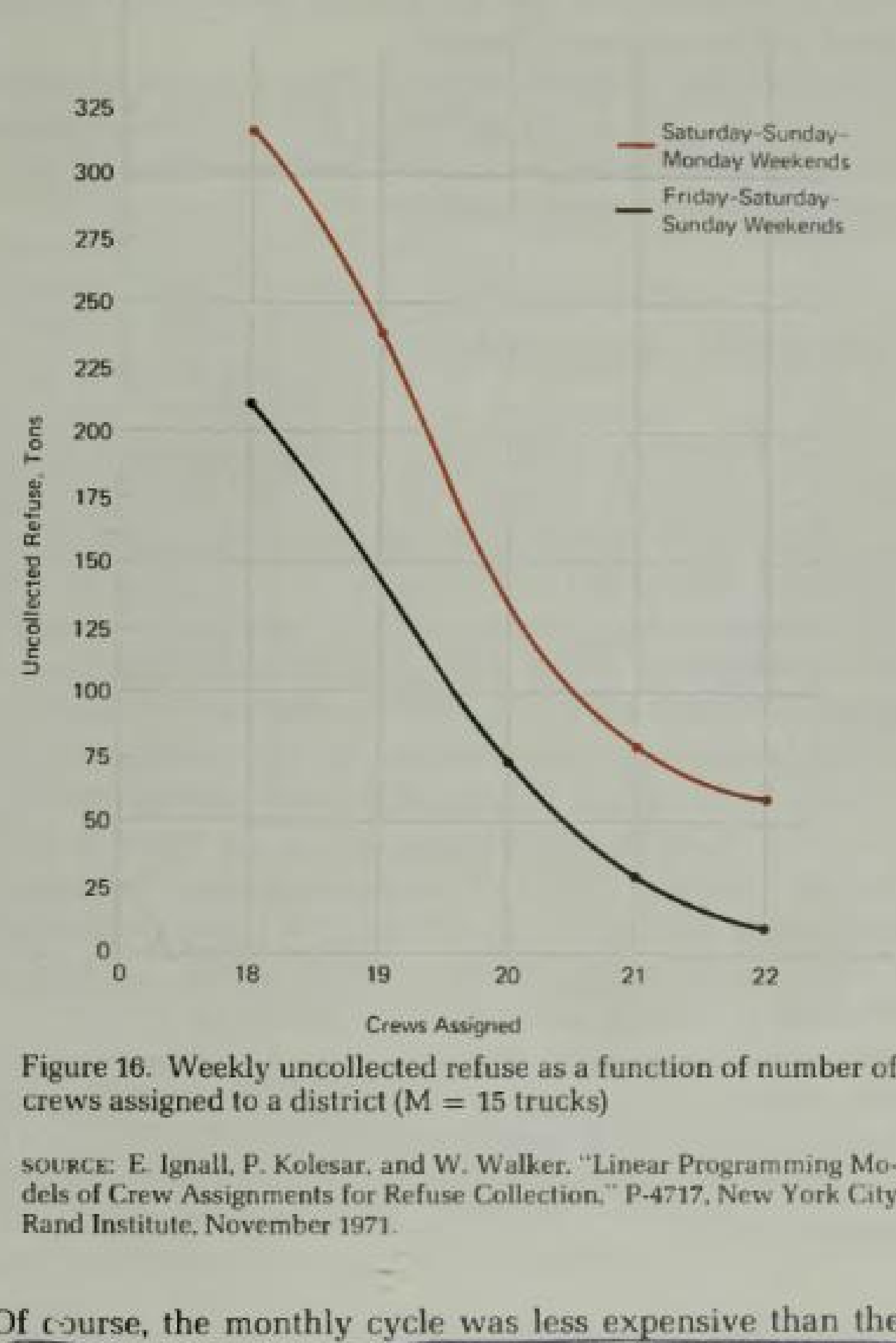


Figure 16. Weekly uncollected refuse as a function of number of crews assigned to a district (M = 15 trucks)

SOURCE: E. Ignall, P. Kolesar, and W. Walker, "Linear Programming Models of Crew Assignments for Refuse Collection," P-4717, New York City Rand Institute, November 1971.

Results. Of course, the monthly cycle was less expensive than the biweekly cycle. However, the simulation model provided the magnitude of the cost differences. Three sources of savings were:

1. Interhouse travel was reduced by almost 50 percent. Interhouse travel accounted for 1.1 to 2.5 hours of the work day.
2. As the days between collection increased, the ratio of tons per route-mile increased, resulting in a higher number of productive or collection hours per week—an economy of scale.
3. The increase in days between collection allowed for a 15 percent reduction in equipment costs and a 12 percent reduction in personnel costs.

The simulation model indicated that the monthly collection cycle would produce an average cost reduction of 13 percent.

The results also indicated that the termite crews were more effective than the hand crews, saving an additional 34 percent of the cost. The changes in schedule, equipment–crew configuration, and policies resulted in an annual cost-reduction of $400,000 for the city of Phoenix.

Other Service System Applications

There are a number of other service systems that have comparable planning and control problems: for example, the post office, bus and rail transportation service, libraries, the criminal justice system, and various kinds of medical service.

Like hospitals and medical services in general, the post office has come under fire because of rising costs and declining service. Since post office activities are labor-intensive, productivity in the post office has not changed greatly, even though salaries have increased by 60 percent in the past ten years or so. The post office is faced with severe load variation; the daily volume may vary by as much as a factor of 8. In addition, because of day-end mailing practices, 40 to 60 percent of the daily "collection and acceptance" mail is received between 4:00 P.M. and 6:00 P.M. These and other complications produce manpower-deployment problems. Hardy, Krajewski, Ritzman, and Vitt [1972] examined the broad spectrum of planning and control activities in the post office and McBride [1972] reviewed the status of research into post office operations.

Morse [1967] addressed some of the planning and control problems of rail-transportation service, and Reitman [1971] presented a computer simulation predicting the performance of a rail system. Blumstein and Larson [1969] presented an analysis of the criminal-justice system, and Morse [1967] presented library operations models.

NONMANUFACTURING SYSTEMS

Like service systems, the nonmanufacturing systems that we shall discuss have no finished product inventories to buffer the productive system from demand variations. Our classification is guided mainly by the relative immediacy of the demand for service; the dividing line is an arbitrary one.

Research and Development

Taubert]1968] developed an aggregate planning cost model for the Search Laboratory in 1968.* The laboratory is housed in a 100,000 square-foot facility and employs a staff of 400. Approximately 300 of the staff are classified as direct technical employees, and the balance function as indirect administrative support for the operations of the laboratory.

The laboratory offers a research capability through its scientific staff and facilities, and widely fluctuating employment could severely impair this capability. The research programs of the laboratory are funded both by the government and the corporation, and an important part of the operating environment is wide fluctuations in government sales and rapid shifts in technology. Thus, the operations planning problem is defined by the need for employment stability on the one hand and wide fluctuations in government sales on the other.

Specifically, the operations planning problem is centered in (1) a monthly decision by the director to determine the size of the scientific staff and administrative staff, as well as (2) the allocation of the scientific staff to government contracts, corporate research programs, and overhead. Overhead charges arise when no contracts or corporate research programs are available for scientists, and they are in addition to the charges that normally are made to overhead for the usual indirect costs. Overhead is used as a buffer to absorb fluctuations in the demand for scientific manpower. The four independent decision variables that are incorporated into the aggregate planning model are as follows:

1. The size of the scientific staff:
   a. WS_t, manpower allocated to government contracts.
   b. WR_t, manpower allocated to corporate research programs.
   c. WO_t, manpower allocated to overhead.
2. WI_t, the size of the administrative support staff.

Cost Model. Figure 17 shows the twelve cost relations that form the components for the cost model of the Search Laboratory. A variety of mathematical relationships are included, such as linear, piecewise linear, constraints, and nonlinear forms. Taubert also built into the model a complete set of equations representing the overhead cost structure that is used to compute the overhead

*The materials dealing with the research laboratory are a direct application of the SDR aggregate and disaggregate planning methods discussed in Chapter 10, but here they are applied to a system without inventories.

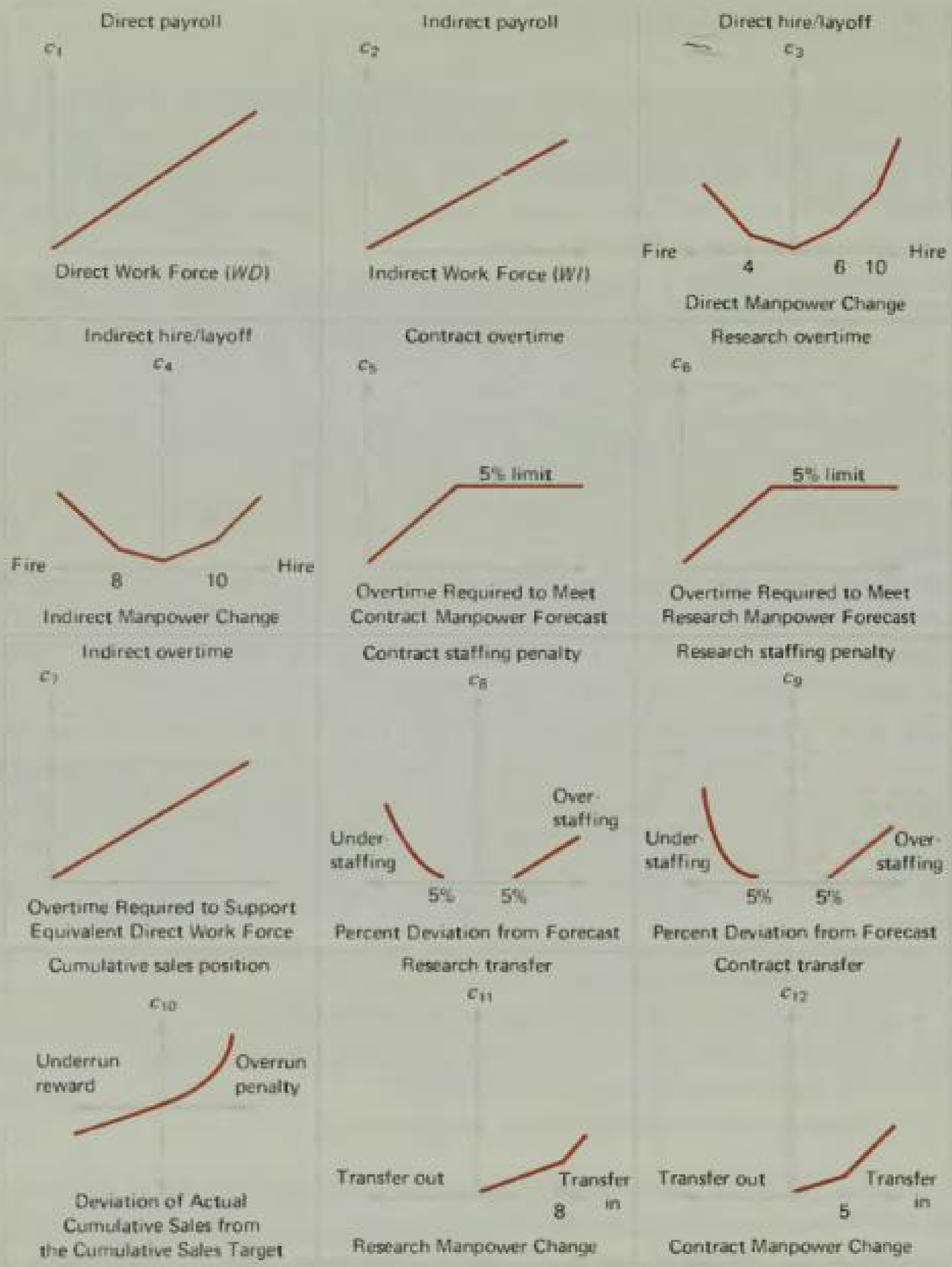


Figure 17. The twelve cost relationships entering the Search Laboratory cost model

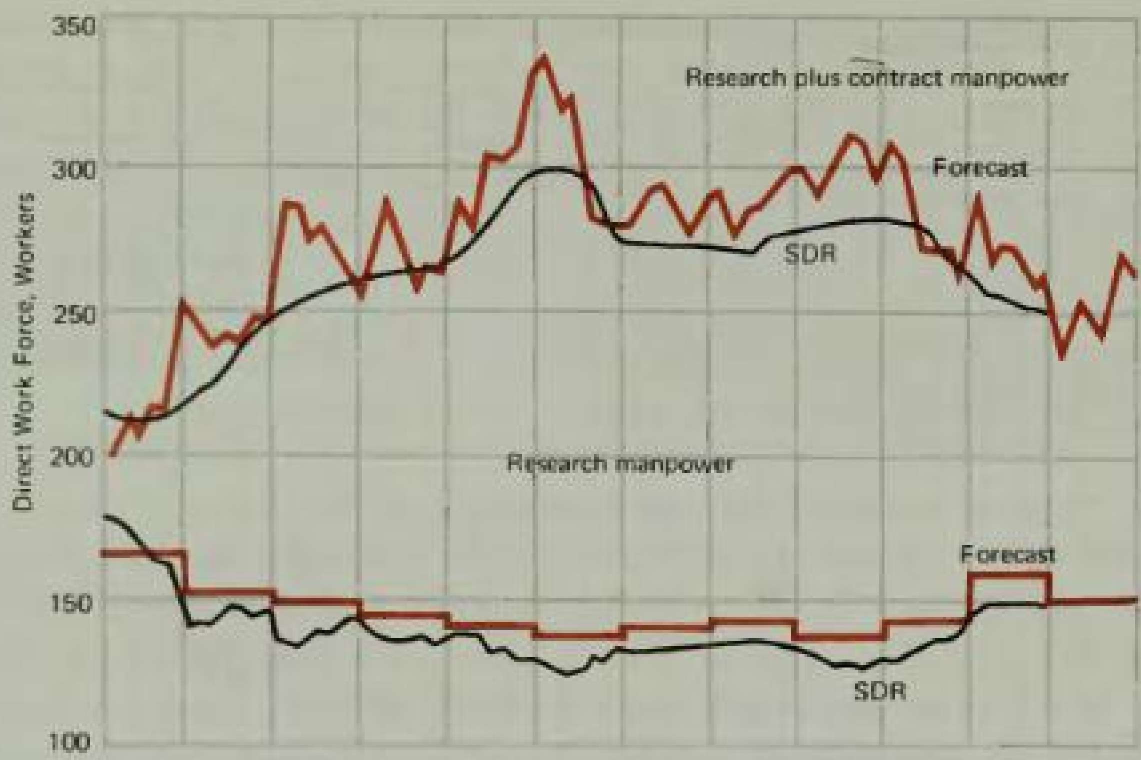
SOURCE: W. H. Taubert, "The Search Decision Rule Approach to Operations Planning" (Ph.D. diss., University of California, Los Angeles, 1968).

rate for any given set of decision variables. The overhead rate then is used to compute the monthly government sales volume, which, in turn, is compared with a cumulative sales target. The inputs to the decision system are monthly forecasts of contract manpower, research manpower, and overhead manpower, and a cumulative sales target that represents the financial plan of the laboratory. The total manpower forecast must be met, and it is part of the director's operations planning problem to determine the best combination of decision variables to accomplish the objective. Failure to meet the manpower requirements increases costs, and this effect also is implemented in the cost model.

Results. Taubert validated the cost model against the financial record of the laboratory over a 5.5 year period. Following the validation, the decision system was operated for each month in the 5.5-year test period. A 6-month planning horizon was used that required SDR to optimize a 24-dimensional response surface (4 decisions per month for a 6-month planning horizon). Figure 18 summarizes the comparative results: Figure 18*a* contrasts SDR decisions on contract-and-research manpower and forecasts, and Figure 18*b* shows a similar comparison of actual management decisions compared with forecasts. The SDR decisions respond much more smoothly to fluctuating manpower forecasts than do actual management decisions.

The costs resulting from the comparison of SDR decisions with actual management decisions indicate that SDR would have produced cost savings. Over the 5.5 year test period, the SDR advantage ranged from a high of 19.7 percent to a low of 5.2 percent, averaging 11.9 percent over the entire test period. The SDR decisions produced lower overhead rates and significant cost reductions in direct payroll, research-programs staffing, sales-target penalities and rewards, and direct hiring costs. It achieved these results largely through more extensive use of overtime.

Following the successful application of SDR to the aggregate planning problem of the Search Laboratory, Taubert disaggregated the decision variables for the size of the scientific staff in six departments. In effect, each department was considered as a miniature laboratory and had its own contract-and-research manpower forecast as a cumulative sales target. Scientific staff had to be allocated to each of the six departments. Some transferring of personnel between departments was allowed, but this practice was limited to represent the fact that scientists generally are not interchangeable, and that they cannot be shifted readily from one department or area of technical expertise to another simply to meet fluctuating manpower requirements. Thus, residual departmental manpower adjustments had to be handled by hiring and layoff decisions. Again, it



(*a*) Comparison of SDR decisions with manpower forecasts

(*b*) Comparison of management decisions with manpower forecasts

Figure 18. Results of SDR decisions for Search Laboratory I compared to forecasts and actual management decisions

SOURCE: W. H. Taubert, "The Search Decision Rule Approach to Operations Planning" (Ph.D. diss., University of California, Los Angeles, 1968).

appeared that SDR was able to produce logical and realistic decisions. The SDR decisions followed the forecast with a smooth response; when faced with downturns in manpower forecasts, these decisions carried members of the scientific staff in overhead for short periods of time.

Higher Education Systems of Planning and Control

In recent years, higher education has been viewed as an operational system and has been the subject of a great deal of research and application. On the broadest level, a federal planning model has been developed to analyze the accessibility to higher education [Huckfeld, 1973]. This model tries to relate the flows of various kinds of funds, the economy, and higher education costs to the demand for higher education and the supply of enrollment spaces.

At the level of university campus operation, Clark, Huff, Haight, and Collard [1973] developed a resource requirements prediction model that uses a cost simulation approach to analyze various institutional alternatives for utilizing limited resources. The computer program generates four types of reports: (1) line-item budgets for various organizational units within an institution; (2) program budget reports that indicate departmental contributions to various instructional programs; (3) summary reports for the institution; and (4) display reports for all parameter data.

Another campus simulation model was developed in Canada by Judy and Levine [1965]. This model builds up the instructional work load for each department for each simulated year, and then it computes required resources, processing (in sequence) enrollment formultaion, resource loading, space requirements, and budgets.

Schroeder [1974] developed a goal programming model for university planning. Lee and Moore [1974] developed an admissions planning model, in the form of goal programming, to optimize conflicting goals for the admission of various categories of students. Lee and Clayton [1972] also developed a goal programming model for academic resource allocation. An excellent survey of work in education is given by Schroeder [1973]. These works either focused on a broad, nationwide educational system level or on the institutional or campus level. The following material deals with the aggregate and detailed planning and scheduling problems of a school of management.

Aggregate Planning in an Academic Environment. Let us examine the broadly based operational plans for a school of management, as developed by Feinberg

[1972] and Geoffrion, Dyer, and Feinberg [1972]. The faculty of the school may be viewed as engaging in several primary activities: formal teaching, school service (for example, administration, curriculum development, etc.), research, and other activities such as student counseling. The formal teaching occurs at three levels: graduate, lower division undergraduate, and upper division undergraduate. The basic planning unit of output is the equivalent course section and all activities are related to that equivalence. "Course releases" are given for administrative activities, curriculum development, research, etc., so that the overall allocation of faculty effort can be planned in terms of equivalent course-section capacity.

Figure 19 summarizes an aggregate planning and scheduling model for such a school. The major features of the model are: the identification of the variables under school control; the identification of the variables that are not controllable by the school; policies, procedures and restrictions; and the multiple criteria and output by which resulting aggregate schedules are judged. The capacity to carry

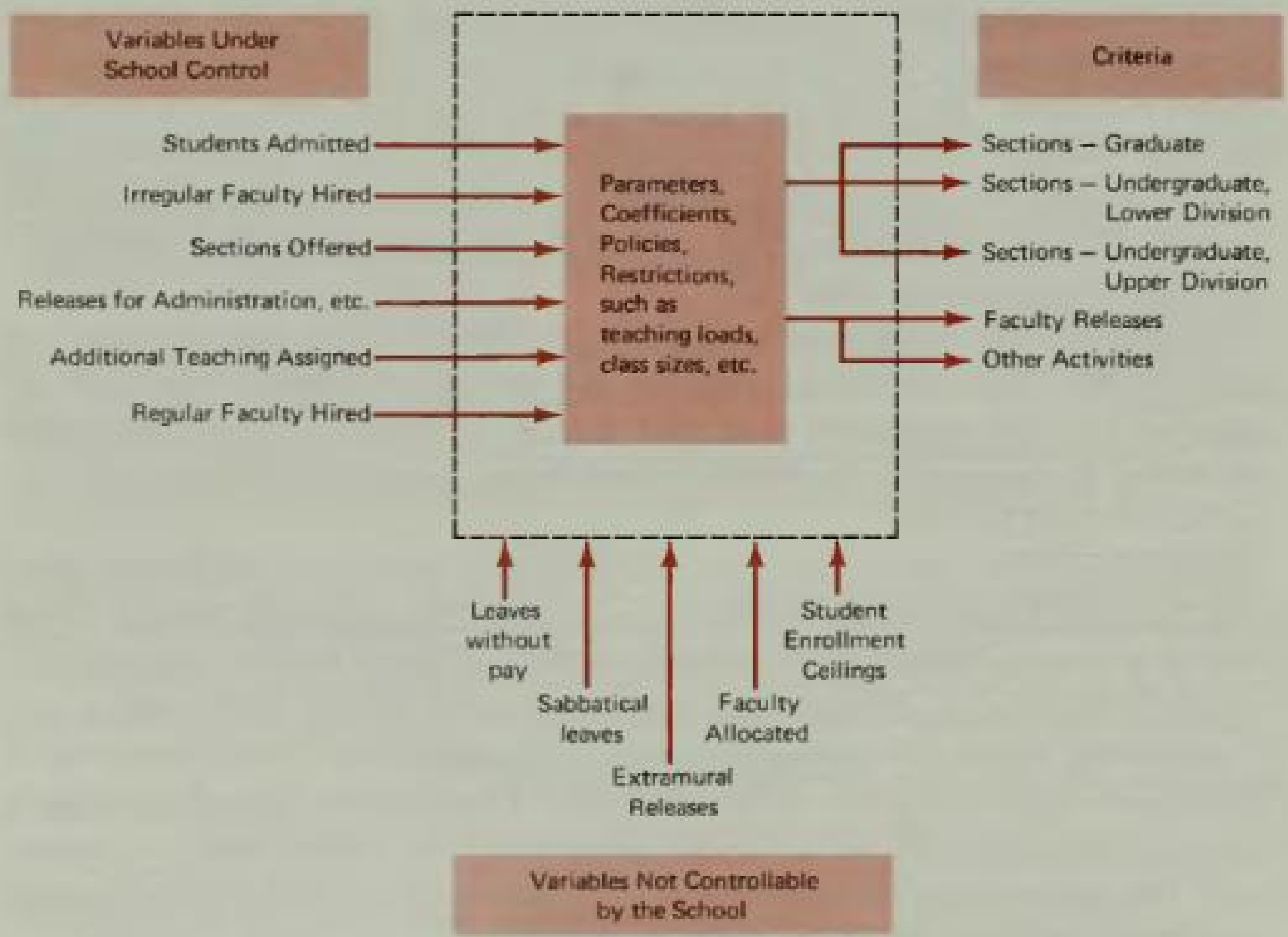


Figure 19. Aggregate planning model for a School of Management

ADAPTED FROM: A. Feinberg. "An Experimental Investigation of An Interactive Approach for Multi-Criterion Optimization With An Application to Academic Resource Allocation" (Ph.D. diss., University of California, Los Angeles, 1972).

on the various activities is complicated by the differing roles of the various kinds of faculty. For example, lecturers provide only teaching service; they are assigned nine quarter courses per academic year, but they are given no administrative, committee, or other assignments. Regular faculty teach five quarter courses per year, but they carry on other activities, such as counseling, committee service, and research. Teaching assistants assist in courses and sometimes teach sections of courses in the ratio of one full-time assistant for six equivalent course sections. University rules permit only faculty with certain qualifications to teach graduate level courses, upper division undergraduate courses, and so on.

Thus, administrators are forced to think in terms of various decision alternatives when they have to fill faculty vacancies and decide on the other variables under control. The number of students admitted can be controlled; the aggregate number of course sections can be varied; and the amount of equivalent, nonteaching, activities can be varied.

The central university administration determines certain upper limits of capacity through its budget allocations for faculty and student enrollment. Leaves of absence and extramural releases reduce the capacity that is available in the planning period.

The criteria and output listed on the right of Figure 19 are the most unique concept of the school model. These outputs represent the aggregate schedule for different levels of courses to be offered, plus faculty time spent in ways other than teaching. However, they are also the dimensions of a multiple-criterion problem, which define the decision maker's preference for a given mix of activities. The decision model is actually interactive between the decision maker and a computational model (in which the decision maker estimates weights for his/her preference for each of the outputs). Given these weights, a mathematical programming algorithm allocates available faculty time to (1) courses at the several levels, (2) releases, (3) research, and (4) additional activities. The computer output of the results then is examined, and the weights are altered if necessary.

The weights express the decision maker's preference for the relative emphasis between criteria—they are his/her estimates of marginal substitution rates between each pair of ~~criteria.~~ For example, they express how many sections of lower division courses the decision maker is willing to give up for an additional graduate section. The approach to the problem attempts to incorporate within the quantitative model dimensions of the multiple criteria for decisions, as diagrammed in Chapter 2, Figure 2. Given the broad allocations of faculty effort in the aggregate schedule, detailed schedules for actual faculty course assignments, as well as correlated classroom and other facility schedules, must be made. Dyer [1973] developed an interactive computer program to help solve the type of multiple-criteria problems encountered in the aggregate planning problem for the school of management.

Detailed Planning and Scheduling. Given the broad allocations of faculty effort in the aggregate schedule, we must develop detailed schedules for actual course assignments as well as for correlated classroom and other facility schedules. Here again, a computer interactive program approach has been effective. Dyer and Mulvey [1976] formulated the problem in terms of a network of teachers, quarters or semesters, courses, programs, etc., with constraints on individual faculty load and course offerings. The general network representation is shown in Figure 20 and the general solution strategy is shown in Figure 21. Individual professors provide data regarding their preference for teaching certain courses, and the entire system is designed to maximize faculty preferences in a network optimization program.

Figure 20 represents the optimization program as a network flow model. The flow on the arcs of the network is in "course section equivalents," one unit of which is defined as the time and effort equal to the actual teaching of one course section. The network nodes either are faculty- or course-related. For each faculty member, there are up to three nodes (in the UCLA format) corresponding to the annual fall, winter, and spring schedules, respectively. There are similar sets of nodes for courses.

The flow on the arcs indicates the lower and upper bounds on teaching for each faculty member. For example, the arc connecting the source and Node 1 is marked with (5, 5), indicating the lower and upper bounds on the annual teaching load in numbers of quarter courses for Instructor 1. Since the lower and upper bounds are the same in this instance, Instructor 1 will be scheduled for exactly five quarter courses. Leading from the node for Instructor 1 are nodes for the three quarters, and the arcs again are marked with lower and upper bounds on the number of courses. Instructor 1's preference is to teach no less than one nor more than two courses in any given quarter.

There are similar sets of nodes for the courses. For example, for course MGT 111, at least one section must be offered in the fall quarter with an upper bound of four sections. In the winter quarter, exactly four sections must be offered; flowing out of the course node, we note that five sections must be offered during the academic year.

The format of the program report has been designed so that we can call for reports as needed; for example, by quarter, by instructor, by curriculum, or by program.

The general solution strategy is summarized in Figure 21, and is described step by step as follows (the step numbers are indicated in Figure 21:

S = Spring
W = winter
F = Fall

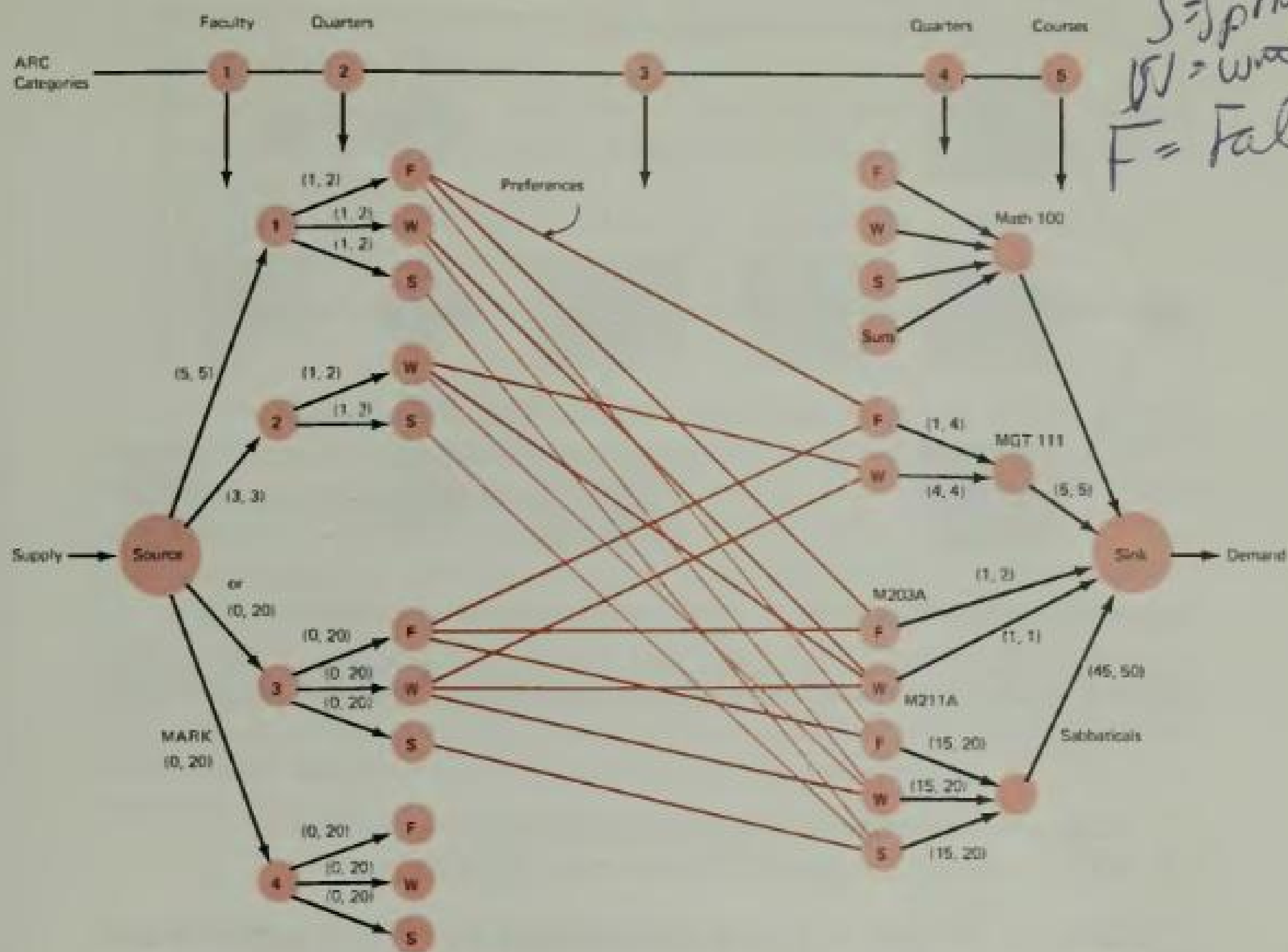


Figure 20. Network representation of the faculty course-scheduling problem

SOURCE: J. S. Dyer and J. M. Mulvey, "An Integrated Optimization/Information System for Academic Departmental Planning," *Management Science* 22(August 1976): 1332–1341.

1. Generate the data required for the approximate model. These data include faculty and course information that have been modified to reflect the administrative policies.
2. Generate the network model and solve for an approximate solution.
3. Determine whether the candidate schedule is a feasible alternative. If not, return to Step 1.
4. Determine whether the candidate schedule is acceptable. If not, go to Step 6.
5. Print the schedule.
6. Decide whether to persist in attempting to improve the candidate schedule by hand. If not, go to Step 8.

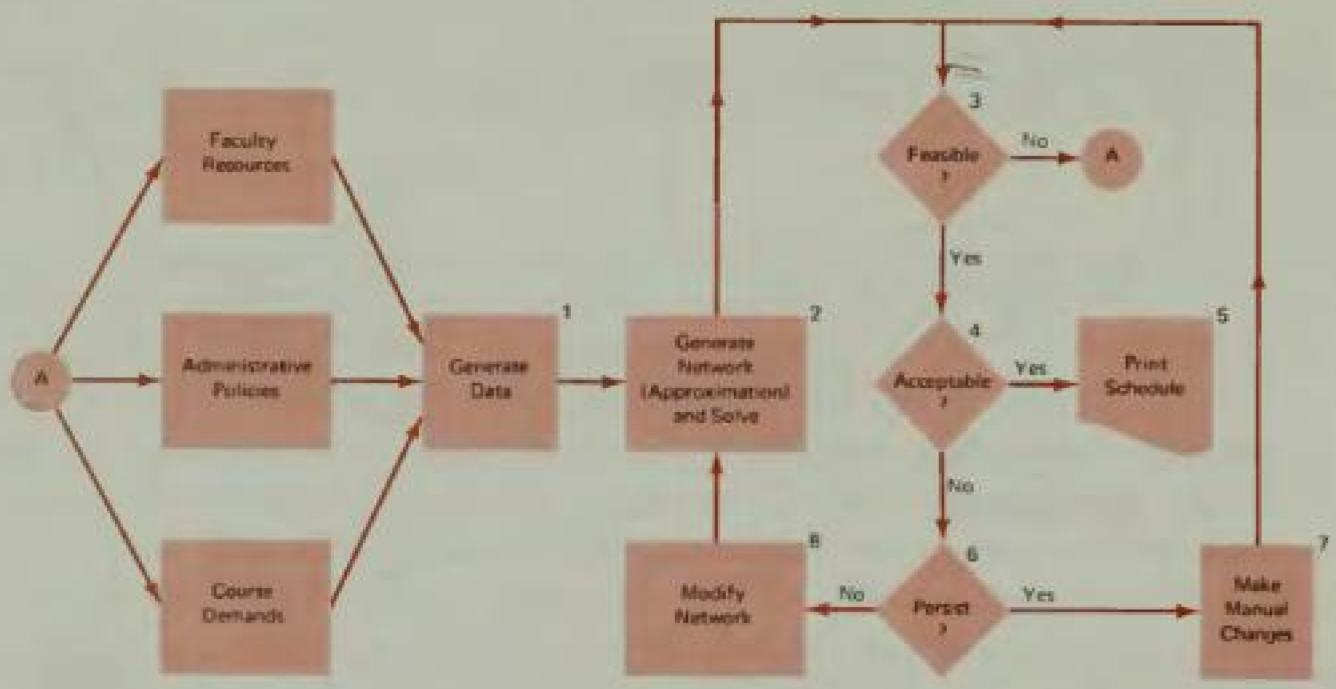


Figure 21. Solution strategy for the faculty course-scheduling problem

SOURCE: J. S. Dyer and M. J. Mulvey. "An Integrated Optimization/Information System for Academic Departmental Planning." *Management Science* 22(August 1976): 1322–1341.

7. Make manual changes in the candidate schedule. Go to Step 3.
8. Make changes in the network formulation. Go to Step 3.

This system was used as a basis for scheduling the 1974–75 academic year. Since the program is interactive, preliminary runs were made, and the results were circulated to curriculum chairmen, program directors, and other administrative personnel. Curriculum chairmen and others consulted with individual faculty members etc. Inputs from these consultations went into the next iteration, and the process was repeated. A solution was decided on after several iterations, and a schedule that still could be changed was circulated to the faculty. Little change was necessary, however, largely because of the program's faculty preference optimizing nature, and because of the interactive nature of the process.

Harwood and Lawless [1975] developed a faculty teaching schedule program in the context of goal programming. In order to identify goals that faculty members considered important in assigning teaching schedules, selected faculty members were polled, and a consensus was generated around the following seven goals (listed in rank order):

1. Consistent observance of expected teaching load (number of courses each term).

2. Granting of individualized requests (no classes at particular times, teaching of particular classes, etc.).
3. Classes assigned so that least possible number of preparations is required.
4. Shortest possible teaching day (least possible time lapse between earliest and latest classes).
5. Maximum assignment of two sections of the same course (for variety).
6. Least possible number of teaching days.
7. Assignment of night classes so that the number of nights assigned is minimal.

Of course, there is some conflict among the preceding goals. When absolute goal attainment in a decision process is impossible, it is common to attempt to minimize the deviation from goals; this minimizing of deviation is exactly what is accomplished through the goal programming technique.

The model requires that some variables be integer values. It takes as inputs the faculty complement and a schedule of classes offered; constraints are specified by the list of goals. The objective function contains weights for each goal's position in the ranking. The solution procedure successively seeks the achievement of goals in their order of priority; in the typical goal programming format, higher priority goals are constraints that cannot be violated.

Planning and Control in Public Utilities

Electric power generating utilities have moved quickly to use interactive planning and control systems. The nature of these planning systems is illustrated by a set of interconnected models that represent the operation of the utility. Figure 22 shows a common modular structure in the industry, in which the following models are interconnected: rate and revenue; operating costs; finance; construction; and corporate consolidation. Operating plans are developed by inputting forecasts and raising "what if" questions through the adjustment of model parameters, in interactive mode (see Ogden [1972] and Dietz and Scavullo [1974].

Employment Planning in Public Utilities. Krajewski and Thompson [1975] developed an aggregate employment planning model for public utilities and applied it within the specific setting of a telephone company. The model has important implications both for management and the public. The implications for management involve improving resource management and taking advantage

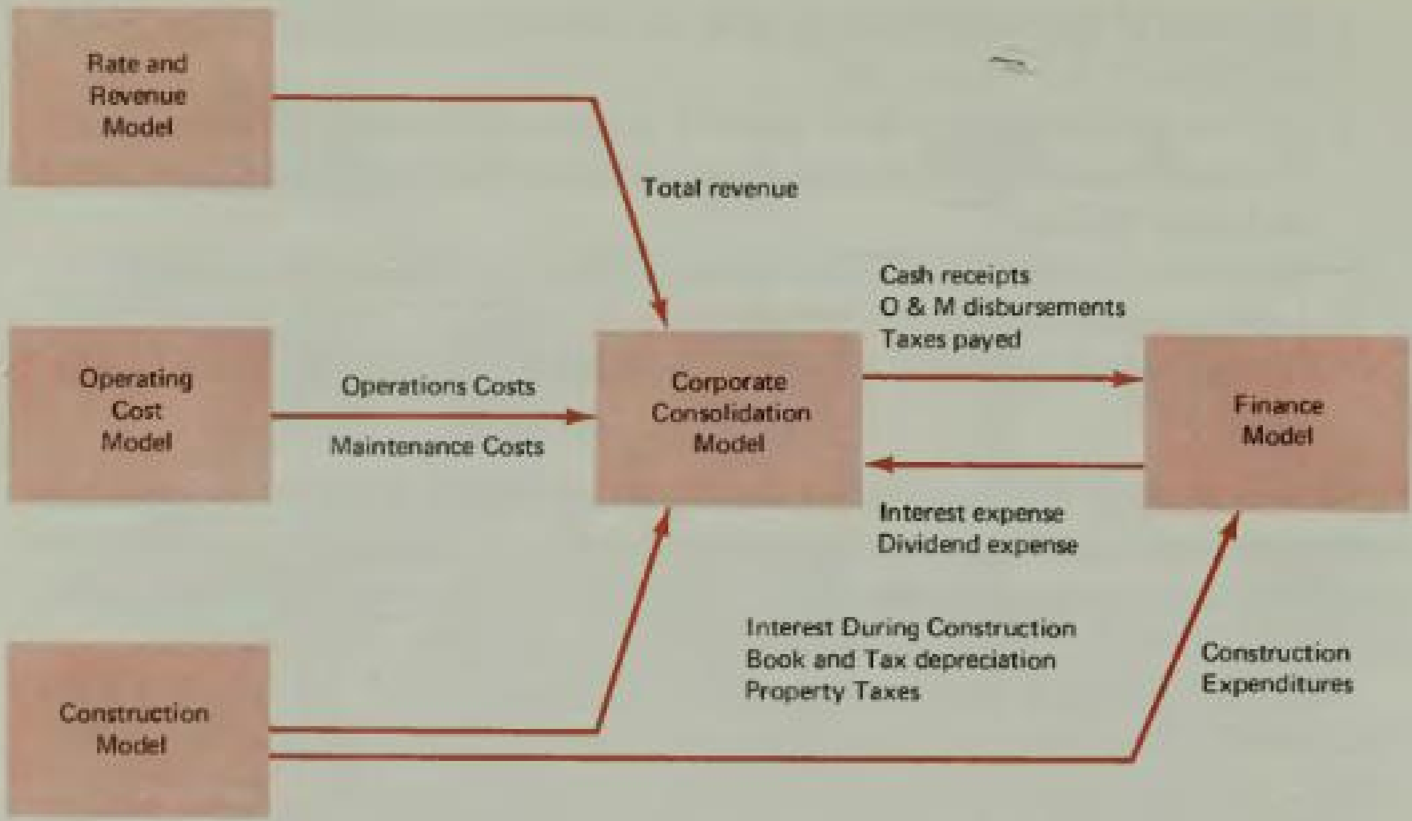


Figure 22. Corporate planning system model linkage for electric power generation utilities

SOURCE: J. Ogden, "What Happens If: A Planning System for Utilities," *Public Utilities Fortnightly*, March 1972 (Courtesy On-Line Decisions, Inc.)

of the infrastructure of costs to maximize returns. The implications for the public involve the allowable prices (rates) they must pay.

In the specific setting of a telephone company, Krajewski and Thompson developed a linear programming model that covered the plant personnel required to install, remove, relocate, repair, and maintain all equipment used for telephones (ranging from the telephone itself to cable and wires necessary for cross-country transmission). Four classes of employees were involved: PBX repair workers, installer-repair workers, cable splicers, and line workers. In general, the four work-groups were regarded as crafts; the only interchangeability was that line workers also could make telephone connections.

The demand for service is subject to some seasonality for installation and removel of phones (summer peak), and weather conditions can produce seasonal demands for service and cable installation and splicing. On the other hand, telephone repair service is closely related to the number of phones in service. Because of the seasonal nature of the aggregate demand, decisions involving the size and timing of hirings, layoffs, and use of overtime work and part time employees are important. In making employment decisions, management can use the available flexibility as a buffer with which to approach seasonality, since the system's output cannot be stored. In dealing with demand variations, management controls two additional variables. However, both variables interact with the employment decisions: they can vary the level of service; and, for construction work they can subcontract.

Krajewski and Thompson developed a linear programming model to minimize aggregate employment related costs over a planning horizon. The costs were: full- and part-time wage costs for regular and overtime work; hiring costs; layoff costs; and subcontracting costs. Within the model, hiring costs were segregated for replacement, a rehire of a person laid off, and hiring of personnel for expansion. The latter required additional capital investment, which, in turn, had rate base implications. The model also contained constraints on employment, production overtime, and backlog or service delay.

Management can use the employment planning model to prepare annual budgets for the plant department and to determine the monthly operating guidelines for subcontracting as well as for hiring, layoff, and overtime according to labor class.

Some of the most interesting results of the Krajewski-Thompson study concern the service delay-versus-cost relationships and the rate of return-versus-cost to public relationships. In a simple example, they showed that the service delay cost curve was quite shallow; the annual cost difference between 0 and a 5-day delay was only $11,200. The reason was found in the model structure, since the service delay–cost tradeoff occurred only through overtime and hiring costs, which were a small fraction of total employment costs.

An expansion of the work force to accommodate demand increases requires supporting capital investment. This investment goes into the *rate base* on which the company is allowed to earn a return. Therefore, Krajewski and Thompson defined an additional cost to expand, involving equity rate of return and depreciation. This cost enters into the balance of costs for employment planning and helps create an incentive to expand the work force in such capital-intensive work groups as line workers. In Krajewski and Thompson's example, for a demand growth of 10 percent, the model with a rate of return yields a 31 percent larger line-worker work force for the same service than when the capital cost is excluded. The net result is a greater total cost of $5,300, which is passed on to the consumer for the same service level. Thus, there is an incentive for the company to expand the line-worker work force and use less overtime and subcontracting in order to optimize its rate base.

WORK SHIFT SCHEDULING

One of the problems that emerges from the study of service and nonmanufacturing systems is the scheduling of manpower to provide service for a highly variable work load (often for twenty-four hours per day, seven days per week). This happened with the scheduling of physicians in the university health service outpatient clinic, the scheduling of nurses, and the scheduling of emergency systems. The work shift scheduling problem particularly concerns the scheduling of telephone operators because of the severe work load variations and the continuous service policies. Recently, considerable work has been done on this problem, and Chapter 15 discusses this work.

Summary

In some manufacturing systems, we can produce products during a seasonal low that are to be used during peak demand periods. Unlike inventoriable products, however, services cannot be stored; therefore, we are somewhat restricted in our ability to smooth the load on a service facility. Conceptually, in service activities, we hold the capability to give service in readiness, to use when it is needed. It is the capability that is stored, rather than the output.

In some instances we may be able to get the customer, patient, or client to do something productive while he/she is waiting. If we can transfer some of the service activity to the person being serviced, costs may be reduced. This may constitute one of the few strategies available for improving productivity in service activities. The idea of allowing the customer to do part of the required service work has become quite accepted in self-service markets, self-service gas stations, cafeterias, etc.

While waiting line models in themselves do not solve complex practical problems, they provide the concepts and insights that enable us to solve such problems. As modules, these models can be used in larger simulation studies. Thus, simulation is an important practical tool for service system design. The examples of the university outpatient clinic study and the hospital system study illustrated this importance. Simulation methods can represent the existing complexities and accept empirical arrival- and service-distributions as well.

Service systems have the added complication of variations in demand (hourly, daily, weekly, etc.) without inventories as a buffer. Emergency systems such as police, fire, and ambulance service systems represent the extreme case, in which virtually no buffering is possible, and the strategies are based on optimum deployment. Health care systems achieve equivalent buffering by scheduling voluntary arrivals to the hospital, to manage the patient-mix. Attempts to deal with the nurse scheduling problem currently involve integrated systems of planning and control that depend on (1) forecasting of load, and (2) buffering, by means of flexible staffs to be assigned where load peaks occur. Mathematical programming allocation systems appear to offer considerable promise.

Operations planning and control in some nonmanufacturing systems (such as the research and development laboratory) are a direct application of industrial methodology, such as SDR. Proposals for aggregate planning in education systems have focused on a methodology for trade-offs between competing values and criteria within the model. Detailed scheduling in education systems has taken the form of network optimization in an interactive process, and goal programming.

An important problem that has emerged from the study of service systems has been the work shift scheduling problem. Although this is common in many service situations such as telephone exchange operators and airline scheduling, it also exists in continuous-manufacturing situations that operate around the clock, but are manned by people working forty hours per week.

Review Questions

1. What general effects result from the fact that we cannot store the output of service and nonmanufacturing systems?

2. Using as an example the outpatient clinic discussed at the beginning of the chapter, what would you say are the unique aspects of service systems, from an operations point of view?

3. How does the service facility module differ from the simple input-process-output block that we used before to describe operations?

4. What mechanisms does the hospital administrator have for smoothing work load?

5. Compare aggregate planning in hospitals with the industrial aggregate planning discussed in Chapter 10.

6. Appraise hospital planning practices in terms of their effectiveness for solving the nurse scheduling problem.

7. What flexibility is available to nursing staff supervisors to meet fluctuations in day-to-day load?

8. What are the effects of pooling nursing resources for two or more wards?

9. Why was it not possible to use the waiting line formulas to help solve the problems of the outpatient clinic?

10. Why is queuing felt to be an unacceptable method for dealing with the overload conditions in medical systems?

11. In the hospital admissions system, what function does the "call list" have? What are the waiting line implications of these facts:
    a. The study of work load variations were not correlated.
    b. When the data from the two wards were combined, the variance was reduced by 30 percent.

12. Account for the large cost differences between actual and optimal schedules in Warner and Prawda's mathematical programming application to nurse scheduling.

13. Characterize the distribution of call for service in ambulance systems. What kinds of variation occur, and how can they be described?

14. How is it possible to provide service capacity to meet the demand for ambulance service as described by Figure 11?

15. Since the demand for police service varies significantly over the day (as shown by Figure 13), what are the possible ways that this short-range variation can be absorbed? Compare the demand for service in hospitals and police protection (Figures 6 and 13). Do the managers of each of these kinds of systems have equivalent mechanisms for controlling the demand or "arrival" process?

16. In the Fetter–Thompson simulation study of hospital operations, what was the general effect of increasing load on facility utilization?

17. In the Fetter–Thompson simulation study, what was the value of induction as a means of leveling facility requirements?

18. Figure 14 shows both daily expected variation and variation between days for fire alarms in New York City. What are the possible ways that the fire department can remain in control and provide the needed service?

19. What are the advantages and disadvantages of traditional deployment policies in fire departments?

20. Describe the manpower replanning problem in the Food and Drug Administration. What are the advantages of the "automated planning tool" in comparison with the FDA's previous planning methods?

21. In the municipal refuse-collection study by Ignall, Kolesar, and Walker:
    a. Which was the constraining resource?
    b. What was the optimum number of crews? Would there be advantages in operating with more crews than those that were defined by the optimum?
    c. What changes in policy would result if the criterion were changed to minimizing uncollected refuse?

22. In the operations planning model for a research laboratory, what is the mechanism for absorbing fluctuations in the demand for scientific manpower?

23. In the operations planning model for a research laboratory, what would be the advantages of disaggregating the decision variables for the size of the scientific staff in each of the six departments?

24. In the aggregate planning model for a school of management, what is the mechanism that allows the decision maker to make trade-offs among conflicting objectives? What are the buffers that can be used to absorb student load fluctuation?

25. In the detailed planning model for a school of management, what alternate criteria might have been used instead of maximizing faculty preferences?

26. In the goal programming study of faculty teaching schedules, six of the seven goals were satisfied in the example. Do you think that it is typical for nearly all the conflicting goals in a resource scheduling problem to be

satisfied simultaneously? Why do you think it was true in the teaching schedule situation?

27. In the employment planning model for public utilities, what buffer does management have for dealing with the seasonal fluctuations in demand?

28. In the public utilities employment planning model, what incentive does management have for expanding work force in capital intensive work groups? Is there a seeming conflict of interest?

29. What is the work shift scheduling problem?

References

Abernathy, W. J.; N. Baloff; and J. C. Hershey. "The Nurse Staffing Problem: Issues and Prospects." *Sloan Management Review* 13(Fall 1971): 87–109.

______; ______; ______; and S. Wandel. "A Three-Stage Manpower Planning and Scheduling Model—A Service-Sector Example." *Operations Research* 21(May–June 1973): 693–711.

Adam, E. E., Jr.; W. L. Berry; and D. C. Whybark. "The Hospital Administrator and Management Science." *Hospital Administration* 19(Winter 1974).

Baker, K. R. "Scheduling a Full-Time Workforce to Meet Cyclic Staffing Requirements." *Management Science* 20(August 1974): 1561–1568.

Blum, E. H. *Development Research of the New York City Fire Project.* R-968. New York: Rand Institute, May 1972.

______. *Urban Fire Protection: Studies of the Operations of the New York City Fire Department.* R-681. New York: Rand Institute, January 1971.

Blumstein, A., and R. Larson. "Models of a Total Criminal Justice System." *Operations Research* 17(March–April 1969): 199–232.

Brigham, F. "On a Congestion Problem in an Aircraft Factory." *Operations Research* 3(1955): 412–428.

Carter, G.; J. M. Chaiken; and E. Ignall. *Response Areas for Two Emergency Units*. R–532–NYC/HUD. New York: Rand Institute, March 1971. Also published in *Operations Research* 2(May–June 1972): 571–594.

———, and E. Ignall. *A Simulation Model of Fire Department Operations: Design and Preliminary Results*. R–632–NYC. New York: Rand Institute, December 1970.

Chaiken, J. M., and R. C. Larson. "Methods for Allocating Urban Emergency Units: A Survey." *Management Science* 19. Part 2 (December 1972): 110–130.

Church, J. G. "Sure Staf: A Computerized Staff Scheduling System for Telephone Business Offices." *Management Science* 20. Part 2 (December 1973): 708–720.

Clark, D. G.; R. A. Huff; M. J. Haight; and W. J. Collard. *Introduction to the Resource Requirements Prediction Model 1.6*. Boulder, Colo.: National Center for Higher Education Management Systems at Western Interstate Commission for Higher Education, 1973.

Conway, R. H. *Operations Research and Hospital Design*. 40th National ORSA Meeting, Honolulu, Hawaii, October 1971.

Dietz, R. V., and R. V. Scavullo. "Industrial Progress: A System to Develop Utility Systems Planning." *Public Utilities Fortnightly* (May 9, 1974).

Drake, A. W.; R. L. Keeney; and P. M. Morse. *Analysis of Public Systems*. Cambridge, Mass.: MIT Press, 1972.

Dyer, J. S. "A Time-Sharing Computer Program for the Solution of the Multiple Criteria Problem." *Management Science* 19(August 1973): 1379–1383.

———, and J. M. Mulvey. "An Integrated Optimization/Information System for Academic Departmental Planning." *Management Science* 22(August 1976): 1332–1341.

Feinberg, A. "An Experimental Investigation of an Interactive Approach for Multi-Criterion Optimization with an Application to Academic Resource Allocation." Ph.D. dissertaion, University of California, Los Angeles, 1972.

Fetter, R. B., and J. D. Thompson. "The Simulation of Hospital Systems." *Operations Research* 13(September–October 1965): 689–711.

Fitzsimmons, J. A. "Emergency Medical Systems: A Simulation Study and Computerized Method for the Deployment of Ambulances." Ph.D. dissertation, University of California, Los Angeles, 1970.

______. "A Methodology for Emergency Ambulance Deployment." *Management Science* 19(February 1973): 627–636.

Flagle, C. D. "Operations Research in the Health Services." *Operations Research* 10(September–October 1962): 591–603.

______. "The Problem of Organization for Hospital Inpatient Care." *Management Sciences: Models and Techniques*. Vol. 2. Paris: Pergamon Press, 1960. Pages 275–287.

______, and J. P. Young. "Application of Operations Research and Industrial Engineering to Problems of Health Services, Hospitals, and Public Health." *Journal of Industrial Engineering* 17(November 1966): 609–614.

Geoffrion, A. M.; J. S. Dyer; and A. Feinberg. "An Interactive Approach for Multi-Criterion Optimization, with an Application to the Operation of an Academic Department." *Management Science* 19 (December 1972): 357–368.

Hardy, S. T.; L. J. Krajewski; L. P. Ritzman; and L. D. Vitt. "Production Planning and Control Activities in the Post Office." *AIIE Technical Papers*. Atlanta: *American Institute of Industrial Engineers*, 1972.

Harwood, G. B., and R. W. Lawless. "Optimizing Organizational Goals in Assigning Faculty Teaching Schedules." *Decision Sciences* 6(1975): 513–524.

Hershey, J.C.; W. J. Abernathy; and N. Baloff. "Comparison of Nurse Allocation Policies—A Monte Carlo Model." Research paper no. 98. Stanford University, Graduate School of Business, August 1972.

Horvath, W. J. "Operations Research in Medical and Hospital Practice." In P. M.

Morse, *Operations Research for Public Systems*. Cambridge, Mass: MIT Press, 1967.

Huckfeld, V. *A Federal Planning Model for Analysis of Accessibility to Higher Education*. Boulder, Colo.: National Center for Higher Education Management Systems at Western Interstate Commission for Higher Education, 1973.

Ignall, E.; P. Kolesar; and W. Walker. *Linear Programming Models of Crew Assignments for Refuse Collection*. P-4717. New York: Rand Institute, November 1971. Also published in *IEEE Transactions on Systems, Man and Cybernetics* (November 1972): 661–666.

Judy, R. W., and J. B. Levine, *A New Tool for Educational Administrators*. Toronto: University of Toronto Press, 1965.

Kolesar, P. *Mathematical Programming Applications in the Analysis of the Deployment and Utilization of Fire-Fighting Resources*. P-4988. New York: Rand Institute, March 1973.

———, and E. H. Blum. "Square Root Laws for Fire Engine Response Distances." *Management Science* 19(August 1973): 1368–1378.

Kolesar, P.J., K. L. Rider, T. B. Crabill, and W. E. Walker, "A Queuing-Linear Programming Approach to Scheduling Police Patrol Cars." *Operations Research* 23(November–December 1975): 1045–1062.

———, and W. E. Walker. "An Algorithm for the Dynamic Relocation of Fire Companies." *Operations Research* 22(March–April 1974): 249–274.

Krajewski, L. J., and L. P. Ritzman. "Dissaggregation in Manufacturing and Service Organizations: Survey of Problems and Research. *Decision Sciences* 8(January 1977).

———, and H. E. Thompson. "Efficient Employment Planning in Public Utilities." *The Bell Journal of Economics and Management Science* (Spring 1975).

Kwak, N. K., P. J. Kuzdrall, and H. H. Schmitz. "The GPSS Simulation of Scheduling Policies for Surgical Patients." *Management Science* 22(May 1976): 982–989.

Larson, R. C. "Improving the Effectiveness of New York City's 911." In A. W.

Drake, R. L. Keeney, and P. M. Morse, *Analysis of Public Systems*. Cambridge, Mass.: MIT Press, 1972.

______. *Urban Police Patrol Analysis*, Cambridge, Mass.: MIT Press, 1972a.

Lee, S. M. "An Aggregative Resource Allocation Model for Hospital Administration." *Socio-Economic Planning Sciences* 7(1973): 381–395.

______, and E. R. Clayton. "A Goal Programming Model for Academic Resource Allocation." *Management Science* 18(April 1972): 395–408.

______, and L. J. Moore. "Optimizing University Admissions Planning." *Decision Sciences* 5(July 1974): 405–414.

Levenson, G. S., and A. J. Tenzer. *The Service Facilities of the Bureau of Fire Communications: A Cost Analysis of a Proposed Consolidation*. RM-5726-NYC. Santa Monica, Ca: The Rand Corporation, September 1968.

Liebman, J. S. *The Development and Application of a Mathematical Programming Model of Personnel Allocation in an Extended Care Nursing Unit*. Operations Research and Industrial Engineering, Johns Hopkins University, 1971.

Maier-Roth, C., and H. B. Wolfe. "Cyclical Scheduling and Allocation of Nursing Staff." *Socio-Economic Planning Sciences* 7(1973): 471–487.

McBride, C. C. "Post Office Mail Processing Operations." In A. W. Drake, R. L. Keeney, and P. M. Morse, *Analysis of Public Systems*, Cambridge, Mass.: MIT Press, 1972.

Miller, P. J., and J. S. Burgess, Jr. "Simulating a Municipal Problem." *Industrial Engineering* (June 1975): 10–14.

Milsum, J. H.; E. Turban; and I. Vertinsky. "Hospital Admission Systems: Their Evaluation and Management." *Management Sciences* 19(February 1973): 646–666.

Morse, P.M. *Operations Research for Public Systems*. Cambridge, Mass.: MIT Press, 1967.

Ogden, J. " 'What Happens If'—A Planning System for Utilities." *Public Utilities Fortnightly* (March 16, 1972).

Press, J. S. *Some Effects of an Increase in Police Manpower in the 20th Precinct of New York City*. R-704-NYC. New York: Rand Institute, October 1971.

Reitman, J. *Computer Simulation Applications*. New York: John Wiley & Sons, 1971. Chapter 8.

Rising, E. J.; R. Baron; and B. Averill. "A Systems Analysis of a University-Health-Service Outpatient Clinic." *Operations Research* 21(September–October 1973): 1030–1047.

Ritzman, L. P., and L. J. Krajewski. "Multiple Objectives in Linear Programming: An Example in Scheduling Postal Resources." *Decision Sciences* 4(July 1973.

Rosenshine, M. "Randomizing Police Patrols." *Proceedings: Operations Management Workshop*. Fifth National Meeting of the American Institute for Decision Sciences. Boston, November 14, 1973.

Rosenthal, B., and R. C. Murphy. "Instant Manpower Replanning is Here." *State Government Administration* (January–February 1974): 6.

Schroeder, R. G. "Resource Planning in University Management by Goal Programming." *Operations Research* 22(July–August 1974): 700–710.

———. "A Survey of Management Science in University Operations." *Management Science* 19(April 1973): 895–906.

Stevenson, K. A. "Emergency Ambulance Transportation." In A. W. Drake, R. L. Keeney, and P. M. Morse, *Analysis of Public Systems*, Cambridge, Mass.: MIT Press, 1972.

Taubert, W. H. "The Search Decision Rule Approach to Operations Planning." Ph.D. dissertation, University of California, Los Angeles, 1968.

Trivedi, V., and D. M. Warner. "A Branch and Bound Algorithm for Optimum Allocation of Float Nurses." *Management Science* 22(May 1976): 972–981.

Vu, V. T., and W. C. Turner. "Systems Design for Rural Refuse Collection." *AIIE Transactions* 8(March 1976): 84–95.

Ward, R. A. "Operational Research in Local Government." In P.M. Morse, *Operations Research for Public Systems*. Cambridge, Mass.: MIT Press, 1967.

Warner, D. M., and J. Prawda. "A Mathematical Programming Model for Scheduling Nurses." *Management Science* 19(December 1972): 411–422.

Chapter 15

Work Shift Scheduling

Shift scheduling is one of the general problems to emerge from the study of service and nonmanufacturing systems. Essentially, shift scheduling involves scheduling manpower to meet some variable demand pattern (usually short-term, such as daily and/or weekly, but also including seasonal variations). This problem emerges within the context of service and nonmanufacturing systems because such systems are characterized by variable short-term demand that cannot be backlogged, and by a system output that is not inventoriable.

We alluded to such problems in Chapter 14, and the range of applications reported in the literature indicates the importance of the problem. For nurse scheduling, see Abernathy, Baloff, and Hershey [1971]; Ahuja and Shepard [1975]; Maier-Roth and Wolfe [1973]; and Rothstein [1973]. For police protection, see Butterworth and Howard [1973]. For airline reservation offices, see Linder [1969]. For supermarkets and retail stores, see Walsh [1974]. For public transportation, see Bennett and Potts [1968]. For telephone exchange operators, see Burman and Segal [1976]; Church [1973]; Harveston Luce, and Smuczynski [1972]; Henderson and Berry [1976, 1977]; Luce [1973, 1974]; and Segal [1974]. For bank tellers, see Mabert and Raedels [1975]. For the post office, see Ritzman, Krajewski, and Showalter [1976].

Similar problems involve the scheduling of employees who work only a forty-hour week in continuous, seven-day manufacturing operations [Baker, 1974; Rothstein, 1972; Walsh, 1974].

NATURE OF THE OPERATOR SHIFT SCHEDULING PROBLEM

Much of the work that has provided good solutions to the shift scheduling problem has occurred within the context of scheduling telephone traffic exchange operators, because of the severity of the problem in the telephone service industry. As many as 350 operators may be involved in a large telephone exchange. Figure 1 shows the extreme demand variation in a typical operator requirements curve for a large exchange. This figure shows the number of operators needed to meet demand by half-hour time increments. At 7:00 A.M., only 7 operators are needed; but by 10:00 A.M., the load builds up rapidly to a requirement of 90, followed by a decline to 62 by 1:00 P.M. and by another peak later in the afternoon. There is a somewhat lower evening peak, which declines to a low requirement after 1:00 A.M.

Forecasting Demand

Standard forecasting techniques for short-term horizons have been used successfully. Church [1973] used an exponential forecasting model that forecasted call volume for each half-hour of the day and for resulting clerical load on a weekly basis. Harveston, Luce, and Smuczynski [1972] analyzed telephone operator demand, using Box-Jenkins methods to forecast demand each half-hour for a week in advance of General Telephone Company of California. Berry, Mabert and Marcus [1975] have applied exponential smoothing to the forecasting of teller window demand at a branch of the Purdue National Bank.

Work Shift Types

Operators work in shifts characterized both by a start-time and an end-time, which begin and end on the hour or the half-hour. The total shift duration may vary from two to nine hours, but regular length tours are seven, eight or nine hours long and may include one half-hour break and two quarter-hour relief

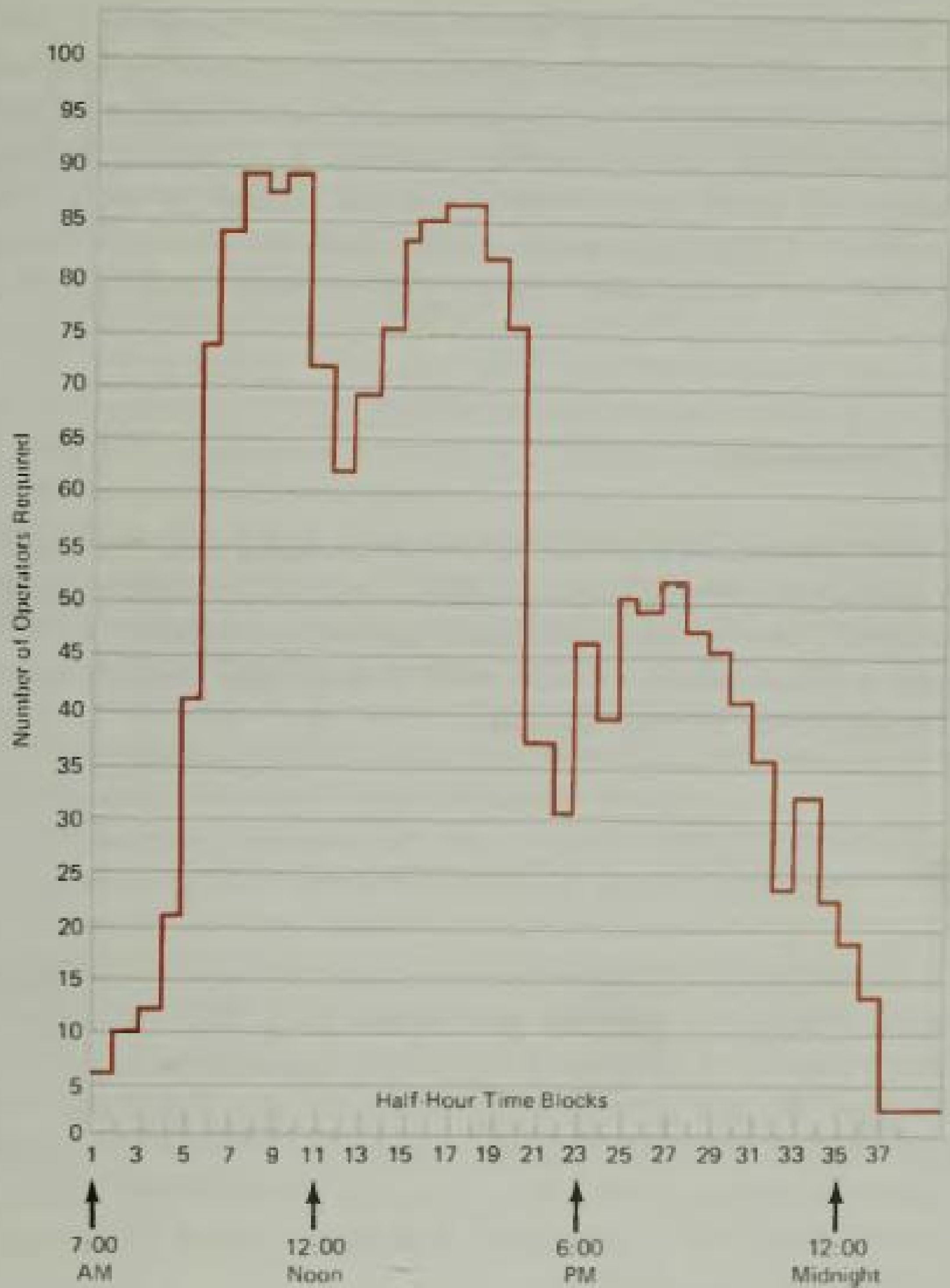


Figure 1. A typical operator requirements curve for a large telephone exchange

SOURCE: M. Segal, "The Operator-Scheduling Problem: A Network Flow Approach," *Operations Research* (July–August 1974): 808–823.

periods. Part time operators may be assigned to a four-hour shift with one quarter-hour relief period, or to a two- or three-hour shift with no relief periods. Split shifts may be used, which consist of two intervals of four hours, each with a quarter-hour relief period that is separated by a two- or three-hour break [Segal, 1974].

Since we define the problem as one of allocating or assigning operators to work shifts in order to meet demand, we then must define a working subset of the preceding types of work shifts. Henderson and Berry [1975] noted that when 60 different 8 hour shift types (involving different work period lengths, relief break timing, and lunch period lengths) can be started in every 15-minute demand period from 6:00 A.M. to 4:00 P.M., the master set of work shift types is a surprising 2,460. We would select the permissible work shift types from the master set.

Problem Formulation

Given demand forecasts and a working subset of shift types, the problem has been formulated as an integer linear programming problem of assigning operators to work shifts at minimum cost, subject to meeting demand in each period. An optimum schedule always would meet demand and provide some slack capacity during certain hours of the day.

SOLUTION TECHNIQUES

At present, integer programming is not a practical technique for solving actual size problems. Therefore, modified procedures have been developed. Segal [1974] of Bell Telephone Laboratories developed a network flow approach that generates integer, but not optimal, assignments. Heuristic solutions were proposed by Harveston, Luce, and Smuczynski [1972] of General Telephone Company of California; by Luce [1973, 1974]; by Church [1973] of Bell Canada; and by Henderson and Berry [1975]. Henderson and Berry [1975a] also proposed an optimal branch and bound solution.

The heuristic approach of Henderson and Berry starts with an integer linear programming formulation that equalizes the cost of assigning operators to shift types, thus minimizing the number of operators rather than the cost. This allows us to compute a theoretical minimum number of operators that is required to meet demand in most cases. We can determine a second theoretical minimum by relaxing the integer restriction on the variables and computing the usual linear programming solution, the results of which are integerized. Henderson and Berry then ran experiments that varied the size of the working subset of shifts, the procedure for selecting the working subset, and the scheduling heuristics. Three scheduling heuristics were used: a linear programming improvement heuristic (LPI), a linear programming vector exchange heuristic (LPVE), and a random start vector exchange heuristic (RSVE).

In two steps, the LPI heuristic transforms the linear programming solution into a feasible solution. First, the noninteger solution variables are raised to the next higher integer value. Next, since the first step produces an excess number of operators, the second step attempts to reduce the number of operators, without creating a solution in which the demand is not satisfied in each period. Starting with the first solution variable, each variable in the solution is reduced by one unit, one at a time. After a variable has been reduced, the solution is checked for feasibility, using the linear programming model. If an infeasible solution results, the variable is returned to its previous solution value before the value of the next variable is reduced.

The LPVE heuristic begins with the solution produced by the LPI heuristic and attempts to improve it by exchanging operators among different types of work shifts. The heuristic considers all possible ways of providing a feasible solution by (1) removing two operators from the solution, and (2) adding one operator. If, at any stage of the solution process, the total number of operators equals the theoretical minimum number of operators produced by the standard linear programming solution, the procedure terminates.

The RSVE heuristic involves two steps. First, a feasible integer solution is generated randomly. Next, the LPVE heuristic is used to improve this starting solution, which involves procedures to guarantee a feasible solution. Beginning with the first demand period, work shifts are selected randomly from the working subset and added to the solution until the demand for the period is satisfied. Next, the supply of operators for the second period is compared with the demand for that period. Additional work shifts are generated randomly until the second period demand is satisfied. This process continues until a complete shift schedule is produced.

Results

The solution quality (percentage above the optimal solution value) of the LPVE heuristic was better than the RSVE heuristic in 83 percent of the problems and better than the LPI heuristic in 48 percent of the problems. When working subsets of fifty or more shifts were selected, slightly better solutions were generated by all three scheduling heuristics, which apparently provided greater flexibility in the scheduling process.

In a later study, Henderson and Berry [1975a] developed a branch and bound algorithm that provides optimum solutions to the integer programming formulation of the work shift problem when the costs of assigning are assumed to be equal. The algorithm was tested on problems that had as many as 100 different shift types and 72 demand periods.

GENERAL TELEPHONE—AN INTEGRATED WORK SHIFT SCHEDULING SYSTEM

The general concepts of work shift scheduling have been applied at the General Telephone Company of California in an integrated, computerized system. The company has used the system since 1973 to schedule approximately 2,600 telephone operators in 43 locations in California. The size of installations ranges approximately from 20 to 220 operators.

The system combines a computerized forecasting system and conversion to operator requirements, the scheduling of tours or shifts, and the assignment of operators to shifts. The resultant system illustrates the company's application of advanced concepts and techniques to solve a difficult problem.*

Demand for Service

The service offered is the telephone exchange: operators are assigned to provide directory assistance, coin telephone customer dialing, and toll call assistance. The standard for service is supplied by the Public Utilities Commission in unusually specific terms: service must be provided at a resource level such that an incoming call can be answered within ten seconds, 89 percent of the time. The difficulty in implementing the response standard lies in the severe demand variability of incoming calls. We indicated this kind of variability in our previous discussion and in Figure 1. Now, however, we will amplify the true nature of the variation.

Figures 2, 3, 4, and 5 show typical call variations during the year, the week, the day, and within a peak hour. Figure 2 shows the annual variation, highlighting the two sharp peaks that include Mother's Day and Christmas Day. The data indicates calls that were made during the busiest hour in each of the 52 weeks. The minimum occurred in the twenty-eighth week (3,200 calls), and the maximum occurred during Christmas (4,400 calls). The peak-to-valley ratio is 1.38.

*The materials in this section are from E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System," *Decision Sciences* 7(October 1976).

Figure 2. Typical distribution of calls during the busiest hour for each week during a year

SOURCE: E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System." *Decision Sciences* 7(October 1976). (Courtesy, General Telephone Company of California)

Translating the seasonal scheduling problem, the company must provide about 38 percent more capacity at Christmas time than in the twenty-eighth week, and in general the summer months involve a somewhat lighter load.

Figure 3 shows daily call load for January 1972 at one location. The weekly pattern is very pronounced, and the Saturday and Sunday call load constitute only about 55 percent of the typical load through the week. While the telephone company offers somewhat lower weekend toll call rates to help smooth the load, the resultant weekly variation still is very large.

Figure 4 shows the half-hourly variation for a typical 24 hour period. Peak call volume is in the 10:30–11:00 A.M. period (2,560 calls), and the minimum occurs at 5:30 A.M. (about twenty calls). The peak-to-valley variation for the typical half-hourly load is 128. Again, the telephone company offers somewhat lower night time toll call rates. Figure 4 suggests the daily problem of scheduling operator shifts to meet the load.

Finally, Figure 5 shows the typical intrahour variation in call load, indicating the number of simultaneous calls by one minute intervals. This variation is random. In fact, continuous tracking of the mean and standard deviation of calls per minute indicates that the standard deviation is equal to the square root of the mean (reasonable practical test for randomness), or that the arrival rates are described by the Poisson distribution. For the sample of Figure 5, $\bar{x} = 15.75$

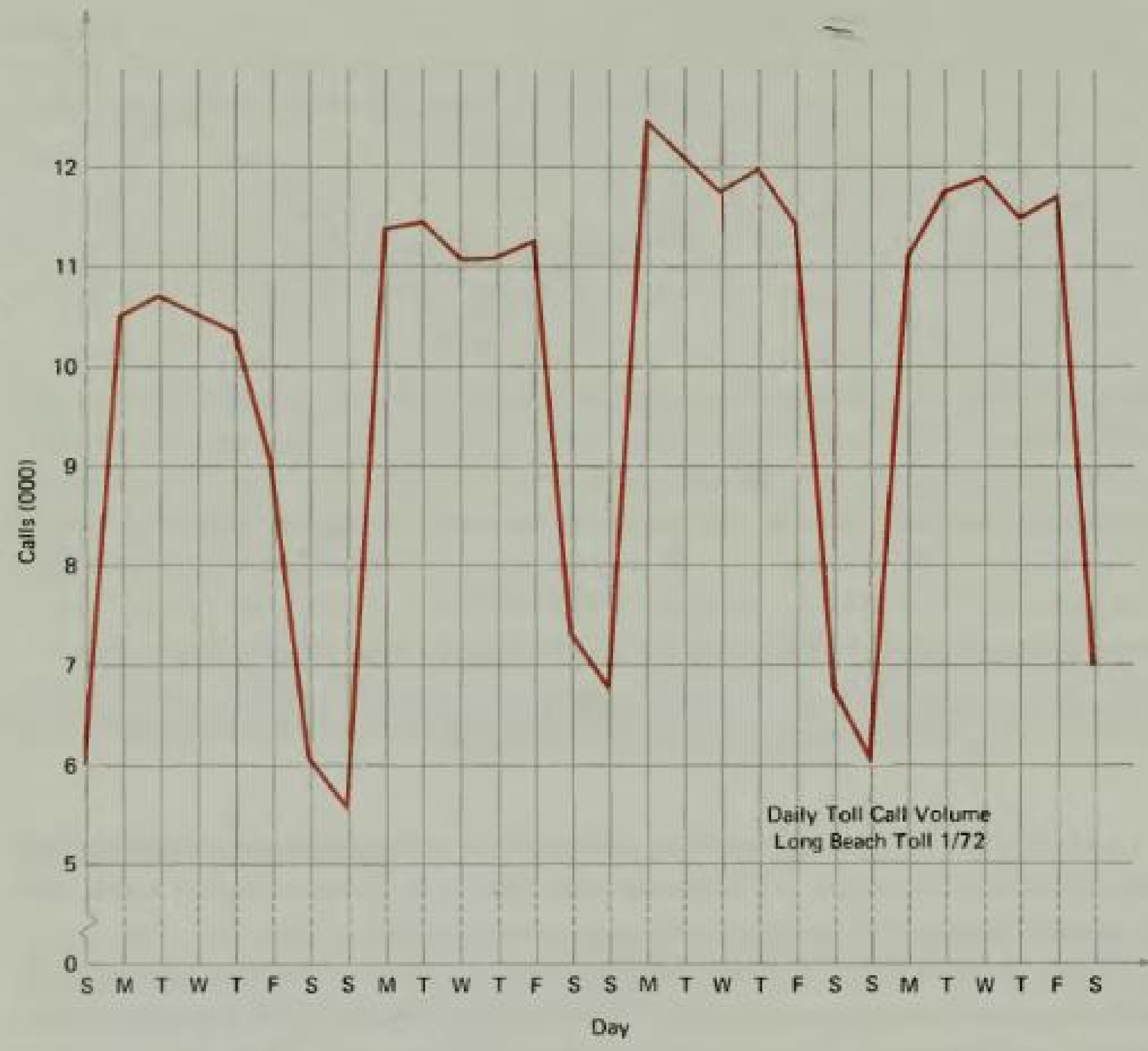


Figure 3. Daily call load for January 1972

SOURCE: E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System," *Decision Sciences* 7(October 1976). (Courtesy, General Telephone Company of California)

calls per minute, $s = 4.85$ calls per minute, and $\sqrt{15.75} = 3.99$. Therefore, the variation within the hour is taken as random. We cannot cope with this variation by planning and scheduling. We must simply accept it and provide enough capacity to absorb the random variations.

The overall situation that results from the typical distributions of Figures 2, 3, 4 and 5 is that a forecastable pattern exists for seasonal, weekly, and daily variation. In addition, the call rate at any selected minute is described adequately by a Poisson distribution.

Figure 4. Typical half-hourly call distribution

SOURCE: E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System," Decision Sciences 7(October 1976). General Telephone Company, 1974.

The Integrated System

Given the description of the demand for service, Figure 6 indicates the system developed at General Telephone. There are basically three cycles of planning and scheduling, which involve information feedback concerning actual experience. The forecast of daily calls is the heart of the system. As we shall see, the forecast considers seasonal and weekly variation as well as trends. The forecast is converted to a distribution of operator requirements by half-hour increments. Based on the distribution of operator requirements, a schedule of tours or shifts is developed, and, finally, specific operators are assigned to tours. This sequence of modules is entirely computerized, as indicated in Figure 6.

Given the operator schedule, there are two additional cycles that operate on a manual basis. First, a schedule for "today" may be affected by unintended

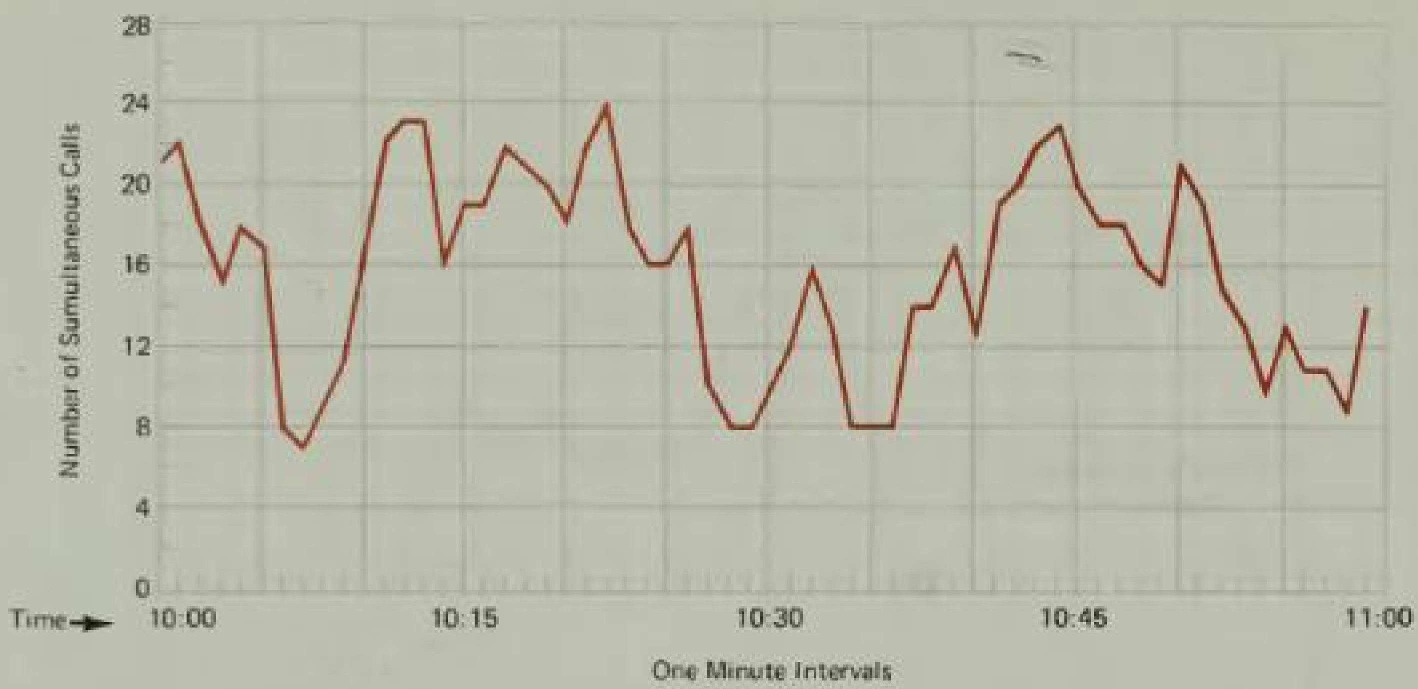


Figure 5. Typical intrahour distribution of calls, 10:00–11:00 A.M.

SOURCE: E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System," *Decision Sciences* 7(October 1976). General Telephone Company, 1974 (Courtesy, General Telephone Company of California)

events, such as operator illness or an emergency increase in call load. Supervisors in local installations cope with such events, and this is the "Intraday Management Cycle" shown in Figure 6. In addition, there is the "Monthly Future Force Cycle," in which management can make higher level adjustments based on reports of actual operations and on forecasts involving particular trend and seasonal factors. The hiring and training of operators is planned in the future cycle, up to 12 months in advance.

Forecasting Demand

The demand forecasting system is based on a Box-Jenkens [1970] model. The system would involve the following major terms if we were attempting to forecast the number of calls at a specific location for next Monday:

$$\begin{aligned}
\text{Calls next Monday} = {} & \text{Calls last Monday} \\
& + \text{Weekly growth at this time last year} \\
& \quad (\text{Monday}_{-52} + \text{Monday}_{-53}) \\
& - \text{error last week} \times \theta \\
& - \text{error 52 weeks ago} \times \phi \\
& + \text{error 53 weeks ago} \times \phi\theta
\end{aligned}$$

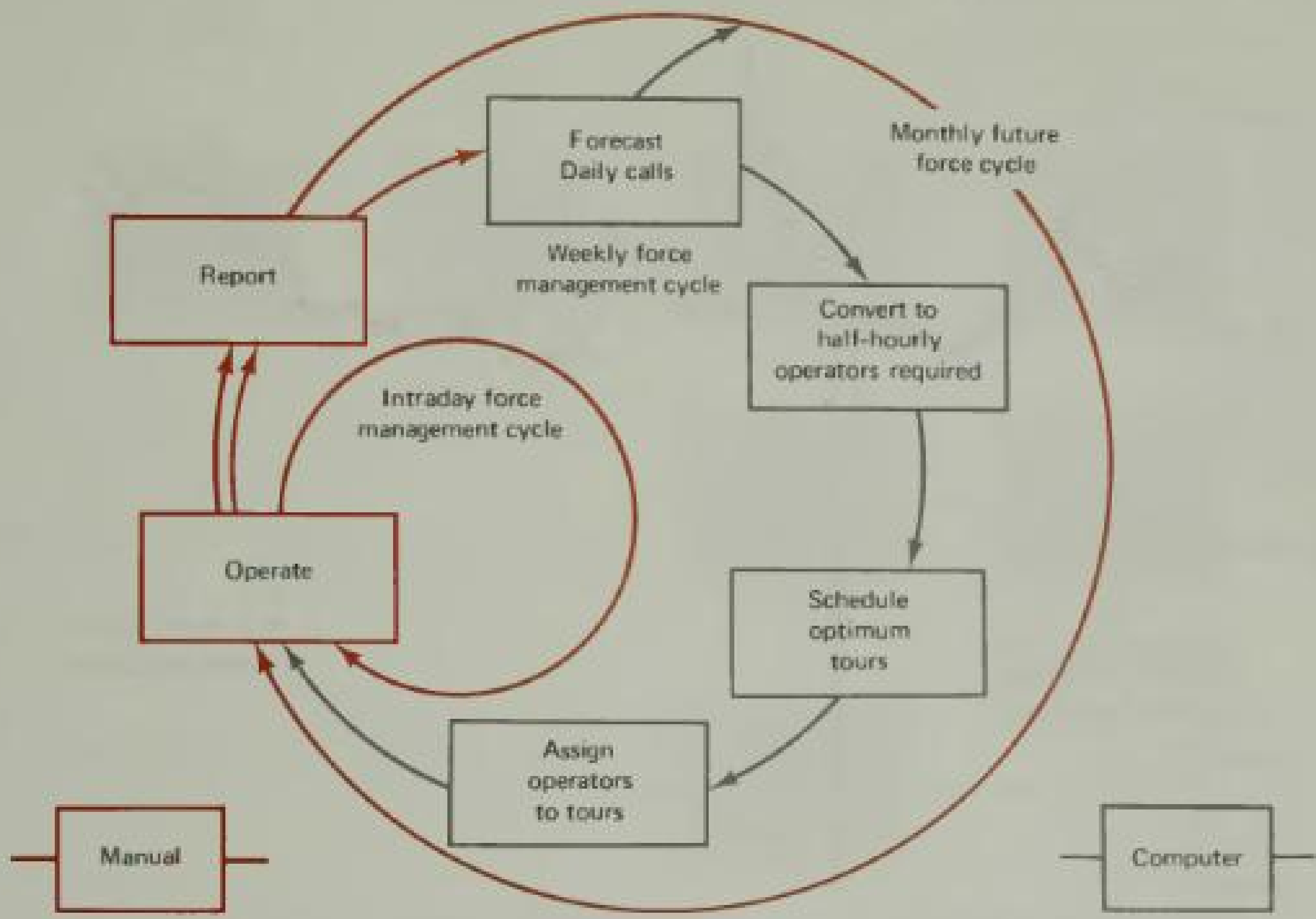


Figure 6. "Force Management System"

SOURCE: E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System," *Decision Sciences* 7(October 1976). General Telephone Company, 1974 (Courtesy, General Telephone Company of California)

where θ is a nonseasonal moving average parameter, and ϕ is a seasonal moving average parameter.

In terms of actual operation, the computer inputs are: last week's calls by day and type of service (toll, assistance, directory service); coefficients (work units per call) for the forecasted week by day and type of service; and board load (productivity) by day for the forecasted week. The computer outputs are forecasts of daily calls for up to five weeks in advance and a translation of the forecast into required board hours by day (also for up to five weeks in advance).

Forecast Errors. Figure 7 shows a typical record of comparison between forecasted and observed numbers of calls for Santa Monica. The uncanny forecast for day 151 is Thanksgiving, and people predictably are more interested in dinner and family affairs than in communication. The average error for the forecasting system as a whole is 3.5 percent.

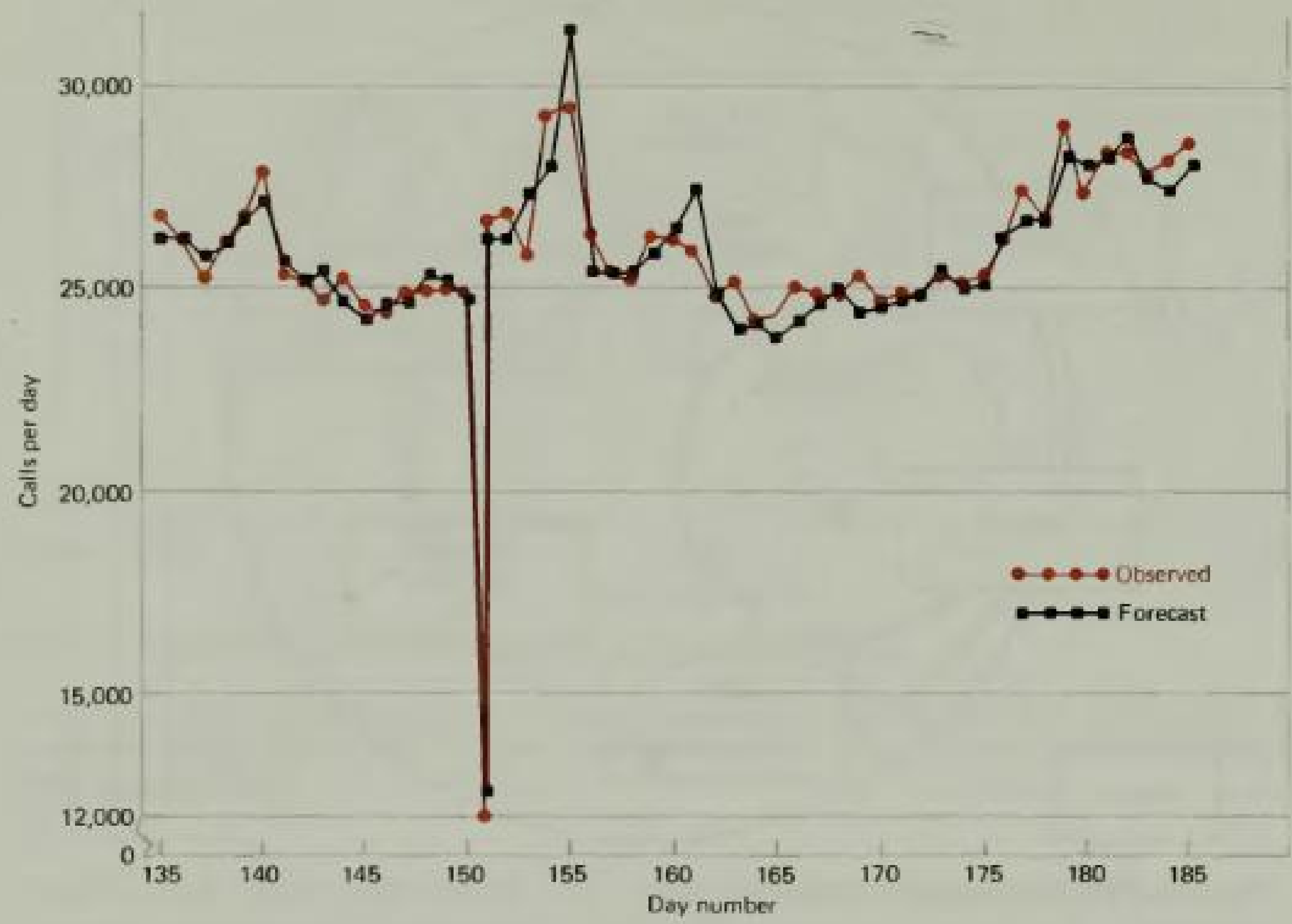


Figure 7. Sample of forecast versus observed numbers of calls at Santa Monica

SOURCE: E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System," *Decision Sciences* 7(October 1976). General Telephone Company, 1974 (Courtesy, General Telephone Company of California)

Conversion to Half-Hourly Operator Requirements

The objective at this point is to produce a daily schedule of operator requirements. The profile formed by the requirements curve is called the "topline," and the program required to generate it is called the "topline program." The program itself produces a printout of half-hourly operator requirements for each day in a week, and Figure 8 shows the formula for the conversion. The parameter that defines the model is average call duration, based on studies of actual times, efficiency, and the response time standard. The response time standard (speed of answer) is a constraint. The result for each day is the information for the topline profile.

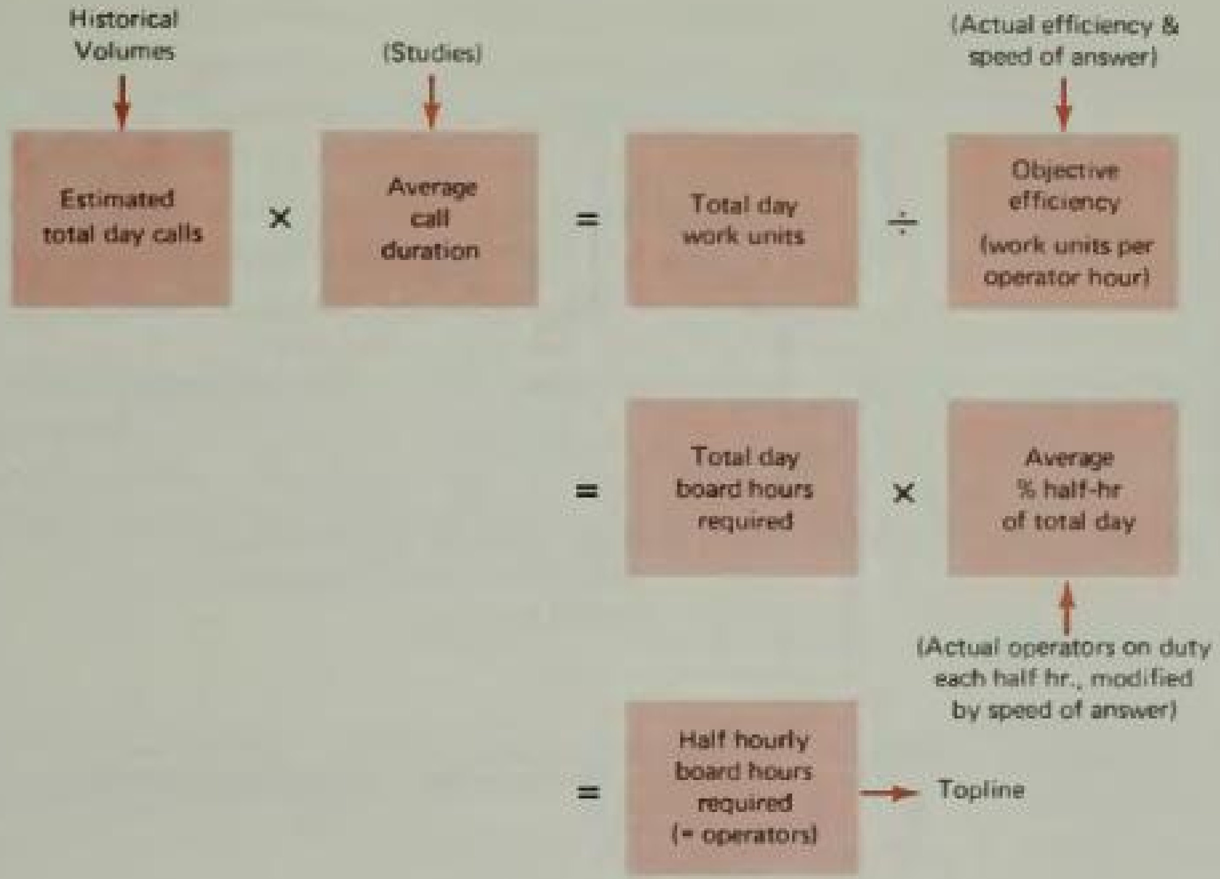


Figure 8. Model for conversion of calls to half-hourly operator requirements (topline)

SOURCE: E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System," *Decision Sciences* 7(October 1976). General Telephone Company, 1974 (Courtesy, General Telephone Company of California)

Actual half-hourly staffing is based on a percent of total daily requirements. Exponential smoothing of each half-hourly percent is used to develop the topline program. A table based on an Erlang queuing model is used to adjust the actual half-hourly staffing to account for the speed of answering.

Scheduling of Shifts

The graphical representation of a topline profile is shown in Figure 9. The problem in assigning tours or shifts involves fitting in shifts so that they aggregate to the topline profile (also shown in Figure 9).

The *Shift Set*. In order to build up shifts so that they aggregate to the topline profile, we need flexibility in shift types; and we get it in the shift lengths and in the positioning of lunch hours and rest periods. A rich shift set provides flexi-

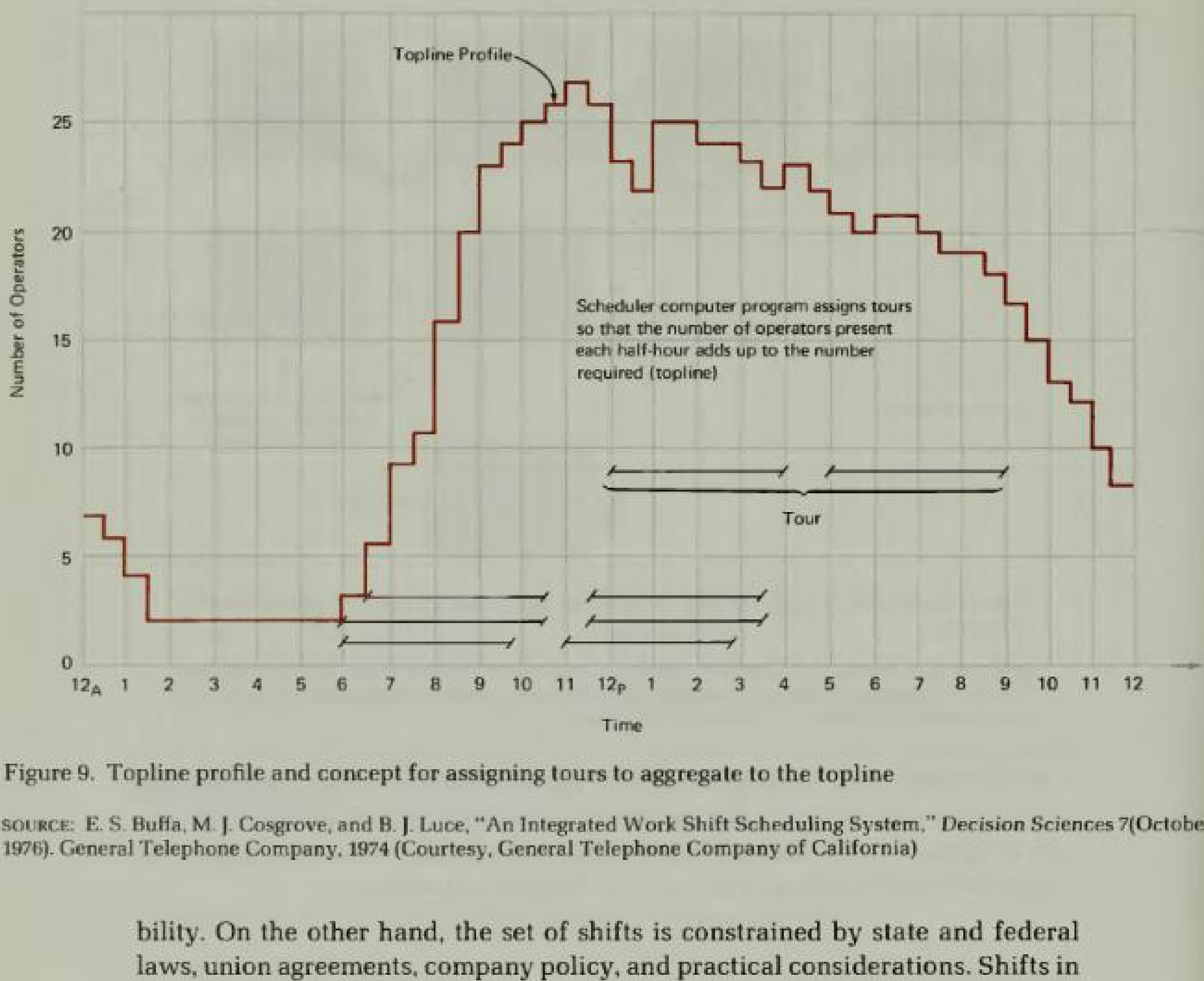


Figure 9. Topline profile and concept for assigning tours to aggregate to the topline

SOURCE: E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System," *Decision Sciences* 7(October 1976). General Telephone Company, 1974 (Courtesy, General Telephone Company of California)

bility. On the other hand, the set of shifts is constrained by state and federal laws, union agreements, company policy, and practical considerations. Shifts in the set are actually selected based on California State restrictions, company policy, and local management input concerning the desireability of working hours by their employees.

Each shift consists of two working sessions separated by a rest period, which may be the lunch period. Each working session requires a 15 minute rest period near the middle of the session. The following rules enumerate the admissible shift set:

1. Shift are 6.5, 7, or 8 hours.
2. Work sessions are in the range of 3 to 5 hours.

3. Lunch periods either are a half-hour or an hour.
4. Split work periods are in the range of 3.5 to 5 hours (split work periods are separated by more substantial nonwork periods).
5. Eight-hour shifts end before 9 P.M.
6. Seven-hour shifts end from 9:30 to 10:30 P.M.
7. Six-and-a-half-hour shifts end at 11:00 P.M. or later.
8. Earliest lunch period is at 10:00 A.M.

The Scheduling Algorithm. Luce [1973] developed a heuristic algorithm for assigning shifts from the approved set so as to minimize the absolute differences between operators demanded by the topline profile in period i, D_i, and the operators provided, W_i, when summed over all n periods of the day; that is:

$$\text{Minimize} \sum_{i=1}^{n} \left|D_i - W_i\right| \tag{1}$$

The strategy is to build up the operator resources in the schedule, one shift at a time, drawing on the universal set of approved shifts. The criterion stated in Equation 1 is used to choose shifts at each step. As the schedule of W_i values is built up, conceptually we attempt to minimize the distance between the demand and the number of operators supplied curves, as illustrated by Figure 10.

At each stage in building up the schedule, some remaining distance between D_i and W_i exists. The criterion for the choice of the next shift is the following test on each alternative shift: add the contributions of the shift to W_i (1 for all working periods and 0 for idle periods, such as lunch and rest periods), and recalculate Equation 1. Choose the shift that minimizes (1). In order to counteract the shorter length shifts that the preceding rule would favor, weight the different shorter shifts by the ratio of the working times. Thus, if the longest shift is 8 hours, then a 7 hour shift would be weighted $8/7 = 1.14$.

As the number of time intervals and shift types increase, the computing cost increases. Luce states that computing costs are moderate when the number of time intervals is less than 100 and the number of shifts is less than 500, and he presents additional refinements that we shall not discuss.

As indicated in Figure 10, the final profiles for D_i and W_i do not coincide perfectly in any real case.

Computer Output. Operators provided by the algorithm will be slightly greater or less than the demand, and the aggregate figures are a measure of the effectiveness of a given schedule. For example, Figure 11 is sample output, and the aggregate statistics are given at the head: total hours required = 295; total hours scheduled by half-hours = 296.50; etc.

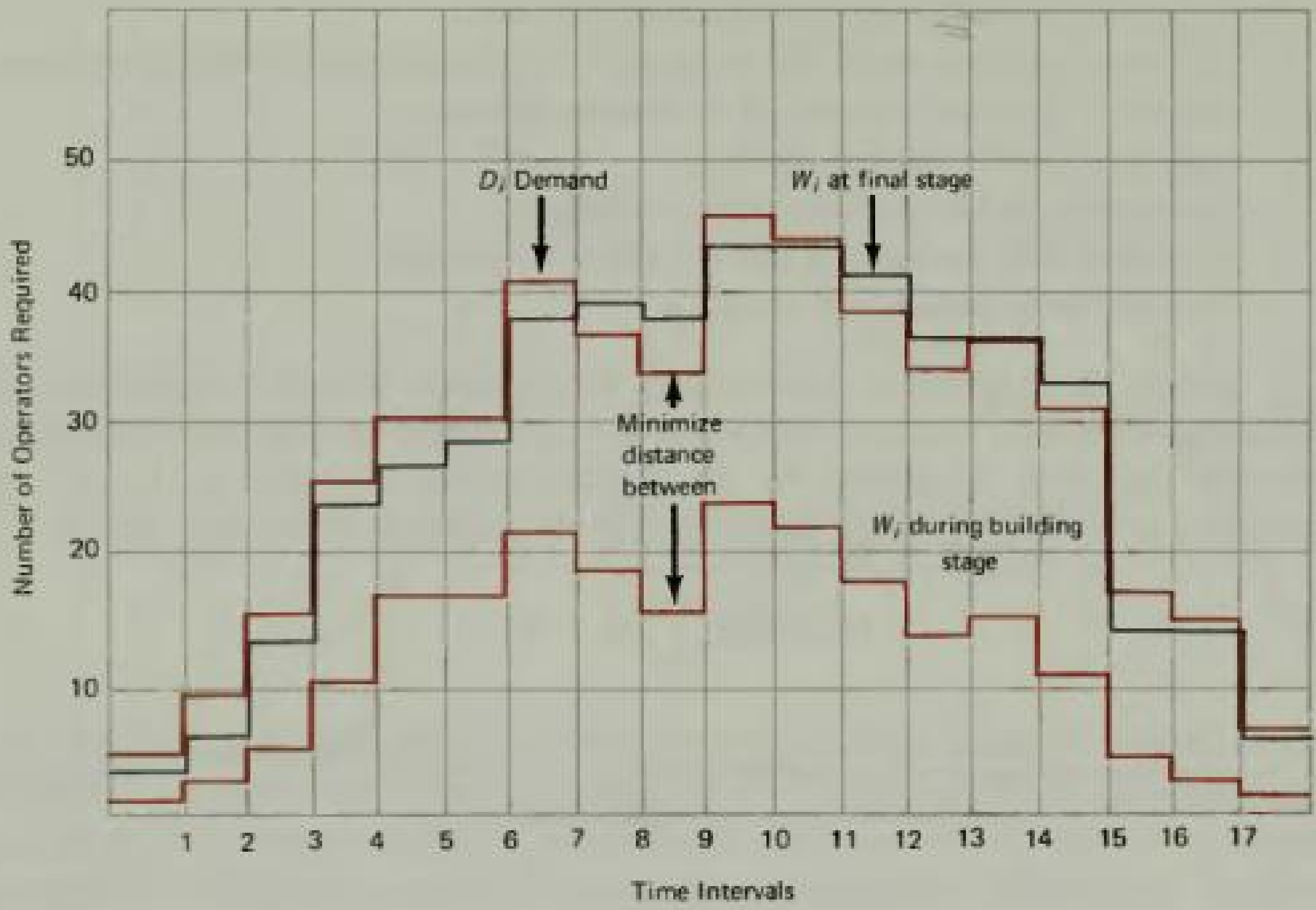


Figure 10. Concept of the schedule building process, using the criterion, minimize $\sum_{i=1}^{n} |D_i - W_i|$

SOURCE: E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System," *Decision Sciences* 7(October 1976). (Courtesy, General Telephone Company of California)

In addition, the output shown in Figure 11 indicates by half-hours, for each hour of the day, the regular and absentee relief (AR) operators required and provided. It also indicates the excess or shortages of operators provided. Note how closely the number of the operators provided conforms to requirements.

The lower half of Figure 11 shows a partial list of shifts actually used, shift length, and the positioning of lunch and rest periods. The system also provides absentee relief (AR) allocations, which are to be used as needed, based on experience factors.

Assigning Operators to Shifts

Given a set of shifts that meets the demand profile, the next step is to assign operators to shifts. The twenty-four hour day, seven-day week operation com-

Figure 11. Sample computer output showing summary scheduling statistics, and tours used

SOURCE: B. J. Luce, Jr., "A Shift Scheduling Algorithm," 44th National Meeting of the Operations Research Society of America, San Diego, November 12–14 1973.

plicates this process: important questions of equity arise regarding the timing of days off and the assignment of overtime work (which carries extra pay). Employee shift and other preferences, as well as seniority status, also must be considered.

Luce [1974] developed a computing algorithm that makes "days off" assignments within the following general rules:

1. Give at least one day off in a week.
2. Days off are one or two.

3. Maximize consecutive days off.
4. If days off cannot be consecutive, maximize the number of work days between days off.
5. Treat weekends separately on a rotational basis in order to preserve equity because
   a. Overtime pay is given for weekend work;
   b. Weekends are the most desirable days off.
6. Honor requests for additional days off on a first come-first assigned basis.

The days off procedure must be carried out to assure that a final feasible schedule will result. Trading off for work days is allowed. The actual assignment of operators to shifts considers employee shift preferences. Each operator makes up a list of shifts in rank order. The list can have different preferences for each day of the week. The order of satisfying preferences is determined on a seniority basis, and in the matching process, assignments are made to the highest ranked shift available for each operator.

The final employee schedule for each day is a computer output (shown in Figure 12). The schedule specifies for each operator the beginning and end of two work periods (separated by lunch), and the time for each of two rest periods.

During 1974, the company realized a net annual savings of over $170,000 in clerical and supervisory costs, as well as achieving a 6 percent increase in work force productivity.

SHIFT SCHEDULING FOR MULTIPLE ACTIVITIES

While the telephone operator scheduling problem deals with homogeneous work content, some situations involve a mix of activities. Church [1973] presented a computerized model for scheduling office staff activities between answering incoming telephone calls and processing paper work. Bell Canada has installed the system on a trial basis. While the problem is apt to be common for telephone companies, it applies equally to the usual office situation (in which personnel divide their time between answering calls and other routine office work), as well as to utilities and service organizations.

In the telephone business office application, the office manager must schedule enough service representatives to take calls throughout the day so that they can give good service to calling customers. The office also is responsible for other activities, such as arranging for telephone installations, investigating annoyance

SCHEDULE OF HOURS SHFD TOLL 8/18/74

NO.	OPERATOR NAME		BEGIN	END	BEGIN	END	1ST	2ND
101 AR	W H BARTON		8.00	11.00	11.30	2.30	8.30	1.15
102	S A MARTIN		8.00	11.00	12.00	3.00	8.30	1.30
103	C E MORTON		7.00	11.30	12.00	3.30	8.45	1.30
104	B L PYROR	SSR	7.00	11.30	12.00	3.30	9.15	2.00
105	B T OLAH		7.00	12.00	1.00	4.00	9.30	2.00
106	M J EASTMON		8.00	12.30	1.00	4.30	10.00	2.15
107	V BECKER		7.45	12.00	1.00	4.45	10.00	2.30
108	D ARNOLD		7.45	12.00	1.00	4.45	10.15	2.45
109	D L LOLLIS	SSR	7.45	12.00	1.00	4.45	10.15	3.00
110	L A MC COLLIN		8.00	12.00	1.00	5.00	10.15	3.00
111 AR	R MC NEELY		8.00	12.30	1.30	5.00	10.15	3.30
112	A L LOWE	*	8.30	1.00	1.30	5.00	10.30	3.00
113	E J PIERSON		8.30	1.00	1.30	5.00	11.00	3.15
114	P A JACKSON		9.00	12.30	1.00	5.30	10.00	3.15
115	B J BAILEY		9.00	1.00	1.30	5.30	10.45	3.15
116	N L FRANCISCO		9.00	1.30	2.30	6.00	11.15	4.30
117	J L CROKEN		9.30	2.30	3.30	6.30	11.45	4.30
118	L J BABCOCK		10.00	2.00	3.00	7.00	12.00	4.30
119 AR	D L DAVIS		10.00	2.00	3.00	7.00	11.45	5.00
120	L J RODRIGUEZ		11.30	3.30	4.00	8.00	1.45	5.30
121	D VECIL		8.30	1.30	5.30	8.30	11.00	7.15
122	T M ARDUINI		12.00	5.00	5.30	8.30	2.15	6.45
123	J J CHASE		12.00	3.30	4.30	9.00	2.00	7.00
124	D L BATTLES		9.30	12.30	5.30	9.30	11.00	7.30
125	R D ORTIZ		9.30	12.30	5.30	9.30	11.15	7.45
126	L R HAGGERTY		9.30	1.00	6.00	9.30	11.30	7.30
127	S C ORTIZ		10.00	1.00	5.30	9.30	11.15	7.45
128	M RODRIGUEZ		10.00	2.00	7.00	10.00	11.30	8.30
129	C I NOE		11.00	3.00	7.00	10.00	1.15	8.15
130 AR	M F GUARINO		2.30	5.30	6.00	10.00	3.30	8.15
131	J ROMERO		11.00	2.00	6.30	10.30	12.15	8.00
132	D FLOYD		11.00	3.00	7.30	10.30	1.00	9.00
133	M C KATSIKAS	TEMP	11.30	3.30	7.30	10.30	1.45	8.45
134	J D MARKLAND		3.00	6.00	6.30	10.30	4.30	8.30
135	D J JACKS JR		3.00	7.00	7.30	10.30	4.45	8.30
136	L R SWARTZ		4.00	7.00	7.30	11.00	5.30	8.45
137	E M KELLEY		4.00	7.00	7.30	11.00	5.30	9.15
138	L D ROWLIN		4.30	7.30	8.00	11.30	5.30	9.30
139	J Y REDMOND		4.30	7.30	8.00	11.30	5.45	9.45
140	D J HAWES		4.30	7.30	8.00	11.30	5.45	10.00
141	S R MC CULLOUGH		4.30	7.30	8.00	11.30	5.45	10.00
142	E J SMITH		4.30	8.00	8.30	11.30	5.45	10.15
143	V H CASTLE		4.30	8.00	8.30	11.30	6.15	10.15
144	W J HERTAN		5.00	8.00	8.30	12.00	6.30	10.30
145 AR	P E LEWIS		5.00	8.30	9.00	12.00	7.00	10.45
146	M WALL		5.00	8.30	9.00	12.00	7.00	10.45
147	C F PERRY		7.00	10.30	11.00	3.00	8.45	12.30
148	K E KELLY		8.00	11.00	11.30	3.00	9.30	1.15
149	C K GILES		8.00	11.30	12.00	3.00	9.45	1.30
150	M L O CONNELL	WTO	11.00	0.0	0.0	6.30	0.0	0.0
151	D D NOLT	WTO	11.30	0.0	0.0	7.00	0.0	0.0
152	B J COWGILL	WTO	11.30	0.0	0.0	7.00	0.0	0.0

Figure 12. Posted operator schedule, specifying two work periods (lunch between), and two rest periods

SOURCE: B. J. Luce, Jr., "Employee Assignment System," Joint ORSA/TIMS National Meeting, Boston, April 22–24, 1974.

calls, and collecting accounts. The basic problem is how to allocate and schedule effort between the two kinds of activities so that personnel can give good customer service and accomplish other work.

The model depends for input on an exponential forecasting model that gives forecasts both for the volume of incoming calls and the clerical work load. The call load forecast must give projected call load for each half-hour, while the clerical load forecast is for the coming week. A queuing model translates the forecasted call load into the number of service representatives available to answer calls. A standard for service is built into the model that enables one to preset the probability that a customer will have to wait more than twenty seconds between talking to the switchboard operator and starting a conversation with a service representative.

The output of the model is the number of service representatives who should be available each half-hour for the next day. A heuristic algorithm then distributes the assignments among the available service representatives. Based on staff preferences, the algorithm spreads the load equally over the office staff, giving several assigned periods in sequence followed by several periods for clerical work, rather than changing rapidly back and forth between the two activities. Both staff and management seem to have accepted the system as an improvement, and the initial test resulted in at least a 4 percent increase in office productivity.

UTILIZING PART-TIME WORKERS

When demand for service varies significantly but follows a fairly stable weekly pattern, the use of part-time employees can give managers added flexibility with which to develop minimum cost schedules. ~~Mabert and Raedels [1976] reported~~ such an application involving eight branch offices of the Purdue National Bank. Typical demand for service in the branches is shown in Figure 13; Mondays and Fridays usually exhibit peak demand. The problem also involved the anticipation of increased demand for service on paydays.

Traditionally, the bank employed only full-time tellers to meet requirements. This policy required staffing to meet peak demand and resulted in relatively poor average staff utilization during the week. Bank management later changed the staffing policy by employing full-time tellers equal to minimum expected demand and by using part-time tellers to staff the peak needs.

Problem Formulation

The basic problem involved determining the minimum cost teller assignments for the eight-branch system and, secondarily, minimizing (1) the number of part-time tellers, and (2) the teller transfers between branches. Although Mabert ~~and Raedels~~ used an integer programming formulation of the problem, the size of the problem limits the use of this technique. The scale of the mathematical programming problem increases rapidly with the number of branches because

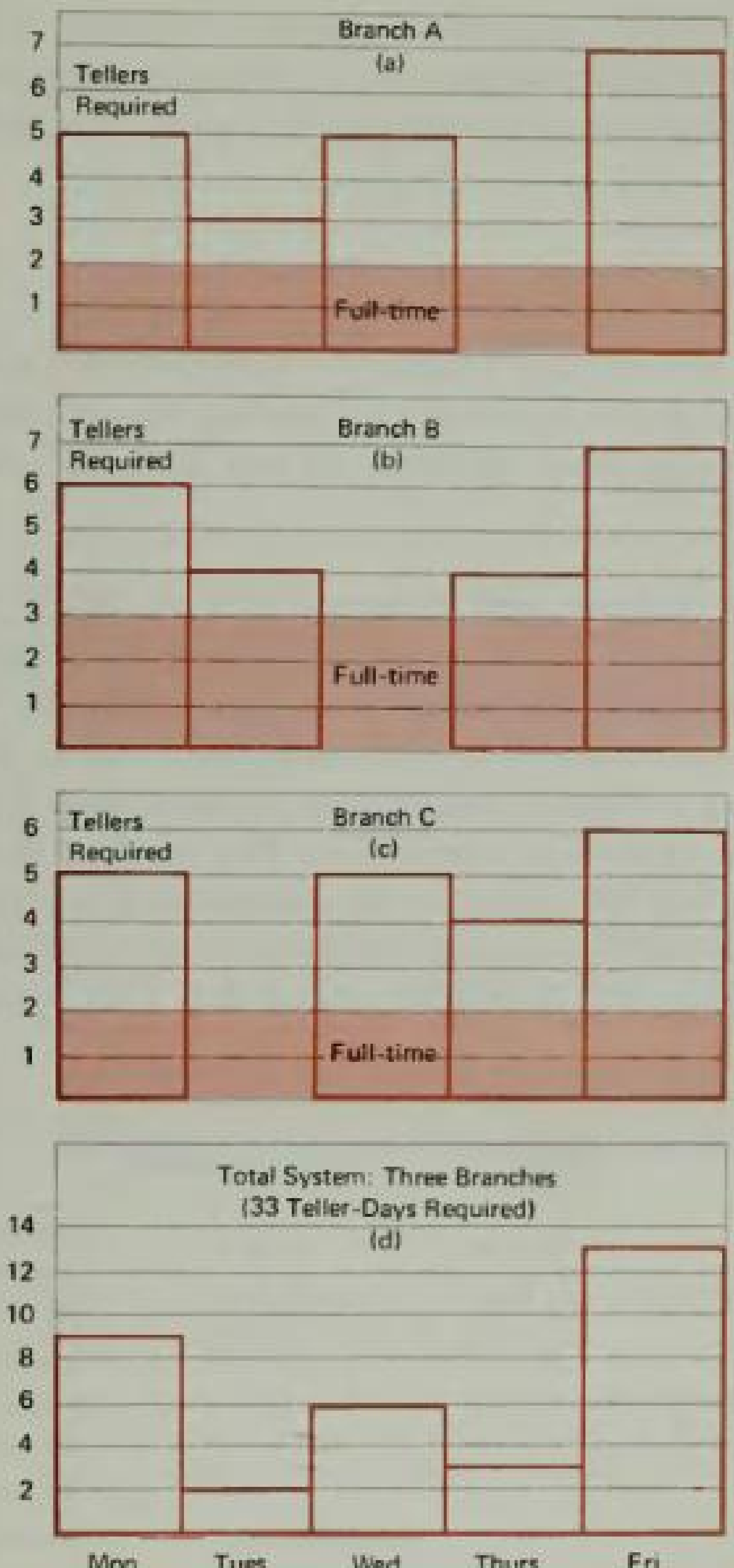


Figure 13. Teller requirements for three branches and aggregate system teller requirements

SOURCE: V. A. Mabert, and A. R. Raedels, "The Detail Scheduling of a Part-Time Work Force: A Case Study of Teller Staffing," *Decision Sciences* 7(October 1976).

of the rapidly increasing number of possible alternate work assignments. Therefore, while the integer programming formulation represented the problem quite well, other techniques were needed to find a practical solution.

Heuristic Solution Methods

For the simplified, three-branch system, the aggregated requirements for part-time tellers only are shown in Figure 13*d*, totaling thirty-three teller days. The development of the aggregated requirements was the first step in both heuristic procedures.

Condensed Integer Problem with Heuristic (*CIPH*). While the multibranch problem was too large to be solved by integer programming methods, when branch requirements were aggregated (as in Figure 13*d*), the smaller problem could be solved economically. The next step, then, was to determine a solution to the aggregate problem, using integer programming, allowing two- and three-day schedules as feasible alternatives. The result of this step, shown in Figure 14, indicates that the aggregate requirements for thirty-three teller days could be met, as well as the daily requirements schedule (shown in Figure 13*d*). Two different three-day, and three different two-day, work schedules were used.

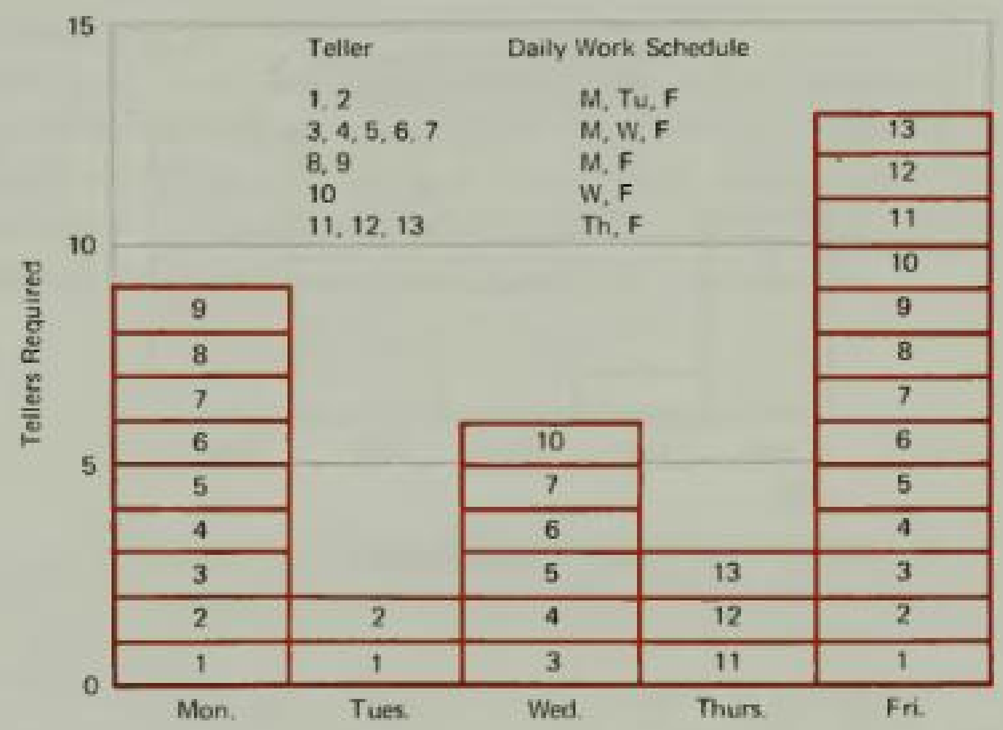


Figure 14. Optimal assignments to meet the aggregate teller requirements shown in Figure 13d

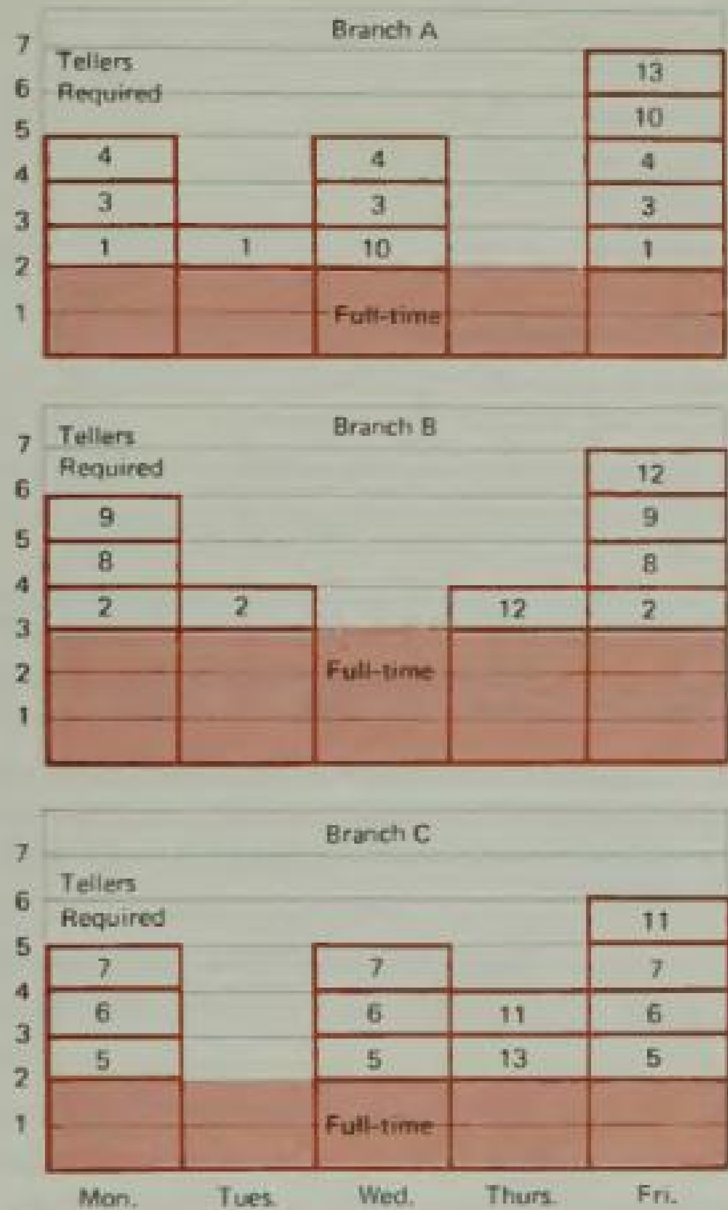


Figure 15. Assignments of tellers to branches, using the CIPH procedure

SOURCE: V. A. Mabert, and A. R. Raedels, "The Detail Scheduling of a Part-Time Work Force:" A Case Study of Teller Staffing." *Decision Sciences* 7(October) 1976).

The aggregate schedule was disaggregated by assigning schedules to branches by means of a heuristic procedure. The procedure first involves assignments within branches until branch requirements are met, as far as possible, without transfers. Then, unassigned tellers are assigned between branches to cover the remaining requirements. For the sample three-branch problem, Figure 15 shows that Teller 13 needed to be transferred between branches *A* and *C* in order to complete the requirements schedule.

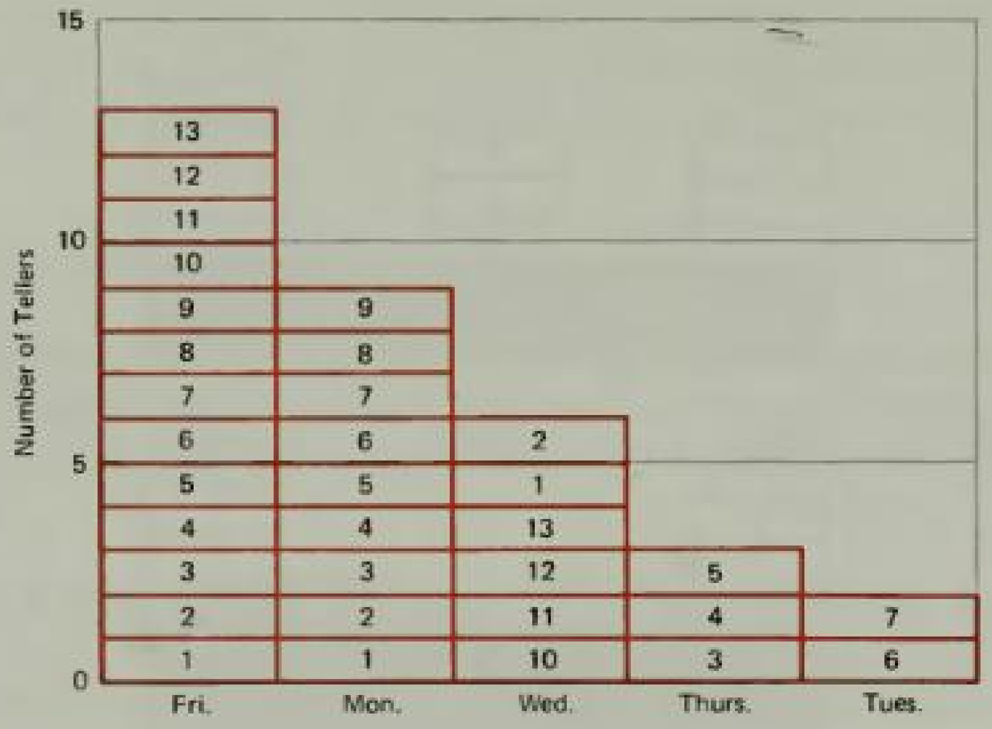


Figure 16. Assignment of tellers to decreasing-demand histogram. Beginning with the highest-demand day, tellers are assigned in sequence from 1 to 13 in cycles until all 33 teller blocks have been assigned.

SOURCE: B. A. Mabert, and A. R. Raedels, "The Detail Scheduling of a Part-Time Work Force: A Case Study of Teller Staffing," Decision Sciences 7(October 1976).

Heuristic Assignment Procedure (HAP). The alternative heuristic procedure starts with a requirements schedule (such as in Figure 13*d*). Since the Friday schedule requires thirteen tellers, this is the feasible minimum number of tellers. A decreasing demand histogram, shown in Figure 16, then is developed from Figure 13*d*. For example:

Friday—13 tellers required
Monday—9 tellers required
Wednesday—6 tellers required
Thursday—3 tellers required
Tuesday—2 tellers required

The thirteen tellers then are assigned in sequence, beginning with Friday and progressing through the daily requirements, according to the preceding sequence of days. The teller numbers simply are repeated until all thirty-three requirement blocks have been assigned a teller number. A feasible set of two- and three-day teller schedules can be abstracted from Figure 16, as follows:

Teller	*Daily Work-Schedule*
1, 2	M, W, F
3, 4, 5	M, Th, F
6, 7	M, Tu, F
8, 9	M, F
10, 11, 12, 13	W, F

Finally, assignments are made to branches following the same procedure that the CIPH approach used.

For the sample three-branch problem, the HAP procedure gives Tellers 12 and 13 schedules that involve transfers between branches.

Results for the Eight-Branch Problem. When the 2 heuristic procedures were applied to the actual 8 branch problem for a 7 week test, both procedures resulted in an identical number of tellers used. However, the average number of transfer assignments for CIPH was 8.8, compared with 9.4 for HAP. On the other hand, the computer execution time was 0.436 seconds for CIPH and only 0.040 seconds for HAP. Statistical tests indicated no significant differences between the results obtained by the two heuristic procedures. However, as noted, the CIPH procedure required approximately 10 times more computer time for solution.

According to Mabert and Raedels, the Purdue National Bank has been using the HAP technique on a manual basis so successfully that the scheduling of part-time tellers who use the technique saves the 8 branch system about $30,000 per year.

Summary

The work shift scheduling problem emerges from the study of service and nonmanufacturing systems, although its application is not confined to such systems. Variable work load during the day and week, as well as the need to assign manpower for continuous seven-day operations of any kind, raises the important question of how to schedule for minimum cost. Such problems may become even more significant as our society begins seriously to contemplate alternative work-time patterns such as the four-day, ten-hour per day, and thirty-hour per week proposals that currently are being considered.

The work shift scheduling problem has been formulated as an integer linear programming problem, but the current state of the art does not permit the solution of real world problems in this format. Heuristic solutions have been used to obtain very good, though not optimal, solutions.

The General Telephone Company Computerized Force Management system is an excellent example of an integrated system that places the work shift scheduling algorithms within context. Significant forecasting problems exist—conversion to the equivalent work units and the scheduling itself—as well as the very important assignment of operators to shifts and the related "days-off" schedule.

When the work involves multiple activities, such as in the Bell Canada example, the work shift scheduling problem becomes more complex because the load for the different activities must be forecast and because the scheduling must be geared to the multiple activity forecasts.

Finally, when the nature of the work allows the use of part-time workers, shift scheduling procedures can be employed effectively to reduce aggregate payroll costs (as indicated in the bank teller study).

A variety of methods have been used in the shift scheduling problem. While the use of mathematical programming has been attempted, the size and complexity of problems usually is too great to be solved by the integer method. Therefore, a variety of heuristic approaches have been used that have yielded practical results in applications.

Review Questions

1. Define the important characteristics of the work shift scheduling problem.

536

2. What are the various work shift types that may be used? Do they represent constraints on scheduling solutions, or do they make the problem easier?

538

3. Describe the formal problem statement or formulation of the work shift scheduling problem. What solution techniques have been proposed?

4. In the Henderson–Berry heuristic solution, define the:
   a. LPI heuristic.
   b. LPVE heuristic.
   c. RSVE heuristic.

5. In the Henderson–Berry study, why should it have been true that better results were obtained when the working subset of shifts was increased?

6. Describe the demand function for the General Telephone Company in terms of seasonal, trend, weekly, daily, and hourly variation.

7. Outline the major steps or modules in the General Telephone Company's Computerized Force Management System. What are the major feedback loops?

40

8. Describe the forecasting system used at General Telephone. How does it take account of trends, seasonals, weekly, daily, and hourly variations?

9. What is the function of the "topline" program in the General Telephone system?

10. Describe the functioning of the scheduling algorithm in the General Telephone System. What is the criterion for successful scheduling?

11. Describe the process for assigning operators to shifts in the General Telephone system.

12. How is the shift scheduling problem for the Bell Canada situation different from that at General Telephone?

13. State the nature of the formulation of the shift scheduling problem in the bank teller study.

574

14. For the Mabert–Raedels bank teller study, define the:
    a. CIPH solution procedure.
    b. Heuristic assignment procedure.

15. Summarize the results achieved by shift scheduling using part time workers in the Purdue National Bank.

References

Abernathy, W. J.; N. Baloff; and J. C. Hershey. "The Nurse Staffing Problem: Issues and Prospects." *Sloan Management Review* 13(Fall 1971): 87–109.

———; ———; ———; and S. Wandel. "A Three-Stage Manpower Planning and Scheduling Model—A service-Sector Example." *Operations Research 21* (May–June 1973): 693–711.

Ahuja, H., and R. Sheppard. "Computerized Nurse Scheduling." *Industrial Engineering* (October 1975): 24–29.

Baker, K. R. "Scheduling a Full-Time Workforce to Meet Cyclic Staffing Requirements." *Management Science* 20(August 1974): 1561–1568.

Bennett, B. T., and R. B. Potts. "Rotating Roster for a Transit System." *Transportation Science* 2(February 1968).

Berry, W. L., V. A. Mabert, and M. Marcus, "Forecasting Teller Window Demand With Exponential Smoothing," Paper No. 536, *Institute for Research in the Behavioral, Economic, and Management Sciences*, Purdue University, November, 1975.

Box, G. E. P., and G. M. Jenkins. *Time Series Analysis, Forecasting, and Control.* San Francisco: Holden-Day, 1970.

Buffa, E. S., M. J. Cosgrove, and B. J. Luce. "An Integrated Work Shift Scheduling System." *Decision Sciences* 7(October 1976).

Burman, D., and M. Segal. "An Operator Scheduling System." ORSA/TIMS National Meeting, March–April 1976.

Butterworth, R. W., and G. T. Howard. "A Method of Determining Highway Patrol Manning Schedules." ORSA 44th National Meeting, November 1973.

Church, J. G. "Sure Staf: A Computerized Staff Scheduling System for Telephone Business Offices." *Management Science* 20. Part 2 (December 1973): 708–720.

Healy, W. E. "Shift Scheduling Made Easy." *Factory* 117(October 1969).

Harveston, M. F.; B. J. Luce; and T. A. Smuczynski. "Telephone Operator Management System—TOMS." ORSA/TIMS/AIIE Joint National Meeting, November 1972.

Henderson, W. B., and W. L. Berry. "Determining Optimal Shift Schedules for Telephone Traffic Exchange Operators." *Decision Sciences* 8(January 1977).

______, and ______. "Heuristic Methods for Telephone Operator Shift Scheduling: An Experimental Analysis." *Management Science* 22(August 1976): 1372–1380.

Hill, A. V., and V. A. Mabert. "A Combined Projection–Causal Approach for Short Range Forecasts." Paper no. 527. Purdue University, Krannert Graduate School of Industrial Administration, Institute for Research in the Behavioral, Economic, and Management Sciences, September 1975.

Jelinek, R. C. "Tell the Computer How Sick the Patients Are and It will Tell How Many Nurses They Need." *Modern Hospital* (December 1973).

Larson, R. C. "Improving the Effectiveness of New York City's 911." In A. W. Drake, R. L. Keeney, and P. M. Morse, *Analysis of Public Systems*. Cambridge, Mass.: MIT Press, 1972.

Linder, R. W. "The Development of Manpower and Facilities Planning Methods for Airline Telephone Reservations Offices." *Operational Research Quarterly* 20(1969): 3–21.

Luce, B. J. "Dynamic Employment Planning Model." ORSA 44th National Meeting, November 1973.

______, "Dynamic Employment Planning Model." ORSA/TIMS Joint National Meeting, April 1974.

______, "Employee Assignment System." ORSA/TIMS Joint National Meeting, April 1974a.

Mabert, V. A., and A. R. Raedels. "The Detail Scheduling of a Part-Time Work Force: A Case Study of Teller Staffing." *Decision Sciences* 7(October 1976).

Maier-Roth, C., and H. B. Wolfe. "Cyclical Scheduling and Allocation of Nursing Staff." *Socio-Economic Planning Sciences* 7(1973): 471–487.

Monroe, G. "Scheduling Manpower for Service Operations." *Industrial Engineering* (August 1970).

Murray, D. J. "Computer Makes the Schedules for Nurses." *Modern Hospital* (December 1971).

Paul, R. J., and R. E. Stevens. "Staffing Service Activities with Waiting Line Models." *Decisions Sciences* 2(April 1971).

Ritzman, L. P.; L. J. Krajewski; and M. J. Showalter. "The Disaggregation of Aggregate Manpower Plans." *Management Science* 22(July 1976): 1204–1214.

Rothstein, M. "Hospital Manpower Shift Scheduling by Mathematical Programming." *Health Services Research* (Spring 1973): 60–66.

———, "Scheduling Manpower by Mathematical Programming." *Industrial Engineering* (April 1972): 29–33.

Segal, M. "The Operator-Scheduling Problem: A Network-Flow Approach." *Operations Research* 22(July–August 1974): 808–823.

Walsh, D. S. "Computerized Labor Scheduling: Supermarkets Jumping on the Bandwagon." 25th Annual Conference and Convention, American Institute of Industrial Engineers, May 1974.

Chapter 16

Maintaining System Reliability

If the output of a productive system maintains standard quality, quantity, and cost, we think of it as being reliable: that is, it continues to do what it was designed to do. In our discussion of operations planning and control systems (see Part IV Introduction, Figure 1), we established basic control loops for performance. Control was accomplished by monitoring the output, comparing it with standards, interpreting differences, and taking action to readjust processes so that they conformed to standards. We also discussed broader level control systems with which to achieve a system optimum.

It is not sufficient to think only in terms of average standards of performance: if the system is erratic in its performance, it is unreliable. It is somewhat harder to maintain system reliability with complex systems than it is with simple ones, and an analogy to machines will explain why.

RELIABILITY OF MACHINES

Complex machines may break down often, even though they are designed and manufactured according to the highest standards. We can view machines as being made up of a sequence of components, each of which performs a function.

For example, when we strike a key on an electric typewriter, an electric switch is closed. This actuates a solenoid, which causes the mechanical action of the type to strike the ribbon, which in turn transfers carbon to the paper. The type linkage returns to its normal position, and the carriage indexes one space to make ready for the next cycle. The number of individual and mechanical-electrical components required to perform correctly in sequence (series) is larger than we might think. If any single component fails to function correctly—due either to breakdown or faulty adjustment—the system as a whole fails to function correctly. We are interested in how this affects system reliability.

If we are dealing with a machine that is made up of $n = 50$ components in series, each with an average reliability of 99.5 percent, the reliability of the machine as a whole is only about 77 percent (see Figure 1). In other words, if the average probability of breakdown of each of the 50 components were $100 - 99.5 = 0.5$ percent, then the probability that the machine might break down because any one of the components broke down is 23 percent. Figure 1 shows us that as complexity increases, in terms of the number of components in series, the reliability of the system as a whole declines very rapidly.

The reliability of a system in which the components are in parallel differs entirely. With parallel components, there are two or more components that, in some way, perform the same function. Therefore, if the system is to fail, both parallel components must fail, thus increasing the system's reliability. Increasing reliability through parallel components and systems is expensive, but when the possible losses are great, it may be justified. In space vehicle systems, for example, parallel systems often are used because of the potential costs of system failure. In productive systems, we may provide parallel paths by having more than one machine that can perform the same operation.

RELIABILITY IN PRODUCTIVE SYSTEMS

There are close parallels between the reliability of machines and of organizations producing services and products. When the number of sequential operations required is small and the reliability of each operation is high, the reliability of the system as a whole can be quite good. When the number of sequential operations is large and an error or defect in quality can come from any one operation, then there is a potential for unreliable performance (see Figure 1). We can extend the parallel to include the functioning of entire organizations. Complexity creates the possibility of unreliable performance, and this is one of the most important reasons why larger, more complex organizations must use for-

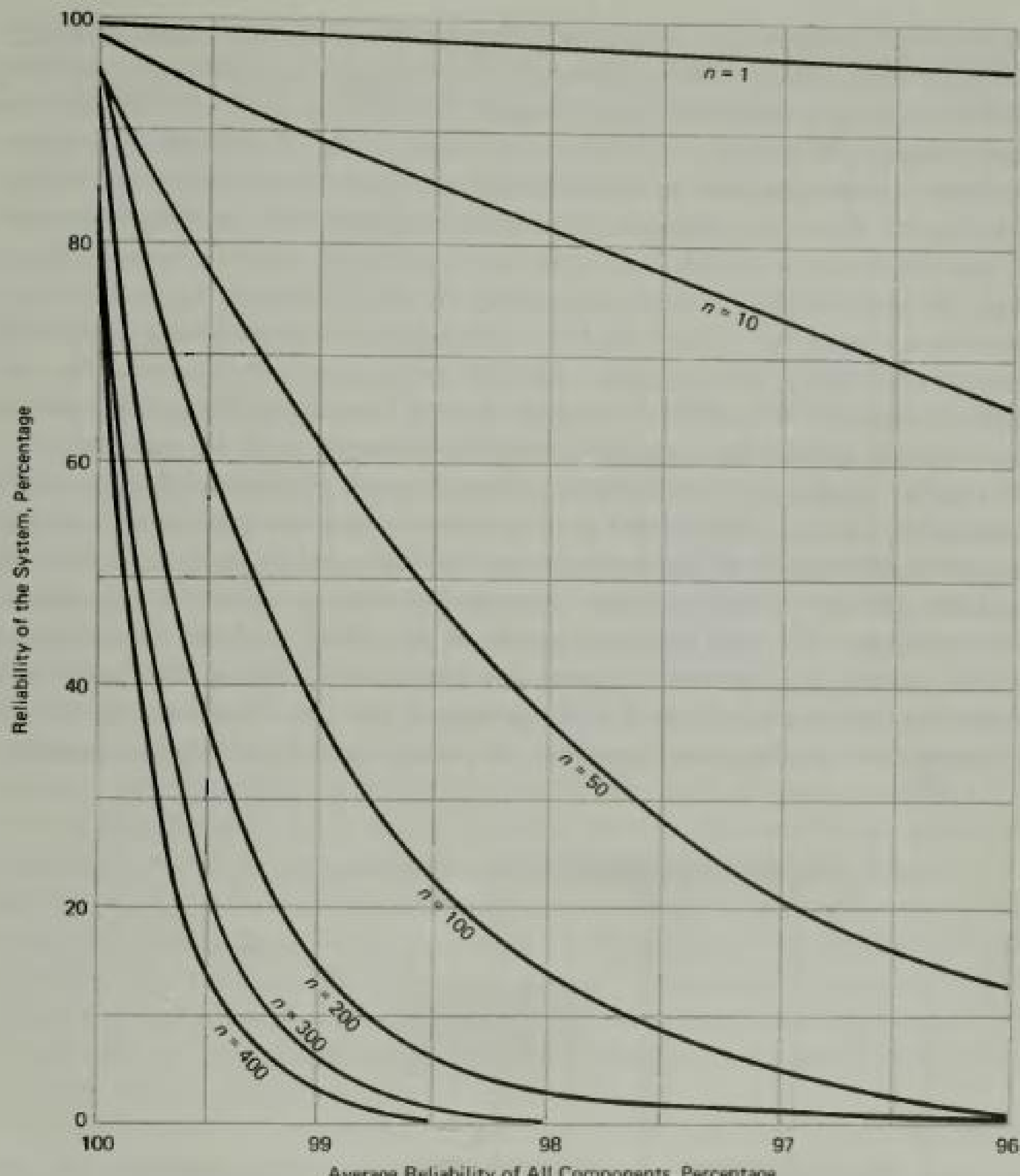


Figure 1. Overall system reliability as a function of complexity (number of components) and component reliability with components in series

SOURCE: R. Lusser, "The Notorious Unreliability of Complex Equipment," *Astronautics* 2(February 1958)

mal procedures in dealing with many kinds of standardized problems. Such organizations become bureaucratic to become more reliable, although the bureaucracy may create problems of its own.

Now let us contrast the nature of reliability for service systems with the reliability for systems that produce physical products in which machines may dominate. One characteristic of machines is that they can perform repetitive operations fairly consistently, reproducing the same activities. Variations in quality and output quantity tend to be minimized. On the other hand, operations dominated by humans tend to exhibit relatively wide variation in measures of performance. When we combine all the wide variations that are found in manual operations into a complex sequence (as is common in some service operations), the maintenance of system reliability becomes difficult. In such service operations, consistent output quality may depend on individuals following exacting procedures, and variation can have important consequences. For example, if a nurse fails to follow the procedure of identifying the patient before administering medication and a mixup occurs, the results can be disastrous.

We try to control the reliability of output through general schemes, as diagrammed in Figure 2. The output quality and quantity are monitored in some way, often through sampling procedures, and the results are compared with standards. (While we generally are interested in quality measures, changes in output quantity also may be symptomatic of reliability problems. Associated costs of quality and quantity control are derivatives.) When the results are interpreted, we may conclude that the processes are out of adjustment or that something more fundamental is wrong, requiring machine repair (or possibly

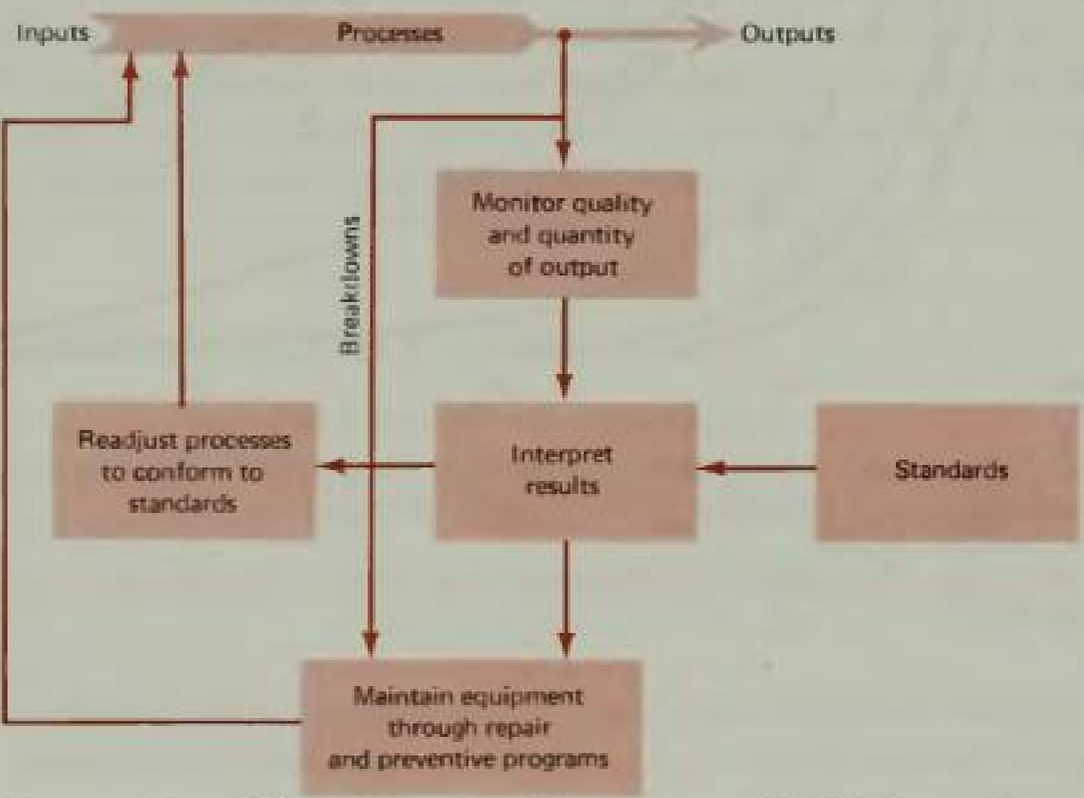


Figure 2. Control loops for maintaining system reliability by monitoring quality and quantity of output

retraining in manual operations.) If equipment actually breaks down, then the maintenance control loop is called directly. Information on output quality and quantity also may be used to form preventive maintenance programs that are designed to anticipate breakdown. Thus, while other important interactions (which we shall discuss) have their effects, the control of system reliability centers on quality control and equipment maintenance.

The Reliability System

Figure 2 suggests the nature of control loops for quality and maintenance control. However, it is local in nature and leaves a great deal unsaid. We must ask: Where did the standards come from? What is the nature of the productive system, and is it appropriate? Figure 3 places the reliability system in context, and refers to some of the materials that we discussed in Chapters 5 through 8 dealing with productive system design. In Chapter 5, we discussed the design of products and services, and it is in these activities that basic objectives for quality are established. The organization must set policies regarding the desired quality in relation to markets and needs, investment requirements, return on investment, potential competition, etc. For profit making organizations, this involves the judgment of where in the market they have a relative advantage.

For nonprofit organizations, policy setting may involve meeting standards set by legislation and/or by the organization to maximize its position. For example, one reason that universities acquire superior academic staffs is to raise funds, improve the physical plant, obtain outstanding students, enhance research output, and attract funds. Hospitals may set high standards of care partially to attract the best known physicians, support fund raising programs, and attract interns with the best qualifications. The post office may set standards for delivery delay that accomodate current budgetary levels.

The policies set by management in Box 1 of Figure 3 provide the guidelines for the design of the organization's products and services. As we discussed in Chapter 5, this design process is an interactive one, in which the productive system design is both considered in and influenced by the design of products and services, as shown in Boxes 2 and 3. For manufacturing systems, the design of products in this interactive fashion is termed *production design*. The interaction affects quality considerations, since equipment capability must be good enough to produce at least the intended quality.

Out of the process for the design of products/services and productive system design comes specifications of quality standards, as shown in Box 4 of Figure 3. Here, we are dealing with a system of quality standards for: materials that are

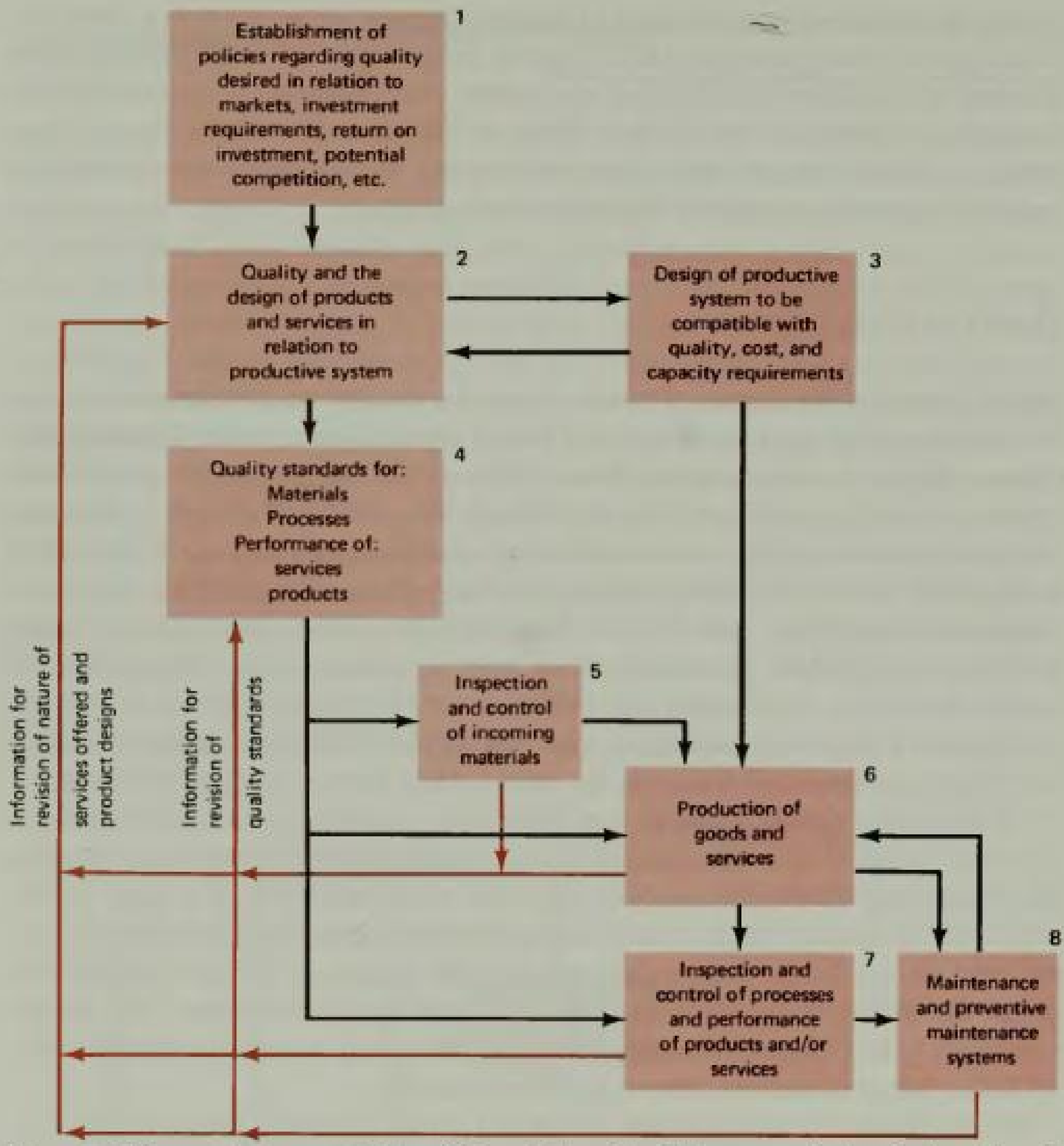


Figure 3. Schematic representation of the relationships between policies, design of products and/or services, design of productive system, and the maintenance of the system reliability for quality and quantities

consumed in processes, as well as raw materials; the standards for the output of processes, such as the specification of dimensions, tolerances, weights, and chemical compositions; and the performance standards for the outputs. The nature of performance standards for products is well known. Manufacturers state (or overstate) the capabilities of their product: the fidelity range of an amplifier, the acceleration of an auto, the waterproof finish of a table surface, etc.

The performance standards of services seem (so far) to be somewhat less formalized. What performance do we expect from postal systems, educational systems, police and fire protection, and medical care systems? This relative lack of formal standards, and the difficulty of defining them, may be the reason for the paucity of our knowledge about the control of the quality of services.

Given standards, however, we can set up controls for incoming materials in Box 5, and for the processes and performance of products and services in Box 7. An interrelated control loop concerns the maintenance of the capabilities of the physical system in Box 8, through repair and preventive maintenance programs.

Secondary control loops are shown in Figure 3 that seek a system optimum. These appear as information flow from Boxes 5, 6, 7, and 8 to Boxes 2 and 4. Their function is to influence the nature of services and products offered, and to help revise quality standards, respectively.

In order to present the concepts that are useful in quality control and maintenance, we shall deal with these subjects separately in the following sections. However, let us bear in mind the overall conceptual framework of Figure 3. Quality control and maintenance are interrelated, and they support each other in the broad objective of maintaining system reliability.

CONTROL OF QUALITY

Our general block diagram for control calls for a measurement system to generate information on which to base control actions. In industry, this is the inspection function. Inspectors make measurements that are called for by the quality standards, thereby separating acceptable from nonacceptable units. However, no control or corrective action is implied. When we link measurement, investigation to determine why an unacceptable product has been produced, and corrective action, we have completed the control loop.

Variation and Control

All processes exhibit variation; and the manager's task is to distinguish between tolerable variation that is representative of the stable system and major changes that result in an unacceptable product. The manager must be aware of the system's inherent capability in order to know when system behavior is abnormal. Thus, since we are dealing with systems that exhibit variation, the manager's control model must be a probabilistic one.

Sampling Information

Because of the ongoing nature of processes and their inherent variability, we must base quality control decisions on samples. First, we cannot usually examine all the data, because the process is continuous, and at best we have access to a sample at a particular point in time. Second, even if the entire universe of data were available, it might be uneconomical to analyze it. Third, measurement and inspection sometimes require destruction of the unit; and fourth, with some products any additional handling is likely to induce defects and therefore should be avoided. Thus, the sampling of information about the state of incoming raw materials and of control of processes is the common approach on which to base decisions and control actions.

Risks and Errors

Because we normally must use samples of data drawn from a system that naturally exhibits variation, we can make mistakes, even in controlled experiments. Figure 4 can help us summarize the nature of errors and risks taken; here, we classify the actual state of the system and the decision taken. The process either is in control or it is not; or, similarly, we have a batch of parts or materials that have been generated by a system that either was or was not in control.

As Figure 4 shows, we can decide either to accept or reject the output. If the process is in control—and if, based on our information, we would reject the output—then we have made an error that is called a *Type I error*. We, the producer, risk making such an erroneous decision on the basis of the proba-

True state of system	Decision	
	Reject output as bad	Accept output as good
Process is in control	Type-I error (Producer's risk)	Correct decision
Process is out of control	Correct decision	Type-II error (Consumer's risk)

Figure 4. Errors and risks in quality control decisions

bilities that are associated with the inherent variability of the system and the sample size. Logically, this risk is called the producer's risk, since—if the decision is made—it is the producer who absorbs the loss.

Similarly, there is a risk that we may accept output as a good product when in fact the process is out of control (see Figure 4). This decision is called a *Type II error*, and since the bad product presumably is shipped, the probability of such a decision is termed the consumer's *risk*. In statistical control models, we can preset the probabilities of Type I and Type II errors, and therefore we can remain in control.

Kinds of Control

Figure 3 shows that, functionally, we control quality by controlling (1) incoming materials, (2) processes at the point of production, and (3) the final performance of products and services, with the maintenance system acting in a supporting role. For some products, quality control of product performance includes the distribution, installation, and use phases.

From the point of view of control methods, we can apply statistical control concepts by sampling a lot of incoming materials to see whether it is acceptable (acceptance sampling), or by sampling the output of a process to keep that process in a state of statistical control (process control).

Acceptance sampling lets us control the level of outgoing quality from an inspection point to ensure that, on the average, no more than some specified percentage of defective items will pass. This procedure assumes that the parts or products already have been produced, and that we wish to set up procedures and decision rules to ensure that outgoing quality will be as specified or better. In the simplest case of acceptance sampling, we draw a random sample of size n from the total lot N and decide, on the basis of the sample, whether or not to accept the entire lot. If the sample signals a decision to reject the lot, the lot either may be subjected to 100 percent inspection, in which bad parts are sorted out, or may be returned to the original supplier (perhaps a vendor or another department within the organization). Parallel acceptance sampling procedures can be used to classify parts as simply good or bad (sampling by attributes) or to make some kind of actual measurement that indicates how good or bad a part is (sampling by variables.)

In process control, we monitor the actual ongoing process that makes the units. This allows us to make adjustments and corrections as soon as they are needed, so that bad units in any quantity are never produced. This procedure is a direct application of the statistical control chart and, as with acceptance sampling, parallel procedures are available for those situations in which sam-

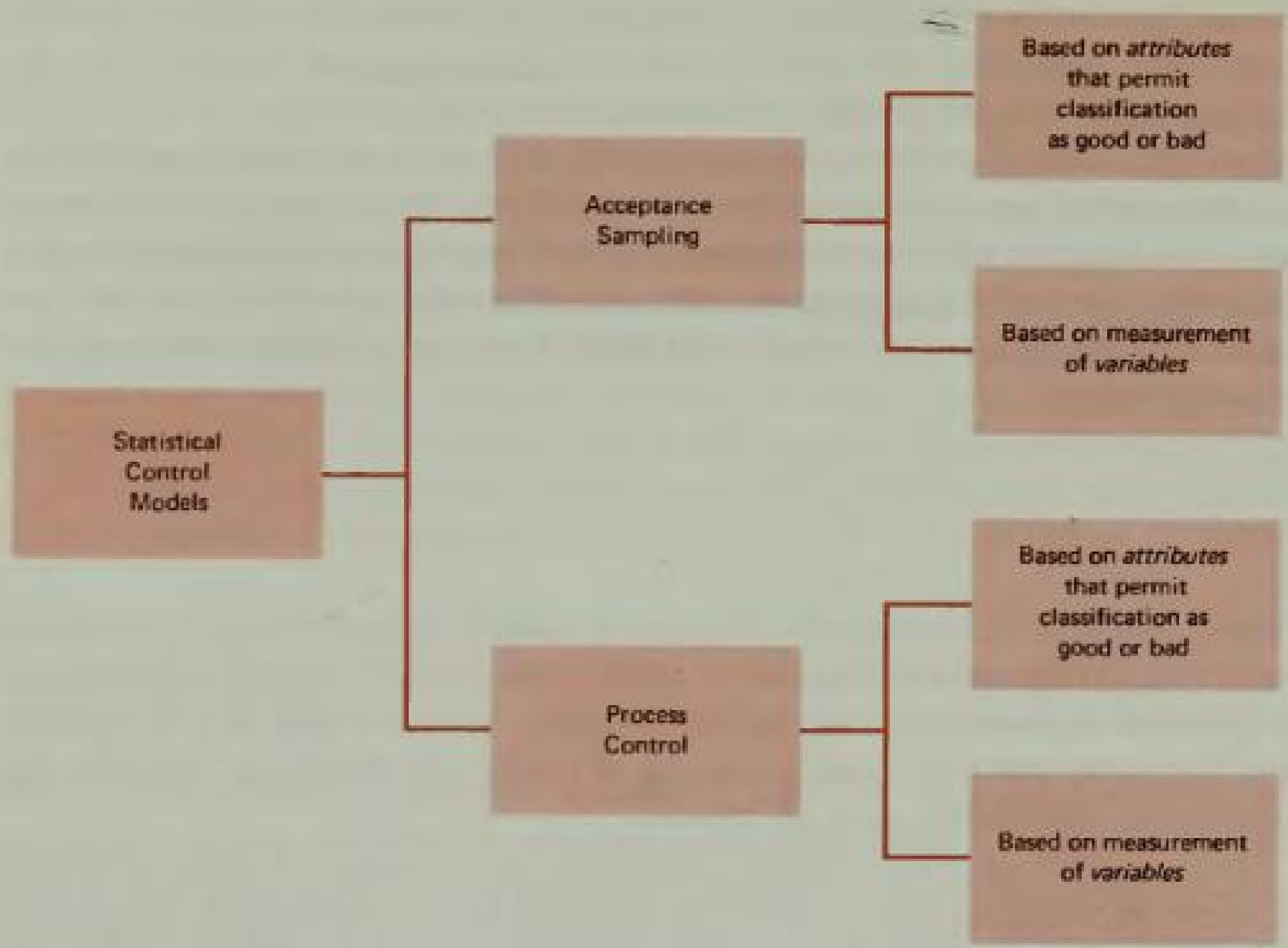


Figure 5. Classification of statistical control models

pling is done by attributes, and for those in which measurements are made of variables that measure quality characteristics. Figure 5 summarizes the classification of statistical control models.

ACCEPTANCE SAMPLING

In general, acceptance sampling is appropriate when:

1. Possible losses made by passing defective items are not great, and the cost of inspection is relatively high. In the limiting situation, this can mean no inspection at all.
2. Inspection requires the destruction of the product; for example, when it is necessary to determine the strength of parts by pulling them apart. We must infer the acceptability of the entire lot from a sample in these instances.
3. Further handling of any kind is likely to induce defects, or when mental or

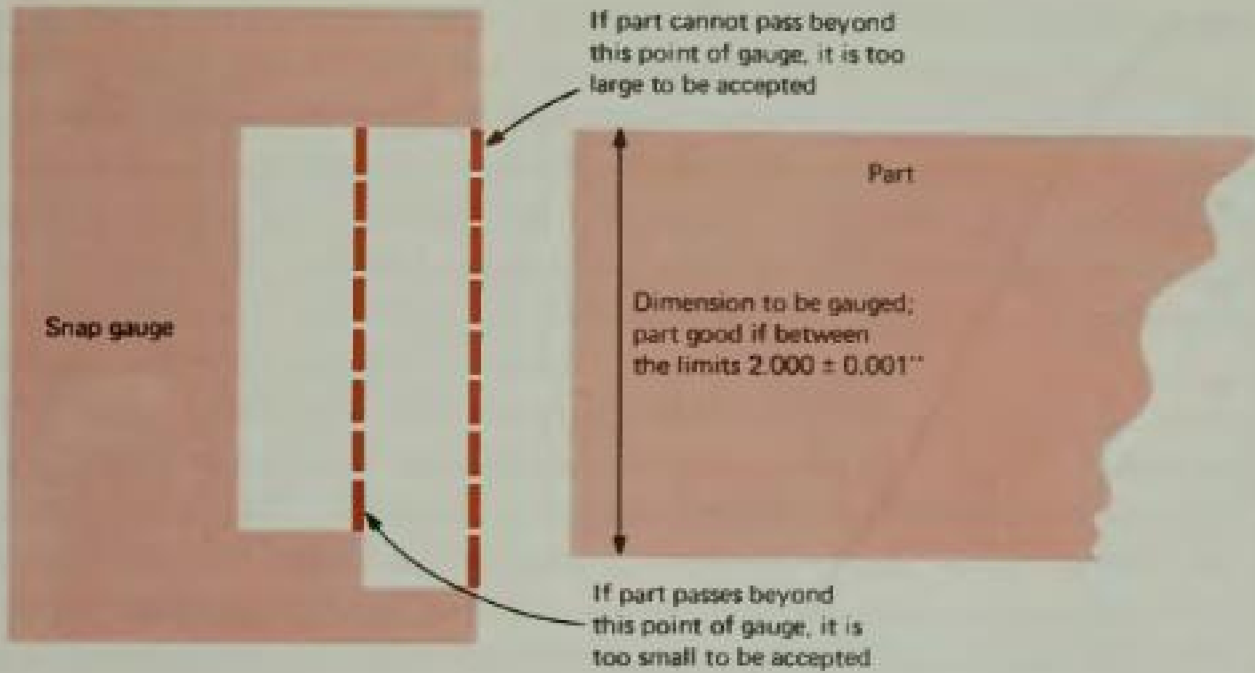


Figure 6. A part measured by a snap gauge. No actual part measurement is recorded, but classification of parts is made as simply good or defective

physical fatigue is an important factor in inspection. In either instance, a sampling plan may actually pass fewer defective items than would 100 percent inspection, and it also is less expensive.

The inspection procedure for attributes sampling results in classifying units as simply good or not good. For part dimensions, this often can be accomplished by the use of snap or plug gauges that incorporate a go-not-go feature, as indicated in Figure 6. If the inspection is for some attribute of appearance, such as surface paint defects, the two-part classification of good or not good also would be made. In all instances, attributes sampling inspection uses criteria to grade products as acceptable or not acceptable. The statistical methods used are based on distributions, such as the binomial distribution or the Poisson distribution. The appropriateness of these and other distributions in specific situations is beyond our scope (see Duncan [1974]).

Operating Characteristic (OC) Curves

To specify a particular sampling plan, we indicate the random sample size n and the number of defectives in the sample c (acceptance number) that is permitted before the entire lot from which the sample was drawn must be rejected. The OC curve for a particular combination of n and c shows how well the plan discriminates between good and bad lots. Figure 7 is an OC curve for a sampling plan with sample size $n = 50$ and acceptance number $c = 1$.

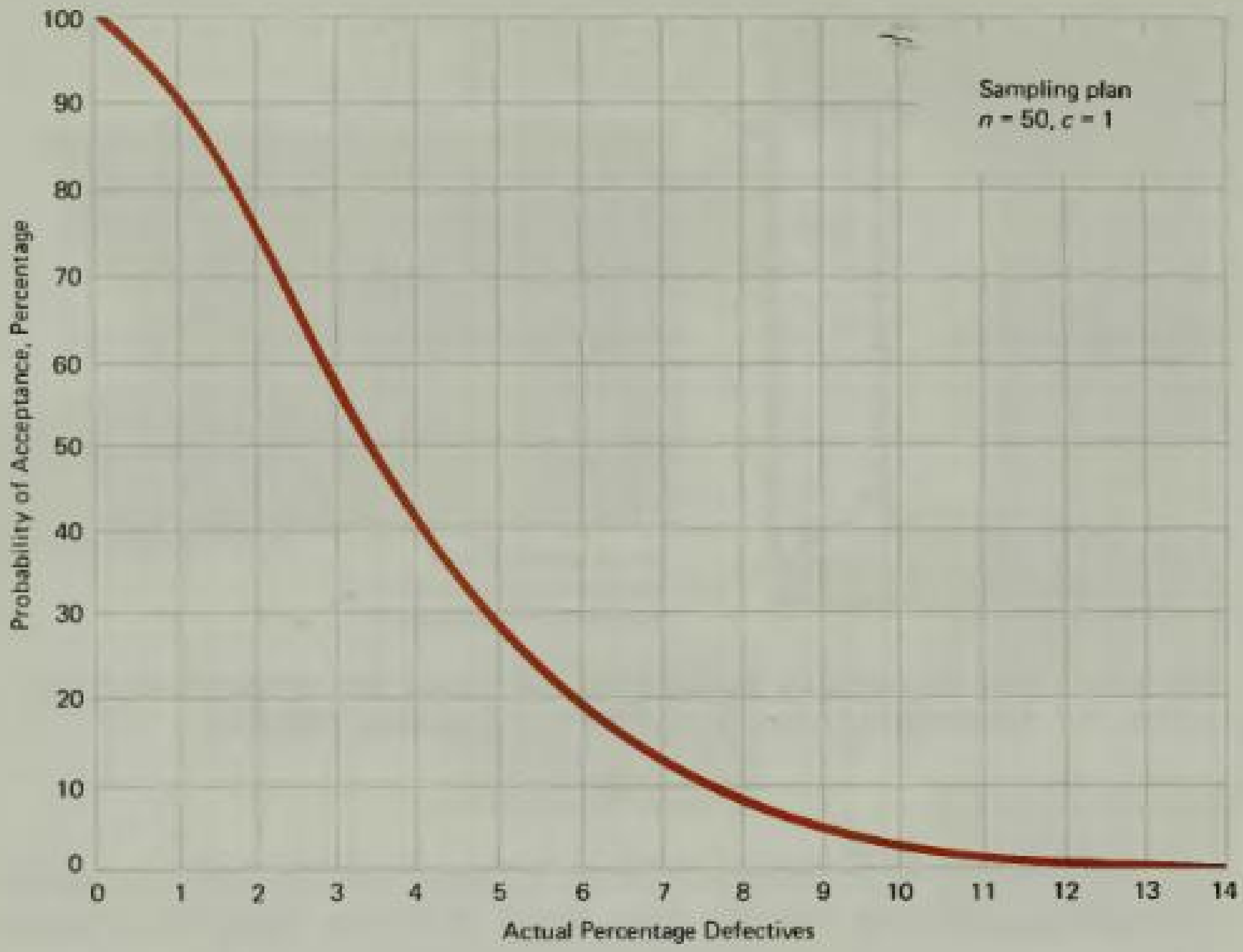


Figure 7. Operating characteristic (OC) curve for a sampling plan with $n = 50$ and $c = 1$

SOURCE: H. F. Dodge, and H. G. Romig, *Sampling Inspection Tables*, 2d ed. (New York: John Wiley & Sons, Inc., 1959)

Figure 7 shows the probability of acceptance of a lot containing various values of percentage defectives. For example, if the actual lot quality were 2 percent, samples of $n = 50$ would accept the lot as satisfactory about 73 percent of the time and reject it about 27 percent of the time. In other words, the probability of finding zero or one defectives in random samples from such a lot is 73 percent, whereas the probability of finding more than one defective is only 27 percent. However, if the actual quality of the lot were somewhat more than 2 percent defective (say 5 percent), the probability of accepting these lots would fall drastically to about 27 percent. This is the preferred situation in a sampling plan. If the actual quality is good, we want the probability of acceptance to be low. Thus, the OC curve shows how well a given plan discriminates.

The discriminating power of a sampling plan depends largely on the sample size, as we might expect; the OC curve gets steeper as the sample size increases. What happens to the OC curve if only the acceptance number changes? Primarily, the level of the OC curve changes: lower acceptance numbers make the plan "tighter"—that is, they hold outgoing quality to lower percentages of defects. By making sampling plans more discriminating (increasing sample sizes) or tighter (decreasing acceptance numbers), we can approach any desired level of outgoing quality, but at increasing inspection costs. This increased inspection effort creates lower probable costs of passing defective items, and at some point the combination of these incremental costs is at a minimum, thus defining an optimum sampling plan for a given situation.

Therefore, to justify a 100 percent sample (a vertical OC curve), the probable losses due to passing bad products would have to be large in relation to inspection costs, and perhaps would result in the loss of contracts and customers. On the other hand, to justify not inspecting, inspection costs would have to be very large in relation to probable losses due to passing defective items. The most usual situation lies between these extremes, where a risk of accepting lots that are, in fact, bad exists. As we pointed out in Figure 4, the first risk commonly is called the producer's risk, and the second is called the consumer's risk.

Industrial Practice

The concepts of acceptance sampling are used widely in industry, both for attributes- and variables-sampling. To specify a sampling plan, we must specify both the producer's and consumer's risks, and the quality levels that are regarded as acceptable and unacceptable. Then we must find a combination of sample size n and acceptance number c that meets these requirements. While the producer's and consumer's risks may be set at any level, and presumably should be set to balance the affected costs, practice has settled largely on a preset producer's risk of $\alpha = 5$ percent and consumer's risk of $\beta = 10$ percent. The mechanics of actually finding specific plans that fit can be accomplished by using standard tables, charts, or formulas [Dodge and Romig, 1959; Duncan, 1974].

Variations and extensions of the acceptance sampling concept are found in double sampling and sequential sampling. In double sampling, an initial sample is taken, and the number of defectives that is found is compared with two acceptance numbers, c_1 and c_2. If the number of defectives is less than c_1, the lot is accepted. If it exceeds the larger acceptance number c_2, the lot is rejected and subjected to 100 percent inspection. If, however, the number of defectives is between c_1 and c_2, a second sample is taken. If the total number of defectives found in the combined sample exceeds c_2, the lot is rejected and inspected 100 percent. If the number of defectives in the combined sample is less than c_2, the lot is accepted. The same general parameters specify a double sampling plan that meets requirements, and again tables, graphs, and formulas are available to the practitioner [Dodge and Romig, 1959].

Sequential sampling plans carry the idea behind double sampling a step further. Small samples are drawn at random, but after each sample is inspected the cumulative results are analyzed, and a decision is made to accept, reject, or take another sample. The sequential sample sizes can be as small as $n = 1$.

PROCESS CONTROL

As we already have noted, any real process exhibits variation. In productive processes, for example, variations occur in dimensions, in the percentage of items that meet specifications, in chemical composition of materials, and in many other ways. In clerical operations, we would expect variation in the number of clerical errors produced. In shipping operations, we would expect variation in transit time. In personnel, we would expect variation in absenteeism and in turnover. These and other situations are representative of what we would expect. In most processes, many causes of variation that stem from several sources contribute to what we term the *normal* variation in the process.

In productive processes, this complex of causes of variation stems from factors such as differences in the performance of machines and workers, minor differences in work methods among workers, variations in materials, and differences in atmospheric conditions. Machines are not capable of perfection. They may contribute to variation because of "play" in their mechanical systems, and changes in atmospheric conditions may add to or subtract from this effect. Employees are subject to all sorts of pressures in their work and private lives that may contribute to variations in work methods and performance. The materials used may vary in dimensions or composition, and may contribute to normal variation in the parameters of the system's output.

Statisticians say that these normal variations stem from random or *chance* causes. In general, we cannot do anything about these random variations without altering the system. For example, we may be able to design a new machine that is capable of greater normal precision, or perhaps we can control the temperature and humidity of the environment surrounding the productive system. These kinds of changes would produce a new normal variation that resulted from a new system of chance causes.

Statistical control theory is designed to separate relatively large causes of variation that are due to some change in the normal pattern from variations that are due to chance causes. These larger individual sources of change are called *assignable* causes. For example, the common assignable causes in a productive process are:

1. Differences among workers.
2. Differences among machines.
3. Differences among materials.
4. Differences due to interaction between any two or among all three of these factors.

A comparable set of possible assignable causes could be developed for any process. For example, assignable causes for variation in absenteeism might be disease epidemics, changes in interpersonal relations at home, or changes in an employee's work situation.

When a process is in a state of control, only chance can cause variations in the number of defects, the size of a dimension, chemical composition, turnover, absentee rates, etc. Statistical control methods allow us to set up standards of expected normal variation due to chance causes, so that when variations due to one or more of the assignable causes are superimposed, we can tell that something basic has changed. Then it is possible to investigate immediately to find the assignable cause of variation and correct it before the nonstandard condition has gone on too long. These mechanisms of statistical control are called *control charts*.

Conceptual Framework for Control Charts

If we begin with a set of measurements taken in sequence concerning a quality characteristic of some process, we may form the data into a distribution and compute the mean and standard deviation of the distribution. If we assume that the data come from a normally distributed parent distribution, we may make precise statements about the probability of occurrence associated with measurements that are a given number of standard deviation units from the mean value. Specifically:

68.7 percent of the values normally fall within $\pm\sigma$
95.45 percent of the values normally fall within $\pm 2\sigma$
99.73 percent of the values normally fall within $\pm 3\sigma$

These percentage values represent the area under the normal curve between the given limits, as shown in Figure 8. These numbers state the probability of occurrence for values that come from the current normal distribution. For example, the chances are 95.45 out of 100 that a measurement taken at random will fall within the $\pm 2\sigma$ limits and only 4.55 that it will fall outside these limits.

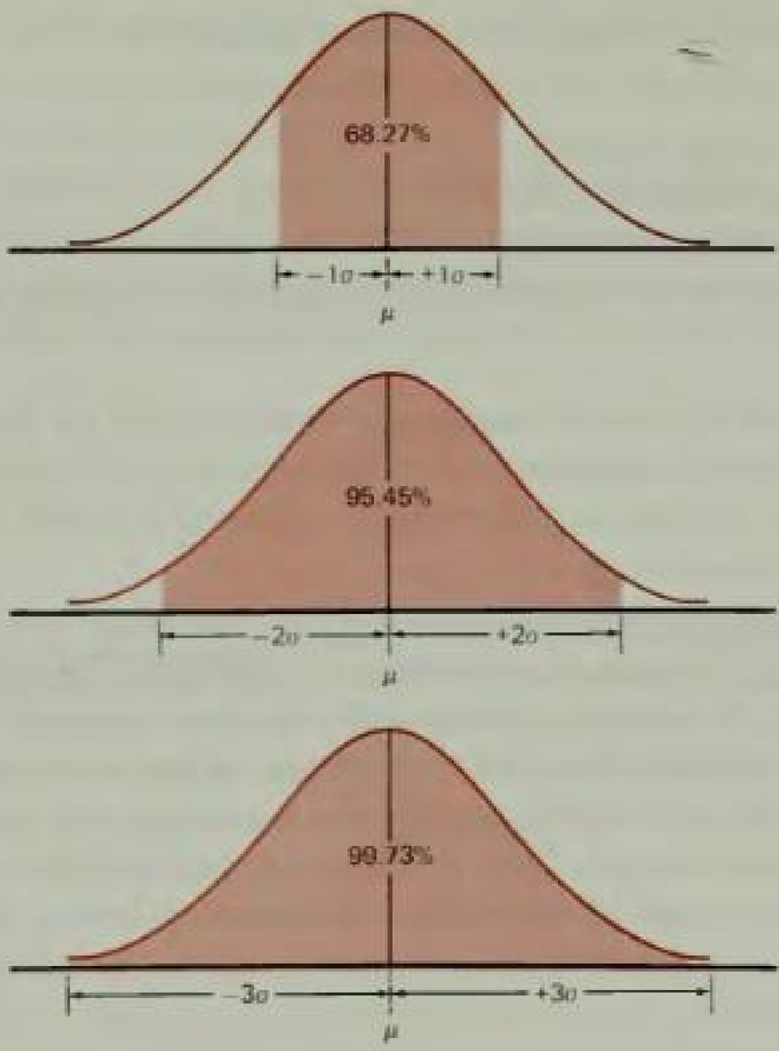


Figure 8. Areas under the normal curve for different sigma limits

These values, as well as decimal values, are available from tables for the normal probability curve. The normal capability of processes, or the natural tolerance, commonly is taken as $\pm 3\sigma$. Although $\pm 3\sigma$ represents common practice, other limits can be used. The selection of the control limits should include a consideration of the cost of letting the process continue to run when it is out of control versus the cost of stopping and adjusting the process when it does not need to be adjusted.

Sampling Distributions

If we take samples of four from the distribution of individual measurements of shaft diameters in Figure 9 and determine an average for each sample, we have a new distribution. We regard each sample mean as an observation, and if we plot the frequency distribution of the sample means, it will have a mean and a

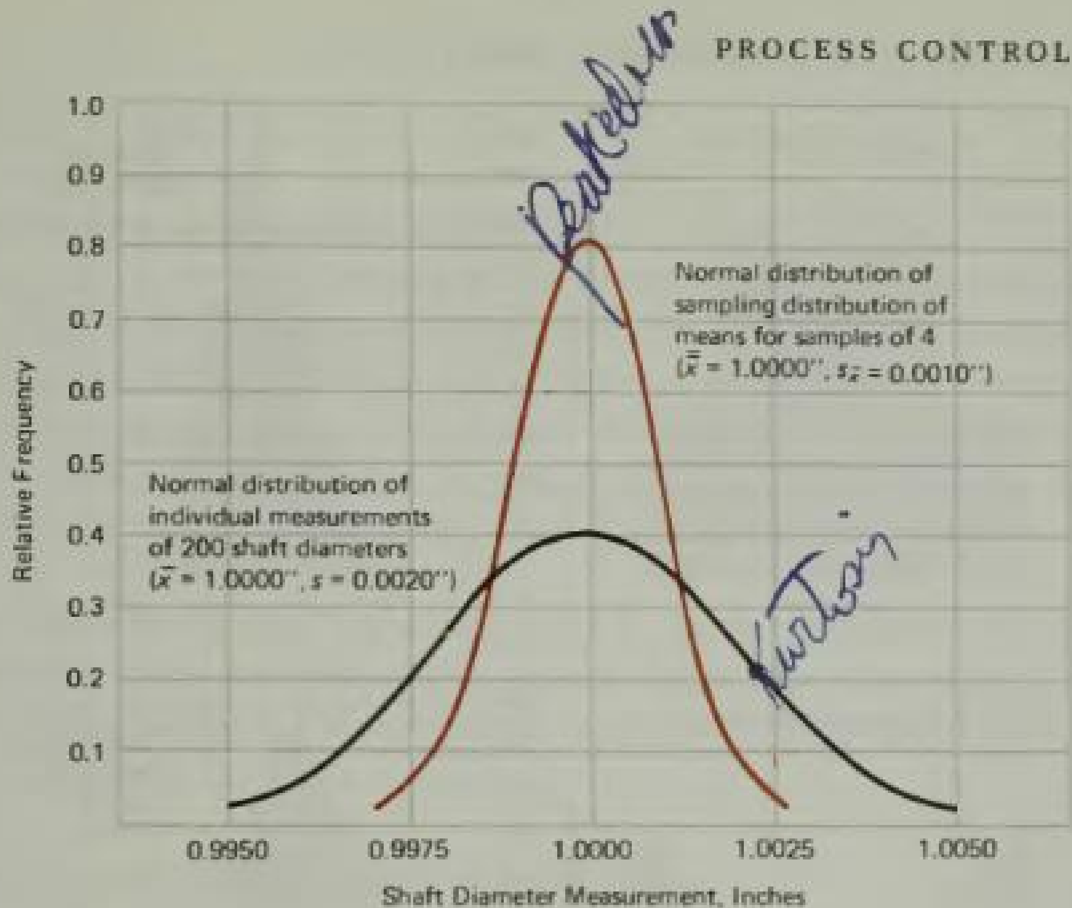


Figure 9. Relation between the distribution for individual observations and the sampling distribution of samples of 4 for shaft diameter data

variance of its own. This distribution is called a *sampling distribution of means* of $n = 4$. To distinguish the *statistics* from those of the distribution of individual observations, we use the notation $\bar{\bar{x}}$ for the grand mean of the sampling distribution, and $s_{\bar{x}}$ for the standard deviation. We expect that $\bar{x}$ and $\bar{\bar{x}}$ will be very nearly the same and that they will be equal in the limit as the number of samples increases.

The variance and standard deviation will be much smaller for the sampling distribution of means, however, since the variation is reduced by the averaging process within each sample. Figure 9 shows the resulting relationship between the two distributions for data on the diameters of 200 shafts. Actually, the relationship between s and $s_{\bar{x}}$ is given by

$$\boxed{s_{\bar{x}} = \sqrt{s^2/n}}$$

where *n* is the size of the sample.

One of the extremely important values in using sample means rather than individual measurements is that *although a universe distribution may depart radically from normality, the sampling distribution of means of random samples will be approximately normal if the sample size is large enough*. This statement, known as the *central limit theorem*, is very important, for it gives us

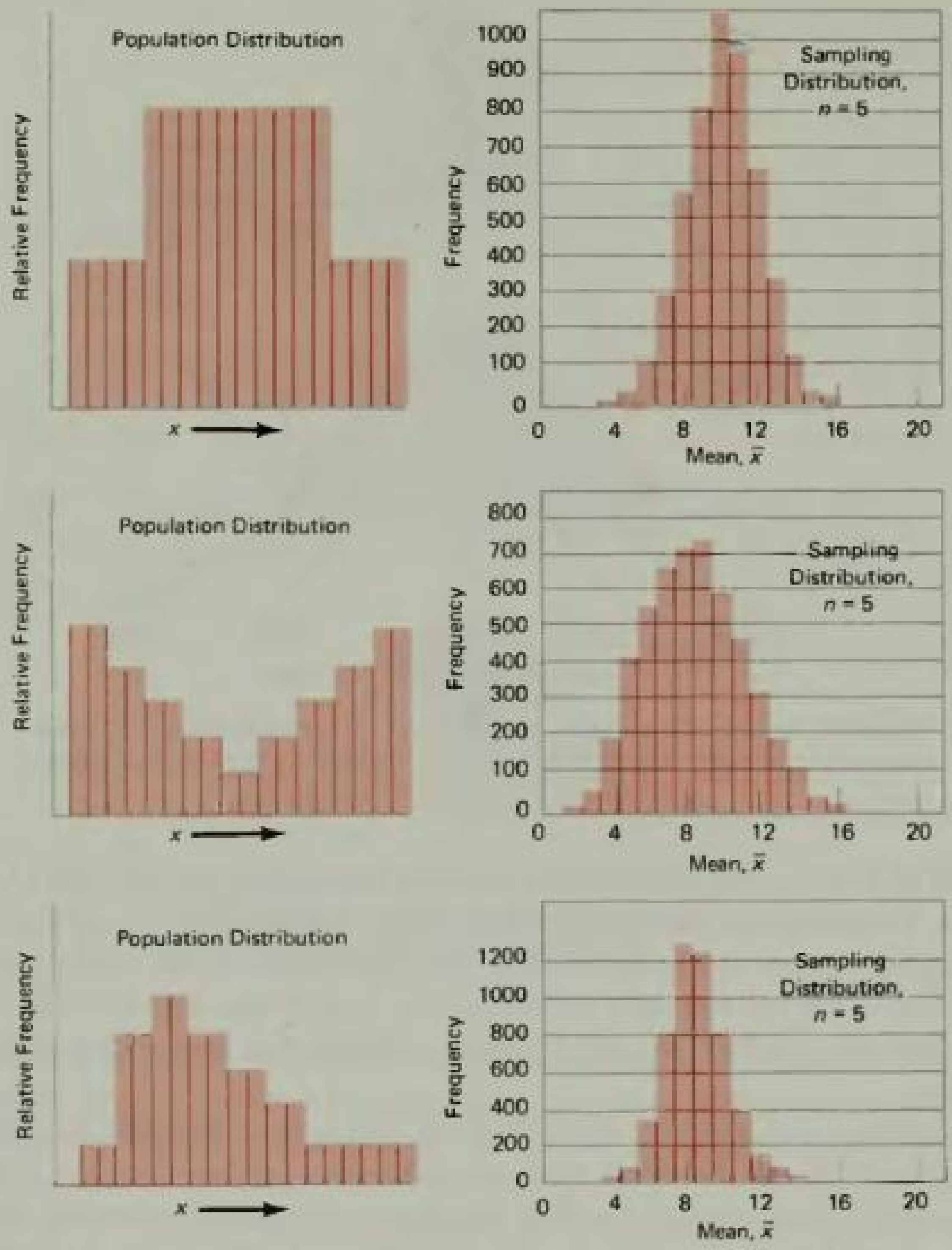


Figure 10. Normality of sampling distributions. The three distributions on the left are populations from which 5000 samples of $n = 5$ were drawn at random. The resulting sampling distributions are shown on the right

some assurance that the probabilities associated with the $\pm 3s_{\bar{x}}$ limits will apply. Figure 10 demonstrates that the deviation from normality can be fairly great; yet with samples as small as five, the distribution of sample means will follow the normal distribution quite closely.

Industrial Practice

Concepts of process control are used widely in industry both for attributes- and variables-control. When control is based on attributes, control charts of a fraction or proportion of defectives are based on the binomial distribution, and are called p-charts. When control is based on some actual measurement, $\bar{X}$-charts are developed on the basis of the normal distribution. Control charts on measures of variability also are used, frequently basing control on the range as a measure of variability [Duncan, 1974; Enrick, 1966; and Grant and Leavenworth, 1972].

Again, while the control limits may be set at any desired levels, the $\pm 3\sigma$ limits have become standard industrial practice. Table 6 in Appendix E provides factors that are useful in the convenient construction of control limits.

CONTROLLING THE QUALITY OF SERVICES

The great contrast in quality control practices between manufacturing and service organizations is the first thing that strikes an observer. The previous material on industrial quality control clearly delineated what to measure and control, as well as sophisticated methodology for accomplishing these ends. But in nonprofit organizations such as those found in government, education, foundations, hospitals and other health care units, and religious organizations, the objectives and outputs seem less well defined, and the control methodology seems relatively crude.

Let us first try to understand why these contrasts exist. The profit motive provides a focus for all kinds of managerial controls, including quality. On the other hand, nonprofit organizations exist to render service, and their success is judged in those terms. Measuring the quality of the services is difficult, in part because the attributes of quality are somewhat more diffuse. For example, how is a quality education measured? Is quality health care measured by the death rate, length of hospital stay, or the treatment process used for specific diseases? Is the quality of police protection measured by the crime rate, feeling of security by citizens, or indexes of complaints of police excesses? (It is ironic that if the size of the police force is increased, crime rates are observed to increase because more of the crimes that are committed are acted on by police and reported by observers.) Is the quality of fire service measured by reaction time, the annual extent of fire damage, or what? In partial answer to these questions, we must note that the quality characteristics of most of these kinds of services are multidimensional and often controversial, and it may be impossible to reduce quality measurement to something like specific dimensions, chemical composition, etc.

Anthony, Dearden, and Vancil [1972] summarize some of the specific problems affecting managerial control of services as follows:

1. *Relationship with the users of services.* In nonprofit organizations, a new "customer" is seen as a problem rather than an opportunity, since he or she places added strain on existing resources. If the customer brings new resources along, usually they barely cover incremental costs.
2. *Lack of competition.* Resources for nonprofit organizations are obtained from government appropriations, endowments, or fees that recover costs, rather than from the competitive marketplace. Thus, the necessity to allocate and use resources efficiently is minimized.
3. *Diffused responsibility.* In many nonprofit organizations, the line of responsibility is not a clear, single chain of delegation of power.
4. *Politics.* In government organizations, there are strong pressures to make short-run decisions that are inconsistent with the optimum use of resources, because of the annual budgeting cycle and the re-election syndrome.
5. *External pressures.* Public review, through news media and political parties, creates frequently erratic and illogical pressure on managers of public organizations.
6. *Legislative restrictions.* Government organizations must operate within statutes that are much more restrictive than the charters and bylaws of corporations, and that often prescribe detailed operating practices that are difficult to change and that may not be efficient.
7. *Behavioral factors.* The general "atmosphere" in a nonprofit organization tends to be less conducive to good management control practice than it is in profit oriented organizations.

Yet with all the above problems affecting the managerial control of services, the clamor for quality of service and its control at intended levels continues, partly through the external pressures just mentioned. In the government sector, planning, programming, and budgeting (PPB) represent a serious attempt to establish a framework within which control might be attained. The major elements of PPB systems usually involve (1) an across-the-board program structure; (2) a multiyear program and financial plan; (3) program analysis; and (4) program updating and review procedures. By establishing a multiyear program and financial plan, PPB makes possible a more rational allocation of resources. Program analysis emphasizes the definition of multiple objectives and benefits, and the establishment of information systems. This format at least makes the establishment of programs to monitor and control the quality of services possible. It still is not clear how much progress actually will result from PPB.

Quality Control in Health Care

Mounting costs, complaints, and the enormous increase in public funds allocated to health care have focused attention on the development of formal mechanisms for assessing quality. In 1972, Congress created the Professional Standards Review Organizations (PSROs) as a means of self regulation in health care. The general thrust of these organizations is to set standards for health care, as well as mechanisms for review to ensure that these standards are maintained in practice. The organizations are intended to be relatively local and, theoretically, physicians who practice medicine in local medical service areas are to be held responsible for the quality of their practice. To date, there has been a great deal of breast beating in the medical literature, but the effect of the PSROs is sure to be the progressive development of standards of care and ways of measuring and controlling the quality of health care.

Development of Health Care Standards. Current efforts to establish health care standards and control mechanisms focus on an evaluation of outcomes (did the patient die, recover, recover at what level, etc.) and inquire into the treatment process that is used if the outcomes lie outside the standardized limits. A general model for evaluating the quality of patient care has been developed by Williamson [1971], which falls into this outcomes-process assessment format. Williamson's format evaluates four factors in the following order:

1. The data required to determine the need for care, specific therapy, and prognosis (diagnostic outcomes).
2. The health status of the patients at a given time period following treatment (therapeutic outcomes).
3. The procedures carried out in order to furnish the physician with facts on which to base his or her diagnosis (diagnostic process).
4. The planning, implementing, and evaluation of therapy (therapeutic process).

The heart of the quality control process is in the establishment of standards for outcome criteria. For example, a criterion for death as an outcome might be established by determining the maximum acceptable case fatality rate for patients with a given health problem. "Peer judgment" offers a practical method for setting standards. The measurement problems for the four elements in the program are significant; however, the data base is available and can be analyzed.

Definition of Risks and Errors. Though the definitions are not couched in decisions theory terms, the medical profession has developed equivalents for Type I and Type II errors, acceptable quality level, unacceptable quality level, and risks. The key definitions are as follows:

> *False-positive*: Individuals receiving care who did not need it; Type I error.
> *False-negative*: Individuals needing care who did not receive it; Type II error.

We then can state a maximum false-positive for a given health problem as something equivalent to acceptable quality level, and we can define the probability of rejecting samples at this level as an equivalent to the producer's risk.

Similarly, we can state a maximum false-negative for a given health problem as the limit of bad quality, and we can state the probability of accepting this poor quality as the equivalent of the consumer's risk. In fact, both risks are absorbed by the patient (consumer) in the medical system, since a false-positive results in patients receiving unneeded, possibly injurious, care for which they are billed, and a false-negative results in patients not receiving needed care.

Process Study and Action. In the Williamson control format, measured findings with established criteria reveal whether detailed study of the medical care process is indicated. Williamson suggests that a 95 percent confidence interval covering the measured findings be used. For example, if the maximum acceptable case fatality rate were set at 5 percent, a measured rate of 10 percent, with confidence limits of 4–15 percent, would not be sufficiently different from the criterion to warrant process study. The result of process study, when it is invoked, might bring the standards into question, or might result in the alteration of health care procedures.

Example 1. A study was made of urinary tract infections diagnosed in a community hospital in the Midwest, involving over 6,000 consecutive admissions. Criteria were established independently by group judgment, taking into account the sensitivity and specificity of methods for detecting urinary tract infections, and the seriousness of implications to the patient of a "missed diagnosis" or a "misdiagnosis." The maximum acceptable percentage of false-negatives was set at 15 percent and of false-positives at 20 percent.

Measurement of actual outcomes revealed that 265 of the 6,145 consecutive admissions probably required urinary tract care; however, 187 of these patients did not receive this care from the regular hospital staff, resulting in a false-negative rate of over 70 percent. Of 110 patients thought by the hospital staff to have urinary tract infections, 32 had negative urine test results, producing a false-positive rate of 29 percent. Process study in this example was indicated because maximum acceptable criteria for both false-negative and false-positive diagnoses were exceeded by measured findings [Gonella, Goran, and Williamson, 1970].

Example 2. Another study of urinary tract infections used the same criteria of maximum acceptable outcomes established for Example 1. Measurement of outcomes was accomplished by an independent study team who examined 133 consecutive new patients who were admitted to the medical clinic. Over three months later, the patients' charts were examined, and recorded results were compared with the findings of the study team. Of eighteen patients requiring urinary tract care, ten were missed by the clinic staff—56 percent. There were no false-positives [Williamson, 1971].

Example 3. A study of heart failure was conducted in an eastern city hospital that was interested in applying the Williamson quality control strategy to the study of patients in the emergency room. The first sample consisted of 113 consecutive admittees who were suspected of having acute coronary artery disease. Criteria were based on staff judgment and set at 5 percent as the maximum acceptable levels for both false-negatives and false-positives in diagnostic results. Measurement of diagnostic outcomes indicated a false-negative rate of 3 percent and a false-positive rate of 0 percent. Process study was not indicated, since the criteria were not exceeded [Williamson, 1971].

Example 4. A study assessing the health status of patients at the end of a one year follow-up (therapeutic outcomes) was made of seventy-five patients among the 113 consecutive emergency room patients who were suspected of having an acute coronary occlusion. The criterion of maximum acceptable fatality rate set by the medical staff by a peer judgment technique was 30 percent. Measurement of outcomes on one year follow-up revealed that 31 percent had died. Process study was not indicated, since the 95 percent confidence limits were not exceeded [Williamson, 1971].

The examples indicate that setting standards for desired outcome, monitoring performance with comparisons to standard, and taking corrective action according to the traditional control loop concept may have valid use in health care systems. While the idea of setting standards for recovery, death rates, etc., may be shocking to some, it is realistic in decision theory terms and may result in improvements of care quality and in the creation of some basis for the control of institutions that currently provide substandard care.

THE MAINTENANCE FUNCTION

Quality control procedures are designed to track characteristics of quality directly and to take action to maintain quality within limits. In some instances, the required action may be equipment maintenance. In that case, the maintenance function would act in a supporting role to keep equipment operating effectively to maintain quality standards as well as the quantitative and cost standards of output.

We can view the problem as a matter of maintaining the reliability of the entire productive system. In general, this reliability can be maintained and improved by:

1. Increasing the size of repair facilities and crews so that average machine downtime is reduced because maintenance crews are less likely to be busy when a breakdown or call for service occurs.
2. Using preventive maintenance where practical so that critical parts are replaced before they fail. It often is possible to do this on second and third shifts, thus avoiding interfering with normal schedules. Whether preventive maintenance is worthwhile or not depends on the distribution of breakdowns and the relation of preventive maintenance time to repair time, as we shall see.
3. Providing for slack in the system at critical stages so that parallel paths are available. This means that excess capacity is maintained so that some machines can be down without affecting the delay costs to any great degree.
4. Making individual components within a machine, or the machines within the system, more reliable through improvements in engineering design. For example, special lubrication systems may be used to extend the life of working parts.
5. Decoupling successive stages of the productive system by inventories between operations, when possible. The resulting independence of operations localizes the effect of a breakdown, so operations preceding and following the machine that is down are less likely to be affected.

Since increasing reliability by any of these means is costly, we can justify it only insofar as the costs are offset by cost reductions in idle labor, scrap, lost business, etc. Method 4 may be regarded as an engineering design problem, coupled with economic analysis. Method 5 interacts with problems of layout and inventory control, which we have discussed in previous chapters. Methods 1, 2, and 3, however, represent ideas that we have not yet discussed.

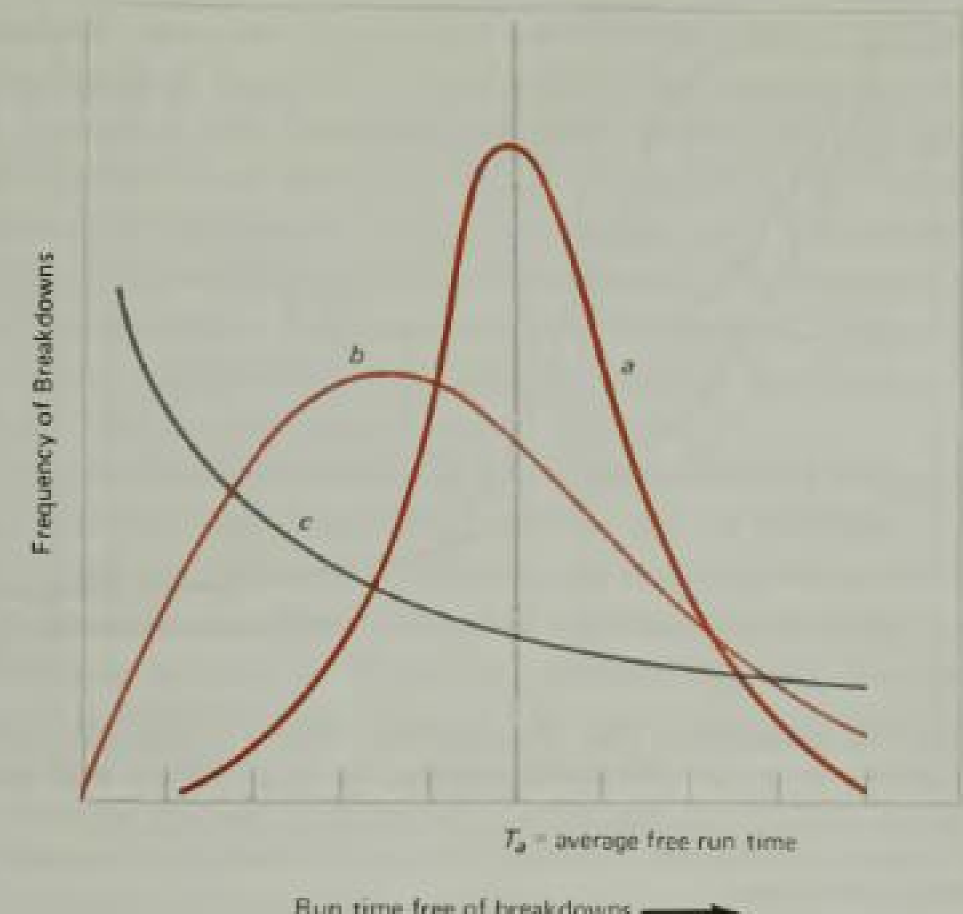


Figure 11. Frequency distribution of run time free of breakdowns representing three degrees of variability in free run time

BREAKDOWN TIME DISTRIBUTIONS

Breakdown time distribution data are basic to the formulation of any general policies concerning maintenance. Breakdown time distributions show the frequency with which machines have maintenance-free performance for a given number of operating hours. Ordinarily, they are shown as distributions of the fraction of breakdowns that exceed a given run time. Breakdown time distributions are developed from distributions of run-time free of breakdowns, as shown in Figure 11.

Figure 12 shows three breakdown time distributions. These distributions take different shapes, depending on the nature of the equipment with which we are dealing. For example, a simple machine with few moving parts would tend to break down at nearly constant intervals following the last repair. That is, it would exhibit minimum variability in breakdown time distributions. Curve *a* of Figure 11 would be fairly typical of such a situation. A large percentage of the breakdowns occur near the average breakdown time, T_a, and only a few occur at the extremes.

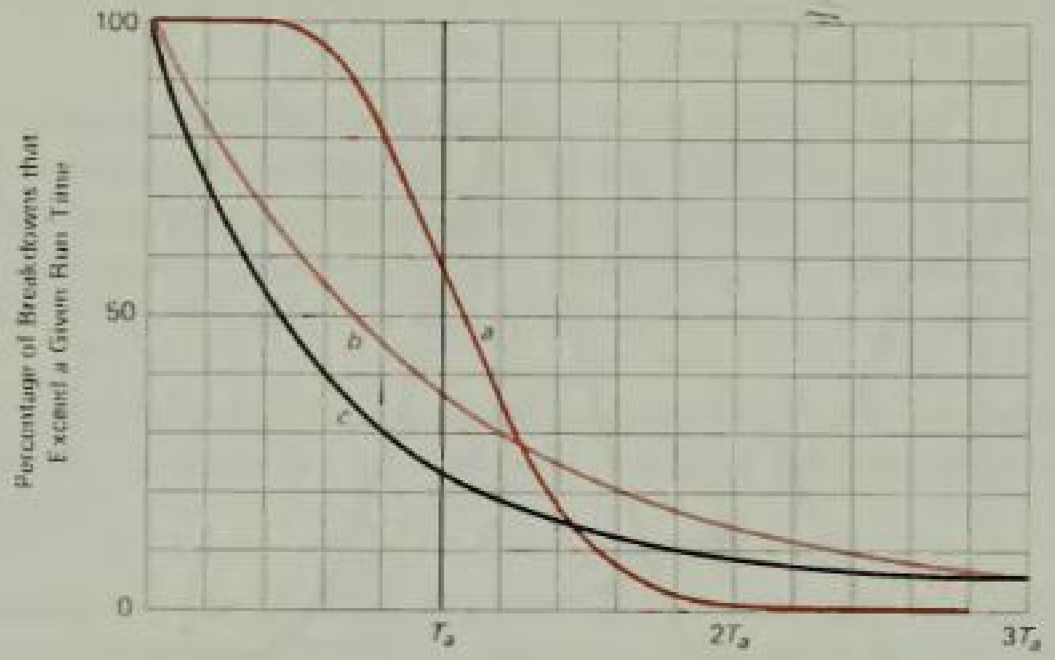


Figure 12. Breakdown time distributions. Curve *a* exhibits low variability from the average breakdown time T_a. Curve *b* is the negative exponential distribution and exhibits medium variability. Curve *c* exhibits high variability; vertical line shows constant breakdown times

ADAPTED FROM: C. M. Morse, *Queues, Inventories, and Maintenance* (New York: John Wiley & Sons, Inc., 1958)

In a more complex machine with many parts, each part would have a failure distribution. When all these parts were grouped together in a single distribution of the breakdown time of the machine for any reason, we would expect to find greater variability because the machine could break down for any one of a number of reasons. Some breakdowns could occur shortly after the last repair, or any time. Therefore, for the same average breakdown time T_a, we would find much wider variability of breakdown time, as in curve *b* of Figure 11.

To complete the picture of representative breakdown time distributions, curve c is representative of distributions with the same average breakdown time T_a, but with wider variability. A large proportion of the breakdowns with a distribution such as curve *c* occurs just after repair; on the other hand, many machines have a long running life after repair. Curve *c* may be typical of machines that require "ticklish" adjustments. If the adjustments are made just right, the machinery may run for a long time; but if not, readjustment and repair may be necessary almost immediately.

In models for maintenance, we normally deal with distributions of the percentage of breakdowns that exceed a given run time, as shown in Figure 12. They are merely transformations of the distributions of free run time typified by those in Figure 11. Taking curve *a* of Figure 11 as an example, we may convert it to the breakdown time distribution of curve *a* in Figure 12 in the following way. If the vertical scale of Figure 11 is converted to the percentage of breakdowns that occur instead of the frequency of breakdowns, we can easily plot the percentage of breakdowns that exceed a given run time. First, we know that all the breakdowns, or 100 percent, exceed an average run time of zero. To obtain curve *a* in Figure 12 we simply subtract successively the percentages that occur at different free run times. We can see by examination of Figure 12 that almost 60 percent of the breakdowns exceeded the average breakdown time T_a, and that very few of the breakdowns occurred after $2T_a$.

In practice, actual breakdown time distributions often can be approximated by standard distributions, three of which are shown in Figure 12. Curve *b* is actually the negative exponential distribution that we discussed in connection with waiting line theory. From this point on, we will refer to the breakdown time distributions plotted in Figure 12, recognizing that the basic data for breakdown time distributions would be accumulated as a distribution of the frequency of free run time, as shown in Figure 11.

PREVENTIVE MAINTENANCE

Assume a preventive maintenance policy for a single machine that provides for an inspection and perhaps replacement of parts after the machine has been running for a fixed time, called the *preventive maintenance period*. The maintenance crew takes an average time, T_m, to accomplish the preventive maintenance. This is the *preventive maintenance cycle*. A certain proportion of the breakdowns will occur before the fixed cycle has been completed. For these cases, the maintenance crew will repair the machine, taking an average time, T_s, for the repair. This is the *repair cycle*. These two patterns of maintenance are diagrammed in Figure 13. The probability of occurrence of the two different cycles depends on the specific breakdown time distribution of the machine and the length of the standard preventive maintenance period. If the distribution

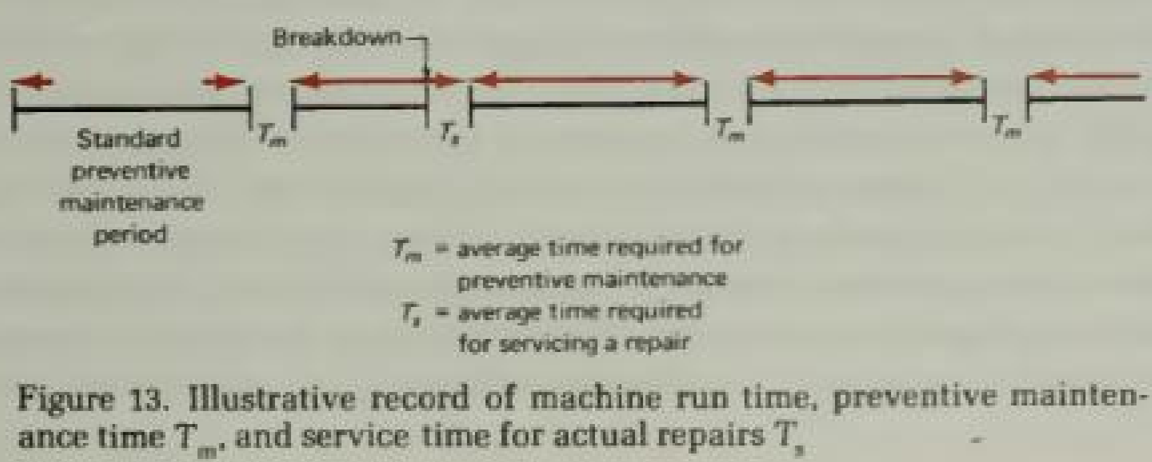


Figure 13. Illustrative record of machine run time, preventive maintenance time T_m, and service time for actual repairs T_s

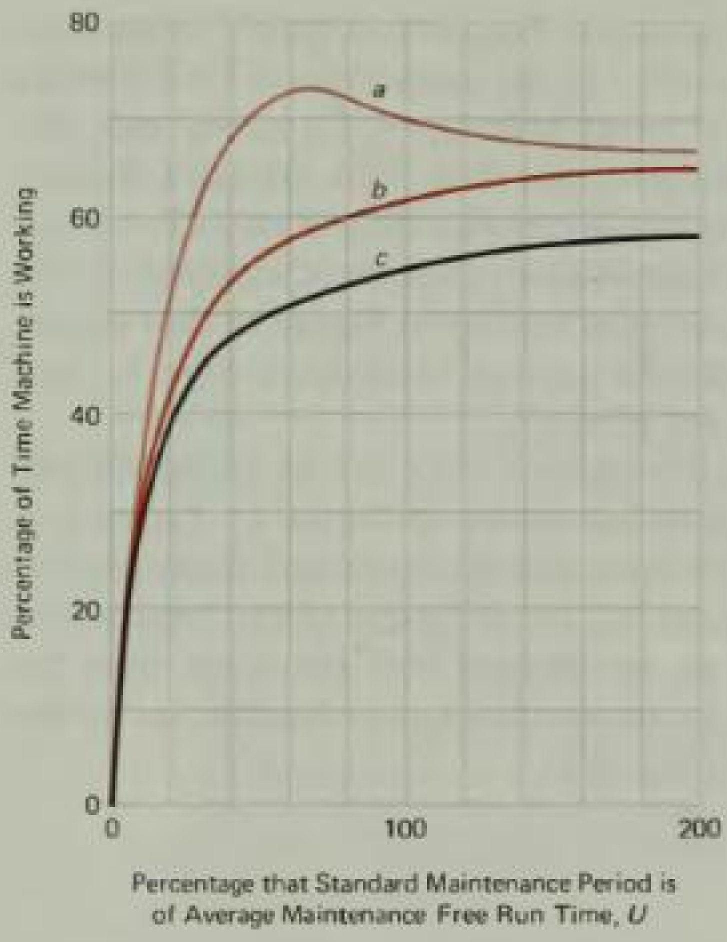


Figure 14. Percentage of time a machine is working for the three distributions of breakdown time shown in Figure 12. Preventive maintenance time, T_m, is 20 percent of T_a; repair time is 50 percent of T_a

ADAPTED FROM: C. M. Morse, *Queues, Inventories, and Maintenance* (New York: John Wiley & Sons, Inc., 1958)

has low variability and the standard period is perhaps only 80 percent of the average run time without breakdowns, T_a, actual breakdown would occur rather infrequently, and most cycles would be preventive maintenance cycles. If the distribution were more variable, for the same standard preventive maintenance period more actual breakdowns would occur before the end of the standard period. Obviously, shortening the standard period would result in fewer actual breakdowns, and lengthening it would have the opposite effect for any distribution.

Assuming that either a preventive maintenance or a repair puts the machine in shape for a running time of equal probable length, Morse [1958] has shown that the percentage of machine running time depends on the ratio of the standard maintenance period and the average run time, T_a, for the breakdown time distribution. Figure 14 shows the relationship between the percentage of time that the machine is working and the ratio of the standard period to average run time, T_a, for the three distributions of breakdown times shown in Figure 12. In general, when the standard period is short (say less than 50 percent of T_a), the machine is working only a small fraction of the time. This is because the machine is down so often owing to preventive maintenance. As the standard period is lengthened, more actual breakdowns occur that require repair. For curves *b* and *c*, this improves the fraction of time during which the machine is running because the combination of preventive maintenance time and repair time produces a smaller total down time.

Curve *a*, however, contains a peak, or an optimum preventive maintenance period, which maximizes the percentage of machine working time. What is different about curve *a*? It is based on the low variability breakdown time distribution from Figure 12. For curve *a*, lengthening the maintenance period beyond about 70 percent of T_a reduces the fraction of machine working time because actual machine breakdowns are more likely. For the more variable distributions of curves *b* and *c* this is not true, because breakdowns are more likely throughout the distributions of these curves than they are in curve *a*. Comparable curves can be constructed showing the percentage of time that the machine is in a state of preventive maintenance, and the percentage of time that the machine is being repaired because of breakdown.

Guides to a Preventive Maintenance Policy

We can make some generalizations about preventive maintenance policy through the concepts we have developed. First, preventive maintenance generally is applicable to machines with breakdown time distributions that have low variability, exemplified by curve *a* of Figure 12. In general, distributions with less variability than the exponential, curve *b*, are in this category because low variability means that we can predict with fair precision when the majority of breakdowns will occur. A standard preventive maintenance period then can be set that anticipates breakdowns fairly well.

Equally important, however, is the relation of preventive maintenance time to repair time. If it takes just as long to perform a preventive maintenance as it does to repair the machine, there is no advantage in preventive maintenance, since the amount of time that the machine can work is reduced by the amount of time that it is shut down for repairs. In this situation, the machine will spend a minimum amount of time being down for maintenance if we simply wait until it breaks down. This is shown by Figure 15, in which we compare the percentage of the time that a machine is working when repair time is greater than preventive maintenance time with the percentage when the two times are equal.

Both curves are based on the low variability distribution of curve *a*, Figure 12. Note that curve *d* exhibits an optimum, but curve *e* does not. For curve *e*, the percentage of time that the machine works continues to increase as the standard maintenance period is lengthened, which results in more repairs and fewer preventive maintenance cycles. Clearly, there is no advantage in preventive maintenance when $T_m \geq T_s$ from the viewpoint of maximum machine work-

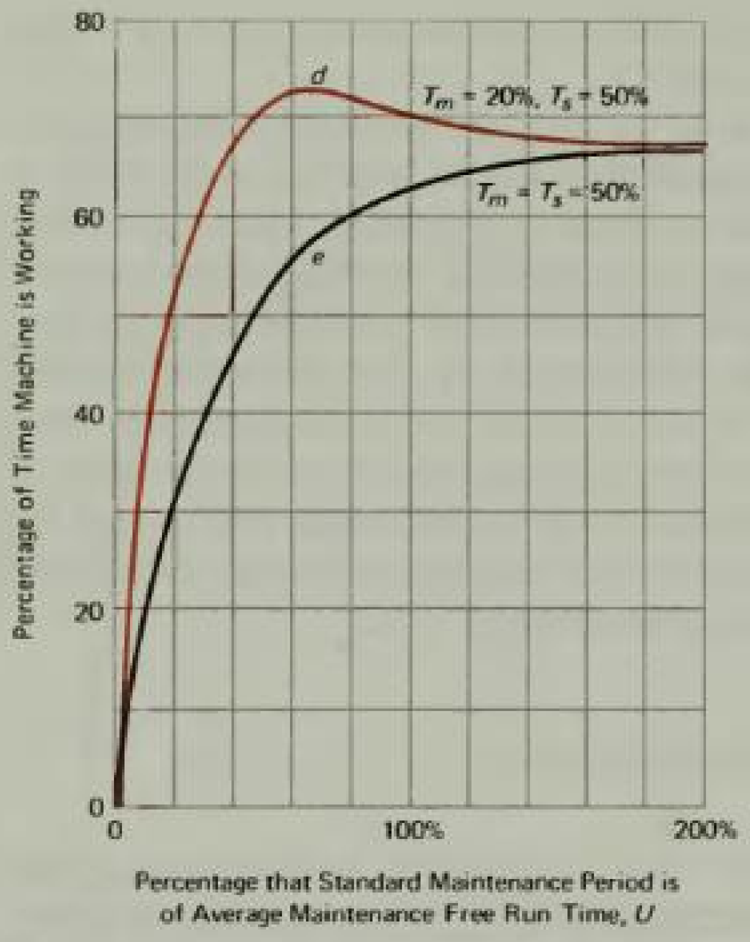


Figure 15. Comparison of percentages of machine working time when preventive maintenance time is less than repair time, curve *d*, and when they are equal, curve e. Curve d is identical with curve a, Figure 14. T_m and T_s are expressed as percentages of the average breakdown time, T_a.

ing time. In summary, preventive maintenance is useful when breakdown time distributions exhibit low variability and when the average time for preventive maintenance is less than the average time for repair after breakdown.

The effect of downtime costs can modify these conclusions, however. Suppose that we are dealing with a machine in a production line. If the machine breaks down, the entire line may be shut down, and very high idle labor costs will result. In this situation, preventive maintenance is more desirable than repair *if* the preventive maintenance can take place during second or third shifts, vacations, or lunch hours, when the line normally is down anyway. This is true even when $T_m \geq T_s$. The determination of the standard preventive maintenance period would require a different, but similar, analysis in which the percentage of machine working time is expressed as a function of repair time only, since preventive maintenance takes place outside of normal work time.

An optimal solution would minimize the total of downtime costs, preventive maintenance costs, and repair costs. The effect of the downtime costs would be to justify shorter standard preventive maintenance periods and to justify making repairs more quickly (at higher cost) when they do occur. There are many situations, however, in which extra manpower on a repair job would not speed it up. In such cases, total downtime might be shortened by overtime on multiple shifts and weekends, with attendant higher costs. Optimal solutions would specify the standard preventive maintenance period, the machine idle time, and the repair crew idle time, striking a balance between downtime costs and maintenance costs.

A special case exists when sales require the plant to operate at full capacity to meet demand. Under these circumstances, reduced machine working time due to repair and preventive maintenance reduces the number of units that can be sold and therefore affects income. Optimal maintenance policy then is weighted even more heavily in favor of shortening standard preventive maintenance periods and using more effort to repair machines quickly when breakdowns occur. Morse [1958, pp. 165–166] has developed basic models that fit this special case by maximizing an income-less-maintenance cost function.

MAINTAINING SEVERAL MACHINES

The single machine situation that we have been discussing contains basic elements of general policy that can be carried over into the multimachine case. However, when several machines must be serviced, our problem more closely resembles the usual waiting line model. If we assume that all machines have the same breakdown time distribution, breakdowns are comparable to arrivals in the waiting line model, and the repair crew is comparable to the service station. As machines break down, the repair crew services them in the average time, T_s, as before. If the crew is already working on a machine, successive machines that break down must wait for service, and the costs associated with downtime grow with the delay. We can reduce the chance that this will happen by increasing the size of the crew, but this solution costs money and increases the amount of time that the crew will be idle while waiting for breakdowns to occur. The problem involves striking a balance between the downtime costs of the machines and the idle time costs of the maintenance crew [Buffa, 1972; Morse, 1958]. Since the arriving population is finite, we may use the methods of the finite queuing model discussed in Appendix D, and the *Finite Queuing Tables* [Peck and Hazelwood, 1958], which are reproduced in part in Table 4 of Appendix E.

Summary

When productive systems involve a network of activities with many required sequences of activities, it is difficult to maintain the reliability of the system as a whole, even though the performance of each individual operation may be

99 percent reliable, because of the fundamental mathematical combination of the reliability of components in series. By contrast, two or more components in parallel that can perform the same function result in systems with increased reliability.

In productive systems, quality control and equipment maintenance systems are used to sustain reliability. The quality control system functions as a primary control loop, and the maintenance system functions as a secondary, supporting, control loop, maintaining reliability in the longer term.

Quality control begins in an enterprise's preproduction planning phases, in which policies regarding market strategies are developed. Quality standards then are developed out of the iterative process of product–service design and productive system design. The system must be designed so that it can produce the set quality levels at reasonable cost.

Monitoring quality levels of output is necessarily a sampling process, since the entire population seldom is available for screening. In general, two kinds of statistical methods are used in quality control: acceptance sampling and process control. These classifications are further divided into methods that are designed for attributes- and variables-control.

Quality control of services is difficult for a variety of reasons that are related to the unique character of services, such as the relationship with clients, lack of competition as a force for control, diffused responsibility, politics, external pressures, and legislative restrictions. Nevertheless, attempts are being made to establish a framework for all kinds of control in government systems through planning, programming, and budgeting (PPB). Legislation has placed great emphasis on quality control in health care systems, and self regulation experiments now are developing. One of the promising approaches for controlling the quality of health care involves outcome evaluation, with inquiry into health care processes when outcomes fall outside of established standards.

The equipment maintenance system supports the general quality control system. Analysis of breakdown time distributions provides guidelines to the development of preventive maintenance policies. In general, these policies are appropriate when breakdown time distributions exhibit low variability, and when the average time for preventive maintenance is less than the average repair time following breakdown. Also, when downtime costs are large, preventive maintenance is preferable to repair if it can be performed when the facilities normally are down anyway.

The multiple machine, multiple repair crew maintenance problem follows the structure of finite waiting line models and can be analyzed with the aid of the *Finite Queuing Tables*, if the assumptions apply. Such analyses can help us determine optimal repair crew sizes and other aspects of the organization of the maintenance function. Simulation studies have been effective when the problem structure is complex and when the usual mathematical assumptions do not apply.

Review Questions

1. If the *reliability* of a productive system means its capability to maintain quality, schedule, and cost standards, which of the following kinds of systems is likely to encounter the greatest quality-reliability problems? schedule-reliability problems? Why?
   a. Skyscraper construction project.
   b. Automobile engine plant.
   c. Restaurant.
   d. Hospital.

565 2. Compare the reliability of systems with components in series versus components in parallel.

3. How does variability of performance affect system reliability?

4. What are the primary- and secondary feedback control loops in the broad reliability system?

5. Define the following terms:
   a. Type I error. 572
   b. Type II error.
   c. Producer's risk. 572
   d. Consumer's risk.

6. What kinds of control can be exercised in maintaining quality standards?

7. What conditions make acceptance sampling appropriate?

8. What is an OC curve, and what information does it convey? 575

9. What is the effect of decreasing acceptance numbers in the OC curve? What practical result does it achieve in terms of outgoing quality levels?

579

10. What are the sources of assignable causes of variation in quality control?

11. What is the conceptual framework for statistical control charts?

12. What is the significance of the fact that the sampling distribution of means can be normal even though the distribution of individual measurements departs radically from normality?

13. What is the relationship between s and $s_{\bar{x}}$?

583

14. What are the criteria of quality in:
    a. Banking service.
    b. The post office.
    c. An institution of higher learning.
    d. A hospital.
    e. The Internal Revenue Service.

15. Discuss some of the problems that affect managerial control of services.

16. In controlling the quality of health care, what is the meaning of a *false-positive*? a *false-negative*?

17. In the Williamson control model for health services, under what conditions do we examine the treatment processes that are used?

18. Select some service system other than health care and see if you can establish some basis for the following:
    a. Standards for the quality of the service, and how it is measured.
    b. The nature of a Type I error.
    c. The nature of a Type II error.
    d. Means used (formal or informal) for controlling the quality of the service.

19. What kinds of costs are associated with machine breakdown?

20. Discuss the general methods by which the reliability of productive systems can be maintained.

21. What is a breakdown time distribution?

589

22. Discuss the types of situations of machine breakdown that are typified by curves *a*, *b*, and *c*, respectively, in Figures 11 and 12.

23. What are the general conditions for which preventive maintenance is appropriate for a single machine?

24. If it takes just as long to perform a preventive maintenance as it does a repair, is there an advantage to preventive maintenance? How can high downtime costs modify this?

25. Compare the problem of maintenance of a bank of machines to the general waiting line model discussed in Chapter 4 and Appendix D.

References

Anthony, R. H.; J. Deardon; and R. F. Vancil. *Management Control Systems.* Homewood, Ill.: Richard D. Irwin, 1972.

Barlow, R., and L. Hunter. "Optimum Preventive Maintenance Policies." *Operations Research* 8(1960): 90–100.

Bovaird, R. L. "Characteristics of Optimal Maintenance Policies." *Management Science* 7(April 1961): 238–254.

Bowker, A. H., and H. P. Goode. *Sampling Inspection by Variables.* New York: McGraw-Hill, 1952.

Buffa, E. S. *Operations Management: Problems and Models.* 3d ed. New York: John Wiley & Sons, 1972. Chapters 15, 19, and 20.

Cunningham, C. E., and W. Cox. *Applied Maintainability Engineering.* New York: John Wiley & Sons, 1972.

Dodge, H. F., and H. G. Romig. *Sampling Inspection Tables.* 2d ed. New York: John Wiley & Sons, 1959.

Duncan, A. J. *Quality Control and Industrial Statistics*. 4th ed. Homewood, Ill.: Richard D. Irwin, 1974.

Enrick, N. L. *Quality Control*.] 5th ed. New York: Industrial Press, 1966.

Gilbert, J. O. W. *A Manager's Guide to Quality and Reliability*. New York: John Wiley & Sons, 1968.

Goldman, A. S., and T. B. Slattery. *Maintainability: A Major Element of System Effectiveness*. New York: John Wiley & Sons, 1964.

Gonella, J. A.; M. J. Goran; and J. W. Williamson et al. "The Evaluation of Patient Care." *Journal of the American Medical Association* 214(1970): 2040–2043.

Gradon, F. *Maintenance Engineering: Organization and Management*. New York: John Wiley & Sons, 1973.

Grant, E. L., and R. S. Leavenworth. *Statistical Quality Control*. 4th ed. New York: McGraw-Hill, 1972.

Hansen, B. L. *Quality Control*. Englewood Cliffs, N. J.: Prentice-Hall, 1963.

Hardy, S. T., and L. J. Krajewski. "A Simulation of Interactive Maintenance Decisions." *Decision Sciences* 6(1975): 92–105.

Harris, D. H., and F. B. Chaney. *Human Factors in Quality Assurance*. New York: John Wiley & Sons, 1969.

Jardine, A. K. S. "Decision Making in Maintenance." *Chartered Mechanical Engineer* 21(January 1974): 35–39.

______. *Maintenance, Replacement and Reliability*. New York: John Wiley & Sons, 1973.

Juran, J. M., and F. M. Gryna. *Quality Planning and Analysis: From Product Development Through Usage*. New York: McGraw-Hill, 1970.

Kirkpatrick, E. G. *Quality Control for Managers and Engineers*. New York: John Wiley & Sons, 1970.

Lusser, R. "The Notorious Unreliability of Complex Equipment." *Astronautics* 2(February 1958): 74–78.

Meister, D., and G. F. Rabideau. *Human Factors Evaluation in System Development*. New York: John Wiley & Sons, 1965.

Morse, P. M. *Queues, Inventories, and Maintenance*. New York: John Wiley & Sons, 1958.

Peck, L. L., and R. N. Hazelwood. *Finite Queuing Tables*. New York: John Wiley & Sons, 1958.

Smith, C. S. *Quality and Reliability: An Integrated Approach*. New York: Pitman, 1969.

Turban, E. "The Use of Mathematical Models in Plant Maintenance." *Management Science* 13(February 1967): 342–358.

Vance, L. L., and J. Neter. *Statistical Sampling for Auditors and Accountants*. New York: John Wiley & Sons, 1956.

Weiss, G. H. A Problem in Equipment Maintenance." *Management Science* 8(April 1962): 266–278.

Williamson, J. W. "Evaluating Quality of Patient Care." *Journal of the American Medical Association* 218(October 25, 1971): 564–569.

Part V

Synthesis

Chapter 17

Production/Operations Management—Summary

Times are changing, and in no place is this fact more evident than in the field of production/operations management. The field is in a state of flux from several viewpoints. The word *production* reflects a broader view than it used to. Today, it deals with the *operations* side of any enterprise, and we may find productive systems in offices, stores, and hospitals, as well as in factories. All these systems involve inputs, some kind of processing, and outputs. Production/operations management deals with decision making within the system. One recent conceptual change is that waste in the system and its possible effect on pollution now are regarded as part of the processing system.

This book has been organized around the problems encountered in the design and operation of productive systems, with a background of applicable analytical methods. Within each problem area, we discussed how that problem was affected by the special conditions of different kinds of productive systems. Let us now try to integrate our knowledge in terms of each of the major kinds of systems: distribution systems; production-distribution systems for high volume products; intermittent systems; large-scale projects; and service and nonmanufacturing systems.

DISTRIBUTION SYSTEMS

In essence, the management of distribution systems is centered in the management of an inventory system, usually a multistage inventory system. The design of the physical distribution system is concerned with the strategic location of inventory points in relation to markets in order to provide the required service at reasonable transportation cost. Logistics dominate the system design in determining whether or not to use intermediate warehouses and, if so, where to locate them.

The operating problems of distribution systems are centered in inventory management. What is the distribution of demand? How do we forecast requirements? When should we replenish inventories? How much inventory should be ordered at one time? What buffer stocks should be maintained to offer the service needed? How do we determine the appropriate service level?

Though the problems are similar for each level in the system, they have a somewhat different emphasis. For example, for the retailer the problem of determining distribution of demand and forecasting demand is local, while at the factory warehouse it may be national or international in scope. On the other hand, the factory warehouse may need a data processing system for a relatively few sizes, types, and styles, whereas the distributor and retailer are likely to be dealing with a much larger number of items. The concept of service and service systems is physical in terms of the inventories required at each level; however, at the retail level, service systems must include not only the availability of inventory but also a service system that is designed around the concept of waiting line analysis.

The major production and operations management problems that confront distribution kinds of organizations are as follows:

1. Determination of the nature of the distribution of demand.
2. Forecast of demand.
3. Determination of how much to order at one time.
4. Determination of when to reorder.
5. Determination of service levels and the size of buffer stocks.
6. Design of data processing systems.

HIGH VOLUME PRODUCTION-DISTRIBUTION SYSTEMS

Adding a production system to the head end of a distribution system changes the focus and adds enormous complexity to the managerial problems. The system under managerial control centers around the productive system and to some degree extends downstream into the distribution system, depending on the particular organization. In some instances, the system under managerial control may include the entire distribution system. One of the great advantages of coupling the production and distribution systems is that inventories become an important trade off to other means of absorbing fluctuations in demand, and the kinds of aggregate planning models discussed in Chapter 10 take on considerable significance.

Forecasting consumer demand is extremely important both in the short and long run. In the short run, forecasts are the basis for raw material procurement and of aggregate planning and scheduling of facilities and labor in order to combine the various sources of capacity in the most economical way. Longer-range forecasts, however, are significant in planning long-range aggregate capacity and its location and layout, as well as the size and location of distribution warehouses.

The process of generating day-to-day schedules for high volume systems involves working within the constraints of the aggregate plan and taking advantage of whatever flexibilty may be left to "cut and try" for a practical schedule. This may include rebalancing lines, adjusting inventory, or using overtime and subcontracting. The nature and design of the information and data-processing system that links measurement of actual demand, forecasting, inventory control, and production control is extremely important.

Though no one would say that the design, planning, scheduling, and control problems of high volume production-distribution systems are simple, nonetheless they are simpler than for intermittent systems.

The major production and operations management problems that confront organizations that produce and distribute high volume standardized products are as follows:

1. Forecasting demand and the behavior of multistage inventory systems.
2. Long range aggregate planning for facilities—plant capacility, sizes, and location, and warehouses sizes and location.
3. Production facility design, involving job and process design as well as physical layout.
4. Aggregate planning and scheduling for facilities and manpower.
5. Raw materials procurement.

6. Day-to-day scheduling and adjustment of output levels as demand becomes known.
7. Design of data processing systems.

INTERMITTENT SYSTEMS

In contrast with the high volume system, intermittent systems key everything to the basic requirement of holding facilities and manpower "in inventory" to supply the needs of a variable demand (in terms of its design, style, and technological requirements). Thus, the jobbing printer holds in readiness equipment and trained personnel who are capable of performing a wide variety of operations to reproduce the printed word. Jobbing machine shops hold in readiness equipment and trained mechanics who can perform a wide variety of operations on various metals of different sizes, types, and designs. If the number of orders falls temporarily, such an organization does not sell its equipment, nor does it fire its skilled mechanics; for it is this capability that is for sale. Similarly, and perhaps even more essential for organizational survival, aerospace and other research and developmental organizations may stockpile engineering and scientific brains because this is the kind of capability that is crucial in obtaining contracts.

Although most often the relationships with the eventual consumer may be direct, the internal complications are tremendous. These complications arise from the custom nature of the fabrication process, in which each order or item requires individual planning and scheduling and follows a unique processing sequence. The typical time delays in the information system occur in the bid and order procedure, the special production planning and scheduling requirements, and the special ordering of materials. The physical flow time involved in actual fabrication and assembly usually is relatively long because of the intermittent nature of physical flow. The inventory problems largely concern raw material and in-process inventories, and the scheduling problem focuses more on the use of individual pieces of equipment than on the factory as a whole as with high volume systems.

The most complex job shop system is the open shop, which is open to job orders from virtually anyone. Under such circumstances, we must forecast, design the physical facility, make aggregate plans, schedule, procure materials, and bid with the greatest uncertainty. The closed job shop is the captive shop of some concern, and it manufactures for its own internal use in its own product line. Its product line usually has a degree of predictablility, although the captive shops also may receive internal one-time orders. Closed job shops produce a largely forecastable line of parts, components, and products. This is an important distinction because, if we know in advance what our product-mix will be, then the problems take on a considerably different hue.

Planning and control methods for intermittent systems are unique, and are termed *requirements planning*. The requirement or demand for most component parts is not well represented by a statistical model. Rather, the demand for these items is *dependent* on the production schedules of primary and secondary products. Because of the demand dependence, the usual assumption of constant demand in EOQ models is not valid, and the determination of production lot-sizes is achieved more economically through periodic or part-period ordering policies.

The major production and operations management problems that confront job shop kinds of organizations may be summarized as follows:

1. Designing and laying out a system to minimize aggregate handling cost.
2. Forecasting demand.
3. Aggregate planning for the use of facilities.
4. Scheduling orders to meet promised delivery dates.
5. Scheduling labor and equipment to minimize combined costs of machine setup, machine downtime, labor overtime and undertime, and in-process inventories.
6. Scheduling equipment to utilize most efficient processes.
7. Procuring materials in economical quantities to mesh with the production schedule.
8. Developing bidding policy and procedure to obtain orders at margins that will achieve a balance between the use of labor and facilities and the desire for profit.

LARGE-SCALE PROJECTS

A structural model for large-scale projects is not very different from a job shop model. The differences lie in the immense complexity and in the extensive time delays and coordination that are required within the system. In planning what must be done, the very complexity of the project makes process sequences extremely important, for operations that are performed out of sequence can cause delays and extra costs. Thus, developing the production plan as a network of operations is the key to managerial control of the project.

Given the network of required operations, a second focus of problems centers on scheduling and effectively using available resources. The generation of the critical path schedule is an important input to the scheduling and control process, and it may be regarded as a first feasible schedule. Knowing the permissible slack or schedule slippage of certain operations gives management flexibility in achieving a practical schedule. This flexibility also can be used to level the labor requirements over the entire project, or to enable the use of limited equipment for several operations in a way that neither conflicts with nor extends the project time.

Although at first the inventory problem seems simple and direct, it is more complex than we might think. Inventory cost in general is directly related to overall project time. To minimize the investment in inventories during the project, material receipt schedules must be coordinated carefully with the schedule of operations. If orders for raw material supply are released in a block at the beginning of the project, a large fraction of inventories will be held for a much larger period than necessary and also probably will create a physical storage problem. If material is wasted or scrapped so that a certain operation cannot be completed, then the entire project may be delayed if the activity is on the critical path. Thus, an important inventory decision problem centers on the quantity of material to be ordered in the first place. A relatively inexpensive item of material could cause idle labor costs, or possible penalties for not meeting project delivery dates.

In summary, the major production and operations management problems of large-scale projects are:

1. Planning a network of operations to accomplish the desired end result.
2. Developing schedules of the network of operations so that the network's critical path schedule meets promised delivery dates.

economy

3. Allocating the use of limited resources of equipment and/or labor in ways that will not interfere with the critical path schedule.
4. Procuring materials by a schedule that minimizes total inventory costs but meets the needs of the critical path schedule.
5. Developing bidding policy and procedures to obtain contracts at margins that will achieve a balance between the use and maintenance of the stockpile of critical resources (engineers, scientists, skilled labor, key facilities, etc.) and the desire for profit.

SERVICE AND NONMANUFACTURING SYSTEMS

Service and nonmanufacturing systems recently have become significant as productive systems, partly because rising costs within such systems have resulted in public scrutiny, and partly because the scale of many of these systems has increased to the point where they cannot be ignored. One important characteristic of service and nonmanufacturing systems is that the output of the system cannot be stored. Therefore, it cannot be used as a buffer to absorb demand fluctuations. This fact, coupled with often severe short-term demand variations in service systems, creates important problems in providing the needed capacity to maintain service at reasonable cost. The work shift scheduling problem has emerged as a generally significant one.

Waiting line models help us understand how service systems function. However, simulation models have been more practical in analyzing complex, real systems. Mathematical programming models have been useful for resource scheduling in refuse collection, academic settings, and medical care systems.

The major operations management problems of service and nonmanufacturing systems are:

1. Forecasting demand for service, particularly short-term demand fluctuations.
2. Designing systems that provide managers with some mechanism for buffering the system from the most severe demand fluctuations, often through the management of the arrival process.
3. Developing aggregate plans for the use of resources in the near term.
4. Designing work shift scheduling systems that gear short-term capacity to the arrival process.
5. Designing detailed scheduling systems that use limited resources of equipment and manpower effectively.
6. Creating job, process, and facility designs for effective operations to provide for needed service levels.

These five major kinds of systems do not represent rigid classifications, since, in practice, systems tend to be combinations of job shop and line or job shop and project, manufacturing as well as service centers, nonmanufacturing as well as manufacturing, etc. However, the classifications do provide a basis for developing concepts and techniques for the typical production line, distribution system, job shop, project, or service system. Analytical methods have evolved around the concept of the pure system and have facilitated the analysis of combination systems (*i.e.*, examining the components and merging the results).

A SUMMARY VIEWPOINT

Throughout this book, we have noted that most problems involve a balance of costs as well as other criteria, resulting in management's need to make trade-offs. Invariably, cost factors affected by a given decision follow different patterns, so that the best solution never involves minimizing one cost factor at the expense of others. This means that we *do* produce some scrap, we *do* have some late orders, we do sometimes run out of stock, and, in service systems, we do have some idle capacity. If we do not, we are controlling some factors too tightly.

We must remember that many of the quantitative techniques are more than simply problem solving techniques. Perhaps their most important function is to help us understand the nature of the system with which we are dealing. For example, waiting line theory clarifies the probabilistic nature of the flow problem. If we understand something of waiting line theory, we may be able to make excellent judgments about how to staff activities in which arrivals of the thing being processed are controlled by random processes. When we understand that many of the more complex processing systems can be described as a network of waiting lines, we may understand why an order seems to take so long to get through such a system.

Mathematical programming provides formal means for allocating scarce resources to competing demands. But there are many kinds of allocation problems in an enterprise, most of which we may not be able to solve formally. However, the structure of the problems, and the answers to the formal problems, can be transferred to all kinds of allocation problems. For one thing, we know that alternative optimum solutions probably exist, as well as many more solutions that are almost as good as the optimal ones. As a result, we know that we have considerable flexibility, and this makes it easier to accommodate a variety of conditions.

We have not attempted to deal with many of the human behavior problems that obviously occur in productive systems. Perhaps it is important to point out, however, that formal models and analysis often take account of human problems even though the human "variables" may not enter directly into the models that are constructed. A good example is the computer model for layout—CRAFT—discussed in Chapter 8. A manager may feel that it is important that two work groups not be separated in a new layout. Before we agree, however, let us appraise this requirement. An excellent solution may exist that accomodates this constraint. At the same time, this constraint may be very costly. The point is that we should make the analysis, rather than specifying conditions without analyzing or appraising them.

Finally, the broader the viewpoint we can take, the more likely we are to avoid suboptimization. Thus, it is not valid for us to consider inventory problems in isolation because inventory levels depend in part on output fluctuations. As our knowledge of production/operations management grows, we shall be able to determine the effects of the interactions of many other variables in the system. We shall be able to take a true "systems" viewpoint.

Part VI

Appendixes

Appendix A

Cost Models

Many of the predictive models that are used ~~by managers are~~ constructed as cost behavior models. Since most of these models assume certainty, it is easy to evaluate them; presumably, we would select the alternative for which the cost model predicts minimum costs. Practical managers may not view these kinds of analyses as models but may refer to cost models as "cost analyses."

In constructing cost data for decision purposes, we should recognize that the objective of cost data that result from accounting differs from the objective of cost data that are constructed for decision making. In general, accounting systems meet the needs of federal and state tax laws, as well as of certain phases of an organization's financial management. When the end-use of cost data is applied to decision making, we must be careful that average costs and allocations of fixed overhead cost items do not mask the actual behavior of costs. We shall be interested in the *net effect* on costs of each alternative that is considered. As we shall see, the implications of these "net effects" criterion shift somewhat, according to the nature of the problem.

FIXED AND VARIABLE COSTS

It is often worthwhile to consider how costs vary with the volume that is produced. We think of such cost items as property taxes, indirect labor, and building depreciation as being nonvariable, or fixed, and we view such items as direct labor, materials, and certain supplies used as being variable with volume. However, this dichotomous way of thinking often is dangerous, because "fixed" costs may sometimes be thought of as varying with volume, and costs commonly thought to be variable may behave like fixed costs. Many cost items may be fixed only over short ranges of volume. Indirect labor is commonly in this class. One material handler may be adequate to a certain volume level above which we must have two handlers. To be sure, our handler is not fully utilized until we approach the critical volume, and when we add the second, both handlers probably will be underutilized. The semivariable nature of such cost elements can affect certain decision problems.

BREAKEVEN ANALYSIS

Breakeven analysis uses the fixed and variable nature of costs to indicate the range of volume that is necessary for profitable operation. If we could divide all costs into those that vary with volume and those that do not, we could compute an average total cost per unit for a given volume. Semivariable costs can be reduced to a fixed component and a variable component. This concept of average unit cost is correct only at one volume of computation, however, since the fixed costs per unit would change as we averaged them over different volumes.

Conceptually, it would be helpful to consider the fixed costs as a total pool of costs that must be covered by net revenue, over and above variable costs, before any profit whatever is made. This point or volume of sales, where total net revenue after variable costs exactly equals the total pool of fixed costs, is called *the breakeven point*: it is the point in the progress of sales where the total revenue exactly covers the total costs, variable and fixed. Below this volume, a loss is recorded; above it, a profit is recorded.

Figure 1 is a diagram of the structure of a simple breakeven chart. Units of volume are plotted on the horizontal scale, and dollars of sales or cost are plotted on the vertical scale. The sales line, which begins at the origin, is a straight line, since dollar sales are assumed to be proportional to units sold. The total cost line intersects the vertical axis at a value equal to the fixed costs and

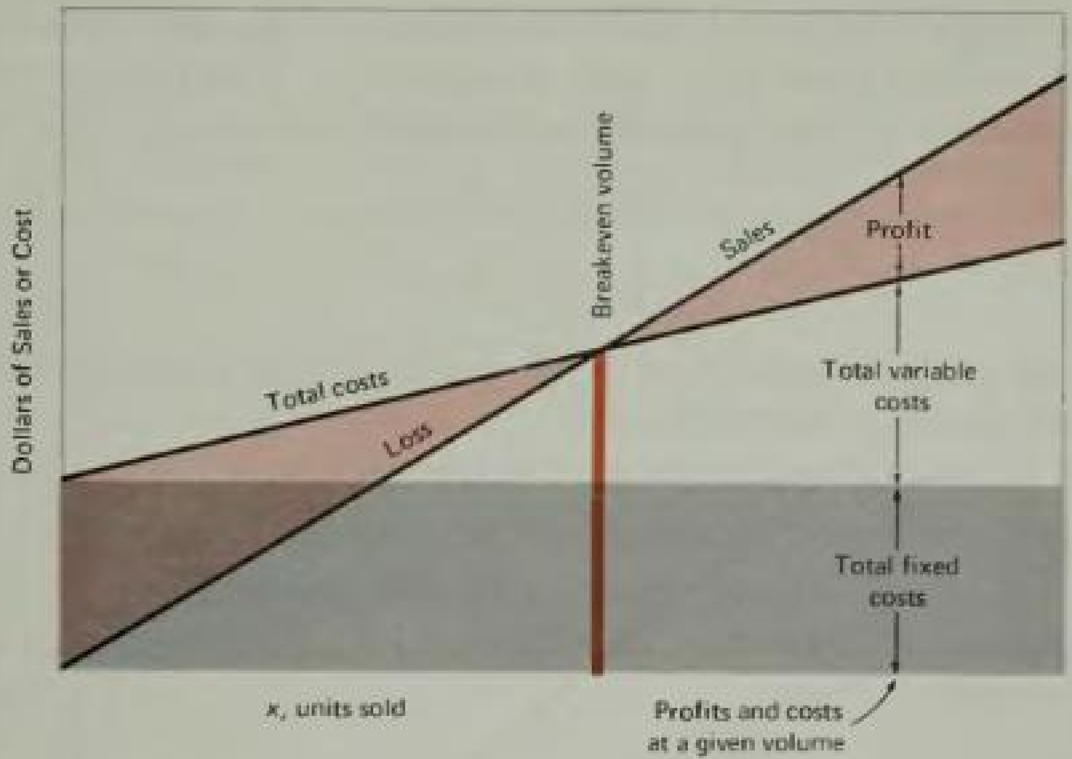


Figure 1. A simple breakeven chart

increases in proportion to the number of units sold. Above the breakeven point, the ratio of profits to sales increases with each unit sold. This is because of the broadened base for the absorption of fixed costs. Contribution is at a fixed ratio, however.

Contribution

Contribution is the difference between sales and variable costs, or the contribution to fixed costs and profit; that is,

$$C = S - V \tag{1}$$

and

$$S = F + V + P \tag{2}$$

where

C = contribution
S = sales
V = variable costs
F = fixed costs
P = profit

Since both S and V vary with volume, C also varies with volume. C can be calculated easily if we know the percentage of the sales dollar that is V. Suppose, for example, that variable costs are 60 percent of the sales dollar and that fixed costs are \$3,000,000. Then, from Equation 1, C is 40 percent. The only costs that have not yet been deducted are the fixed costs, so:

$$C = F + P \tag{3}$$

or

$$P = C - F$$

Now we can compute profit at any level of sales. If total sales are \$10,000,000, then C is 40 percent, or \$4,000,000, and:

$$P = C - F = \$4{,}000{,}000 - \$3{,}000{,}000 = \$1{,}000{,}000$$

If total sales are \$8,000,000, C is 40 percent, or \$3,200,000, and:

$$P = C - F = \$3{,}200{,}000 - \$3{,}000{,}000 = \$200{,}000$$

The contribution concept allows us to compute total profit at various levels of sales rather easily.

Construction of Breakeven Charts

Although the theory behind breakeven charts is simple, it is not so simple to obtain good data from which to develop the chart, because the line between fixed and variable costs is not definite. We may not be safe in taking armchair classifications as valid. We may deduce that certain direct labor costs ought to be variable, but are they really? They may contain a fixed element. The point is that, to construct an accurate breakeven chart from cost elements, we may have to do a great deal of prior work to establish the actual behavior of cost elements in relation to volume. Good breakeven charts require excellent cost accounting systems.

Another way to approach the problem is by means of the scatter diagram of costs. Data on total costs are plotted for several years, and an average line is drawn in. The presumption is that the different years will represent different volumes, thus enabling us to infer a total cost line in relation to volume. The point at which the line crosses the vertical axis is an estimate of the fixed costs. Let us take the following data shown in Figure 2 and plot them.

weakness of this

The result looks good and is easy to obtain. We must question, however, whether the relation of costs to volume might represent inflationary tendencies during the period more than it represents how costs vary with volume. If this is a possibility, the data may be deflated to remove the effect of rising costs and prices.

Year	Total Costs (dollars)	Sales (dollars)	Year	Total Costs (dollars)	Sales (dollars)
1973	1.45 million	1.75 million	1975	2.20 million	3.10 million
1974	1.70 million	2.20 million	1976	2.30 million	3.50 million

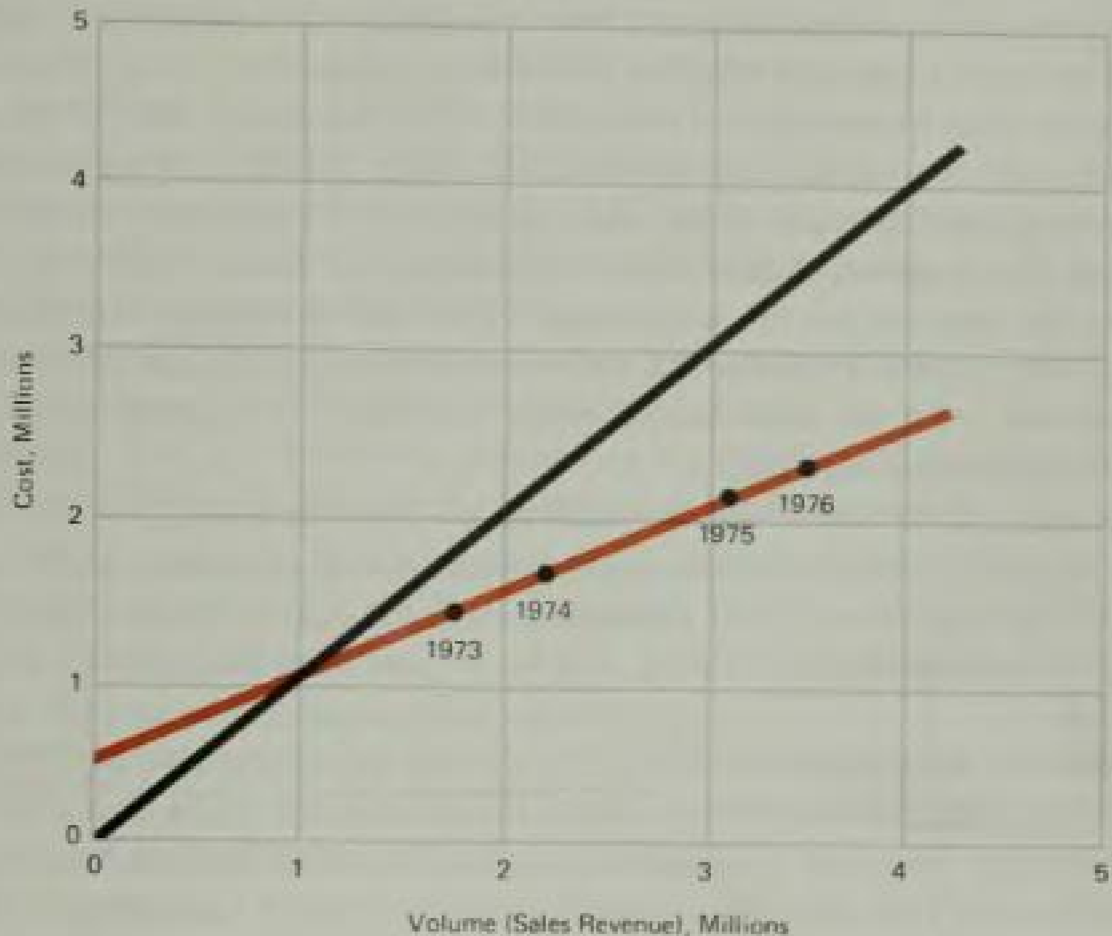


Figure 2. Scatter plot of total costs versus sales volume

3 It also is important that the data for the several years be fairly representative of a single set of conditions of technology, product mix, and cost performance. If major technological changes occurred during the period, we would not expect the several years to yield consistent data on which to base the construction of a breakeven chart.

INCREMENTAL COSTS

Incremental costs are costs that vary with the alternative courses of action that are considered. They are the "net effects" that result. (The terms *incremental* *costs*, *marginal costs*, *out-of-pocket costs*, and *differential costs* are used somewhat interchangeably.)

We should recognize that decision making in an on-going situation always must take account of the status quo. Existing conditions may favor one alternative. An example of this is the common "make-versus-buy" decision. If we currently are purchasing an item and are considering the possibility of making it instead, we should question whether we have the capacity to do so. If we do, the incremental costs of making the item will be only the direct costs of labor and materials, plus any actual net additions to other costs, such as power and supplies. The machinery, building, and supervisory and executive staffs already exist; their cost does not change if we manufacture the item. Therefore, we dare not base our decision on the accountant's concept of average manufacturing cost. We only need to consider the net incremental costs. If available capacity does not exist, the net incremental costs will have to include the costs of providing the needed capacity.

Conversely, if we currently are making an item that we propose to buy instead, we shall not rid ourselves of the average manufacturing cost of the item. The building, supervisory and executive staff, etc., still remain. Incremental costs that are eliminated ordinarily will be far less than the average manufacturing cost.

The concept of incremental cost often reveals a shifting line between fixed and variable costs. For example, consider the direct labor costs on a continuous production line. The sum of all labor costs for the entire line can be visualized fairly well as varying with the number of units produced (assuming that if the line is stopped, workers either are sent home or assigned other jobs). On the other hand, if we view individual operations on the line, we find that their labor costs generally are not subject to managerial control. It may appear that one of these operations could be improved by the use of a better tool, better motion patterns, etc. If improvements are installed, we probably shall find that labor costs per unit on that operation do not go down (are fixed) because that operation is paced by the line. The only individual operation on the line that can have variable direct labor costs is the "bottleneck" or limiting operation. If that operation can be improved, then direct labor costs for the entire line will decrease in proportion to the improvement, but not beyond the limits imposed by the next most limiting operation.

There may be as many cost patterns within an organization as there are alternative plans. The important point to keep in mind is that incremental costs are associated with whole plans or alternatives, so average costs almost never are good estimates of a decision's net-cost effects. Following are two examples that illustrate incremental cost behavior in some types of problems.

Example 1

A company selling about $100,000 annually has three factory employees. The company specializes in plating operations on a jobbing basis. Current conditions in the labor market make employment of part-time help difficult, and work load variations would make this undesirable anyway. One man, who is the foreman, has the essential knowledge and skill required for the various plating operations. He is paid $300 per week. The other two men are general laborers, each paid $240 per week. The three men work forty hours per week. Production could rise considerably before a fourth worker would be necessary and also would have to fall considerably before the company could get along with only the foreman and one laborer.

For a certain special plating process, it has been the practice to route the work to an outside plater for a particular operation that demands special equipment. The company now is considering purchasing the special equipment so that the operation could be performed in the home plant. Cost figures are:

Cost of outside processing per piece	$0.50
Expected volume of work on special equipment next year	10,000 pieces
Estimated hours per piece spent by:	
Foreman	0.04 hour
Laborers	0.075 hour
Value of power used by new equipment per piece	$0.02
Average value of materials used	$0.02
Burden rate per hour of labor	100 percent

There is extra room in which to locate the new equipment, which would cost $2,000, installed. What should the company do?

In this situation, labor is a fixed cost and can be ignored. The conventional burden rate is irrelevant. The only new overhead costs are the power and the cost of the equipment itself. The only pertinent direct cost is the material. The total incremental cost of the two alternative plans for next year is: .02+.02
.04 × 10,000 =
400

Outside processing $5,000
In-plant processing $2,000 + 400 = $2,400

The economics of the situation obviously favor in-plant processing. However, this is not the proper way to handle the equipment cost item, for we have ignored opportunity cost of alternative investments and the fact that the equipment has a probable life greater than a year. We shall discuss these capital costs later in this appendix.

Example 2

A job order tool and die shop is operating at essentially full capacity. The usual estimating procedure for new work involves an analysis of the customer's drawings and specifications to determine the materials required and the sequence and type of machining operations needed. Standard time data for the machine operations then are referred to in order to obtain estimates of the number of labor hours required. The resultant estimate of labor hours is multiplied by $6, the total material costs are added, and, finally, 15 percent is added for profit. The $6 labor rate is composed of an average hourly pay rate to workers of $4 per hour, plus burden at 50 percent of direct labor hours. The burden rate commonly is adjusted every six months to reflect current conditions. It includes the usual items of overhead: depreciation, supplies, *overtime premium*, administrative and selling expense, etc.

A new order currently is being considered. The estimating procedure yields $1,000 for material and 1,750 labor hours. The total bid including profit is $13,225. Should the company take the order if the customer agrees to the price and delivery?

The usual accounting practice of putting overtime premium in as overhead works very well for computing profit and loss as well as average costs. For evaluating this contract, however, we need incremental costs. Since the plant is operating at full capacity, we probably can assume operation beyond the breakeven point; therefore, all fixed costs already have been covered. On the other hand, overtime labor would be required to do the job. Total out-of-pocket costs would be material plus labor plus variable overhead. Therefore, the marginal income would be $13,225 − 1,000 − (1,750 × 4.00 × 1.5) − variable overhead, or $1,725 − variable overhead.

13225 - 1000 - 10500
13225 - 11500 = 1725

CAPITAL COSTS

Capital costs affect decision problems in production/operations management whenever a physical asset or expenditure is involved that provides a continuing benefit or return. From an accounting point of view, the original capital expenditure must be recovered through the mechanism of depreciation and must be deducted from income as an expense of doing business. The number of years over which the asset is depreciated, and the allocation of the total amount to each of these years (i.e., whether depreciation is straight-line or some accelerated rate) represent alternative strategies that are directed toward tax policy. We must remember that all these depreciation terms and allocations are arbitrary and have not been designed from the point of view of cost data for decision making, relative to operations. Our interest will be in capital cost data, which will help us decide among alternative courses of action involving different systems of production, different degrees of mechanization, the replacement of existing assets, etc.

Let us begin our analysis with a very simple example, which we shall treat simply. We have just installed a piece of equipment that performs a highly specialized operation for us. The installation was a custom job that fit our particular situation. Because of the specialized nature of the equipment and the custom installation, the equipment is of no value to anyone else; its salvage value is zero.

As with many kinds of productive equipment, the useful physical life of our equipment can be prolonged almost indefinitely by maintenance and repair, so it is difficult to say yet what the life of the equipment will be. Since the equipment has no salvage value, we must face the fact that the entire $10,000 seems to go down some sort of economic sink the minute the equipment becomes ours. We say that the $10,000 is a sunk cost, meaning simply that it is gone forever, regardless of what we may list as the "book value" of the equipment. Since the $10,000 is "sunk," it is completely irrelevant to any future decisions because no future decision can affect it.

The cost of owning the equipment (in contrast with the costs of operating and maintaining it) is simply $10,000. We hope to spread this total over a period of time so that the *average* annual cost of ownership will not be too great. In five years, the average cost of owning the equipment is only $10,000/5 = $2,000 per year. In ten years, the average annual cost is down to $1,000 per year. Regardless of how long we keep the equipment, this cost of owning is irrelevant because it is a past cost. The only future costs that we shall incur for this equipment are the costs of operating and maintaining it. Once the machine is installed, these are the only costs that are subject to managerial control by future decisions.

Let us assume that two men are required to operate the equipment at $4,000 per year per man. Maintenance costs are expected to be $2,000 the first year, thereafter increasing by $200 per year. Figure 3 shows the total of these costs, diagrammed in relation to time. We assume that there are no other pertinent costs. Since the original cost of the equipment is sunk and the only future costs are the operating and maintenance costs, we are in a position to see under what conditions we would consider this setup to be obsolete and would replace it. We would want to replace it any time that we could find a functionally equivalent setup that could offer a total average annual cost of owning plus operation and maintenance that fell below the cost curve of Figure 3. Let us assume that during the *fourth* year of operation of this setup (which we shall call the present setup), a new equipment design is developed.

Figure 3. Operating and maintenance costs, showing rising costs of maintenance

This *new equipment design* has an important advantage over the old one. Because it is more automatic, it can be operated by one person. The more automatic features require an additional $1,000 maintenance effort, however; therefore the first-year operating and maintenance cost is expected to be $4,000 operating labor plus $3,000 maintenance. This total of $7,000 is expected to increase by $200 per year as before, because of mounting maintenance charges. The improved design costs $12,000, installed. Now we want to see whether the average annual total of the costs of owning plus operating and maintaining the improved equipment design are less than the $10,600 current annual expense (fourth year) for the present setup. First, let us note that while we ignore the original sunk cost of the present setup, we cannot ignore the installed cost of the improved design. It is still a future cost. We have not bought it yet, so the installed cost must be considered.

To see the picture clearly, let us plot, on the same chart, the average annual cost of owning (capital cost) and the annual operating and maintenance costs for the improved equipment design. These costs are shown in Figure 4, where we also show the total of the two costs year by year. The total cost curve is developed simply by adding the cost of owning to the operating and maintenance costs for each year. The total cost is very high during the first two years, it reaches a minimum during the eighth year, and then it begins to rise again. It is high in the early years because the annual average costs of owning the equipment are very high during those years. It begins to rise again (but slowly) after

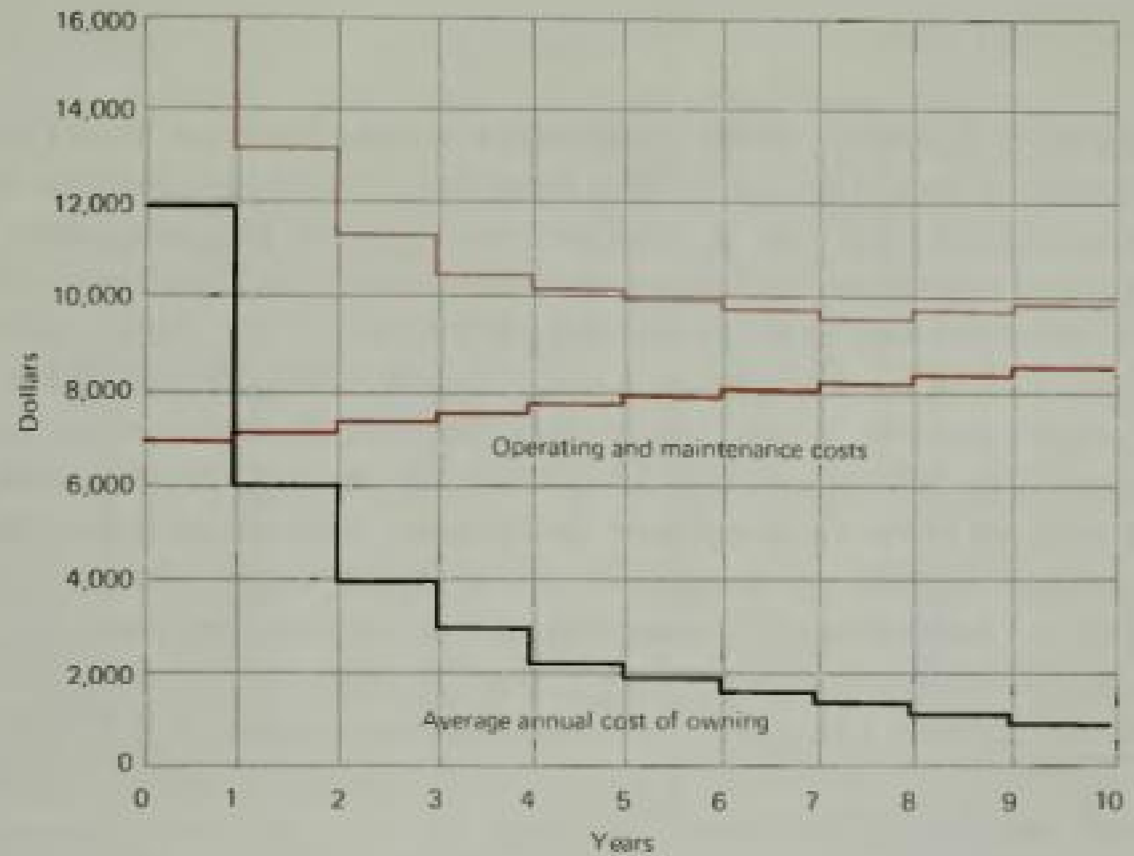


Figure 4. Year-by-year average costs for a proposed replacement machine costing $12,000 initially and having no salvage value at any time

the eighth year because of the influence of rising maintenance costs. Now let us examine the significance of the total cost curve for this simplified example. For the present setup, the total incremental cost for the fourth year is $10,600, as we already have noted—a sum that we expect will become larger in future years. However, the total average annual cost for the proposed setup will be less than $10,600 after its fourth year. It seems clear that a decision to replace the present setup is needed.

The situation that we have assumed is quite clear-cut. However, we may ask: "What criteria for comparison are we using here?" The answer is that we are looking at the best possible future cost performance for both the present and proposed setups. The best cost performance possible for the present design is $10,600, which is this year's operating and maintenance cost. The best cost performance possible for the proposed design is achieved if it is held in service for eight years; its total cost would average only $9,900 per year for the entire eight-year span. As long as the best performance of the proposed design (represented by the minimum of the total cost curve) is less than the best performance of the present design, it would be economical to make the switch. Of course, we must temper this statement by recognizing important intangible values and the accuracy of the cost estimates.

Opportunity Costs

We have been assuming a fairly simple situation, in which we have chosen to ignore certain factors. Let us introduce some additional ideas. Suppose that we are discussing an asset that is used for more general purposes, such as an over-the-road semitrailer truck. Let us assume that we own such a truck and that the first question before us is: "How much will it cost us to *own* this truck for one more year?" These costs of owning, or capital costs, cannot be derived from the organization's ordinary accounting records. The cost of owning it for one more year depends on its current value. If the truck can be sold on the second hand market for \$5,000, this is a measure of its economic value. Since it has value, we have two basic alternatives: we can sell it for \$5,000 or we can retain it. If we sell, the \$5,000 can earn interest or a return in an alternative investment. If we keep the truck, we forego the return, which then becomes an *opportunity cost* of holding the truck one more year. Similarly, if we keep the truck, it will be worth less a year from now, so there is a second opportunity cost, which is measured by the fall in salvage value during the year.

The loss of opportunity to earn a return and the loss of salvage value during the year are the costs of continued ownership. They are opportunity costs rather than costs paid out; nevertheless, they can be quite significant in comparing alternatives that require different amounts of investment. There is one more possible component of capital cost for the next year if the truck is retained—the cost of possible renewals or "capital additions" that are necessary to keep the truck operating. We are not thinking of ordinary maintenance here but of major overhauls, such as a new engine or an engine overhaul that extends the physical life for some time. In summary, the capital costs, or costs of owning the truck for one more year, are:

1. Opportunity costs:
   a. Interest on opening salvage value.
   b. Loss in salvage value during the year.
2. Capital additions or renewals required to keep the truck running for at least an additional year.

By assuming a schedule of salvage values, we can compute the year-by-year capital costs for an asset. This is done in Table 1 for a truck that cost \$10,000 initially, and for which the salvage schedule is indicated. The final result is the projected capital cost that is incurred for each year. If we determine the way in which operating and maintenance costs increase as the truck ages, we can plot a set of curves of yearly costs. The combined capital plus operating and maintenance cost curve will have a minimum point, as before. This minimum of the combined cost curve defines the best cost performance year in the life of the

TABLE 1
YEAR-BY-YEAR CAPITAL COSTS FOR A SEMITRAILER TRUCK, GIVEN A SALVAGE SCHEDULE (INTEREST AT 10%)

Year	*Year-End Salvage Value*	*Fall in Salvage Value during Year*	*Interest on Opening Salvage Value*	*Capital Cost, Sum of Fall in Value, and Interest*
New	$10,000	—	—	—
1	8,300	$1,700	$1,000	$2,700
2	6,900	1,400	830	2,230
3	5,700	1,200	690	1,890
4	4,700	1,000	570	1,570
5	3,900	800	470	1,270
6	3,200	700	390	1,090
7	2,700	500	320	820
8	2,300	400	270	670
9	1,950	350	230	580
10	1,650	300	195	495

equipment. Beyond that year, the effect of rising maintenance costs more than counterbalances the declining capital costs. Note that such a plot is different from Figure 4, in which we plotted annual average costs instead of yearly costs.

Obsolescence and Economic Life

What is the effect of obsolescence on the cost of owning and operating a machine? By definition, when a machine is obsolete, an alternative machine or system that is more economical to own and operate exists. Clearly, the existence of the new machine does not cause any increase in the cost of operating and maintaining the present machine. Those costs already are determined by the design, installation, and condition of the present machine. However, the existence of the new machine causes the salvage value of the present setup to fall and, therefore, induces an increased capital cost. Thus, for assets in technologically dynamic classifications, the salvage value schedule falls rapidly in anticipation of typical obsolescence rates. Economic lives are very short. On the other hand, when the rate of innovation is relatively slow, salvage values hold up fairly well.

Table 2 compares year-by-year capital costs for two machines that initially cost $10,000 but have different salvage schedules. The value of Machine 1 holds the best; Machine 2 has more severe obsolescence reflected in its salvage schedule. The result is that capital costs in the initial years are greater for Machine 2 than for Machine 1. The average capital costs for the first five years are:

Machine 1	$1,913
Machine 2	$2,198

Therefore, if the schedules of operating expenses for the two machines were identical, Machine 1 would seem more desirable. However, since the timing of the capital costs is different for the two machines, it would be better, for comparative purposes, if we could adjust all figures to their equivalent present values.

Present Values

Since money has a time value, future expenditures and opportunity costs will have different present values to us. What do we mean by the time value of money? Since money can earn interest, $1,000 in hand now is equivalent to $1,100 a year from now if the present sum can earn interest at 10 percent. Similarly, if we must wait a year to receive $1,000 that is due now, we should expect not $1,000 a year from now but $1,100. When the time spans involved are extended, the appropriate interest is compounded and its effect becomes much larger. The timing of payments and receipts can make an important difference in the value of various alternatives.

Let us illustrate this point briefly and more precisely before returning to the

TABLE 2

COMPARISON OF CAPITAL COSTS FOR TWO MACHINES COSTING $10,000 INITIALLY BUT WITH DIFFERENT SALVAGE SCHEDULES (INTEREST AT 10%)

Machine 1				*Machine 2*			
Year-End Salvage Value	*Fall in Value During Year*	*Interest at 10% on Opening Value*	*Capital Cost*	*Year-End Salvage Value*	*Fall in Value During Year*	*Interest at 10% on Opening Value*	*Capital Cost*
$10,000	—	—	—	$10,000	—	—	—
8,330	$1,670	$1,000	$2,670	7,150	$2,850	$1,000	$3,850
6,940	1,390	833	2,223	5,100	2,050	715	2,765
5,780	1,160	694	1,854	3,640	1,460	510	1,970
4,820	960	578	1,538	2,600	1,040	364	1,404
4,020	800	482	1,282	1,860	740	260	1,000
3,350	670	402	1,072	1,330	530	186	716
2,790	560	335	895	950	380	133	513
2,320	470	279	749	680	270	95	365
1,930	390	232	622	485	195	68	263
1,610	320	193	513	345	140	49	189

example of the two machines. We know that if a principal sum P is invested at interest rate i, it will yield a future total sum S in n years hence if all the earnings are retained and compounded. Therefore, P in the present is entirely equivalent to S in the future by virtue of the compound amount factor. That is:

$$S = P(1 + i)^n \tag{4}$$

where $(1 + i)^n$ = the compound amount factor for interest rate i and n years.

Similarly, we can solve for P to determine the present worth of a sum to be paid n years hence. That is:

$$P = \frac{S}{(1 + i)^n} = S \times PV_{sp} \tag{5}$$

where PV_{sp} = the present value of a single payment S to be made n years hence with interest rate i. Therefore, if we were to receive a payment of \$10,000 in ten years, we should be willing to accept a smaller but equivalent sum now. If interest at 10 percent were considered fair and adequate, that smaller but equivalent sum would be:

$$P = 10{,}000 \times 0.3855 = \$3{,}855$$

since

$$\frac{1}{(1+0.10)^{10}} = PV_{sp} = 0.3855$$

Now let us return to the example of the two machines. The capital costs for each machine occur by different schedules because of different salvage values. If all future values were adjusted to the present as a common base time, we could compare the totals to see which investment alternative was advantageous. We have done this in Table 3, where we have assumed an operating cost schedule in Column 2, determined combined operating and capital costs in Columns 5 and 6, and listed present values in Columns 8 and 9. The present value of the entire stream of expenditures and opportunity costs is \$32,405 for Machine 1. The net difference in present values for the two machines is shown at the bottom of Table 3. Since the operating cost schedule was identical for both machines, the difference reflects differences in the present worth of capital costs. Obviously, the method allows for different operating cost schedules as well.

There are some difficulties with the methods just described. First, we have assumed that the schedule of salvage values is known, which, more often than not, is not true. Second, at some point in the life of the machines it becomes economical to replace them with identical models. Therefore, a chain of identical machines should be considered for comparative purposes; the machine is replaced in the year in which operating and capital costs are exactly equal to the interest on the present worth of all future costs. The essence of this latter

TABLE 3
PRESENT VALUE OF CAPITAL AND OPERATING COSTS FOR THE TWO MACHINES FROM TABLE 2
(SCHEDULE OF OPERATING COSTS IS THE SAME FOR BOTH MACHINES (INTEREST AT 10%))

Year	Operating Cost	Capital Costs (from Table 2)		Combined Operating and Capital Costs		Present Worth Factor for Year Indicated**	Present Worth of Combined Costs for Year Indicated	
		Machine 1	Machine 2	Machine 1	Machine 2		Machine 1	Machine 2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	$3.000	$2.670	$3.850	$5.670	$6.850	0.909	$5.154	$6.227
2	3.200	2.223	2.765	5.423	5.965	0.826	4.490	4.939
3	3.400	1.854	1.970	5.254	5.370	0.751	3.946	4.033
4	3.600	1.538	1.404	5.138	5.004	0.683	3.509	3.418
5	3.800	1.282	1.000	5.082	4.800	0.621	3.156	2.981
6	4.000	1.072	716	5.072	4.716	0.565	2.866	2.665
7	4.200	895	513	5.095	4.713	0.513	2.614	2.418
8	4.400	749	365	5.149	4.765	0.467	2.405	2.225
9	4.600	622	273	5.222	4.873	0.424	2.214	2.066
10	4.800	513	189	5.313	4.989	0.386	2.051	1.926
Totals							$32.405	$32.898

Machine 1. present worth of all future values is total of column (8) less present worth of tenth-year salvage value, i.d. $32.405 − 1610* × 0.386 = 32.405 − 621 = $31.784.
Machine 2. $32.898 − 345* × 0.386 = 32.898 − 133 = 32.765.
*Tenth-year salvage values from Table 2.
**From Table 1 (Appendix E)

statement is that we are seeking a balance between this year's costs (operating and capital costs) and opportunity income from disposal (interest on the present worth of all future costs). When the opportunity income from disposal is the greater of the two, replacement with the identical machine is called for. Most common criteria for comparing alternative capital investments attempt to circumvent these problems by: (1) assuming an economic life and (2) assuming some standard schedule for the decline in value of the asset. We shall now consider some of these criteria.

COMMON CRITERIA FOR COMPARING ECONOMIC ALTERNATIVES

Some of the common criteria used to evaluate proposals for capital expenditures and compare alternatives involving capital assets are: (1) present values; (2) average investment; (3) rate of return; and (4) payoff period.

Present Value Criterion

Present value methods for comparing alternatives take the sum of present values of all future out-of-pocket expenditures and credits over the economic life of the asset. This figure is compared for each alternative. If differences in revenue also are involved, their present values also must be accounted for. Table 1 in Appendix E gives the present values for single future payments or credits; Table 2 in Appendix E gives present values for annuities for various years and interest rates. An annuity is a sum that is received or paid annually. The factors in Table 2 of Appendix E convert the entire series of annual sums to a single sum in the present for various interest rates and years. We shall use the notation PV_a for the present value factor of an annuity.

An Example

Suppose we are considering a machine that costs \$15,000, installed. We estimate that the economic life of the machine is eight years, at which time its salvage value is expected to be about \$3,000. For simplicity's sake, we take the average operating and maintenance costs to be \$5,000 per year. At 10 percent interest, the present value of the expenditures and credits is:

15
15—
5×8= 40—
55—
-3
52—

Initial investment	$\$15{,}000 \times PV_{sp} = 15{,}000 \times 1.000 =$	15,000
Annual operating and maintenance costs	$5{,}000 \times PV_a = 5{,}000 \times 5.335 =$	26,675
	10% 8yrs	41,675
Less credit of present value of salvage to be received in eight years	$3{,}000 \times PV_{sp} = 3{,}000 \times 0.467 =$	1,401
	this is today's equivalency 10% 8yrs ←	\$40,274

today next 8yrs.

The net total of \$40,274 is the present value of the expenditures and credits over the eight-year expected life of the machine. The initial investment is already at present value; that is, the present value factor is 1.000. The annual costs of operation and maintenance are an eight-year annuity, so the entire stream of annual costs can be adjusted to present value by the multiplication of PV_a from Table 2 (Appendix E). Finally, the present value of the salvage is deducted. This total could be compared with comparable figures for other alternatives over the same eight-year period. If another alternative machine is estimated to have a different economic life (perhaps four years), then, if we wanted to make the present value totals comparable, we would compare two cycles of the four-year machine with one cycle of the eight-year machine. If the operating and maintenance costs increased as the machine aged, the present value of the expenditure in each year would be determined separately by PV_{sp}.

Average Investment Criterion

Average investment methods estimate an average annual cost of owning plus operating and maintaining an asset. The average annual capital costs are approximated by average salvage loss plus interest on the average investment, assuming that the decline in value of the asset is on a uniform or straight-line basis. Figure 5 shows the assumed structure for the decline in value of an asset and the calculation of average investment. Thus, if a machine cost \$10,000 and was estimated to have a ten-year life with a salvage value at the end of that time of \$1,000, the capital costs are approximated by:

Depreciation

$$\text{Average annual salvage loss, } \frac{10{,}000 - 1{,}000}{10} = \$900$$

offset for all the years

Annual interest on average investment, at 10 percent

opportunity cost

$$\frac{(10{,}000 + 1{,}000)0.10}{2} = 5{,}500 \times 0.10 = \$550$$

$$\text{Average annual capital cost} = \$900 + \$550 = \$1{,}450$$

only for straight line depreciation

If operating and maintenance costs were estimated to average \$12,500 per year over the ten-year machine life, the total average annual cost for comparison would be 1,450 + 12,500 = \$13,950. Differences in annual revenue among alternatives can be accounted for in the operating costs. Comparable calculations for alternative machines would form the basis for comparison. For a strictly economic comparison, the alternative presenting the lowest average annual cost would be selected.

Rate of Return Criterion

One common method of evaluating new projects or comparing alternative courses of action is to calculate a rate of return, which then is judged for adequacy. Usually, no attempt is made to consider interest costs, so the resulting figure is referred to as the "unadjusted" rate of return (i.e., unadjusted for interest values). It is computed as follows:

$$\text{Unadjusted rate of return} = \frac{100\,(\text{net monetary operating advantage}-\text{amortization})}{\text{average investment}}$$

The net monetary advantage reflects the algebraic sum of incremental costs of operation and maintenance and possible differences in revenue. If the rate computed is a "before-tax" rate, then the amortization

$$\frac{55000 + 10000}{2}$$

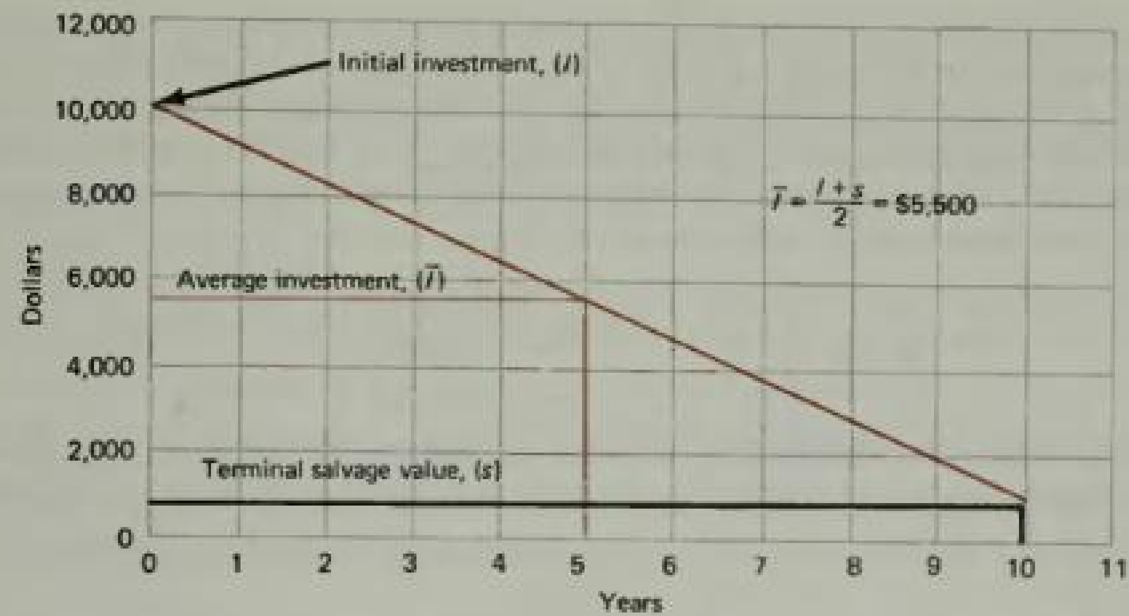


Figure 5. Relationship of initial investment, terminal salvage value and average investment for "average investment methods"

$$\left(\frac{\text{Incremental investment}}{\text{Economic life}}\right)$$

is subtracted and the result is divided by average investment and multiplied by 100 to obtain a percentage return. If an "after-tax" rate is sought, the net increase in income taxes due to the project is subtracted from the net monetary advantage, and the balance of the calculation is as it was before. Obviously, the adequacy of a given rate of return changes drastically if it is being judged as an after-tax return.

An Example

We assume that new methods have been proposed for the line assembly of a product that had been assembled in a central area, each assembly being completed by one individual. The new methods require the purchase and installation of conveyors and fixtures that cost $50,000, installed, including the costs of relayout. The new line assembly methods require five fewer assemblers. After the increased maintenance and power costs are added in, the net monetary operating advantage is estimated as $20,000 per year. Economic life is estimated at five years. The unadjusted before-tax return is:

$$\frac{20{,}000 - \dfrac{50{,}000}{5}}{\dfrac{50{,}000}{2}} \times 100 = 40 \text{ percent}$$

if depreciation added in addition

$58— +$5T00 ?

depreciation

The after-tax return requires that incremental taxes be deducted. Incremental taxable income will be the operating advantage less increased allowable tax depreciation. Assuming straight-line depreciation and an allowed depreciation term of eight years, incremental taxable income is \$20,000 less \$50,000/8, or \$20,000 − \$6,250 = \$13,750. Assuming an income tax rate of 50 percent, the incremental tax due to the project is \$6,875. The after-tax return is therefore:

$$\frac{20{,}000 - 6{,}875 - 10{,}000}{25{,}000} \times 100 = \frac{3{,}125 \times 100}{25{,}000} = 12.5 \text{ percent}$$

Whether or not either the before- or after-tax rates calculated in this example are adequate must be judged in relation to the risk involved in the particular venture and the returns possible through alternative uses of the capital.

Payback

Payoff Periods

The payoff period is the time required for an investment to "pay for itself" through the net operating advantage that would result from its installation. It is calculated as follows:

$$\text{Payoff period in years} = \frac{\text{Net investment}}{\text{Net annual operating advantage after taxes}}$$

The payoff period for the conveyor installation that we discussed previously is

$$\frac{\$50{,}000}{\$13{,}125} = 3.8 \text{ years}$$

It is the period of time for the net after-tax advantage to equal exactly the net total amount invested. Presumably, after that period, "it is all gravy;" the \$13,125 per year is profit since the invested amount has been recovered. We may ask: "If the economic life of the equipment is five years and 10 percent is regarded as an appropriate rate of after-tax return for the project, what *should* the payoff period be?" Obviously, the period for both capital recovery and return is the five-year economic life. The period that recovers capital only, but that also allows enough time in the economic life to provide the return, will be somewhat shorter and will depend on the required rate of return. The payoff period is another interpretation that can be given to the present value factors for annuities, PV_a, given in Table 2 (Appendix E). As an example, for an economic life of five years and a return rate of 10 percent, $PV_a = 3.791$ from Table 2 Appendix E). This indicates that capital recovery takes place in 3.791 years. The equivalent of 10 percent compound interest takes place in 5.000 − 3.791 = 1.209 years.

Therefore, any of the PV_a values in Table 2 (Appendix E) for a given economic life in years and a given return rate indicate the shorter period in years required to return the investment. Or, more simply, they give the payoff period directly.

The proper procedure would be to estimate economic life and to determine the applicable return rate. We would determine from the present value tables the payoff period associated with these conditions. Then we would compute the actual payoff period of the project in question and compare it with the standard period from the tables. If the computed period is less than, or equal to, the standard period, the project meets the payoff and risk requirements that are imposed. If the computed value is greater than the table value, the project would earn less than the required rate.

Problems

1. A nursing home operator bases her income and costs on the number of rooms she has filled. Each room has an income earning potential of \$350 per month. Variable costs are food, materials, and supplies, plus nursing care. Food, materials, and supplies average \$150 per month per patient, and nursing care averages \$4 per day per patient. By using a corps of permanent nurses plus nurses on monthly call, the manager can keep nursing costs quite variable. The pool of fixed costs is \$5,000 per month. The home has 100 rooms, and currently 75 of them are occupied.

   What is the monthly contribution to profit and overhead? At what occupancy level does the manager break even? What would her profit be at 90 percent occupancy?

2. For the nursing home in Problem 1, what would be the price per room per month at which the operator could break even with a 75 percent occupancy?

3. A machine shop operator has three kinds of machines, all of which do "turning" operations. However, the machines have different degrees of mechanization and, with special tools, the productivity of the three machines is quite different for the same job. The out-of-pocket fixed costs to set up each machine plus the cost of special tools, as well as variable costs, are as follows:

Machine	Fixed Costs, Set Up Labor plus Special Tooling	Variable Costs, Labor, Supplies, Power, etc., per Unit
A	$ 5.00	$00.19
B	30.00	00.10
C	70.00	00.06

What are the breakeven volumes between Machines A and B, and C?

$TC_A = 5 + .19X$ $\quad TC_B = 30 + .1X$ $\quad TC_C = 70 + .06X$

4. A university operates its own printing plant in conjunction with a variety of printed materials, such as catalogs, announcements, special pamphlets, and academic journals. There is also a university press that publishes hardbound and softbound books, and these books are printed by the university printing department. The university administration is considering using outside printers to print department and school announcements and catalogs in order to reduce the long lead times associated with the internal printing department. As a typical example, the catalog of the graduate school of management is put out to bid and the bid is compared with the internal cost breakdown. The lowest outside bidder quotes $2,000 for 10,000 copies, with delivery guaranteed in three weeks. The cost analysis for internal printing with delivery in three months is as follows:

Paper	$ 150		
Composition labor	1,000		
Press labor	200		
Direct costs		$1,350	
Press room overhead, 200 percent of direct labor		2,400	
Total press room costs			$3,750
General and administrative overhead, 25 percent of total press room costs			937.50
Total costs			$4,687.50

The university president is astounded by the difference between the outside and inside bids. Should the outside bid be accepted? Why or why not?

5. The Stamped Metal Products Company (StaMCo) is considering the

manufacture of a special item for the air force. After considerable analysis, the company has identified the following costs:

Material	\$0.5 per unit
Tooling	20 hours
Unit direct labor	0.4 hour
Labor rate	\$2 per hour
Factory burden (materials + labor)	115 percent of labor
General and administrative burden	20 percent of factory cost
Target profit	10 percent of total cost

Construct a breakeven chart on a 10,000 unit basis with a unit sales price of \$3.

6. A foreman in an automatic screw machine department uses screw machines of essentially two sizes for all work done. One size is for small parts and the other is for larger parts. Dimensions of the part dictate when it must be run on the large machines. In general, the large machine costs more originally, uses more power, and operates more slowly than the small one. These differences are summarized below for a typical order.

	Small Screw Machine	Large Screw Machine
Original cost	\$18,000	\$23,000
Setup time	6 hours	6 hours
Total run-time (lot of 10,000 parts)	50 hours	65 hours
Value of energy	\$10	\$13

Labor is paid at \$3.25 per hour, and the departmental burden-rate is 100 percent of labor cost.

The foreman finds that he is operating to capacity with the small machines but has additional small parts to run. These parts also can be run on the large machines but, as we noted before, at higher cost. He also can schedule overtime on the small machines to accomplish the work. Assume that the job in question is the typical one for which data are given. Which machine should the foreman use? Why?

7. A trucking firm owns a five-year old truck that it is considering replacing. The truck can be sold for \$5,000, and blue book values indicate that this

salvage value would be $4,000 one year from now. It also appears that the trucker would need to spend $500 on a transmission overhaul if he were to keep the truck. What are the trucker's projected capital costs for next year? Interest is at 10 percent.

8. What is the present value of the salvage of a machine that can be sold ten years hence for $2,500? Interest is at 10 percent.

9. What is the future value in twenty-five years of a bond that earns interest at 10 percent and has a present value of $10,000?

10. What interest rate would a $10,000 bond have to earn to be worth $50,000 in ten years?

11. At 8 percent interest, how many years will it take money to double itself?

12. What is the present value of an income stream of $1,500 for fifteen years at 10 percent interest?

13. What is the value of an annuity of $2,000 per year for ten years at the end of its life? Interest is at 10 percent.

14. The proud owner of a new automobile states that she intends to hold her car only two years in order to minimize repair costs, which she feels, should be near zero during the initial period. She paid $4,000 for the car new, and blue book value schedules suggest that it will be worth only $2,000 two years hence. She normally drives 10,000 miles per year, and she estimates that her cost of operation is $0.10 per mile. What are her projected capital costs for the first two years, if interest of 6 percent represents a reasonable alternate investment for her?

15. Suppose that we are considering the installation of a small computer to accomplish internal tasks of payroll computation, invoicing, and other routine accounting. The purchase price is quoted as $300,000 and salvage value five years later is expected to be $100,000. The operating costs are expected to be $100,000 per year, mainly for personnel to program, operate, and maintain the computer. What is the present value of the costs to own and operate the computer over its five-year economic life? The value of money in the organization is 15 percent.

16. Using the average investment criterion as a basis for comparison, at what annual lease cost would we break even for the data on the computer in Problem 15? If the machine is leased, maintenance is furnished, reducing annual operating costs to $70,000.

17. An aggressive marketer of a new office copier has made its machine available for sale as well as lease. The idea of buying a copying machine seems revolutionary, but less so when we examine our present costs, which come to $6,500 per year for lease plus per copy charges of 2¢ per page. If we own a machine, the cost of paper and maintenance is projected to be $1,500 per year. The new copier costs $10,000, installed, and is assumed to have an economic life of five years and a salvage value of $2,000 (assume 50,000 pages per year).
    a. What is the projected unadjusted rate of return if we install the copier?
    b. If incremental taxes for the project are $1,000, what is the adjusted rate of return?

18. What is the actual payoff period for the office copier project discussed in Problem 17? If interest is 10 percent, what should the minimum payoff period be to make the investment economically sound? Does the office copier project meet the payoff standard?

TR = TC

References

Anthony, R. N. *Management Accounting*. Rev. ed. Homewood, Ill.: Richard D. Irwin, 1970.

Grant, E. L., and W. G. Ireson. *Principles of Engineering Economy*. 5th ed. New York: Ronald Press, 1970

Reisman, A. *Managerial and Engineering Economics*. Boston: Allyn and Bacon, 1971.

______, and E. S. Buffa. "A General Model for Investment Policy." *Management Science* 8(April 1962).

Terborgh, G. *Business Investment Policy*. Washington, D.C.: Machinery and Allied Products Institute, 1958.

Thuesen, H. G.; W. J. Fabrycky; and G. J. Thuesen. *Engineering Economy*. 4th ed. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Appendix B

Linear Programming

The normative or optimizing models that have been of the greatest significance to managers are linear optimization models and their associated powerful solution technique known as linear programming. Linear optimization models have been applied in a wide variety of industries, such as steel, oil refining, utilities and meat packing, as well as in a variety of nonprofit activities, such as education, refuse collection, and health care. In addition, the kinds of applications have included all sorts of resource allocation problems such as long-range financial planning, aggregate capacity planning, portfolio planning and selection, plant location, production planning and scheduling, political districting, corporate financial planning, warehousing and distribution, and air- and water-pollution control. Obviously, linear optimization models are not simply theoretical concepts that must be applied only to toy problems. They are powerful, useful managerial models.

NATURE OF LINEAR OPTIMIZATION MODELS

Linear optimization models are characterized by linear mathematical expressions. In addition, they usually are deterministic in nature, i.e., they do not take account of risk and uncertainty. The parameters of the model are assumed to be known with certainty. Finally, as we shall see, we use linear programming most often when we are attempting to allocate some limited or scarce resource so that we can make decisions that use the particular resource to optimize (either minimize or maximize) a stated criterion.

Elements of the Model Building Process

To develop a linear optimization model, we use the following process:

1. Define the decision variables.
2. Define the objective function, Z, a linear equation involving the decision variables, which identifies our objective in the problem solving effort. This equation predicts the effects on the objective of choosing different values of the decision variables.
3. Define the constraints as linear expressions involving the decision variables. The constraints specify the restrictions on the decisions that can be made. Alternatives can be generated by selecting values for the decision variables that satisfy these constraints.

FORMULATION OF A TWO-PRODUCT MODEL

Let us assume a simple situation for a plant that manufactures washing machines and dryers. The major manufacturing departments are the Stamping Department, the Motor and Transmission Department, and the final assembly lines for the washer and dryer. The Stamping Department fabricates a large number of metal parts for both products, and the Motor and Transmission Department produces the power units for both products. Monthly departmental capacities are:

Stamping Dept.	10,000 washers or	10,000 dryers
Motor and Transmission Dept.	16,000 washers or	7,000 dryers
Washer Assembly Dept.	9,000 washers	–
Dryer Assembly Dept.	–	5,000 dryers

In other words, the same facilities for the two products are used for stamping and for motor and transmission. The Stamping Department can produce parts for 10,000 washers per month or 10,000 dryers per month, as well as combination

-3 v

amounts of washers and dryers. We have a similar situation with the Motor and Transmission Department, but the final assembly lines are separate for the two products.

First, let us set up the restrictions of the problem. We shall denote the number of washers by x and the number of dryers by y. The simplest restrictions are those imposed by the two final assembly lines. They indicate that the number of washers per month must be less than or equal to 9,000 and that the number of dryers must be less than or equal to 5,000. These are plotted in Figure 1. The shaded areas indicate those parts of the graph that are eliminated as feasible solutions to the problem because of assembly line capacity limitations. Any solution to the problem must be a combination of a number of washers and dryers that falls within area *abcd*.

The other restrictions to the problem are only slightly different. The Stamping Department capacity limits production to 10,000 washers per month, 10,000 dryers per month, or any comparable combination of washer and dryer production. Satisfactory combinations are 8,000 washers and 2,000 dryers, 5,000 each, etc. This restriction is shown in Figure 2 as a straight line that restricts further the space allowing feasible solutions. Figure 2 also shows the limitations

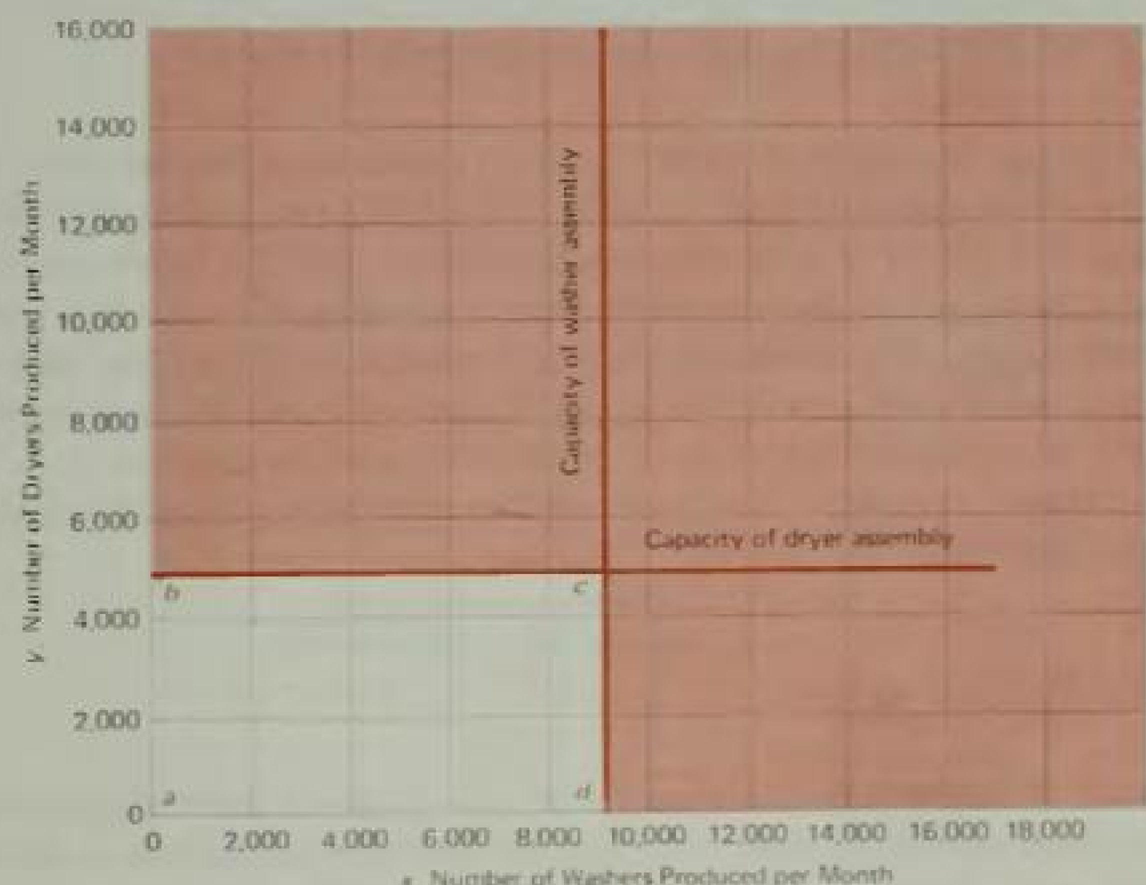


Figure 1. Graphic illustration of limitations imposed by the assembly-line capacity of washers and dryers

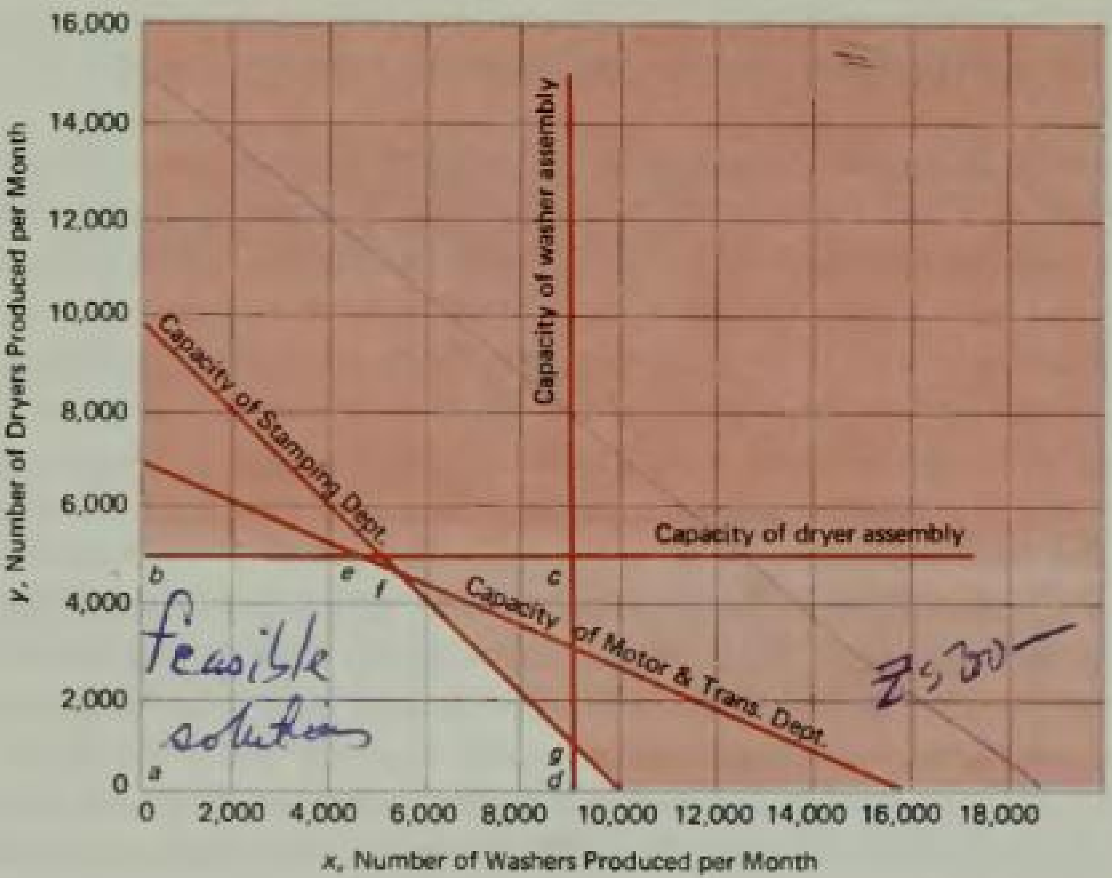


Figure 2. Remaining limitations imposed by capacities of the Stamping Department and Motor and Transmission Department. The area enclosed by *abefgd* includes all feasible solutions to the problem

imposed by the capacities of the Motor and Transmission Department, represented by another straight line that goes through the two points (x = 0, y = 7,000) and (x = 16,000, y = 0). Combinations of washer and dryer production that are equivalent fall on this straight line, as before; for example, washers 8,000, dryers 3,500. We now know that the combination of washer and dryer production that we are seeking lies within the area *abefgd*. All other combinations have been eliminated by the stated restrictions of the problem.

With this general background, let us now formulate the washer-dryer problem as a linear optimization model.

Definition of Decision Variables

Since some of the facilities are shared and each product's costs and profits are different, we must decide how to utilize the available capacities in the most profitable way. Dryers seemingly are more profitable. However, when the manager tried producing the maximum amount of dryers within market limitations, using the balance of his capacity to produce washers, he found that such an allocation of departmental capacities resulted in poor profit performance. He now feels that some appropriate balance between the two products is best, and he wishes to determine the monthly rates of production for each product.

Thus, the decision variables are:

x = the number of washers to be produced per month.
y = the number of dryers to be produced per month.

Definition of the Objective Function

The physical plant and basic organization exists and represents the fixed costs of the enterprise. Since we know that these costs are irrelevant to the production scheduling decision, we ignore them. The manager, however, has obtained price and variable cost information and has computed the contribution to profit and overhead per unit, as shown in Table 1. He wishes to maximize profit, and the contribution rates are related linearly to his objective. Therefore, the objective function that he wishes to maximize is the sum of the total contribution from Washers x ($90x$) plus the total contribution from Dryers y ($100y$), or

$$\text{Maximize } Z = 90x + 100y$$

Definition of Constraints

In general, the constraints were defined as the departmental capacities and were illustrated graphically in Figures 1 and 2. However, now we must state these constraints in mathematical form.

TABLE 1
SALES PRICES, VARIABLE COSTS, AND CONTRIBUTIONS FOR WASHERS AND DRYERS

	Sales price, p	*Variable costs, v*	*Contribution to profit and overhead, $C = p - v$*
Washers, x	\$400	\$310	\$90
Dryers, y	450	350	100

Stamping Department Constraint. The equation for the line describing the Stamping Department capacities that is shown in Figure 2 takes the form $y = mx + b$, where m is the slope of the line and b is the y intercept. Since the slope is -1 and the y intercept is 10,000, the equation is:

$$y = -x + 10{,}000$$

Since we wish to express all the combinations of x and y that fall below the line, the constraint is stated as:

$$x + y = \leq 10{,}000$$

Motor and Transmission Department Constraint. In a similar way, the equation for the line describing the capacities for the Motor and Transmission Department that is shown in Figure 2 is:

$$y = -7/16x + 7{,}000$$

Transposing and reducing the fraction to a decimal, and expressed as a constraint:

$$0.4375x + y \leq 7{,}000$$

Assembly Line Constraints. The mathematical statements for the assembly line constraints are simple, since the washer assembly line can produce no more than 9,000 washers and the dryer assembly line has a capacity of 5,000 dryers. Therefore:

$$x \leq 9{,}000$$
$$y \leq 5{,}000$$

Minimum Production Constraints. The minimum production for each product is 0; therefore:

$$x \geq 0$$
$$y \geq 0$$

The Linear Optimization Model

We now can summarize a statement of the linear optimization model for the two-product washer-dryer company in the standard linear programming format, as follows:

Maximize $Z = 90x + 100y$
Subject to:
$x + y \leq 10{,}000$ (Stamping Department)

$0.4274x + y \leq 7{,}000$	(Motor and Transmission Department)
$x \leq 1{,}000$	(Washer Assembly)
$y \leq 5{,}000$	(Dryer Assembly)
$x, y \geq 0$	(Minimum production)

SOLUTION AND INTERPRETATION FOR THE WASHER-DRYER MODEL

We will assume that we have a mechanism for solving linear optimization models when they are formulated in the standard format that we just showed. Indeed, linear programming computing codes commonly are available in both interactive mode (using a time-share terminal) and batch mode (for large-scale linear programming problem solutions). To use either kind of computing program to solve linear optimization models, we must present the problem to the "black box" in the precise form required. This input format usually is more user oriented in interactive time-share systems, and we shall use one of these kinds of programs to illustrate the solution to our washer-dryer problem.*

Many computer programs for linear programming are available, and the instructions for each individual program are unique to that program, whose documentation indicates exactly how to provide input. The form of the computer output may vary from program to program, but the content is similar. (The computer input for the interactive program used is shown in Figure 3*a*.

Computer Output. Figure 3*b* shows the solution output. First, the terminal prints the optimum value of the objective function, $946,666.667. In other words it states that $Z = 946.667$ in the objective function for an optimal solution.

Next, the terminal prints the values of the variables in the optimum solution. Note that scientific notation is used—that is, the value of each variable is followed by "E" and some number. This means that the number preceding the E is to be multiplied by that number of 10s. For example, E1 means multiply by 10, E2 by 100, etc. E0 indicates that the multiplier is "1" or simply that the value of the variable needs no modification.

Let us consider only the first two variables listed in the solution, which we have named WASH and DRY. The solution states that their optimal values are 5,333.3 and 4,666.7, respectively. This is point *f* in Figure 2, the point where the capacity constraint lines for the Stamping Department and the Motor and

*See J. W. Buckley, M. R. Nagarai, B. L. Sharp, and J. W. Schenck, *Management Problem-Solving with APL*, Melville Publishing Company, Los Angeles, 1974.

```
      LPENTER
ENTER THE NAME OF THIS PROJECT  WASHER-DRYER PROBLEM
MAXIMIZE OR MINIMIZE:  MA
OBJECTIVE FUNCTION:Z=90WASH+100DRY
ENTER CONSTRAINT EQUATIONS, (STRIKE JUST A CARRIAGE RETURN TO STOP INPUT)
[001] WASH+DRY ≤ 10000
[002] .4375WASH+DRY ≤ 7000
[003] WASH ≤ 9000
[004] DRY ≤ 5000
                              (a)


      LPRUN

                      WASHER-DRYER PROBLEM

THE OPTIMAL VALUE OF THE OBJECTIVE FUNCTION IS:    946666.667

                      THE VARIABLES IN THE SOLUTION ARE

VARIABLE  WASH     AT LEVEL    5.3333E3
          DRY                  4.6667E3
          SLK3                 3.6667E3
          SLK4                 3.3333E2

                              (b)
```

Figure 3. a) Computer input, and (b) computer solution to the WASHER-DRYER PROBLEM

Transmission Department intersect. Using the solution values for WASH and DRY, we insert them in the objective function and compute Z:

$$Z = 90 \times 5{,}333.3 + 100 \times 4{,}666.7 = 946{,}667$$

This checks with the optimal value of Z given by the computer solution.

Let us check one further bit of logic: if the solution to our problem is at the intersection of the two capacity constraint equations, then we should be able to solve the equations for the two lines simultaneously to determine those values of WASH and DRY that are common to the equations. First, let us use the equation for the Stamping Department constraint and solve for y, the number of dryers:

$$y = 10{,}000 - x \tag{1}$$

We then substitute this value of y in the constraint equation for the Motor and Transmission Department:

$$y = 7{,}000 - 7/16x \tag{2}$$

solving simultaneously:

$$10{,}000 - x = 7{,}000 - 7/16x$$
$$9/16x = 3{,}000$$
$$x = 5{,}333.3 \text{ washers}$$

Using $x = 5{,}333.3$ in Equation 1:

$$y = 10{,}000 - 5{,}333.3 = 4{,}666.7 \text{ dryers}$$

Of course, this agrees with the approximate values that we could read from the graph of Figure 2, as well as with the computer solution shown in Figure 3.

We also can verify the solution graphically by plotting various values of the objective function on the graph of constraints. At a total contribution of $450,000, for example, the objective function is:

$$90x + 100y = 450{,}000$$

when $\quad x = 0, y = 4{,}500$

and when $\quad y = 0, x = 5{,}000$

The line passing through these two points is shown as the $450,000 line in Figure 4. This line defines all combinations of washer and dryer production that yield a total contribution of 450,000. Now let us choose a larger value of contribution, perhaps $630,000. The $630,000 line plotted in Figure 4 defines all combinations of washer and dryer production that yield this total contribution. Note that this line is parallel to the $450,000 line. If we again increase contribution to $900,000 and plot its line in Figure 4, we see that we shall be limited in the size of total contribution by the point *f*, which defines the combination of washer and dryer production that produces the largest possible contribution within the space of feasible solutions. As we noted before, point *f* is the intersection of the two lines that define stamping capacity and motor and transmission capacity, and the solution is determined by the simultaneous solution of these two lines.

Slack Variables. As we have noted, the solution is at point *f*, where the two constraint equations for stamping capacity and motor and transmission capacity intersect. Another way to interpret this fact is that this solution completely utilizes these two departments—there is no residual slack in their capacities. This is important because any of the other feasible solutions in the polygon *abefgd* of Figure 2 would have involved some slack capacity in one or both of these two departments. If there had been slack capacity for either of these two departments in the optimum solution, that fact would have been indicated in the computer output for the optimum solution.

Now, note that the computer gave us the value of variables for which we did

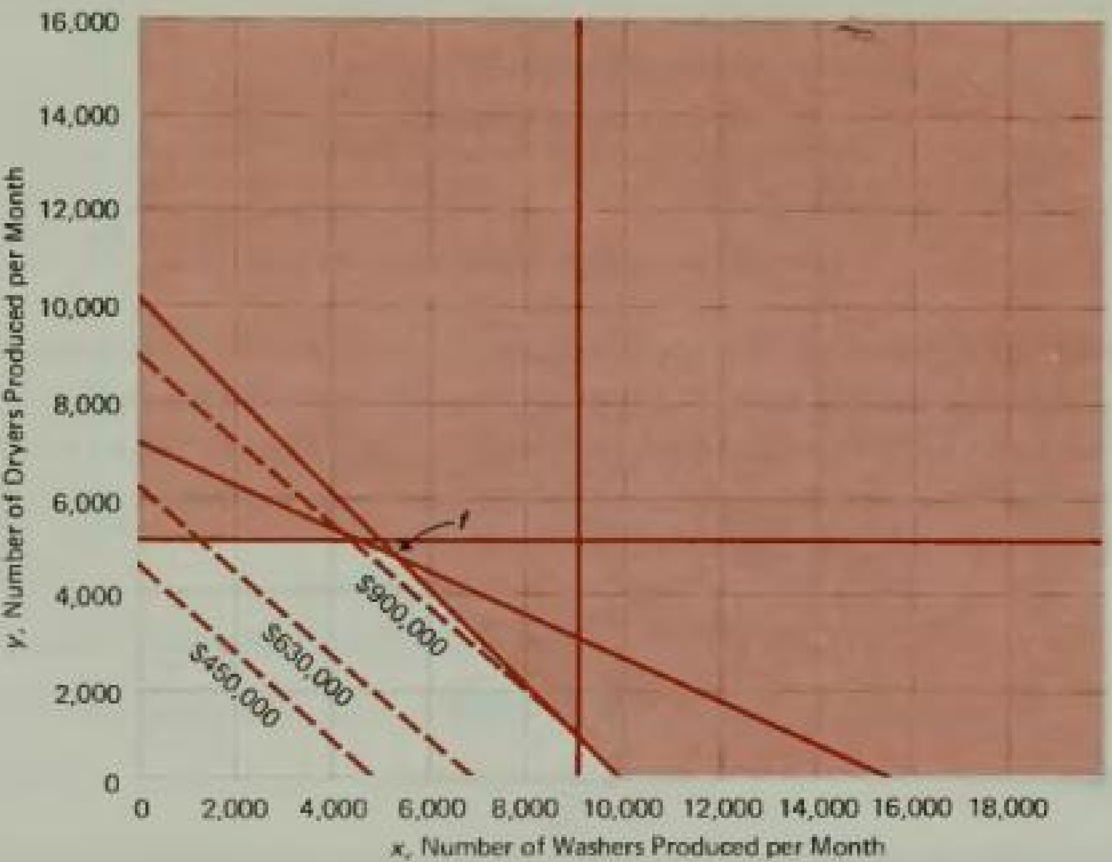


Figure 4. Contribution lines plotted to show the effect of larger and larger contributions. The contribution lines appear as parallel lines. The maximum possible contribution within the area of permitted solutions is defined by a line through point *f*

not ask explicitly, *SLK*3 and *SLK*4. These are the slack values related to Constraints 3 and 4, the assembly line constraints. Constraint 3, WASH $\leq$ 9,000, was the washer assembly line capacity. The solution shows us that if we were producing according to the optimum solution, there would be slack capacity of 3,666.7 units in the washer assembly line. Similarly, the value of *SLK*4 indicates that there is slack capacity of 3,333.3 units in the dryer assembly line.

These interpretations of the optimum solution to the washer–dryer production problem are rather simple. The important point is that equivalent interpretations of more complex problems are a straightforward extension of these ideas. The solution will state the combination of variables that optimizes the objective function. Some, but not all, of the constraints will be the controlling ones, and there will be slack in some of the resources (i.e., they will not all be fully utilized). In our example, the slack was in the use of the assembly lines for the two products. However, if the capacity of the dryer assembly line had been only 4,000 units, it would have become one of the controlling ("tight") contraints (as may be seen from either Figure 2 or 4), and there would have been some slack capacity in the Motor and Transmission Department.

Sensitivity Analysis and Interpretation of Results

If we wanted only the solution to the problem—the optimal combination of variables, the value of slack variables, and the optimum value of the objective function—we could stop at this point by answering "NO" to the next question typed out by the terminal, "DO YOU WISH SENSITIVITY ANALYSIS?" However, additional valuable information is available to the decision maker, who can obtain it simply by answering "YES" to the question, as we have done in Figure 5 for the washer–dryer problem.

While the optimum solution states what to do now, given the objective function and the constraints, the sensitivity analysis raises questions about opportunities and what possibly could or should be done to improve the solution to the managerial problem.

Figure 5 presents the sensitivity analysis in tabular form, first for each constraint and then for the prices (contributions) for each product. For each constraint, a "SHADOW" (price), the "LB" (lower bound of the right-hand side of the constraint), "CURRENT" (current value of the right-hand side), and "UB" (upper bound of the right-hand side) are listed. At first these terms may appear complex; but let us define their meanings in the context of our washer–dryer example.

Shadow Prices. The shadow prices indicate the value of a marginal unit in the right-hand side of the constraint. For example, recall the meaning of the first constraint for the Stamping Department (WASH + DRY ≤ 10,000). It states that the total available capacity is for 10,000 units (either washers or dryers). What would be the marginal value (in the objective function) of one additional unit of capacity? Figure 5 gives the answer as $82.22. If the capacity of the Stamping Department were 10,001 units, the extra unit would add $82.22 to ~~the total~~

```
DO YOU WISH SENSITIVITY ANALYSIS?YES

                    SHADOW        LB           CURRENT      UB
CONSTRAINT 1     8.2222E1      9.5714E3     1.0000E4     1.2063E4
           2     1.7778E1      4.9375E3     7.0000E3     7.1875E3
           3     0.0000E0      5.3333E3     9.0000E3     7.2370E75
           4     0.0000E0      4.5667E3     5.0000E3     7.2370E75

PRICE      WASH                4.3750E1     9.0000E1     1.0000E2
           DRY                 9.0000E1     1.0000E2     2.0571E2
 → END ←
```

Figure 5. Sensitivity analysis for the WASHER-DRYER PROBLEM

contribution. Conversely, if the capacity were only 9,999 units, this amount would be subtracted from the total contribution.

Now observe that the shadow price for the capacity of the Motor and Transmission Department is \$17.78. The marginal value of capacity in the Stamping Department is 4.6 times that for the Motor and Transmission Department. The shadow prices tell the manager that the opportunity provided by increasing stamping capacity is relatively large, thus allowing the appraisal of expansion proposals for both departments.

The shadow prices for Constraints 3 and 4 (assembly lines) are 0 because these constraints do not limit us in the current solution. As we noted before, if dryer assembly line capacity were only 4,000 units, then it would become one of the controlling constraints. Although the optimum solution would change, the shadow price for Constraint 4 would become some positive value, indicating a marginal value to increasing capacity.

Lower, Current, and Upper Bounds. We just defined shadow prices as the value of marginal units of resources. But for what ranges are these marginal rates valid? Can we increase capacity for stamping to two or three times its present capacity and expect to obtain an additional \$82.22 per unit in the objective function? No—there are limits, and the bounds tell us exactly what they are. Taking the Stamping Department capacity as an example, it currently is 10,000 units, as shown in Figure 5 under the "CURRENT" column. However, we see that the shadow price is valid in the range of 9,571 and 12,063 units. If we could increase stamping capacity to 12,063 units, we would obtain an additional 82.22 × 2,063 = \$169,620 in total contribution. We would be able to increase contribution by (169,620 × 100)/946,667 = 17.9 percent. On the down side, if we had a decrease in stamping capacity, perhaps due to machine breakdowns, and capacity fell to the lower bound of 9,571 units, we would lose \$35,272 in total contribution. The interpretation for the bounds on the capacity of the Motor and Transmission Department is similar.

Now let us examine the significance of the bounds on the capacities for assembly lines. Take Constraint 4, the dryer assembly line capacity, for example. Its lower bound is 4,666.7 units. A shadow price of 0 applies if capacity falls to that level (if the constraint is ineffective in that range). But as we already have noted, if capacity falls below 4,666.7 units, the contraint becomes one of those controlling the solution.

The upper bound for Constraint 4 is listed as 7.2370E75. This is the code for infinity in this particular linear programming computer code. In effect, there is no upper bound.

Price Sensitivity. The contribution rates in the objective function generally are termed "prices." Recall that the contribution for a unit of WASH was $90.00 and $100.00 for a unit of DRY, and these are shown as the "CURRENT" values in Figure 5. But what if "prices" change? Would the changes affect the solution? The upper and lower bounds for prices shown in Figure 5 indicate the range of prices (contribution rates) for which the optimum solution is valid. For example, the contribution rate for WASH could be anywhere in the range of $43.75 to $100.00, and the optimum amount of WASH and DRY still would be as indicated in the present solution (i.e., produce 5,333.3 washers and 4,666.7 dryers). Of course, the total contribution would change because the contribution rate changed, but the *optimal decision* would remain the same.

There is a practical significance to the price sensitivity. For example, the manager might estimate contribution for WASH at $90.00; however, these kinds of figures seldom are absolutely precise. Suppose that the contribution is somewhere in the $85.00–$95.00 range. In this case, the same solution applies. As a result, the use of a rough estimate for the contribution rate is adequate, and we should not spend additional time and money to refine the estimate. Thus, the bounds help indicate how we should allocate time and money to refine cost information. If the bounds are tight, it may be worthwhile to be precise, but if they are loose we would gain nothing by attempting to improve the estimates.

This section has emphasized the formulation of linear programming problems and the interpretation of results, assuming a computer technique for solution. Now let us look inside the simplex solution technique.

SIMPLEX SOLUTION

We shall develop the basic simplex method of solution by means of a simplified problem. A manufacturer has two products, I and II, both of which are made in two steps by Machines *A* and *B*. The process times per hundred for the two products on the two machines are (setup times are negligible):

Product	Machine *A*	Machine *B*
I	4 hours	5 hours
II	5 hours	2 hours

For the coming period, Machine *A* has available 100 hours and *B* has available 80 hours.

The contribution for Product I is \$10 per 100 units; for Product II, \$5 per 100 units. The manufacturer is in a market upswing and can sell as much as he can produce of both products for the immediate period ahead. We wish to determine how much of Products I and II he should produce to maximize his contribution.

Formulation of the Problem

We adopt a system of notation. We let x_I be the amount in 100s of Product I, and x_{II} the amount in 100s of Product II that will be produced. We are limited only by the available hours on Machines *A* and *B*, so we know that the sum of the times spent producing the two products on the two machines cannot exceed 100 hours for Machine *A* and 80 hours for Machine *B*. We can express this symbolically as:

$$\text{Machine } A \quad 4x_I + 5x_{II} \leq 100 \tag{3}$$

$$\text{Machine } B \quad 5x_I + 2x_{II} \leq 80 \tag{4}$$

Since the total contributions, which we want to be as large as possible, depend only on the amounts of the two products produced, we can state:

$$\text{Maximize } Z = 10x_I + 5x_{II} \tag{5}$$

This is our *objective function*. We want to find the combination of values for x_I and x_{II} that fits into the restrictions imposed by the manufacturing times for the two products, and the total available time (as expressed by Equations 3 and 4) and in addition makes the total contribution (as expressed by Equation 5) a maximum.

We want to recognize that it is entirely possible that Machines *A* and *B* could have idle time. If W_A is the idle time on Machine *A* and W_B the idle time on Machine *B*, Equations 3 and 4 become:

$$4x_I + 5x_{II} + W_A = 100 \tag{6}$$

$$5x_I + 2x_{II} + W_B = 80 \tag{7}$$

Initial Solution

Let us rearrange Equations 6 and 7 by placing the variables x_I, x_{II}, W_A, and W_B at the heads of columns, with only the coefficients below in columns. Also, let us move the numbers on the right-hand side of the equation to the left. This is done in Table 2.

TABLE 2
PROBLEM EQUATIONS IN TABLEAU FORM

	x_I	x_{II}	W_A	W_B
100	4	5	1	0
80	5	2	0	0

We have two equations and four unknowns. A maximum of two of these variables can have positive values, and at least two of them must be zero. This gives us the key to establishing an initial solution, which we then may use as a basis for improvement. If we start out by assuming that x_I and x_{II} both are zero, it is easy to see by Equations 6 and 7 that W_A must be 100 and W_B must be 80. Admittedly, this solution is a very poor one because it says that all the available time is idle, but it fits the problem equations and allows us to start. We place this initial or trivial solution in the matrix array, as shown in Table 2. This form of a linear optimization model is called the "simplex tableau." Note that we have placed the coefficients of the objective function above the array in Table 3. The stub identifies the variables in the solution and shows their values. To the left in the stub, we have the contribution rates of the variables in the objective function. Neither W_A nor W_B are in the objective function, so zeros are placed in this column. This makes good sense; if the two machines are completely idle, the contribution is zero. The value of the objective function at this point is:

$$10(0) + 5(0) + (0)W_A + (0)W_B = 0$$

Before proceeding, let us name the various parts of the tableau. Figure 6 shows the nomenclature of the various parts of the tableau. The solution stub always

TABLE 3
INITIAL SOLUTION SHOWN IN TABLEAU

			10	5	0	0 ← Coefficients of objective function added
			x_I	x_{II}	W_A	W_B
0	W_A	100	4	5	1	0
0	W_B	80	5	2	0	1

(Answer shown here; variables not in this stub are zero)

(These numbers show the contribution rates of the variables W_A and W_B in the objective function)

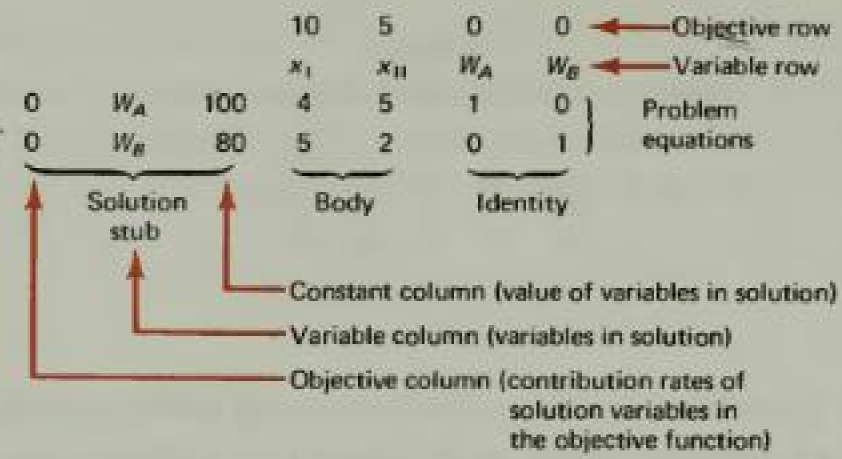


Figure 6. Nomenclature of the simplex tableau

will contain three columns. The body and identity will vary in size, depending on the particular problem.

To improve the initial solution, we must have a measure of the potential improvement that we would make in the objective function by bringing some of the variables into the solution that currently are zero, instead of the variables that currently are in the solution. We shall develop an *index row*, which will be placed just below the present initial tableau. These index numbers will appear under the constant column, the body, and the identity. They are calculated from the formula:

Index number =
Σ (numbers in column) $\times$ (corresponding number in objective column)
$-$ (number in objective row at head of column)

For our problem, the index row numbers are:

(1) Index number for constant column

$$= (100 \times 0 + 80 \times 0) - 0 = 0$$

(2) Index number for first column of body

$$= (4 \times 0 + 5 \times 0) - 10 = -10$$

(3) Index number for second column of body

$$= (5 \times 0 + 2 \times 0) - 5 = -5$$

(4) Index number for first column of identity

$$= (1 \times 0 + 0 \times 0) - 0 = 0$$

TABLE 4
INITIAL SIMPLEX TABLEAU SHOWING INDEX ROW

			10	5	0	0	
			x_I	x_{II}	W_A	W_B	
0	W_A	100	−4	5	1	0	
0	W_B	80	5	2	0	1	
		0	−10	−5	0	0	←Index row

(5) Index number for second column of identity

$$= (0 \times 0 + 1 \times 0) - 0 = 0$$

We now place the index numbers in the initial simplex tableau (see Table 4).

We see that the index row is merely the objective row preceded by minus signs. This occurs only when the objective column contains all zeros.

The larger the negative number, the greater the potential improvement per unit of the new variable to be introduced. If all the numbers under the body and identity of the index row were zero or positive, no further improvement could be obtained, which would indicate that the solution presented in the stub was an optimal solution.

We see from Table 4 that the column headed by the variable x_I has the greatest improvement potential, so we select it as the *key column*. This selection means that the variable x_I will be introduced into the solution in favor of W_A or W_B. To determine whether x_I will replace W_A or W_B, we must select a *key row*. To do this, *we divide each number in the constant column by the corresponding positive nonzero number in the key column*. The resulting quotients are compared, and the *key row* is selected as the row yielding the smallest nonnegative quotient. For our problem, the quotients are:

$$\text{First row } \frac{100}{4} = 25$$

$$\text{Second row } \frac{80}{5} = 16 \text{ (key row)}$$

The smallest nonnegative quotient is 16, computed for the second row, and it becomes the key row. The number that is common to both the key column and the key row is designated the *key number*. Table 5 shows the tableau, with the key column, row, and number identified.

Let us pause for a moment to examine our rationale. We know that x_I will be introduced into the solution, and we wish to know its maximum value consistent with both problem equations, assuming that none of the variables can take

TABLE 5
INITIAL SIMPLEX TABLEAU WITH KEY COLUMN, ROW, AND NUMBER IDENTIFIED

				10	5	0	0
				x_I	x_{II}	W_A	W_B
	0	W_A	100	4	5	1	0
Key row →	0	W_B	80	5	2	0	1
			0	−10	−5	0	0

Key number — Key column

on negative values. In the first equation, x_I would be the largest possible when x_{II} and W_A were zero; that is:

$$4x_I + 5x_{II} + W_A = 100$$

or

$$x_I = \frac{100}{4} = 25$$

In the second equation, x_I would be the largest possible when x_{II} and W_B were zero; that is:

$$5x_I + 2x_{II} + W_B = 80$$

or

$$x_I = \frac{80}{5} = 16$$

Thus, the second equation is the one that limits x_I. It can be no bigger than 16, and this fact dictates the selection of the second row as the key row.

Once we have selected the key column and row, we can prepare a new tableau representing an improved solution. The first step in developing the new tableau is to calculate the coefficients for the *main row*. This main row appears in the same relative position in the new tableau as the key row in the preceding one. It is computed by dividing the coefficients of the key row by the key number. Table 6 shows this development. The variable and its objective number from the head of the key column, x_I and 10, are placed in the stub of the main row replacing W_B and 0 from the previous tableau. The balance of the objective and variable columns in the stub is copied from the previous tableau; the new table, developed to this point, now appears as Table 7.

Now all the remaining coefficients in the new tableau, including the constant column, the body, the identity, and the index row, can be calculated to complete the tableau by the following formula:

TABLE 6
SIMPLEX TABLEAU WITH MAIN ROW OF NEW TABLE

1st tableau			10	5	0	0	
			x_I	x_{II}	W_A	W_B	
0	W_A	100	4	5	1	0	
0	W_B	80	5	2	0	1	
		0	−10	−5	0	0	
2nd tableau							
		16	1	$\frac{2}{5}$	0	$\frac{1}{5}$	← Main row

$$\text{New number} = \text{old number} - \frac{\begin{pmatrix}\text{corresponding}\\ \text{number of}\\ \text{key row}\end{pmatrix} \times \begin{pmatrix}\text{corresponding}\\ \text{number of}\\ \text{key column}\end{pmatrix}}{\text{key number}}$$

(1) First row constant column

$$\text{New number} = 100 - \frac{80 \times 4}{5} = 36$$

(2) First row, first column of body

$$\text{New number} = 4 - \frac{5 \times 4}{5} = 0$$

(3) Index row, constant column

$$\text{New number} = 0 - \frac{80 \times (-10)}{5} = 160$$

TABLE 7
SIMPLEX TABLEAU WITH VARIABLE AND OBJECTIVE COLUMNS OF STUB COMPLETED

1st tableau			10	5	0	0
			x_I	x_{II}	W_A	W_B
0	W_A	100	4	5	1	0
0	W_B	80	5	2	0	1
		0	−10	−5	0	0
2nd tableau						
0	W_A					
10	x_I	16	1	$\frac{2}{5}$	0	$\frac{1}{5}$

TABLE 8
SIMPLEX TABLEAU WITH FIRST ITERATION COMPLETED

1st tableau			10	5	0	0
			x_I	x_{II}	W_A	W_B
0	W_A	100	4	5	1	0
0	W_B	80	5	2	0	1
		0	−10	−5	0	0
2nd tableau						
0	W_A	36	0	$^{17}/_5$	1	$-^4/_5$
10	x_I	16	1	$^2/_5$	0	$^1/_5$
		160	0	−1	0	2

The remaining coefficients can be calculated in the same way and the completed improved solution is shown in Table 8, with a new index row that shows any new possibilities for improvement.

The solution at this stage is:

$$x_I = 16$$
$$x_{II} = 0$$
$$W_A = 36$$
$$W_B = 0$$

The values in the constant column of the stub indicate that x_{II} and W_B are zero, since they are not in the stub at all. The value of the objective function (contribution) for this solution is \$160, which is given in the constant column, index row.

However, Table 8 shows that the solution still can be improved, since a (-1) appears in the index row under the variable x_{II}. Since it is the only negative number in the index row, it is selected as the key column for the next iteration. The key row is selected in the same way as before. The two quotients are:

$$\text{First row} \quad \frac{36}{17/5} = \frac{180}{17} = 10.59$$

$$\text{Second row} \quad \frac{16}{2/5} = 40$$

The first row has the smallest nonnegative quotient, so it is selected as the key row. A new main row is calculated, as before, by dividing the coefficients in the main row by the key number. The new variable, x_{II}, and its objective number are entered in the stub, and the new numbers in the body, identity, and index row are computed as before. The remaining variable and its objective number are copied from the preceding iteration tableau; Table 9 shows the new solution.

TABLE 9
SIMPLEX TABLEAU, SECOND AND FINAL ITERATION COMPLETED

1st tableau			10	5	0	0
			x_I	x_{II}	W_A	W_B
0	W_A	100	4	5	1	0
0	W_B	80	5	2	0	1
		0	−10	−5	0	0
2nd tableau						
0	W_A	36	0	17/5	1	−4/5
10	x_I	16	1	3/5	0	1/5
		160	0	−1	0	2
3rd tableau						
5	x_{II}	180/17	0	1	5/17	−4/17
10	x_I	200/17	1	0	−3/17	5/17
		2900/17	0	0	5/17	30/17

The new solution in Table 9 is optimal, since no further improvement is indicated in the index row. The values of the variables for the optimal solution are:

$$x_I = 200/17$$
$$x_{II} = 180/17$$
$$W_A = 0$$
$$W_B = 0$$

If these values are inserted in Equations 6 and 7, we find that they check exactly. The solution indicates the Products I and II should be produced in the amounts shown by x_I and x_{II} (in 100s) to produce a maximum contribution of $C = 2{,}900/17 = \$170.59$. For each iteration, the value of the objective function is shown in the constant column, index row. It was $160 for the second solution and $2{,}900/17 = \$170.59$ for the optimal solution. The solution is unique; that is, no other combination of x_I and x_{II} will yield a contribution figure that is as high as this.

Procedure Summary

1. *Formulate the problem and the objective function.*
2. *Develop the initial simplex tableau*, including the initial trivial solution and the index row numbers. The index row numbers in the initial tableau are calculated by the formula:

$$\text{Index number} = \sum \begin{pmatrix} \text{numbers} \\ \text{in} \\ \text{column} \end{pmatrix} \times \begin{pmatrix} \text{corresponding} \\ \text{number in} \\ \text{objective} \\ \text{column} \end{pmatrix} - \begin{pmatrix} \text{number in} \\ \text{objective} \\ \text{row at head} \\ \text{of column} \end{pmatrix}$$

3. *Select the key column*, the column with the most negative index number in the body or the identity.
4. *Select the key row*, the row with the smallest nonnegative quotient, which is obtained by dividing each number of the constant column by the corresponding positive, nonzero number in the key column.
5. *The key number* is at the intersection of the key row and key column.
6. *Develop the main row of the new table.*

$$\text{Main row} = \frac{\text{key row of preceding table}}{\text{key number}}$$

   The main row appears in the new tableau in the same relative position as the key row of the preceding table.
7. *Develop the balance of the new tableau.*
   a. The variable and its objective number at the head of the key column are entered in the stub of the new tableau to the left of the main row, replacing the variable and objective number from the key row of the preceding tableau.
   b. The remainder of the variable and objective columns are reproduced in the new tableau exactly as they were in the preceding tableau.
   c. The balance of the coefficients for the new tableau are calculated by the formula:

$$\text{New number} = \text{old number} - \frac{\begin{pmatrix} \text{corresponding} \\ \text{number of} \\ \text{key row} \end{pmatrix} \times \begin{pmatrix} \text{corresponding} \\ \text{number of} \\ \text{key column} \end{pmatrix}}{\text{key number}}$$

8. *Repeat (iterate) steps 3 through 7c* until all the index numbers (not including the constant column) are positive. An optimal solution then results.

The interpretation of the resulting optimum solution is as follows: the solution appears in the stub. The variables shown in the variable column have values that are shown in the corresponding rows of the constant column. The value of the objective function is shown in the constant column, index row. All variables not shown in the stub are equal to zero.

Checking the Work

The simplest and most effective check of the work in progress is to establish a check column to the right of the simplex tableau. The numbers in the check column are simply the algebraic sum of all the numbers in a given row, beginning with the constant column and adding all the coefficients to the right. This check column can be established for each row, including the index row. All transformations of the numbers in the check column are the same as for any of the other numbers in the table. After transformation, the algebraic sum of the row coefficients should equal the transformed number in the check column. If it does not, an error has been made, and it should be traced and corrected before proceeding.

Degeneracy in the Simplex

Degeneracy in the simplex solution can be recognized at the time that the key row is being selected. If a tie exists between two or more rows for the smallest nonnegative quotient at that time, then the problem is degenerate. Table 10 shows this situation for a simple tableau that is slightly different from the example that we have been using. Dividing the constant column by the corresponding number in the key column, we have:

$$\text{Row 1} \quad 100/4 = 25$$
$$\text{Row 2} \quad 125/5 = 25$$

A tie exists, and there is the possibility that if the wrong row is selected as the key row, the variable in the stub of the other row may disappear. The problem may begin to cycle at this point, preventing the achievement of an optimum solution. The degeneracy is resolved by the following procedure.

TABLE 10
DEGENERACY IN THE SIMPLEX TABLEAU

			10	5	0	0
			x_I	x_{II}	W_A	W_B
0	W_A	100	4	5	1	0
0	W_B	125	5	2	0	1
		0	−10	2	0	0

1. Divide each element in the tied rows by the key column number in that row.
2. Compare the resulting ratios column by column, from left to right, first in the identity and then in the body.
3. The first comparison that yields unequal ratios breaks the tie.
4. The key row is the row that has the algebracially smaller ratio.

Applying this procedure to the degenerate problem of Table 10, we begin our paired comparison in the identity. The ratios in column W_A are:

$$\begin{array}{ll} \text{First row} & \frac{1}{4} \\ \text{Second row} & 0 \end{array}$$

The second row yields the algebraically smaller ratio; therefore, it is selected as the key row. The regular simplex procedure then is resumed. This procedure for resolving degeneracy is general and can be applied at any stage of the solution to any size tableau.

This completes the essential procedure of the simplex solution. There is a great deal more to know about the simplex method than is shown here, however. Our illustrative problem involved restrictions (i.e., the hours were limited to certain maximum values). In problems involving requirements (e.g., hours must be greater than or equal to certain specified values), equalities (e.g., hours must be exactly equal to specified values), or approximations, the slack variables must be handled correctly to prepare the model for simplex solution. All these and other topics pertinent to linear programming can be found in books devoted to the study of this subject.

Review Questions

1. Outline the model building process that is used to develop a linear optimization model.

2. A chemical manufacturer produces two products, x and y. Each product is manufactured by a two-step process that involves blending and mixing in Machine *A*, and packaging on Machine *B*. The two products complement each other, since the same production facilities can be used for both products, thus achieving better utilization of these facilities.

   Since the facilities are shared and each product's costs and profits are different, we must figure out how to utilize the available machine time in the most profitable way. Chemical x seemingly is more profitable. However, when the manager tried producing the maximum amount of Chemical x within market limitations, using the balance of machine time to produce Chemical y, he found that such an allocation of machine time resulted in poor profit performance. He now feels that some appropriate balance between the two products is best and wishes to determine the production rates for each product per two-week period.

What are the decision variables for a linear optimization model that is designed to solve the chemical manufacturer's problem?

3. Referring to the problem stated in Question 2, the manager has obtained the following price and cost information on the two chemicals.

	Sales Price	Variable Costs
Chemical x	\$360	\$290
Chemical y	\$460	\$400

The manager wishes to maximize profit in his decision.

What is the objective function for the chemical manufacturer's problem?

4. Further data for the chemical manufacturer's problem stated in Questions 2 and 3 relate to the process-time for the two products on the mixing machine (Machine *A*) and the packaging machine (Machine B), as follows:

Product	Machine *A*	Machine B
x	4 hours	4 hours
y	5 hours	2 hours

For the upcoming two-week planning period, Machine *A* has available 90 hours and Machine B has available 80 hours. Furthermore, market forecasts indicate that the maximum sale of Chemical x is 18 units and of y is 20 units.

Define the constraints for a linear optimization model for the chemical production problem.

5. Make a complete mathematical statement for the linear optimization model developed in Questions 2, 3, and 4.

6. Plot the constraints for the linear optimization model, stated in Question 5, on a graph. Define the feasible solution space.

```
LP RUN
                          CHEMICAL PRODUCTION

THE OPTIMAL VALUE OF THE OBJECTIVE FUNCTION IS:  1476.00

                    THE VARIABLES IN THE SOLUTION ARE

VARIABLE  X        AT LEVEL    1.8000E1
          Y                    3.6000E0
          SLK2                 8.0000E 1
          SLK4                 1.6400E1

DO YOU WISH SENSITIVITY ANALYSIS?  YES

                      SHADOW       LB          CURRENT      UB
CONSTRAINT    1     1.2000E1    7.2000E1    9.0000E1    9.2000E1
              2     0.0000E0    7.9200E1    8.0000E1    7.2370E75
              3     2.2000E1    9.0000E0    1.8000E1    1.8333E1
              4     0.0000E0    3.6000E0    2.0000E1    7.2370E75

PRICE         X                 4.8000E1    7.0000E1    7.2370E75
              Y                 0.0000E0    6.0000E1    8.7500E1
  → END ←
```

Figure 7. Computer solution and sensitivity analysis for the chemical production problem

7. Using the same graph of the problem constraints developed in Question 6, plot objective function lines for $Z = \$1,200$, and $Z = \$1,300$.

   Based on the graph that you have developed, what is the optimum solution?

8. Figure 7 shows the computer output and a sensitivity analysis of the solution to the chemical production problem.
   a. What would be the marginal value of one additional unit of capacity for Machine *A*? for Machine *B*?
   b. Suppose that the manager wished to increase the capacity of Machine *A* to 100 hours. Is the present solution valid? That is, do the indicated shadow prices apply?
   c. Suppose that the demand for Chemical x, which was estimated as 18, was in error and is actually 12. Is the present solution still valid?
   d. Suppose that costs increase for Chemical *y* from \$400 to \$415. Is the present solution still valid?

9. What is the function of slack variables in the simplex solution to linear optimization models?

10. What is shown in each of the following in the simplex tableau?

a. Objective row.
b. Variable row.
c. Objective column.
d. Variable column.
e. Constant column.

11. Outline the procedure for the simplex solution of linear optimization models.

Problems

1. Once upon a time, Lucretia Borgia invited fifty enemies to dinner. The pièce de resistance was to be poison. In those crude days, only two poisons were on the market, Poison X and Poison Y. Before preparing the menu, however, the remarkably talented young lady considered some of the restrictions placed on her scheme:
   a. If she used more than one-half pound of poison, the guests would detect it and refuse to eat.
   b. Lucretia's own private witch, a medieval version of the modern planning staff, once propounded some magic numbers for her in the following doggerel:

      One Y and X two,
      If less than half,
      Then woe to you.

   c. Poison X will kill 75 people per pound, and Poison Y will kill 200 people per pound.
   d. Poison X costs 100 solid gold pieces per pound, and Poison Y costs 400 solid gold pieces per pound.

   After devising a menu to cover up the taste of the poison, Lucretia found that she was very short of solid gold pieces. In fact, unless she were very careful, she would not be able to have another scheduled poisoning orgy that month. So she called in her alchemist, a very learned man, and told him about her problem. The alchemist had little experience in solving problems of this type, but he was able to translate the four restrictions into mathematical statements.

(1) $X + Y \leq 1/2$
(2) $2X + Y \geq 1/2$
(3) $75X + 200Y \geq 50$
(4) $100X + 400Y = \text{cost}$

Assist the alchemist in solving this problem, using graphic methods. The penalty for failure will be an invitation to next month's dinner.

2. A company makes four products x_1, x_2, x_3, and x_4, which flow through four departments: drill, mill, lathe, and assembly. The hours of department time required by each of the products per unit are:

	Drill	Mill	Lathe	Assembly
x_1	3	0	3	4
x_2	7	2	4	6
x_3	4	4	0	5
x_4	0	6	5	3

The unit contributions of the four products and hours of availability in the four departments are:

Product	Contribution
x_1	$ 9
x_2	18
x_3	14
x_4	11

Department	Hours Available
Drill	70
Mill	80
Lathe	90
Assembly	100

Formulate the problem as a linear optimization model to determine optimal product-mix.

3. A company makes five products, x_1, x_2, x_3, x_4, and x_5, which flow through five departments: blanking, forming, straightening, brazing, and assembly. The following processing times are required:

Time Required

Product	Blanking (hours)	Forming (hours)	Straightening (hours)	Brazing (hours)	Assembly (hours)
x_1	1	1	2	0	1
x_2	1	0	½	1	2
x_3	2	3	0	½	1
x_4	0	2	2	1	2
x_5	½	1	½	2	1

The contribution of each product is as follows:

Product	Contribution
x_1	\$8 per unit
x_2	\$9 per unit
x_3	\$13 per unit
x_4	\$15 per unit
x_5	\$7 per unit

Time available in the various departments, and the estimated incremental cost of idle time, are:

Department	Time Available (hours)	Estimated Incremental Cost of Idle-Time per Hour
Blanking	115	\$12
Forming	100	8
Straightening	140	3.50
Brazing	90	30
Assembly	110	12

Formulate the problem in the simplex tableau.

4. A refinery operating in Nebraska uses four crude oils: Oklahoma, West Texas, Wyoming, and Pennsylvania. These crudes have different delivered prices, as indicated below. The refinery makes four basic end products: regular gasoline, high-test gasoline, diesel fuel, and fuel oil. The catalytic cracking and reforming characteristics of the refinery dictate a limited and different product-mix for the different crudes. Data are as follows:

Crude	Delivered Price per Gallon, Cents	Optimal Throughput for Each Crude			
		Regular Gas (%)	High-Test Gas (%)	Diesel Fuel (%)	Fuel Oil (%)
Oklahoma	7	30	10	40	20
West Texas	6	20	10	60	10
Wyoming	5	10		30	60
Pennsylvania	9	30	50	20	
Present market requirement, gallons	–	2,000	1,500	2,800	3,300

Formulate the problem in a simplex tableau with the objective of minimizing crude costs.

5. A company makes three products, x, y, and z, out of seven materials. The material requirements and contributions for each of the three products are as follows:

Product	Unit Material Requirements						
	P_1	P_2	P_3	P_4	P_5	P_6	P_7
x	3		2		1		
Alternative x		3		4			1
y		2	1			2	
Alternative y	1			2	3		
z				4		2	1
Alternative z			2		6		

Product	Unit Contribution
x	$6
y	$5
z	$9

The amounts of each material that are available are:

p_1	100 units
p_2	115 units
p_3	135 units
p_4	90 units
p_5	85 units
p_6	140 units
p_7	170 units

Formulate the problem as a linear optimization model.

6. Mesa Plastics, Inc. is a bulk producer of sheet plastic material, which it sells in three sizes (thicknesses). It has two plants located on the same site. Plant *B* was designed after Plant *A* and was specifically built to produce Sizes 1 and 2 economically, since these two sizes had the largest demand. However, Plant *B* is less economical than *A* for Size 3. Time requirements for the three products in the two plants, and the variable hourly costs and time availability for Plants *A* and *B*, are:

Size	Hours per 100 Pounds	
	Plant *A*	Plant *B*
1	0.25	0.20
2	0.40	0.25
3	0.35	0.40
Variable costs/hour	\$250	\$300
Maximum available hours per week	100	100

Sales revenue and maximum demand for the three products are:

Size	Sales Revenue/ 100 Pounds	Maximum Demand per Week, 100 Pounds
1	\$100	310
2	\$120	300
3	\$150	125

Management is considering how production should be allocated for the upcoming period to the two plants in order to maximize contribution, and the following optimal simplex solution has been obtained in tableau form:

			37.5	40	20	45	62.5	30	0	0	0	0	0
			x_{1A}	x_{1B}	x_{2A}	x_{2B}	x_{3A}	x_{3B}	W_A	W_B	W_1	W_2	W_3
0	W_A	10	0	0	0.0875	0	0	0.15	1	1.25	−0.25	−0.3125	−0.35
40	x_{1B}	125	0	1	−1.25	0	0	2	0	5	0	−1.25	0
37.5	x_{1A}	185	1	0	1.25	0	0	−2	0	−5	1	1.25	0
45	x_{2B}	300	0	0	1	1	0	0	0	0	0	1	0
6.25	x_{3A}	125	0	0	0	0	1	1	0	0	0	0	1
		33,250	0	0	21.875	0	0	37.5	0	12.5	37.5	41.875	62.5

What interpretation of the solution would you give management, particularly regarding possible future profits and how they could be obtained?

References

Bierman, H.; C. P. Bonini; and W. H. Hausman. *Quantitative Analysis for Business Decisions*. 5th ed. Homewood, Ill.: Richard D. Irwin, 1977.

Buckley, J. W.; M. R. Nagarai; B. L. Sharp; and J. W. Schenck. *Management Problem-Solving with APL*. Los Angeles: Melville Publishing Company, 1974.

Buffa, E. S., and J. S. Dyer. *Management Science/Operations Research: Model Formulation and Solution Methods*. Santa Barbara, Ca.: Wiley/Hamilton, 1977.

Levin, R. I., and C. A. Kirkpatrick. *Quantitative Approaches to Management*. 3d ed. Englewood Cliffs, N. J.: Prentice-Hall, 1975.

Machol, R. E. *Elementary Systems Mathematics: Linear Programming for Business and the Social Sciences*. New York: McGraw-Hill, 1976.

McMillan, C. *Mathematical Programming*. 2d ed. New York: John Wiley & Sons, 1975.

Reinfeld, N. V., and W. R. Vogel. *Mathematical Programming*. Englewood Cliffs, N. J.: Prentice-Hall, 1958.

Thierauf, R. J., and R. C. Klekamp. *Decision Making through Operations Research*. 2d ed. New York: John Wiley & Sons, 1975.

Appendix C

Linear Programming— Distribution Methods

Distribution methods of linear programming developed around the classic problem of how to distribute goods from a set of origin points (perhaps factories) to multiple destinations (perhaps warehouses) at a minimum cost. The basic problem can be formulated and solved in the simplex format, but the special methodology that we discuss in this chapter is simpler, easier to understand, and computationally faster.

STEPPING STONE METHOD FOR DISTRIBUTION PROBLEMS

We will begin our survey of distribution (transportation) type problems by using solution methods that commonly are termed *stepping stone methods*. The nature of the problems and solution allows us to have close contact with all phases of the solution and gives us a "feel" for what we are doing and why.

To illustrate, let us assume that distribution situation indicated by Figure 1. Here, we have three factories located in Chicago, Detroit, and Atlanta, which produce some identical products. The distribution system of the organization has established five major distribution points that serve various market areas in Milwaukee, Cincinnati, Des Moines, Buffalo, and New York City. The three

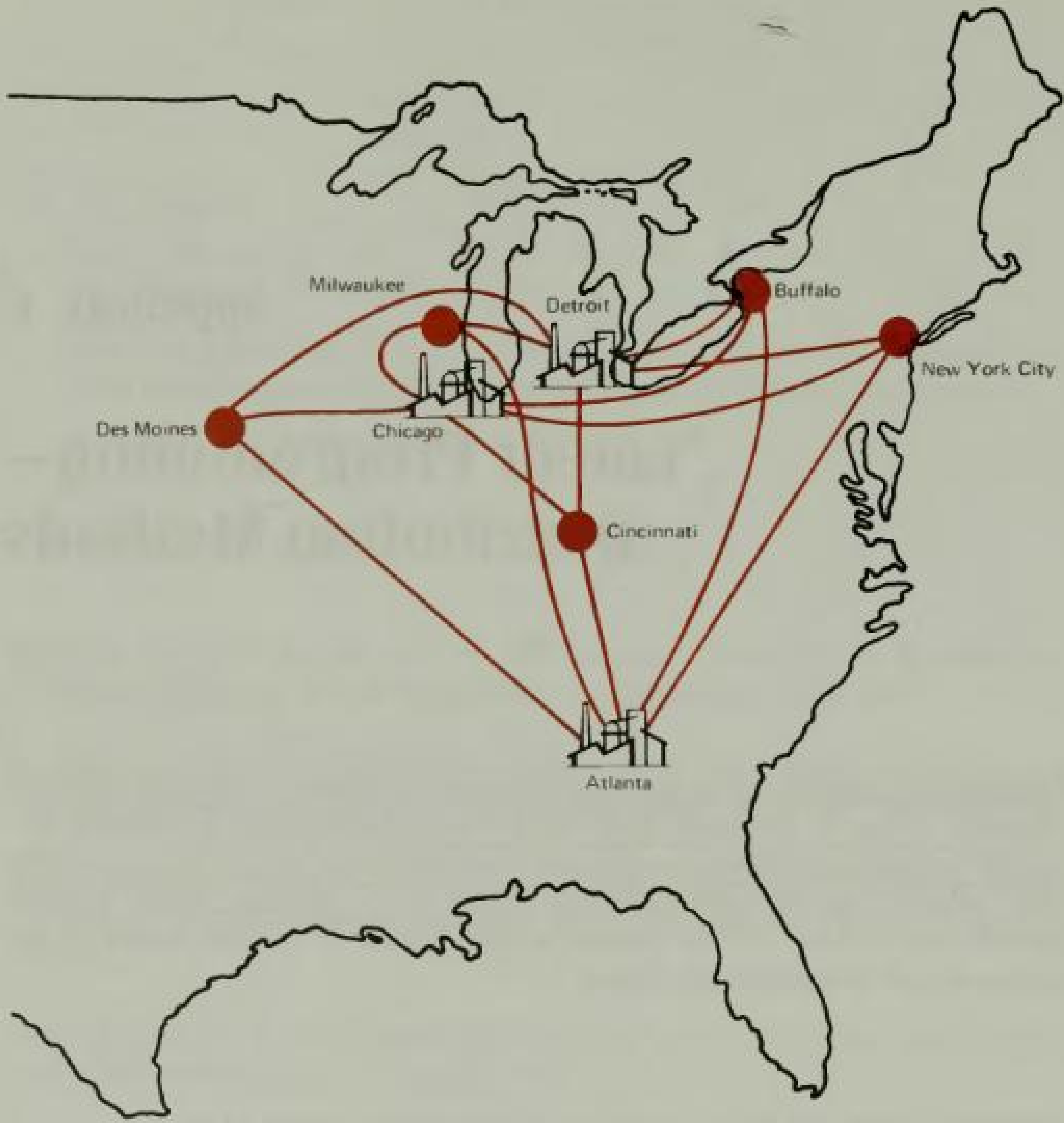


Figure 1. Geographical locations of factories and distribution points

factories have capacities that determine the availability of product, and the market demand in the five major areas determines the requirements to be met. The problem involves determining an allocation of the available product at the three factory locations to the five distribution points in a way that meets the demands and minimizes the cost of distribution for the entire system.

Data for our illustrative problem are shown in Table 1. We see that there are 19,000 units available at the Detroit plant, 28,000 at Chicago, and 25,000 at Atlanta, or a total of 72,000. Similarly, demands in the five market areas are

TABLE 1
SUMMARY OF QUANTITIES OF PRODUCT AVAILABLE AND REQUIRED, AND DISTRIBUTION COSTS PER THOUSAND
(Distribution or Transportation Matrix)

From Factories \ To Dist. Points	Milwaukee (V)	Cincinnati (W)	Des Moines (X)	Buffalo (Y)	New York City (Z)	Available at Factories 1000s
Detroit (A)	42	42	44	40	44	19
Chicago (B)	34	42	40	46	48	28
Atlanta (C)	46	44	42	48	46	25
Required at Dist. Points 1000s →	11	13	7	17	24	72

indicated in the bottom row of the table and also total 72,000 units, although, as we shall see later, this is not a necessary requirement for solution.

These figures of units available and required are commonly termed the *rim conditions*. Table 1 shows the distribution costs per thousand units for all combinations of factories and distribution points. These figures are shown in the small boxes. For example, the distribution cost between the Chicago plant and the Milwaukee distribution point is $34 per thousand units. For convenience in notation, we have labeled the plants *A*, *B*, and *C* and the distribution points *V*, *W*, *X*, *Y*, and *Z*. Table 1 commonly is termed the *distribution or transportation matrix*. Our measure of effectiveness is distribution cost, and we wish to distribute the 72,000 units available from factories *A*, *B*, and *C* to the distribution points *V*, *W*, *X*, *Y*, and *Z* so as to minimize the total distribution cost, within the restrictions imposed by units available and required.

An Initial Solution

We shall begin by assigning the various units that are available in an arbitrary way, ignoring the distribution costs. (We shall see later how to establish good initial solutions.) We begin in the upper left-hand corner of the matrix (the so-called northwest corner), and we note that *A* has 19 (thousand) units available and that *V* needs 11 (thousand). We assign the 11 from *A* to *V*. (See Table 2:

TABLE 2
NORTHWEST INITIAL SOLUTION

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	$\boxed{42}$ 11	$\boxed{42}$ 8	$\boxed{44}$	$\boxed{40}$	$\boxed{44}$	19
B	$\boxed{34}$	$\boxed{42}$ 5	$\boxed{40}$ 7	$\boxed{46}$ 16	$\boxed{48}$	28
C	$\boxed{46}$	$\boxed{44}$	$\boxed{42}$	$\boxed{48}$ 1	$\boxed{46}$ 24	25
Required 1000s →	11	13	7	17	24	72

Total distribution cost:

AV,	11 × 42 =	462
AW,	8 × 42 =	336
BW,	5 × 42 =	210
BX,	7 × 40 =	280
BY,	16 × 46 =	736
CY,	1 × 48 =	48
CZ,	24 × 46 =	1104
		$3176

circled numbers represent assigned product. For example, 11 in box *AV* means 11,000 units to go from *A* to *V*.) We have not yet used up *A*'s supply, so we move to the right under column *W* and assign the balance of *A*'s supply, 8, to *W*. Looking at the requirements for *W*, we note that it has a total requirement of 13, so we drop down to row *B* and assign the balance of *W*'s requirements, 5, from *B*'s supply of 28. We then move to the right again and assign a portion of the balance of *B*'s supply of 7 to X. We continue in this way, stair-stepping down the matrix until all the arbitrary assignments have been made, as in Table 2. The total distribution cost, $3176, is calculated in Table 2, below the matrix. Note that we have 7 squares with assignments (n rows + m columns − 1) and 8 open squares (i.e., squares without assignments).

Improving the Initial Solution

Is the initial northwest corner solution shown in Table 2 the best possible? We can answer this by successively examining the open squares to see whether we

TABLE 3
EVALUATION OF SQUARE BV FOR POSSIBLE IMPROVEMENT

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	42 (−) 11	42 8 (+)	44	40	44	19
B	34 (+)	42 5 (−)	40 7	46 16	48	28
C	46	44	42	48 1	46 24	25
Required 1000s →	11	13	7	17	24	72

Evaluation of square BV: for shifting one unit to BV, change in cost is:

$$+34 - 42 + 42 - 42 = -8$$

Since this is a net improvement in cost, increase the assignment for BV to the maximum possible. This is limited by the assignment for BW which can be reduced by only 5. Maximum improvement at BV is, therefore, $5(-8) = -40$. New total distribution cost $= 3176 - 40 = \$3136$.

can reduce total distribution cost by shifting assignments to these squares. When we have made all possible shifts in assignments of this nature, we can be certain that we have an optimum solution. Let us examine such a procedure. First, we want to be sure that any shift that we make conforms to the restrictions of available and required units, as shown in the rim conditions of the distribution matrix. Let us select the first open square in the first column, square BV. If we were to add 1 (thousand) unit each to BV and AW and subtract the same amount from AV and BW, we still would satisfy the restrictions of availability and requirements. Table 3 shows us that such a shift would be advantageous because we would be shifting from higher cost routes to lower cost routes. We would be adding 1,000 units each to routes BV and AW at a cost of $34 + 42 = \$76$, and subtracting 1,000 units each from AV and BW at a saving of $42 + 42 = \$84$, or a net decrease of $84 - 76 = \$8$ per 1,000 units.

Since we have found an advantageous shift, we want to take advantage of it by increasing the assignment at BV to the maximum. The shift in assignments is limited to five units, however, because that is the existing assignment at BW, and we cannot reduce it below zero. Therefore, the maximum improvement in the solution that we can effect at BV is 5,000 units at a distribution cost saving of \$8 per thousand, or \$40. The new total distribution cost is now \$3,136. We make the

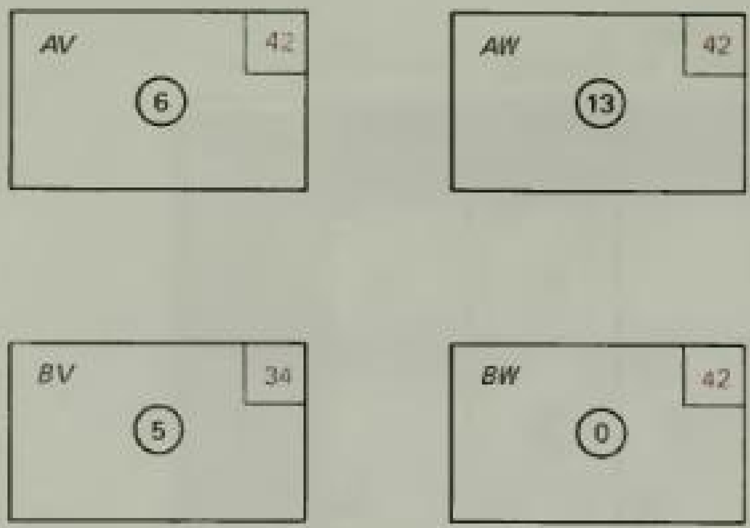


Figure 2. Resulting assignments to the four squares affected by the evaluation of square BV

changes in assignment indicated, and Figure 2 indicates the resulting assignments to the four squares affected.

We now proceed systematically through the table, column by column, evalating each open square and making shifts in assignments when they are advantageous. However, the pattern that is required to evaluate a given open square is not necessarily a rectangular one, as in Tables 3 and 4. We must have a closed path that starts at the square to be evaluated, with right angle turns only at squares that already have assignments, and proceeds either clockwise or counterclockwise. Squares may be skipped to get to the corners, as illustrated in Table 4. No diagonal movement is permitted. Beginning at the square being evaluated, we assign a plus sign, since we propose to add a load, and alternate with minus and plus signs as we go around the closed path. There is only one closed path to evaluate for a given open square if the initial arbitrary solution has been established properly. By adhering to the closed path idea that we have expressed, we ensure that proposed shifts in assignments do not violate the restrictions of the rim conditions (availability and requirements), since, in each row and column, we add and subtract the same amount. Therefore, the totals are unaffected by the shift in assignments that we make.

We proceed systematically through the table, evaluating each open square, column by column or row by row, and making shifts in assignments when they are advantageous. The next open square for evaluation in column V is CV. Square CV is evaluated in Table 4, which shows us that if we were to make the shift in allocations indicated, there would be a net increase in total distribution cost. Therefore, the changes are not made. Proceeding to column W, we pass over square BW, since the evaluation of square BV by Table 3 indicates immediately that BW will only increase total cost.

The next square is CW, which is evaluated in Table 5. CW illustrates a more complex closed path for the evaluation of an open square. In each row and

TABLE 4

EVALUATION OF SQUARE *CV* FOR POSSIBLE IMPROVEMENT

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	42 6	42 13	44	40	44	19
B	34 (−) 5	42	40 7	46 16 (+)	48	28
C	46 (+)	44	42	48 1 (−)	46 24	25
Required 1000s →	11	13	7	17	24	72

Evaluation of square CV: for shifting one unit to *CV*, change in cost is:
$+46 - 34 + 46 - 48 = +10$.

This would result in a net increase in total cost and therefore the changes are not made.

column affected, there is a plus sign and a minus sign; thus, the rim restrictions are not violated by the proposed change in assignments. Table 5 indicates that *CW* also would result in a net increase in total costs; therefore, the changes are not made. As we proceed systematically through the table, evaluating each open square, we may find that squares that previously indicated no improvement later may yield improvement because of subsequent changes that are made. This process is continued until all open squares show no further improvement.

An Optimal Solution. At this point, an optimal solution has been obtained (shown in Table 6). The total distribution cost required by the optimal solution is $2,986, which is $190 less than the original northwest corner solution. This reduction in total distribution cost can be accomplished by the column-by-column evaluation of thirteen open squares, five of which yield improvement. The first time through the table, no improvement is obtained from *CV*, *BW*, *CW*, *CX*, *AZ*, and *BZ*, but squares *BV*, *AX* and *AY* yield improvement. The second time through the table, *BW* and *CW* yield improvement. At this point, the solution is optimal, because reevaluation of all open squares shows no further improvement possible. Table 6 shows the resulting optimal solution and indicates the evaluations of the open squares by the small figures in the lower left-hand corners of the squares. All open squares show that distribution costs either would increase or not change at all by further shifts in assignments.

TABLE 5

EVALUATION OF SQUARE *CW* FOR POSSIBLE IMPROVEMENT

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	42; (+) 6	42; 13 (−)	44	40	44	19
B	34; (−) 5	42	40; 7	46; 16 (+)	48	28
C	46	44; (+)	42	48; 1 (−)	46; 24	25
Required 1000s →	11	13	7	17	24	72

Evaluation of square *CW*: for shifting one unit to *CW*, change in cost is:

$+44 - 48 + 46 - 34 + 42 - 42 = +8$

This would result in a net increase in total cost and therefore the changes are not made.

Alternate Optimal Solutions. The fact that open squares *CX* and *AZ* in Table 6 have zero evaluations is important and gives us flexibility in determining the final plan of action. These zero evaluations allow us to generate other solutions that have the same total distribution cost as the optimal solution generated in Table 6. To take an example, since open square *CX* has a zero evaluation, we

TABLE 6

AN OPTIMAL SOLUTION

(Evaluation of All Squares Without Assignments in this Table Results in No Further Improvement)

From \ To	Milwaukee (*V*)	Cincinnati (*W*)	Des Moines (*X*)	Buffalo (*Y*)	New York City (*Z*)	Available 1000s ↓
Detroit (*A*)	42; +8	42; 2	44; +4	40; 17	44; 0	19
Chicago (*B*)	34; 11	42; 10	40; 7	46; +6	48; +4	28
Atlanta (*C*)	46; +10	44; 1	42; 0	48; +6	46; 24	25
Required 1000s →	11	13	7	17	24	72

Total distribution cost = $2986.

TABLE 7
ALTERNATE BASIC OPTIMUM SOLUTION

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	42 +8	42 2	44 +4	40 17	44 0	19
B	34 11	42 11	40 6	46 +6	48 +4	28
C	46 +10	44 0	42 1	48 +6	46 24	25
Required 1000s →	11	13	7	17	24	27

may make the shifts in allocations indicated by its closed path and generate the alternate optimal solution shown in Table 7. Similarly, we could generate another basic optimal solution by shifting assignments to open square *AZ*.

Degeneracy in Distribution Problems. Another aspect of the mechanics of developing a solution is the condition known as degeneracy. Degeneracy occurs in distribution problems when, in shifting assignments to take advantage of a potential improvement, more than one of the existing assignments go to zero. Degeneracy also can occur in an initial solution that does not meet the $m + n - 1$ requirement for the number of allocations. Examination of the problem in Table 8 shows that degeneracy is about to happen. The problem in Table 8 is only slightly different from the example that we have been using, that is, the requirements for *V* and *X* have been changed. This problem was set up in the usual way, and an initial northwest corner solution was established. The open squares were evaluated column by column, as before, and changes in assignments were made when they indicated potential improvement.

In Table 8, we are evaluating square *AX* by the closed path pattern. Potential improvement is indicated, since a unit of allocation reduces transportation costs by $4 per thousand. We wish to press this advantage to the maximum by shifting as much as possible to *AX*. We are limited, however, by both squares *AV* and BX, each of which has an allocation of 6,000 units assigned to it. When the shift in assignment is made, both *AV* and *BX* go to zero. This is shown in the resulting matrix of Table 9. We now have only six allocations instead of the seven we had before, and we do not meet the restriction on the stepping stone method of solution that we stated earlier: that the number of allocations must be $m + n - 1$.

TABLE 8

EVALUATION OF SQUARE AX PRODUCES DEGENERACY

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	[42] (−) 6	[42] 13	[44] (+)	[40]	[44]	19
B	[34] (+) 6	[42]	[40] 6 (−)	[46] 16	[48]	28
C	[46]	[44]	[42]	[48] 1	[46] 24	25
Required 1000s →	12	13	6	17	24	72

The practical effect of this is that several of the open squares, namely AV, BW, CW, BX, CX, AY, and AZ, cannot be evaluated in the usual way because a closed path cannot be established for them.

We can resolve the degeneracy, however, by regarding one of the two squares in which allocations have disappeared as an allocated square with an extremely small allocation, which we shall call an ϵ allocation. This is illustrated in Table 10. Conceptually, we shall regard the ϵ allocation as infinitesimally small, so that it does not affect the totals indicated in the rim. The ϵ allocation, however, does

TABLE 9

PROBLEM NOW DEGENERATE: SQUARES AV, BW, CW, BX, CX, AY, and AZ, AND AZ CANNOT BE EVALUATED

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	[42]	[42] 13	[44] 6	[40]	[44]	19
B	[34] 12	[42]	[40]	[46] 16	[48]	28
C	[46]	[44]	[42]	[48] 1	[46] 24	25
Required 1000s →	12	13	6	17	24	72

TABLE 10
DEGENERACY RESOLVED BY USE OF THE ε ALLOCATION

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	42 ε	42 13	44 6	40	44	19
B	34 12	42	40	46 16	48	28
C	46	44	42	48 1	46 24	25
Required 1000s →	12	13	6	17	24	72

make it possible to meet the $m + n - 1$ restriction on the number of allocations so that evaluation paths may be established for all open squares. The ε allocation is simply manipulated as though it were no different from the other allocations.

If, in subsequent manipulations, the ε allocation square is the one that limits shifts in assignments, it simply is shifted to the square being evaluated, and the usual procedures are then continued. This is illustrated in Table 11, where we are attempting to evaluate square AZ by the closed path shown. A potential improvement of $8 per 1,000 units is indicated, but the limiting allocation at a negative square is the ε allocation. The net effect of adding and subtracting the ε

TABLE 11
SHIFT OF ε ALLOCATION WHEN IT IS LIMITING

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	42 (−) ε	42 13	44 6	40	44 ε (+)	19
B	34 (+) 12	42	40	46 (−) 16	48	28
C	46	44	42	48 (+) 1	46 24 (−)	25
Required 1000s →	12	13	6	17	24	72

TABLE 12

DISAPPEARANCE OF THE ϵ ALLOCATION WHEN IT FALLS AT A POSITIVE CORNER OF AN EVALUATION PATH

From \ To	V	W	X	Y	Z	Available 1000s
A	42	42 13	44 (−) 6	40	44 (+) ϵ	19
B	34 12	42	40	46 16	48	28
C	46	44	42 (+)	48 1	46 24 (−)	25
Required 1000s →	12	13	6	17	24	72

allocation around the closed path is to move the ϵ allocation from square AV to square AZ. The procedure then is continued as before, until an optimal solution is obtained.

As the procedure continues, the ϵ allocation may disappear. This is illustrated in Table 12, in which we are evaluating the open square CX. Potential improvement of $4 per 1,000 units is indicated, and here we are limited not by the ϵ allocation but by the allocation of 6,000 units at AX. In making the adjustments, we add and subtract 6,000 units around the closed path according to the signs indicated, and the result is that the ϵ allocation at AZ becomes 6,000 units. We now have seven squares with positive allocations, and the ϵ allocation no longer is needed. In carrying through the solution of larger-scale problems, we may find that degeneracy appears and disappears in the routine solution of a problem, or more than one ϵ allocation exists. Also, optimal solutions may be degenerate.

Unequal Supply and Demand

Now we will consider how to handle a problem in which supply and demand are not equal. Suppose, for example, that supply exceeds demand. This situation can be handled by creating a dummy distribution point to receive the extra three units. The nonexistent distribution point is assigned zero distribution costs, since the product never will be shipped. The optimal solution then assigns the available units in the most economical way to the real distribution points and assigns the balance to the dummy department.

When demand exceeds supply, we can resort to a modification of the same technique. In this instance, we create a dummy factory to take up the slack. Again zero distribution costs are assigned for the dummy factory, since the product never will be shipped. The solution then assigns the available product to the distribution points in the most economical way. The solution shows which distribution points should receive "short" shipments in order to minimize total distribution costs.

TECHNIQUES FOR SIMPLIFYING PROBLEM SOLUTION

First, we can simplify the arithmetic complexity considerably by using two methods. A little thought about the example that we used will convince us that it is the cost differences that are important in determining the optimal allocation, rather than their absolute values. Therefore, we can reduce all costs by a fixed amount, and the resulting allocation will be unchanged. In our illustrative example, we may subtract 34 from all distribution cost values so that the numbers with which we must work are of such magnitude to allow many evaluations to be accomplished by inspection. Another simplification in the arithmetic may be accomplished by expressing the rim conditions in the simplest terms. For example, in our illustration, we expressed the rim conditions in thousands of units, thus enabling us to work with two digit numbers only.

Getting an Advantageous Initial Solution

The northwest corner initial solution actually is not used a great deal in practice, since ordinarily it is a rather poor solution, which involves a number of steps to develop an optimal solution. Placing the lowest cost cell in the northwest corner gives us an advantageous start. The usual procedure is to start with some solution by inspecting the most promising routes and entering allocations that are consistent with the rim conditions. In establishing such an initial solution, the only rules to be observed are that there must be exactly $m + n - 1$ allocations, and it must be possible to evaluate all open squares by the closed path methods that we discussed earlier.

If the initial solution turns out to be degenerate, it is simple to increase the allocations to the exact number required by resorting to the ϵ allocation. There are a number of short-cut methods that commonly are used, such as row minimum, column minimum, matrix minimum, and VAM. Although they all have merits, we shall discuss VAM in some detail because it seems particularly valuable for hand computation of fairly large-scale problems. Of course, computer solutions should be used for large-scale problems, if at all possible.

Vogel's Approximation Method (VAM)

VAM facilitates a very good initial solution, which usually is the optimal solution. The technique is a simple one, and it considerably reduces the amount of work that is required to develop a solution. As an example, we shall use the same problem that we used to illustrate the stepping stone method. Table 13 shows the distribution matrix with the distribution costs all reduced by the constant amount, $34. The steps in determining an initial VAM solution are as follows:

1. *Determine the difference between the two lowest distribution costs for each row and each column.* This has been done in Table 13, and the figures at the heads of columns and to the right of the rows represent these differences. For example, in column V, the three distribution costs are 8, 0, and 12. The two lowest costs are 8 and 0, and their difference is 8. In row *A*, the two lowest distribution costs are 6 and 8, or a difference of 2. The other figures at the heads of the columns and to the right of the rows have been determined in a similar way.

TABLE 13

DISTRIBUTION MATRIX WITH INITIAL VAM ROW AND COLUMN DIFFERENCES SHOWN

	8	0	2	6	2		
From \ To	*V*	*W*	*X*	*Y*	*Z*	Available 1000s	
A	8	8	10	6	10	19	2
B	0	8	6	12	14	28	6
C	12	10	8	14	12	25	2
Required 1000s →	11	13	7	17	24	72	

TABLE 14

FIRST *VAM* ASSIGNMENT SATISFIES *V*'S REQUIREMENT
(Row and Column *VAM* Differences are Recalculated)

	↓ 8	0	2	6	2		
From \ To	*V*	*W*	*X*	*Y*	*Z*	Available 1000s ↓	
A	8 X	8	10	6	10	19	2
B	0 11	8	6	12	14	28	6 2
C	12 X	10	8	14	12	25	2
Required 1000s →	11	13	7	17	24	72	

2. *Select the row or column with the greatest difference.* For the example that we are using, the row or column with the greatest difference is column V, which has a difference of 8.

3. *Assign the largest possible allocation within the restrictions of the rim conditions to the lowest cost square in the row or column selected.* This has been done in Table 14. Under column V, the lowest cost square is *BV*, which has a cost of 0, and we have assigned 11 units to that square. The 11 unit assignment is the largest possible because of the restriction imposed by the number required at distribution point V.

4. *Cross out any row or column that has been completely satisfied by the assignment just made.* For the assignment just made at *BV*, the requirements for V are entirely satisfied. Therefore, we may cross out the other squares in that column, since we can make no future assignments to them. This is shown in Table 14.

5. *Recalculate the differences as in Step 1, except for rows or columns that have been crossed out.* This has been done in Table 14, where row *B* is the only one affected by the assignment just made.

TABLE 15

SECOND *VAM* ASSIGNMENT SATISFIES Y'S REQUIREMENT
(Row and Column *VAM* Differences are Recalculated)

	$\cancel{8}$	0	2	↓ $\cancel{6}$	2		
From \ To	*V*	*W*	*X*	*Y*	*Z*	Available 1000s ↓	
A	8 X	8	10	6 (17)	10	19	2
B	0 (11)	8	6	12 X	14	28	$\cancel{6}$ 2
C	12 X	10	8	14 X	12	25	2
Required 1000s →	11	13	7	17	24	72	

6. *Repeat Steps 2 to 5 until all assignments have been made.*
   a. Column Y now exhibits the greatest difference; therefore, we allocate 17 units to *AY*, since it has the smallest distribution cost in column Y. Since

TABLE 16

THIRD *VAM* ASSIGNMENT

	$\cancel{8}$	0	↓ 2	$\cancel{6}$	2		
From \ To	*V*	*W*	*X*	*Y*	*Z*	Available 1000s ↓	
A	8 X	8	10 X	6 (17)	10	19	2
B	0 (11)	8	6 (7)	12 X	14	28	$\cancel{6}$ $\cancel{2}$ 6
C	12 X	10	8 X	14 X	12	25	2
Required 1000s →	11	13	7	17	24	72	

TABLE 17

FOURTH AND FIFTH *VAM* ASSIGNMENTS

	8	0	2	6	2		
From \ To	*V*	*W*	*X*	*Y*	*Z*	Available 1000s ↓	
A	[8] X	[8] 2	[10] X	[6] 17	[10] X	19	2 –
B	[0] 11	[8] 10	[6] 7	[12] X	[14] X	28	6 2 6 –
C	[12] X	[10]	[8] X	[14] X	[12]	25	2
Required 1000s →	11	13	7	17	24	72	

Y's requirements are completely satisfied, the other squares in that column are crossed out. Differences are recalculated. This entire step is shown in Table 15.

b. The recalculated differences now show five of the columns and rows with a difference of 2. The lowest cost square in any column or row is BX, which has a cost of 6. We assign 7 units to BX, which completely satisfies the requirements at X. Table 16 shows the allocation of 7 units at BX, the crossing out of the other squares in column X, and the recalculation of cost differences for the remaining rows and columns.

c. Row *B* now shows a cost difference of 6, and we allocate 10 units to the low cost square *BW*, as shown in Table 17. This completes row *B*. Recalculated cost differences now show that all remaining cost differences in rows and columns are 2. The lowest cost square available is *AW*, so we allocate 2 units there to complete row *A*. This step is also shown as part of Table 17.

d. The last two allocations at *CW* and *CZ* are made by inspection of the rim conditions (shown in Table 18). An evaluation of the open squares in Table 18 shows that this solution is optimal and is identical to the optimal solution shown in Table 6.

TABLE 18
FINAL ASSIGNMENTS AT CW AND CZ BALANCE
WITH RIM RESTRICTIONS AND YIELD VAM
INITIAL SOLUTION WHICH IS OPTIMAL

From \ To	V	W	X	Y	Z	Available 1000s ↓
A	8 X	8 2	10 X	6 17	10 X	19
B	0 11	8 10	6 7	12 X	14 X	28
C	12 X	10 1	8 X	14 X	12 24	25
Required 1000s →	11	13	7	17	24	72

Problems

1. A company has factories at *A*, *B*, and *C*, which supply warehouses at *D*, *E*, *F*, and *G*. Monthly factory capacities are 70, 90, and 115, respectively. Monthly warehouse requirements are 50, 60, 70, and 95, respectively. Unit shipping costs are as follows:

	To			
From	D	E	F	G
A	\$17	\$20	\$13	\$12
B	\$15	\$21	\$26	\$25
C	\$15	\$14	\$15	\$17

Determine the optimum distribution for this company to minimize shipping costs.

2. A company with factories at *A*, *B*, and *C* supplies warehouses at *D*, *E*, *F*, and *G*. Monthly factory capacities are 20, 30, and 45, respectively. Monthly warehouse requirements are 10, 15, 40, and 30, respectively. Unit shipping costs are as follows:

	To			
From	D	E	F	G
A	\$6	\$10	\$6	\$8
B	\$7	\$9	\$6	\$11
C	\$8	\$10	\$14	\$6

Determine the optimum distribution for this company to minimize shipping costs.

3. A company has factories at *A*, *B*, and *C*, which supply warehouses *D*, *E*, *F*, and *G*. Monthly factory capacities are 300, 400, and 500, respectively. Monthly warehouse requirements are 200, 240, 280, and 340, respectively. Unit shipping costs are as follows:

	To			
From	D	E	F	G
A	\$7	\$9	\$9	\$6
B	\$6	\$10	\$12	\$8
C	\$9	\$8	\$10	\$14

Determine the optimum distribution for this company to minimize shipping costs.

4. A company has factories at *A*, *B*, and *C* which supply warehouses at *D*, *E*, *F*, and *G*. Monthly factory capacities are 160, 150, and 190, respectively. Monthly warehouse requirements are 80, 90, 110, and 160, respectively. Unit shipping costs are as follows:

	To			
From	D	E	F	G
A	\$42	\$48	\$38	\$37
B	\$40	\$49	\$52	\$51
C	\$39	\$38	\$40	\$43

Determine the optimum distribution for this company to minimize shipping costs.

5. A company has factories at *A*, *B*, *C*, and *D*, which supply warehouses at *E*, *F*, *G*, *H*, and *I*. Monthly factory capacities are 200, 225, 175, and 350, respectively. Monthly warehouse requirements are 130, 110, 140, 260, and 180, respectively. Unit shipping costs are as follows:

	To				
From	E	F	G	H	I
A	$14	$19	$32	$9	$21
B	$15	$10	$18	$7	$11
C	$26	$12	$13	$18	$16
D	$11	$22	$14	$14	$18

Determine the optimum distribution for this company to minimize shipping costs. (Hint: Use VAM for initial solution.)

6. A company has factories at *A*, *B*, and *C*, which supply warehouses at *D*, *E*, *F*, and *G*. Monthly factory capacities are 250, 300, and 200, respectively for regular production. If overtime production is utilized, the capacities can be increased to 320, 380, and 210, respectively. Incremental unit overtime costs are $5, $6, and $8 per unit, respectively. The current warehouse requirements are 170, 190, 230, and 180, respectively. Unit shipping costs between the factories and warehouses are:

	To			
From	D	E	F	G
A	$8	$9	$10	$11
B	$6	$12	$9	$7
C	$4	$13	$3	$12

Determine the optimum distribution for this company to minimize costs.

7. A company with factories at *A*, *B*, *C*, and *D* supplies warehouses at *E*, *F*, *G*, and *H*. Monthly factory capacaties are 100, 80, 120, and 90, respectively, for regular production. If overtime production is utilized, the capacities can be increased to 120, 110, 160, and 140, respectively. Incremental unit overtime costs are $5, $2, $3, and $4, respectively. Present incremental profits per unit, excluding shipping costs, are $14, $9, $16, and $27, respectively, for

regular production. The current monthly warehouse requirements are 110, 70, 160, and 130, respectively. Unit shipping costs are as follows:

	To			
From	E	F	G	H
A	\$3	\$4	\$5	\$7
B	\$2	\$9	\$6	\$8
C	\$4	\$3	\$8	\$5
D	\$6	\$5	\$4	\$6

Determine the optimum distribution for this company. (Hint: This problem requires that you maximize profits. What simple change in the procedures makes it possible to maximize rather than minimize?)

References

Bierman, H.; C. P. Bonini; and W. H. Hausman. *Quantitative Analysis for Business Decisions*. 5th ed. Homewood, Ill.: Richard D. Irwin, 1977.

Buffa, E. S., and J. S. Dyer. *Management Science/Operations Research: Model Formulation and Solution Methods*. Santa Barbara, Ca.: Wiley/Hamilton, 1977.

Levin, R. L., and C. A. Kirkpatrick. *Quantitative Approaches to Management*. 3d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

McMillan, C. *Mathematical Programming*. 2nd ed. New York: John Wiley & Sons, 1975.

Reinfeld, N. V., and W. R. Vogel. *Mathematical Programming*. Englewood Cliffs, N. J.: Prentice-Hall, 1958.

Thierauf, R. J., and R. C. Klekamp. *Decision Making through Operations Research*. 2d ed. New York: John Wiley & Sons, 1975.

Appendix D

Waiting Lines

In discussing systems concepts in Chapter 3, we presented the input-transformation-output module as the basic component of productive systems. The concepts on which waiting line analyses are built use this basic module, focusing on the special characteristics of input and transformation that commonly are found in service systems. The special characteristics occur in the probabilistic nature of service systems in which the input (parts, clients, customers, etc.) arrive for service randomly, and the time for processing is variable. Therefore, predicting the output of service systems depends on the complex interplay between arrival times and service times. The concepts and techniques of waiting line analysis provide insight into how service systems function, as well as a framework by which to analyze and develop policies and procedures for the design and operation of service systems. Thus, we must understand waiting line theory in order to understand service systems.

WAITING LINE MODELS OF SERVICE SYSTEMS

The original work in waiting line theory was done by A. K. Erlang, a Danish telephone engineer. Erlang started his work in 1905 in an attempt to determine the effect of fluctuating demand (arrivals) on the utilization of automatic dial

TABLE 1
WAITING LINE MODEL ELEMENTS FOR SOME COMMONLY KNOWN SITUATIONS

	Unit Arriving	*Service or Processing Facility*	*Service or Process Being Performed*
Ships entering a port	Ships	Docks	Unloading and loading
Maintenance and repair of machines	Machine breaks down	Repair crew	Repair machine
Assembly line, not mechanically paced	Parts to be assembled	Individual assembly operations or entire line	Assembly
Doctor's office	Patients	Doctor, staff, and facilities	Medical care
Purchase of groceries at a supermarket	Customer with loaded grocery carts	Check-out counter	Tabulation of bill, receipt of payment and bagging of groceries
Auto traffic at an intersection or bridge	Automobiles	Intersection or bridge with control points such as traffic lights or toll booths	Passage through intersection or bridge
Inventory of items in a warehouse	Order for withdrawal	Warehouse	Replenishment of inventory
Job shop	Job order	Work center	Processing

equipment. Since the end of World War II, Erlang's work has been extended and applied to a variety of situations which, it now is recognized, are described by the general waiting line model. Table 1 shows the waiting line model elements for a number of commonly known situations.

Structure of Waiting Line Models

There are four basic waiting line structures that describe the general conditions at the service facility. The simplest structure, shown in Figure 1*a*, is our basic module and is called the single channel, single phase case. There are many illustrations in practice of the simple module: the cashier at a restaurant, any single "window" operation in a post office or bank, or a one-man barber shop. If the number of processing stations is increased but still draws on a single waiting line, we have the multiple channel, single phase case shown in Figure 1*b*.

A simple assembly line or a cafeteria line has a number of service facilities in tandem and is called the single channel, multiple phase case, shown in Figure 1c. Finally, the multiple channel, multiple phase case might be illustrated by two or more parallel assembly lines, as shown in Figure 1*d*. Combinations of any or all of the basic four structures also could exist in networks of queues.

The analytical methods for waiting lines divide into two main categories for

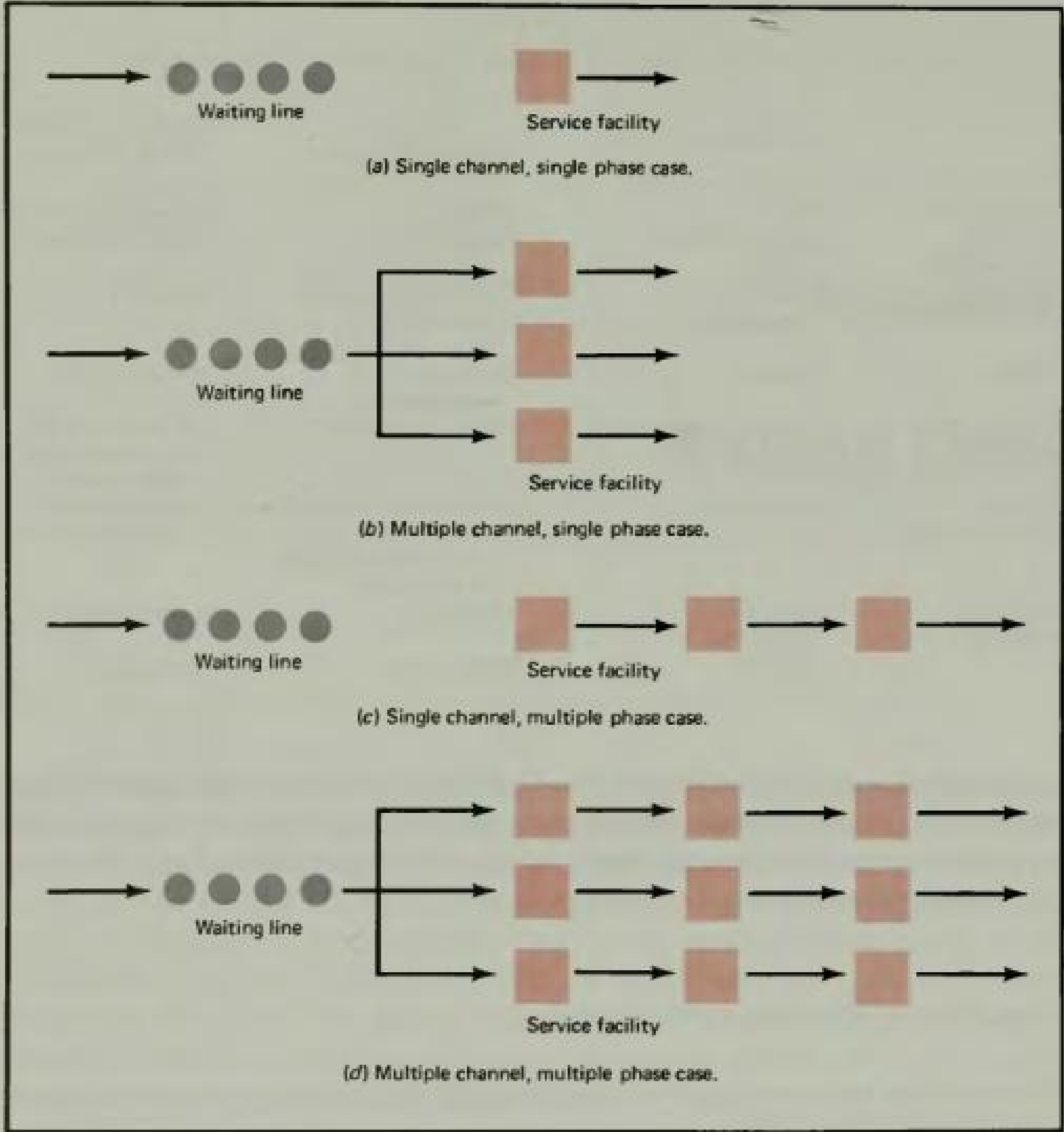


Figure 1. Four basic structures of waiting-line situations

any of the basic structures for Figure 1, depending on the size of the source population of the inputs. When the source population is very large—and in theory, at least, the length of the waiting line could grow without fixed limits —the applicable models are termed *infinite*. On the other hand, when the arriving unit comes from a source that has a fixed upper limit, the applicable models are termed *finite*. For example, if we are dealing with the maintenance of a bank of twenty machines, and if a machine breakdown represents an arrival, the maximum waiting line is twenty machines waiting for service, and a finite model is needed. If, on the other hand, we are operating an auto repair shop, the source population of breakdowns is very large, and an infinite model would be appropriate. We shall discuss both infinite and finite models.

There are other variations in waiting line structures that are important in certain applications. For ~~example, the~~ "queue discipline" implied in Figure 1 is first come–first served. Obviously, there are many other possibilities involving priority systems. For example, in an outpatient clinic, emergencies and patients with appointments are taken ahead of walk-in patients. In job shop scheduling systems, there has been a great deal of experimentation with alternate priority systems. Because of mathematical complexity, simulation has been the common mode of analysis for ~~systems involving queue disciplines~~ other than first come–~~first~~ served.

Finally, the nature of the distributions of arrival and service is a significant structural characteristic of waiting line models. Some mathematical analysis is available for distributions that follow the Poisson or Erlang process (with some variations) or when arrivals or service are constant. If distributions differ from these or are empirical, simulation is likely to be the practical mode of analysis.

We shall organize our discussion of waiting lines in terms of infinite and finite models.

INFINITE WAITING LINE MODELS

Since we shall not discuss all possibilities of infinite models, it will be useful to restrict our thinking to situations involving the first come–first served queue discipline and the Poisson distribution of arrivals. Initially we will deal with the single channel, single phase case (our basic service facility module), and later we will discuss the multiple channel case.

Poisson Arrivals

It has been demonstrated that the Poisson distribution function represents arrival rates in a large number of real world situations. It is a discrete function—i.e., it deals with whole units of arrivals—and negative values, and fractions of workers, products, or machines do not have meaning. The Poisson distribution function is given by:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where λ is the average arrival rate, and
x is any arrival rate.

For example, if $\lambda = 4$, then the probability of an arrival rate of $x = 6$ per hour is:

$$f(6) = \frac{4^6 e^{-4}}{6!} = \frac{4{,}096 \times 0.0183}{720} = 0.104$$

The Poisson distribution for an average arrival rate of $\lambda = 4$ (as well as other values of λ) is shown in Figure 2. The Poisson distribution typically is skewed to the right. The distribution is simple in that the standard deviation is expressed solely in terms of the mean, $\sigma_\lambda = \sqrt{\lambda}$.

The evidence is great that the Poisson distribution represents arrival rates in many applications. Many empirical studies have validated the Poisson arrival rate in general industrial operations, traffic flow, and various service operations. Although we cannot say that all arrival rate distributions are described adequately by the Poisson, we can say that it usually is worth checking to see whether they are. If so, then a fairly simple analysis may be possible. It is logical that arrivals may follow the Poisson distribution when many factors affect arrival time, since the Poisson distribution corresponds to completely random arrivals. This means that each arrival is independent of other arrivals, as well as of any condition of the waiting line. The practical question is whether or not the Poisson distribution is a reasonable approximation to reality.

Poisson Arrivals—Service Time Distribution Not Specified

Since Poisson arrivals are common, a useful model is one that depends on Poisson arrivals but accepts any service time distribution. We assume that the mean service rate is greater than the mean arrival rate, otherwise the system would be unstable and the waiting line would become infinitely large. The queue discipline is first come–first served, and arrivals wait for service (i.e., they neither fail to join the line nor leave it because it is too long). Under these conditions, the expected length of the waiting line is:

$$L_q = \frac{(\lambda\sigma)^2 + (\lambda/\mu)^2}{2(1 - \lambda/\mu)} \tag{2}$$

where L_q is the expected length of the waiting line
λ is the mean arrival rate from a Poisson distribution
μ is the mean service rate
σ is the standard deviation of the distribution of service times

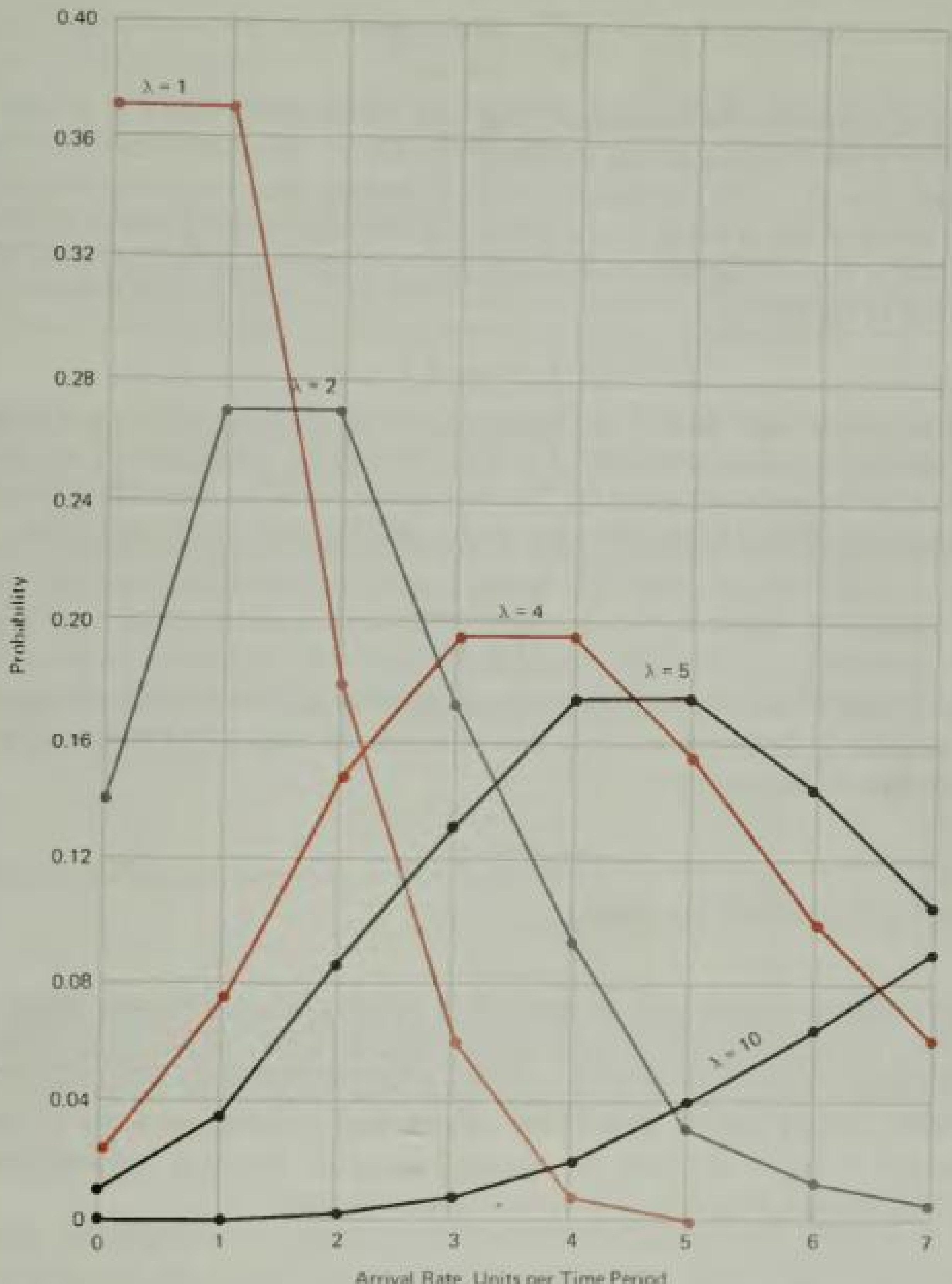


Figure 2. Poisson distributions for several mean arrival rates

Because the ratio λ/μ is itself a useful concept, since it represents the average utilization of the service facility, we define $\rho = \lambda/\mu$. In the general sense, the ratio λ/μ represents the flow intensity of the system. Substituting for ρ in Equation 2 leads to:

$$L_q = \frac{(\lambda\sigma)^2 + \rho^2}{2(1 + \rho)} \tag{3}$$

Since ρ represents the fraction of time that the service facility is in use, by analogy it also represents the expected number of individuals or units being served. Also, $(1 - \rho)$ is the fraction of service facility idle time, or the fraction of time when no one is being served. Since ρ is the number being served, the total number in the waiting line, plus the expected number being served, is the total number in the system, L:

$$L = L_q + \rho \tag{4}$$

Similar simple logic leads to the expected waiting time in line W_q, and time in the system, including service W. The reciprocal of the mean arrival rate is the mean time between arrivals $(1/\lambda)$. The multiplication of the mean time between arrivals and the line length gives the waiting time or the time in the system:

$$W_q = L_q/\lambda \tag{5}$$

$$W = L/\lambda \tag{6}$$

Equations 4, 5, and 6, are useful relationships. The general procedure would be to compute L_q from either 2 or 3, and compute the values of L, W_q, and W as needed given the value of L_q.

Service Time Distributions

While there is considerable evidence that arrival processes tend to follow the Poisson distributions, as we have indicated, service time distributions seem to be much more varied in their nature. This is why the previous model involving Poisson arrivals and an unspecified service time distribution is so valuable. Equation 2, let us compute the queuing statistics, knowing only the mean service rate and the standard deviation of service time.

The negative exponential distribution has been one of the prominent models for service time, and in some instances the assumption has been shown to be valid. Nelson's study [1959] of distributions of arrivals and service times in a Los Angeles jobbing machine shop did *not* indicate that the exponential model fit the actual service time distributions adequately for all the machine centers. The machine centers in which service time distributions seemed adequately described by the negative exponential were turret lathes, milling machines, boring machines, and the wheelabrater. In general, the instances in which the exponential hypothesis was reasonable occurred in the machine centers with the largest mean processing time.

Nelson's study indicated that other mathematical distributions, such as the hyperexponential and the hyperErlang, described the actual distributions better.

Other evidence indicates that, in some cases, the negative exponential distribution fits. For example, the service time at a tool crib was nearly exponentially distributed [Brigham, 1955]. The distribution of local telephone calls that were not made from a pay station also has been shown to be exponential.

Model for Poisson Input and Negative Exponential Service Times

Since the negative exponential dsitribution is described completely by its mean value, the standard deviation being equal to the mean, we can describe this model as a special case of Equation 2. The mean of a negative exponential distribution is the reciprocal of the mean service rate; that is, $1/\mu$. Therefore, $\sigma = 1/\mu$, and if we substitute this value for the standard deviation in Equation 2 and simplify, we obtain:

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} \tag{7}$$

Also, the probability of n units in the system is:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \tag{8}$$

The other relationships between L_q, L, W_q, and W, expressed by Equations 4, 5, and 6, hold for the negative exponential service time distribution as well as for the distribution free case. For the sake of simplicity, many individuals prefer to use Equation 2 or 3, using the appropriate value of σ to reflect the special case.

Constant Service Times. While constant service times are not common in actual practice, it is a reasonable assumption in cases where a machine processes arriving items by a fixed time cycle. Also, constant service times represent a boundary, or lower limit, on the value of σ in Equation 2. Therefore, constant service time is a special case of Equation 2, which can be reduced to simpler form by inserting $\sigma = 0$. The resulting equation is:

$$L_q = \frac{\lambda^2}{2\mu(\mu - \lambda)} \tag{9}$$

The relationships between L_q, L, W_q, and W expressed by Equations 4, 5, and 6 hold for the constant service time distribution as well as for the distribution free case.

For comparison, assume a situation in which $\lambda = 8$ per hour, $\mu = 12$ per hour. Then the comparison of queue statistics when $s = 0.1$ hours and $s = 0$ is:

	$s = 0.1$ hours	$s = 0$ hours
L_q	1.63	0.67
L	2.30	1.34
W_q	12.23 min.	5.0 min.
W	17.25 min.	10.0 min.

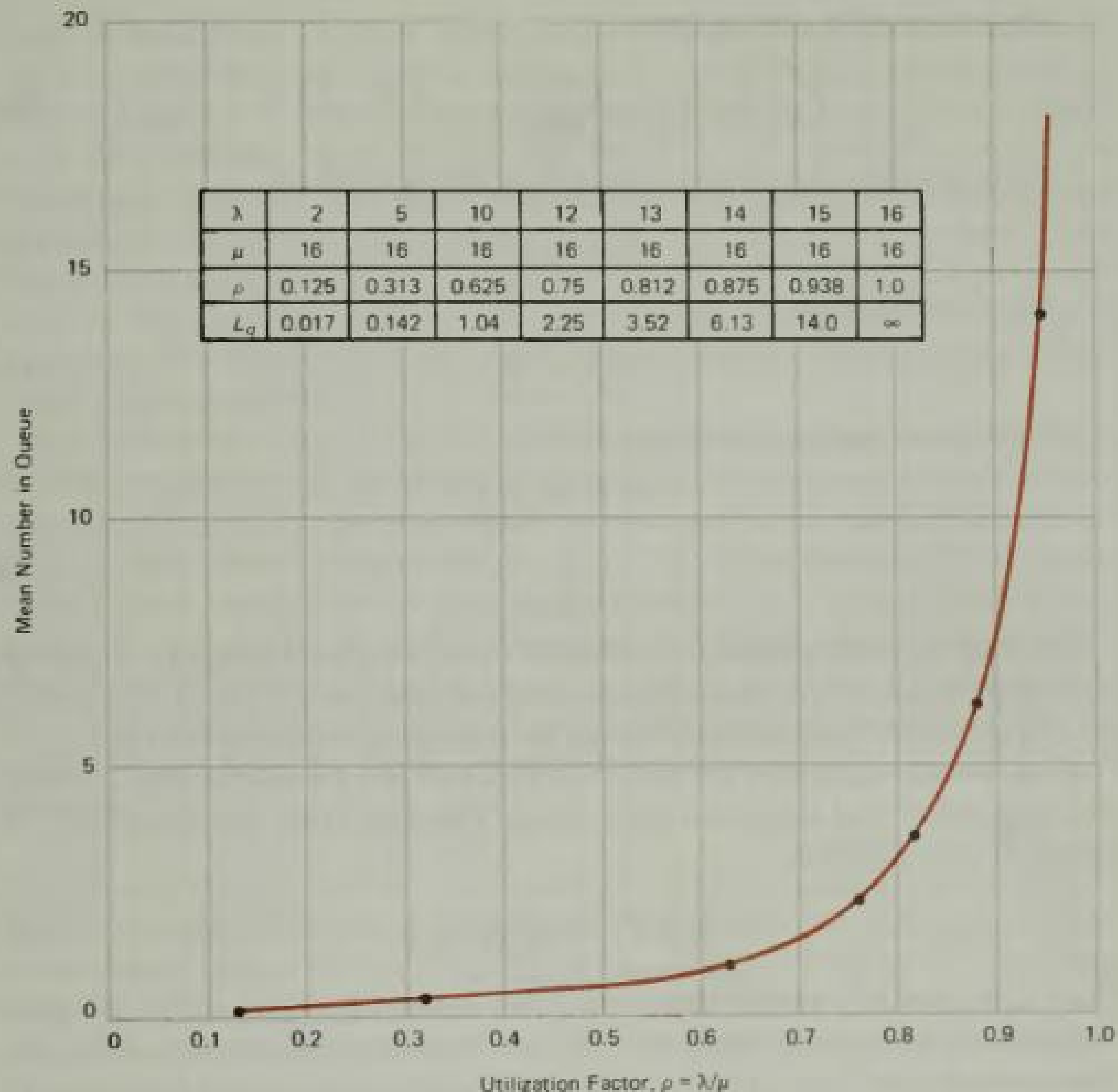
Obviously, constant service times reduce the queuing effects considerably.

Relationship of Q Length to Utilization. Recall that the ratio of $\lambda/\mu = \rho$ represents the service facility utilization. If $\lambda = \mu$, then $\rho = 1$, and theoretically the service facility is used 100 percent of the time. But let us see what happens to the length of the queue as ρ varies from zero to one. Figure 3 summarizes the result for Poisson input and exponential service times. As ρ approaches unity, the number waiting in line increases rapidly and approaches infinity. We can see the truth of this statement by examining Equations 2, 3, 7, and 9 for L_q. In all cases, the denominator goes to zero as ρ approaches unity and the value of L_q becomes infinitely large.

We see now that one of the requirements of any practical system is that $\mu > \lambda$; otherwise we cannot have a stable system. If units are arriving faster on the average than they can be processed, the waiting line and waiting time will increase continuously, and no steady state will be able to be achieved. This simple fact also indicates that there is a value to idle time in the service facility. We must trade off service facility time against the need to give good service.

Multiple Channel, Single Phase Case

In the multiple channel case, we assume the conditions of Poisson arrivals, exponential service times, and first come–first served queue discipline. The effective service rate $M\mu$ must be greater than the arrival rate λ, where M is the number of channels. The facility utilization factor now becomes $\rho = \lambda/M\mu$. First, we must calculate P_0, the probability that there are zero units in the system (service facility idle), since the basic formulas all involve P_0 in their simplest forms. Substituting $r = \lambda/\mu$ for flow intensity:



λ	2	5	10	12	13	14	15	16
μ	16	16	16	16	16	16	16	16
ρ	0.125	0.313	0.625	0.75	0.812	0.875	0.938	1.0
L_q	0.017	0.142	1.04	2.25	3.52	6.13	14.0	∞

Figure 3. Relationship of queue length to the utilization factor p.

$$P_0 = \frac{1}{\left[\sum_{n=0}^{M-1} \frac{r^n}{n!}\right] + \left[\frac{r^M}{M!\left(1 - \frac{r}{M}\right)}\right]} \tag{10}$$

where n is an index for the number of channels, the calculation of the term $\sum_{n=0}^{M-1} \frac{r^n}{n!}$ being the sum of $\frac{r^n}{n!}$ for all the numbers of channels ranging from $n = 0$ to $n = M - 1$.

Then the formulas parallel to the single channel case are:

Mean number in waiting line:

$$L_q = \frac{(r)^{M+1}}{(M-1)!(M-r)^2} \cdot P_0 \tag{11}$$

Mean number in system, including those being serviced:

$$L = L_q + r \tag{12}$$

Mean waiting time:

$$W_q = \frac{L_q}{\lambda} \tag{13}$$

Mean time in system, including service:

$$W = W_q + \frac{1}{\mu}$$

$$= \frac{L}{\lambda} \tag{14}$$

Although the relationships are somewhat more complex, especially for a large number of channels, we can see that we need calculate only L_q, and L, W_q, and W may be calculated quite simply through the interrelationships that exist.

Wagner [1975] calculates the probability that all servers or channels are busy (the probability that there will be a delay), $P(Busy)$. Then, the value of L_q is simple to compute from:

$$L_q = P(Busy)\left(\frac{r}{M-r}\right) \tag{15}$$

Then L, W_q, and W are easily computed from Equations 12, 13, and 14. We have computed L_q for various values of M, the number of service channels, and r, the flow intensity, in Table 3 of Appendix E. Of course, $P(Busy)$ can be computed from Equation 15, using the table value of L_q.

As an example, assume that we expand facilities, adding a second service channel. We want to determine the effect on average waiting time, using the basic data: $\lambda = 8$ per hour, $\mu = 12$ per hour, and now $M = 2$. From Table 3 of Appendix E, for $M = 2$ and $r = \lambda/\mu = 8/12 = 0.67$, we find by interpolating that $L_q = 0.085$ units in line. Then $W_q = L_q/\lambda = 0.085/8 = 0.0106$ hours or 0.064 minutes. Compare these results with the single channel solution for exponential service time of $W_q = 10.0$ minutes. Obviously, adding the second channel eliminates the waiting problem. Overall utilization of the facilities declines from $\rho = \lambda/\mu = 0.067$ to $\rho = \lambda/M\mu = 0.34$.

The *Effect of Increasing the Numbers of Channels*. To compare the various effects of increasing the number of channels, we can refer to Table 3 of Appendix E. If, for example, we were faced with a situation in which the flow intensity was $r = 0.9$ for the single channel case, then L_q would be approximately 8 from Table 3, Appendix E. Adding a second channel reduces the average line length to $L_q = 0.23$. Adding a third channel reduces it to $L_q = 0.03$. The effects are surprisingly large; we can obtain disproportionate gains in line length, and waiting time, by increasing the number of channels. We can see intuitively from Figure 3 that this might be true, since queue length (and waiting time) begins to increase very rapidly at about $\rho = 0.8$. A rather small increase in the capacity of the system (decrease in ρ) at these high loads can produce a large decrease in line length and waiting time.

It is interesting to see whether doubling the capacity within the same service facility is equivalent to doubling the capacity by adding parallel service channels. As a basis for comparison, assume that $r = 0.8$ for the single channel case ($M = 1$). From Table 3 of Appendix E, $L_q = 3.2$. If we double capacity within the single service channel, service rate doubles and $r = 0.4$. From Table 3 of Appendix E, $L_q = 0.2666$. If, on the other hand, we double capacity by adding a second parallel channel, then for $M = 2$, $r = 0.8$, $L_q = 0.1533$. The reduction in line length (and waiting time) is 91.7 percent by doubling capacity within the single channel; however, it is 95.2 percent by doubling capacity through parallel channels.

The Effect of Pooling Facilities. A large manufacturing concern with a 100 acre plant has a well established medical facility, which is located at the plant offices at the eastern edge of the property. The plant has grown from east to west over the years, and currently, travel time to the medical facility is so great that management is considering dividing the facility. The second unit is to be established near the center of the west end of the plant. A study was made of weighted travel times for the present single facility and for the proposed two-facility system. The result indicates that average travel time for the present large medical facility is 15 minutes. The volume averages 1,000 visits per week, or 250 man-hours for travel time. The two-facility plan would reduce the average travel time to 8 minutes, or 133 man-hours per week.

The question now is: What will happen to waiting time in the waiting rooms? For the one large facility, $\lambda = 25$ per hour, and average service time is 20 minutes, or $\mu = 3$ per hour. There are 10 physicians who handle the load. From Table 3 of Appendix E, $L_q = 2.45$, $W_q = 2.45 \times 60/25 = 5.88$ minutes per person, or 98 man-hours per week. Therefore, travel plus waiting time is $250 + 98 = 348$ man-hours per week.

The plan is to divide the medical staff for the two facilities, and it is assumed that the load will divide equally. Thus, comparable data are: $\lambda = 12.5$ per hour per facility, $\mu = 3$ per hour, $M = 5$, and $r = 4.2$. From Table 3 of Appendix E, $L_q = 3.3269$, and $W_q = 16.0$ minutes per person or 267 man-hours per week. The travel plus waiting time for the dual facility plan is, therefore, $133 + 267 = 400$ man-hours per week, compared with 348 for the single large facility. Other alternatives could be computed, probably involving an increased medical staff.

The waiting time for the single large facility is 5.88 minutes per person, compared with 16.0 minutes per person for the two-facility plan. Obviously, one large facility can give better service than the two smaller facilities. If we visualize the two decentralized facilities functioning side by side, we can see intuitively why waiting time increases. If Facility 1 were busy and had patients waiting while at the same time Facility 2 happened to be idle, someone from the Facility 1 waiting room could be serviced immediately by Facility 2, thereby reducing average waiting time. In this situation, the two facilities are drawing from one waiting line. When they are physically decentralized, the facilities must draw on two independent waiting lines, and the idle capacity of one cannot be used by the waiting patients of the other.

Costs and Capacity in Waiting Line Models

While many decisions concerning service systems may hinge on physical factors of line length, waiting time, and service facility utilization, very often system designs will depend on comparative costs for alternatives. The costs involved are commonly the costs of providing the service versus the waiting time costs. In some instances, the waiting time costs are objective, as when the enterprise employs both the servers and those waiting. The company medical facility that we just discussed constitutes such a case. The company absorbed all the travel time and waiting time costs, as well as the cost of providing the service. In such an instance, a direct cost minimizing approach can be taken, balancing the waiting costs or the time-in-system costs against the costs of providing the service.

When the arriving units are customers, clients, or patients, the costs of making them wait is less obvious. If they are customers, excessive waiting may cause irritation, loss of goodwill, and, eventually, sales. However, goodwill cannot be gauged directly. In public service operations and other monopoly situations, the valuation of waiting cost may be even more tenuous because the individual cannot make alternative choices. In all situations in which objective costs cannot be balanced, we may have to set a standard for waiting time—for example, to adjust capacity to keep average waiting time at or below some stated number of minutes.

FINITE WAITING LINE MODELS

Many practical waiting line problems that occur in productive systems have the characteristics of finite waiting line models. This is true whenever the population of machines, workers, or items that may arrive for service is limited to a relatively small finite number. The result is that we must express arrivals as a unit of the population rather than as an average rate. In the infinite waiting line case, the average length of the waiting line is effectively independent of the number in the arriving population; however, in the finite case, the number in the queue may represent a significant proportion of the arriving population, thereby affecting the probabilities associated with arrivals.

The computations for the resulting mathematical formulations are somewhat more difficult than for the infinite queue case. Fortunately, however, *Finite Queuing Tables* (Peck and Hazelwood, 1958) are available that make problem solving very simple. Although there is no definite number that we can point to as a dividing line between finite and infinite applications, the finite queuing tables have data for populations from 4 up to 250, and these may be taken as a general guide. We have reproduced these tables for populations of 5, 10, 20, and 30 in Appendix E, Table 4, to illustrate their use in the solution of finite queuing problems. The tables are based on a finite model for exponential times between arrivals and service times, and on a first come–first served queue discipline.

Use of the *Finite Queuing Tables*

The tables are indexed first by N, the size of the population. For each population size, data are classified by X, the service factor (comparable to the utilization factor in infinite queues), and by M, the number of parallel channels. For a given N, X, and M, three factors are listed in the tables: D (the probability of a delay; that is, if a unit calls for service, the probability that it will have to wait), F (an efficiency factor, used to calculate other important data), and the mean number in the waiting line, L_q. To summarize, we define the factors just expressed, plus those that may be calculated, as follows:

$N =$ population (number of machines, customers, etc.)

$\mu =$ mean service rate

$\lambda =$ mean arrival rate *per population unit*

$X =$ service factor $= \dfrac{\lambda}{\lambda + \mu}$

$M =$ number of service channels

$D =$ probability of delay (probability that if a unit calls for service, it will have to wait)

F = efficiency factor
L_q = mean number in waiting line $= N(1 - F)$
W_q = mean waiting time $= \frac{1}{\mu X}\left(\frac{1-F}{F}\right)$
H = mean number of units being serviced $= FNX = L - L_q$
J = mean number of units running $= NF(1 - X)$
$M - H$ = average number of servers idle

The procedure for a given case is as follows:

1. Determine the mean service rate μ, and the mean arrival rate λ, based on data or measurements of the system being analyzed.
2. Compute the service factor $X = \lambda/(\lambda + \mu)$.
3. Locate the section of the tables listing data for the population size N.
4. Locate the service factor calculated in Step 2 for the given population.
5. Read the values of D, F, and L_q for the number of channels M, interpolating between values of X when necessary.
6. Compute values for W_q, H, and J as required by the nature of the problem.

Example. A hospital ward has thirty beds in one section, and the problem centers on the appropriate level of nursing care. The hospital management believes that patients should have immediate response to a call at least 80 percent of the time because of possible emergencies. Calls from patients follow a Poisson distribution and average almost 19 per hour for the thirty patients. The service time is approximated by a negative exponential distribution, and mean service time is five minutes.

The hospital manager wishes to staff the ward to give service so that 80 percent of the time there will be no delay. Nurses are paid \$5 per hour,and one concern involves the cost of idle time at this level of service. Finally, as a base of comparison, the manager wishes to know how much more patients will have to pay for the 80 percent criterion compared with a 50 percent service level for immediate response, which is the current policy.

The solutions to the problems posed by the hospital manager are developed through finite queuing models. A finite model is required because the maximum possible queue is thirty patients waiting for nursing care.

In terms of the finite queuing model for this situation, the mean service time is $T = 5$ minutes ($\mu = 0.2$ per minute, or 12 per hour), the mean time between calls is $U = 95$ minutes per patient ($\lambda = 0.0105$ per minute, or 0.632 per hour), and therefore the service factor is $X = \lambda/(\lambda + \mu) = 0.632/12.632 = 0.05$. Scanning the *Finite Queing Tables* (Table 4 of Appendix E) under Population $N = 30$, and $X = 0.05$, we seek data for the probability of a delay of $D = 0.20$, since we wish to establish service so that there will be no delay 80 percent of the time. The closest we can come to providing this level of service is with $M = 3$ nurses, and with corresponding data from Table 4 of Appendix E of $D = 0.208$, $F = 0.994$, and $L_q = 0.18$. Note that we must select an integer number of service channels (nurses).

The cost of this level of service is the cost of employing three nurses or $5 \times 3 = \$15$ per hour, or \$360 per day, assuming day and night care. The average number of calls waiting to be serviced will be $L_q = 0.18$, and the mean waiting time will be:

$$W_q = \frac{1}{\mu X}\left(\frac{1-F}{F}\right) = \frac{1}{0.2 \times 0.05}\left(\frac{1-0.994}{0.994}\right) = 0.6 \text{ minutes}$$

The waiting time due to queuing effects is negligible, as intended.

The average number of patients being served will be $H = FNX = 0.994 \times 30 \times 0.05 = 1.49$, and the average number of nurses idle will be $3 - 1.49 = 1.51$. The equivalent value of this idleness is $1.51 \times 5 \times 24 = \$181.20$ per day.

Finally, the number of nurses needed to provide immediate service 50 percent of the time is $M = 2$ from Appendix E, Table 4 ($D = 0.571$, $F = 0.963$, and $L_q = 1.11$). The average waiting time under this policy is $W_q = 3.84$ minutes. The average cost to patients of having the one additional nurse to provide the higher level of service is \$5 per hour or \$120 per day. Divided among thirty patients, the cost is \$4 per patient per day.

Review Problems

1. Given a Poisson distribution of arrivals with mean of $\lambda = 5$ per hour, what is the probability of an arrival of $x = 4$ per hour? What is the probability of the occurrence of fifteen minutes between arrivals?

2. The barber of a one-man shop finds that sometimes customers are waiting, but sometimes he has nothing to do and therefore can read sports magazines. He would rather keep a steady pace when he is working, in hopes of getting blocks of time for reading and keeping up on sports. Hoping to improve his situation, he keeps records for several weeks and finds that an average of one customer per hour comes in for haircuts (Poisson distribution). It takes him an average of twenty minutes per haircut, and the standard deviation of his sample of service times is five minutes.

a. What is the average number of customers waiting for service?
b. What is the average customer waiting time?
c. How much time does the barber have in which to read sports magazines?
d. The other barber in town has fallen ill, and our barber's business increases to an average of two customers per hour. What happens to the average number of customers waiting, the average waiting time, and the time available for reading sports magazines?
e. After practicing at home on his children, the barber finds that he can reduce both the hair cutting time to ten minutes and the variance to virtually zero by using a bowl and only electric clippers (no scissors). Will this solve the waiting time problem during the period in which the other barber is ill? How would it affect the barber's available reading time?
f. With the other barber ill, our barber finds that his customers are screaming for better service, and that his reading time is available only in small increments. He feels under great pressure, but is afraid to try the technological improvements that he has developed. How can he solve his problem?

3. A taxi cab company has four cabs that operate out of a given taxi stand. Customer arrival rates and service rates are described by the Poisson distribution. The average arrival rate is ten per hour, and the average service time is twenty minutes. The service time follows a negative exponential distribution.
   a. Calculate the utilization factor.
   b. From Table 3 of Appendix E, determine the mean number of customers waiting.
   c. Determine the mean number of customers in the system.
   d. Calculate the mean waiting time.
   e. Calculate the mean time in the system.
   f. What would be the utilization factor if the number of taxi cabs were increased from four to five?
   g. What would be the effect of the proposed change in f on the mean number in the waiting line?
   h. What would be the effect of reducing the number of taxi cabs from four to three?

4. A stenographer has five persons for whom she performs stenographic

services. Arrival rates are represented adequately by the Poisson distribution, and service times follow the negative exponential distribution. The arrival rate is five jobs per hour. The average service time is ten minutes.

a. Calculate the mean number in the waiting line.
b. Calculate the mean waiting time.
c. Calculate the mean number of units being serviced.
d. What is the probability that an individual bringing work to the stenographer will have to wait?

5. Trucks arrive at the truck dock of a wholesale grocer at the rate of eight per hour, and the distribution of arrivals is described by the Poisson. The loading and/or unloading time averages five minutes, and the estimate of the standard deviation of service time is $s = 6$ minutes. Truckers are complaining that they must spend more time waiting than unloading. Compute the queue statistics. Do they support the truckers' claim?

6. The wholesale grocer in Problem 5 knows that he probably could solve the problem by expanding the truck dock so that two trucks could be handled simultaneously. However, since this would require a large capital expenditure and disruption of operations during construction, he notes instead the very large standard deviation of service time. On investigation, he finds that some orders involve uncommon items that are not stored in a systematic manner, resulting in a great deal of search time. The grocer revamps the storage system so that all items can be located easily. This reduces the standard deviation to three minutes. What effect does the smaller standard deviation have on average trucker waiting time? How much would the standard deviation have to be reduced in order to cut average waiting time to five minutes? How do these results compare with the results that would be obtained by doubling the size of the truck dock?

7. Consider the large manufacturing concern that has an internal medical facility, discussed under "the effect of pooling facilities" in the text of the appendix. Using the basic data given there, recall that there were ten physicians whom we will assume are paid $3,000 per month, or about $6,928 per week for the ten physicians (medical malpractice insurance premiums are covered by the company). Let us also assume that the average hourly wage of the employees coming to the medical facility is $5 per hour. Computations for the single central facility yield an average

travel time of 250 $\times$ 5 = \$1,250 per week, and a waiting time cost of 98 $\times$ 5 = \$490 per week. The total weekly cost is \$8,668, including physicians' salaries.

a. With the central facility only, how many physicians on the staff will minimize affected costs?
b. Under the dual facility concept, the travel, waiting, and physicians' costs are \$5 (133 + 267) + 6,928 = \$8,928 per week. For the central facility, the comparable figure is \$8,668 per week. Would increased capacity in either or both of the dual facilities improve affected costs?

8. The manager of a large bank has the problem of providing teller service for customer demand, which varies somewhat during the business day from 10 A.M. to 4 P.M. He has a total capacity of six windows and can assign unneeded tellers to other useful work. He also wishes to give excellent service, which he defines in terms of customer waiting time as less than or equal to one minute. To give the best service for any situation, he has arranged the layout so that customers form one waiting line, from which the customer at the head of the line goes to the first available teller. The arrival pattern is as follows:

   10:00 A.M.–11:30 A.M. = 1.8 customers per minute
   11:30 A.M.– 1:30 P.M. = 4.8 customers per minute
   1:30 P.M.– 3:00 P.M. = 3.8 customers per minute
   3:00 P.M.– 4:00 P.M. = 4.6 customers per minute

   The arrivals follow a Poisson process; i.e., the mean value of arrivals varies but is always from a Poisson distribution. The average service time is one minute, and the distribution of service times follows the negative exponential.

   In order to maintain the service standard, how many tellers' windows must be open during each of the time periods?

References

Brigham, F. "On a Congestion Problem in an Aircraft Factory." *Operations Research* 3(1955):412–428.

Buffa, E. S. *Operations Management: Problems and Models*. 3d ed. New York: John Wiley & Sons, 1972. Chapters 14 and 15.

Cox. D. R., and W. L. Smith. *Queues*. New York: John Wiley & Sons, 1961.

Nelson, R. T. "An Empirical Study of Arrival, Service Time, and Waiting Time Distributions of a Job Shop Production Process." *Research Report no. 60. Management Sciences Research Project*, University of California–Los Angeles, 1959.

Panico, J. A. *Queuing Theory*. Englewood Cliffs, N. J.: Prentice-Hall, 1969.

Peck, L. G., and R. N. Hazelwood. *Finite Queuing Tables*. New York: John Wiley & Sons, 1958.

Prabhu, N. U. *Queues and Inventories*. New York: John Wiley & Sons, 1965.

Saaty, T. L. *Elements of Queuing Theory*. New York: McGraw-Hill, 1961.

______. "Resumé of Useful Formulas in Queuing Theory." *Operations Research* 5(April 1957):161–200.

Takács, L. *Introduction to the Theory of Queues*. New York: Oxford University Press, 1962.

Thierauf, R. J., and R. C. Klekamp. *Decision Making Through Operations Research*. 2d ed. New York: John Wiley & Sons, 1975.

Wagner, H. M. *Principles of Operations Research*. 2d ed. Englewood Cliffs, N. J.: Prentice-Hall, 1975.

Appendix E—

Tables

TABLE 1
PV_{sp}, PRESENT-VALUE FACTORS FOR FUTURE SINGLE PAYMENTS

Years Hence	1%	2%	4%	6%	8%	10%	12%	14%	15%	16%	18%	20%
1	0.990	0.980	0.962	0.943	0.926	0.909	0.893	0.877	0.870	0.862	0.847	0.833
2	0.980	0.961	0.925	0.890	0.857	0.826	0.797	0.769	0.756	0.743	0.718	0.694
3	0.971	0.942	0.889	0.840	0.794	0.751	0.712	0.675	0.658	0.641	0.609	0.579
4	0.691	0.924	0.855	0.792	0.735	0.683	0.636	0.592	0.572	0.552	0.516	0.482
5	0.951	0.906	0.822	0.747	0.681	0.621	0.567	0.519	0.497	0.476	0.437	0.402
6	0.942	0.888	0.790	0.705	0.630	0.564	0.507	0.456	0.432	0.410	0.370	0.335
7	0.933	0.871	0.760	0.665	0.583	0.513	0.452	0.400	0.376	0.354	0.314	0.279
8	0.923	0.853	0.731	0.627	0.540	0.467	0.404	0.351	0.327	0.305	0.266	0.233
9	0.914	0.837	0.703	0.592	0.500	0.424	0.361	0.308	0.284	0.263	0.225	0.194
10	0.905	0.820	0.676	0.558	0.463	0.386	0.322	0.270	0.247	0.227	0.191	0.162
11	0.896	0.804	0.650	0.527	0.429	0.350	0.287	0.237	0.215	0.195	0.162	0.135
12	0.887	0.788	0.625	0.497	0.397	0.319	0.257	0.208	0.187	0.168	0.137	0.112
13	0.879	0.773	0.601	0.469	0.368	0.290	0.229	0.182	0.163	0.145	0.116	0.093
14	0.870	0.758	0.577	0.442	0.340	0.263	0.205	0.160	0.141	0.125	0.099	0.078
15	0.861	0.743	0.555	0.417	0.315	0.239	0.183	0.140	0.123	0.108	0.084	0.065
16	0.853	0.728	0.534	0.394	0.292	0.218	0.163	0.123	0.107	0.093	0.071	0.054
17	0.844	0.714	0.513	0.371	0.270	0.198	0.146	0.108	0.093	0.080	0.060	0.045
18	0.836	0.700	0.494	0.350	0.250	0.180	0.130	0.095	0.081	0.069	0.051	0.038
19	0.828	0.686	0.475	0.331	0.232	0.164	0.116	0.083	0.070	0.060	0.043	0.031
20	0.820	0.673	0.456	0.312	0.215	0.149	0.104	0.073	0.061	0.051	0.037	0.026

TABLE 2
PV_a, PRESENT-VALUE FACTORS FOR ANNUITIES

Years (n)	1%	2%	4%	6%	8%	10%	12%	14%	15%	16%	18%	20%
1	0.990	0.980	0.962	0.943	0.926	0.909	0.893	0.877	0.870	0.862	0.847	0.833
2	1.970	1.942	1.886	1.833	1.783	1.736	1.690	1.647	1.626	1.605	1.566	1.528
3	2.941	2.884	2.775	2.673	2.577	2.487	2.402	2.322	2.283	2.246	2.174	2.106
4	3.902	3.808	3.630	3.465	3.312	3.170	3.037	2.914	2.855	2.798	2.690	2.589
5	4.853	4.713	4.452	4.212	3.993	3.791	3.605	3.433	3.352	3.274	3.127	2.991
6	5.795	5.601	5.242	4.917	4.623	4.355	4.111	3.889	3.784	3.685	3.498	3.326
7	6.728	6.472	6.002	5.582	5.206	4.868	4.564	4.288	4.160	4.039	3.812	3.605
8	7.652	7.325	6.733	6.210	5.747	5.335	4.968	4.639	4.487	4.344	4.078	3.837
9	8.566	8.162	7.435	6.802	6.247	5.759	5.328	4.946	4.772	4.607	4.303	4.031
10	9.471	8.983	8.111	7.360	6.710	6.145	5.650	5.216	5.019	4.833	4.494	4.192
11	10.368	9.787	8.760	7.887	7.139	6.495	5.988	5.453	5.234	5.029	4.656	4.327
12	11.255	10.575	9.385	8.384	7.536	6.814	6.194	5.660	5.421	5.197	4.793	4.439
13	12.134	11.343	9.986	8.853	7.904	7.103	6.424	5.842	5.583	5.342	4.910	4.533
14	13.004	12.106	10.563	9.295	8.244	7.367	6.628	6.002	5.724	5.468	5.008	4.611
15	13.865	12.849	11.118	9.712	8.559	7.606	6.811	6.142	5.847	5.575	5.092	4.675
16	14.718	13.578	11.652	10.106	8.851	7.824	6.974	6.265	5.954	5.669	5.162	4.730
17	15.562	14.292	12.166	10.477	9.122	8.022	7.120	6.373	6.047	5.749	5.222	4.775
18	16.398	14.992	12.659	10.828	9.372	8.201	7.250	6.467	6.128	5.818	5.273	4.812
19	17.226	15.678	13.134	11.158	9.604	8.365	7.366	6.550	6.198	5.877	5.316	4.844
20	18.046	16.351	13.590	11.470	9.818	8.514	7.469	6.623	6.259	5.929	5.353	4.870

TABLE 3
VALUES OF L_q FOR $M = 1 - 15$, AND VARIOUS VALUES OF $r = \lambda/\mu$.
POISSON ARRIVALS, NEGATIVE EXPONENTIAL SERVICE TIMES

	Number of Service Channels: M														
r	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.10	0.0111														
0.15	0.0264	0.0008													
0.20	0.0500	0.0020													
0.25	0.0833	0.0039													
0.30	0.1285	0.0069													
0.35	0.1884	0.0110													
0.40	0.2666	0.0166													
0.45	0.3681	0.0239	0.0019												
0.50	0.5000	0.0333	0.0030												
0.55	0.6722	0.0449	0.0043												
0.60	0.9000	0.0593	0.0061												
0.65	1.2071	0.0767	0.0084												
0.70	1.6333	0.0976	0.0112												
0.75	2.2500	0.1227	0.0147												
0.80	3.2000	0.1523	0.0189												
0.85	4.8166	0.1873	0.0239	0.0031											
0.90	8.1000	0.2285	0.0300	0.0041											
0.95	18.0500	0.2767	0.0371	0.0053											
1.0		0.3333	0.0454	0.0067											
1.2		0.6748	0.0904	0.0158											
1.4		1.3449	0.1778	0.0324	0.0059										
1.6		2.8444	0.3128	0.0604	0.0121										
1.8		7.6734	0.5320	0.1051	0.0227	0.0047									
2.0			0.8888	0.1739	0.0398	0.0090									
2.2			1.4907	0.2770	0.0659	0.0158									
2.4			2.1261	0.4305	0.1047	0.0266	0.0065								
2.6			4.9322	0.6581	0.1609	0.0426	0.0110								
2.8			12.2724	1.0000	0.2411	0.0659	0.0180								
3.0				1.5282	0.3541	0.0991	0.0282	0.0077							
3.2				2.3856	0.5128	0.1452	0.0427	0.0122							
3.4				3.9060	0.7365	0.2065	0.0631	0.0189							
3.6				7.0893	1.0550	0.2947	0.0912	0.0283	0.0084						
3.8				16.9366	1.5184	0.4114	0.1292	0.0412	0.0127						

TABLE 3 CONTINUED

VALUES OF L_q FOR $M = 1 - 15$, AND VARIOUS VALUES OF $r = \lambda/\mu$.
POISSON ARRIVALS, NEGATIVE EXPONENTIAL SERVICE TIMES

	Number of Service Channels M														
r	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4.0					2.2164	0.5694	0.1801	0.0590	0.0189						
4.2					3.3269	0.7837	0.2475	0.0827	0.0273	0.0087					
4.4					5.2675	1.0777	0.3364	0.1142	0.0389	0.0128					
4.6					9.2885	1.4867	0.4532	0.1555	0.0541	0.0184					
4.8					21.6384	2.0708	0.6071	0.2092	0.0742	0.0260					
5.0						2.9375	0.8102	0.2786	0.1006	0.0381	0.0125				
5.2						4.3004	1.0804	0.3680	0.1345	0.0492	0.0175				
5.4						6.6609	1.4441	0.5871	0.1779	0.0663	0.0243	0.0085			
5.6						11.5178	1.9436	0.6313	0.2330	0.0883	0.0330	0.0119			
5.8						26.3726	2.6481	0.8225	0.3032	0.1164	0.0443	0.0164			
6.0							3.6828	1.0707	0.3918	0.1518	0.0590	0.0224			
6.2							5.2979	1.3967	0.5037	0.1964	0.0775	0.0300	0.0113		
6.4							8.0768	1.8040	0.6454	0.2524	0.1008	0.0398	0.0153		
6.6							13.7692	2.4198	0.8247	0.3222	0.1302	0.0523	0.0205		
6.8							31.1270	3.2441	1.0533	0.4090	0.1666	0.0679	0.0271	0.0105	
7.0								4.4471	1.3471	0.5172	0.2119	0.0876	0.0357	0.0141	
7.2								6.3135	1.7288	0.6521	0.2677	0.1119	0.0463	0.0187	
7.4								9.5102	2.2324	0.8202	0.3364	0.1420	0.0595	0.0245	0.0097
7.6								16.0379	2.9113	1.0310	0.4211	0.1789	0.0761	0.0318	0.0129
7.8								35.8956	3.8558	1.2972	0.5250	0.2243	0.0966	0.0410	0.0168
8.0									5.2264	1.6364	0.6530	0.2796	0.1214	0.0522	0.0220
8.2									7.3441	2.0736	0.8109	0.3469	0.1520	0.0663	0.0283
8.4									10.9592	2.6470	1.0060	0.4288	0.1891	0.0834	0.0361
8.6									18.3223	3.4160	1.2484	0.5286	0.2341	0.1043	0.0459
8.8									40.6824	4.4806	1.5524	0.6501	0.2885	0.1298	0.0577
9.0										6.0183	1.9368	0.7980	0.3543	0.1603	0.0723
9.2										8.3869	2.4298	0.9788	0.4333	0.1974	0.0899
9.4										12.4189	3.0732	1.2010	0.5287	0.2419	0.1111
9.6										20.6160	3.9318	1.4752	0.6437	0.2952	0.1387
9.8										45.4769	5.1156	1.8165	0.7827	0.3588	0.1673
10.0											6.8210	2.2465	0.9506	0.4352	0.2040

TABLE 4
FINITE QUEUING TABLES

POPULATION 5 X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
POPULATION 5					.090	2	.044	.998	.010	.170	3	.017	.999	.005
						1	.350	.960	.200		2	.145	.985	.075
.012	1	.048	.999	.005	.095	2	.049	.997	.015		1	.611	.853	.735
.019	1	.076	.998	.010		1	.368	.955	.255	.180	3	.021	.999	.005
.025	1	.100	.997	.015	.100	2	.054	.997	.015		2	.161	.983	.085
.030	1	.120	.996	.020		1	.386	.950	.250		1	.638	.836	.820
.034	1	.135	.995	.025	.105	2	.059	.997	.015	.190	3	.024	.998	.010
.036	1	.143	.994	.030		1	.404	.945	.275		2	.177	.980	.100
.040	1	.159	.993	.035	.110	2	.065	.996	.020		1	.665	.819	.905
.042	1	.167	.992	.045		1	.421	.939	.305	.200	3	.028	.998	.010
.044	1	.175	.991	.045	.115	2	.071	.995	.025		2	.194	.976	.120
.046	1	.183	.990	.050		1	.439	.933	.335		1	.689	.801	.995
.050	1	.198	.989	.055	.120	2	.076	.995	.025	.210	3	.032	.998	.010
.052	1	.206	.988	.060		1	.456	.927	.365		2	.211	.973	.135
.054	1	.214	.987	.065	.125	2	.082	.994	.030		1	.713	.783	1.085
.056	2	.018	.999	.005		1	.473	.920	.400	.220	3	.036	.997	.015
	1	.222	.985	.075	.130	2	.089	.993	.035		2	.229	.969	.155
.058	2	.019	.999	.005		1	.489	.914	.430		1	.735	.765	1.175
	1	.229	.984	.080	.135	2	.095	.993	.035	.230	3	.041	.997	.015
.060	2	.020	.999	.005		1	.505	.907	.465		2	.247	.965	.175
	1	.237	.983	.085	.140	2	.102	.992	.040		1	.756	.747	1.265
.062	2	.022	.999	.005		1	.521	.900	.500	.240	3	.046	.996	.020
	1	.245	.982	.090	.145	3	.011	.999	.005		2	.265	.960	.200
.064	2	.023	.999	.005		2	.109	.991	.045		1	.775	.730	1.350
	1	.253	.981	.095		1	.537	.892	.540	.250	3	.052	.995	.025
.066	2	.024	.999	.005	.150	3	.012	.999	.005		2	.284	.955	.225
	1	.260	.979	.105		2	.115	.990	.050		1	.794	.712	1.440
.068	2	.026	.999	.005		1	.553	.885	.575	.260	3	.058	.994	.030
	1	.268	.978	.110	.155	3	.013	.999	.005		2	.303	.950	.250
.070	2	.027	.999	.005		2	.123	.989	.065		1	.811	.695	1.525
	1	.275	.977	.115		1	.568	.877	.615	.270	3	.064	.994	.030
.075	2	.031	.999	.005	.160	3	.015	.999	.005		2	.323	.944	.280
	1	.294	.973	.135		2	.130	.988	.060		1	.827	.677	1.615
.080	2	.035	.998	.010		1	.582	.869	.655	.280	3	.071	.993	.035
	1	.313	.969	.155	.165	3	.016	.999	.005		2	.342	.938	.310
.085	2	.040	.998	.010		2	.137	.987	.065		1	.842	.661	1.695
	1	.332	.965	.175		1	.597	.861	.695	.290	4	.007	.999	.005

Adapted from: L. G. Peck and R. N. Hazelwood, *Finite Queuing Tables* (New York: John Wiley & Sons, 1958).

TABLE 4 Continued
FINITE QUEUING TABLES

POPULATION 5–10

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
	3	.079	.992	.040	.440	4	.037	.996	.020		1	.998	.308	3.460
	2	.362	.932	.340		3	.238	.960	.200	.700	4	.240	.960	.200
	1	.856	.644	1.780		2	.652	.807	.965		3	.678	.815	.925
.300	4	.008	.999	.005		1	.969	.451	2.745		2	.950	.568	2.160
	3	.086	.990	.050	.460	4	.045	.995	.025		1	.999	.286	3.570
	2	.382	.926	.370		3	.266	.953	.235	.750	4	.316	.944	.280
	1	.869	.628	1.860		2	.686	.787	1.065		3	.763	.777	1.115
.310	4	.009	.999	.005		1	.975	.432	2.840		2	.972	.532	2.340
	3	.094	.989	.055	.480	4	.053	.994	.030	.800	4	.410	.924	.380
	2	.402	.919	.405		3	.296	.945	.275		3	.841	.739	1.305
	1	.881	.613	1.935		2	.719	.767	1.165		2	.987	.500	2.500
.320	4	.010	.999	.005		1	.980	.415	2.925	.850	4	.522	.900	.500
	3	.103	.988	.060	.500	4	.063	.992	.040		3	.907	.702	1.490
	2	.422	.912	.440		3	.327	.936	.320		2	.995	.470	2.650
	1	.892	.597	2.015		2	.750	.748	1.260	.900	4	.656	.871	.645
.330	4	.012	.999	.005		1	.985	.399	3.005		3	.957	.666	1.670
	3	.112	.986	.070	.520	4	.073	.991	.045		2	.998	.444	2.780
	2	.442	.904	.480		3	.359	.927	.365	.950	4	.815	.838	.810
	1	.902	.583	2.085		2	.779	.728	1.360		3	.989	.631	1.845
.340	4	.013	.999	.005		1	.988	.384	3.080					
	3	.121	.985	.075	.540	4	.085	.989	.055	POPULATION 10				
	2	.462	.896	.520		3	.392	.917	.415					
	1	.911	.569	2.155		2	.806	.708	1.460	.016	1	.144	.997	.03
.360	4	.017	.998	.010		1	.991	.370	3.150	.019	1	.170	.996	.04
	3	.141	.981	.060	.560	4	.098	.986	.070	.021	1	.188	.995	.05
	2	.501	.880	.600		3	.426	.906	.470	.023	1	.206	.994	.06
	1	.927	.542	2.290		2	.831	.689	1.555	.025	1	.224	.993	.07
.380	4	.021	.998	.010		1	.993	.357	3.215	.026	1	.232	.992	.08
	3	.163	.976	.120	.580	4	.113	.984	.080	.028	1	.250	.991	.09
	2	.540	.863	.685		3	.461	.895	.525	.030	1	.268	.990	.10
	1	.941	.516	2.420		2	.854	.670	1.650	.032	2	.033	.999	.01
.400	4	.026	.997	.015		1	.994	.345	3.275		1	.285	.988	– .12
	3	.186	.972	.140	.600	4	.130	.981	.095	.034	2	.037	.999	.01
	2	.579	.845	.755		3	.497	.883	.585		1	.302	.986	.14
	1	.952	.493	2.535		2	.875	.652	1.740	.036	2	.041	.999	.01
.420	4	.031	.997	.015		1	.996	.333	3.335		1	.320	.984	.16
	3	.211	.966	.170	.650	4	.179	.972	.140	.038	2	.046	.999	.01
	2	.616	.826	.870		3	.588	.850	.750		1	.337	.982	.18
	1	.961	.471	2.645		2	.918	.608	1.960	.040	2	.050	.999	.01

TABLE 4 Continued
FINITE QUEUING TABLES

POPULATION 10

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
	1	.354	.980	.20	.085	3	.037	.999	.01		2	.437	.947	.53
	1	.371	.978	.22		1	.692	.883	1.17		1	.919	.680	3.20
.042	2	.055	.999	.01		2	.196	.988	.12	.145	4	.032	.999	.01
.044	2	.060	.998	.02	.090	3	.043	.998	.02		3	.144	.990	.10
	1	.388	.975	.25		2	.216	.986	.14		2	.460	.941	.59
.046	2	.065	.998	.02		1	.722	.867	1.33		1	.929	.662	3.38
	1	.404	.973	.27	.095	3	.049	.998	.02	.150	4	.036	.998	.02
.048	2	.071	.998	.02		2	.237	.984	.16		3	.156	.989	.11
	1	.421	.970	.30		1	.750	.850	1.50		2	.483	.935	.65
.050	2	.076	.998	.02	.100	3	.056	.998	.02		1	.939	.644	3.56
	1	.437	.967	.33		2	.258	.981	.19	.155	4	.040	.998	.02
.052	2	.082	.997	.03		1	.776	.832	1.68		3	.169	.987	.13
	1	.454	.963	.37	.105	3	.064	.997	.03		2	.505	.928	.72
.054	2	.088	.997	.03		2	.279	.978	.22		1	.947	.627	3.73
	1	.470	.960	.40		1	.800	.814	1.86	.160	4	.044	.998	.02
.056	2	.094	.997	.03	.110	3	.072	.997	.03		3	.182	.986	.14
	1	.486	.956	.44		2	.301	.974	.26		2	.528	.921	.79
.058	2	.100	.996	.04		1	.922	.795	2.05		1	.954	.610	3.90
	1	.501	.953	.47	.115	3	.081	.996	.04	.165	4	.049	.997	.03
.060	2	.106	.996	.04		2	.324	.971	.29		3	.195	.984	.16
	1	.517	.949	.51		1	.843	.776	2.24		2	.550	.914	.86
.062	2	.113	.996	.04	.120	4	.016	.999	.01		1	.961	.594	4.06
	1	.532	.945	.55		3	.090	.995	.05	.170	4	.054	.997	.03
.064	2	.119	.995	.05		2	.346	.967	.33		3	.209	.982	.18
	1	.547	.940	.60		1	.861	.756	2.44		2	.571	.906	.94
.066	2	.126	.995	.05	.125	4	.019	.999	.01		1	.966	.579	4.21
	1	.562	.936	.64		3	.100	.994	.06	.180	5	.013	.999	.01
.068	3	.020	.999	.01		2	.369	.962	.38		4	.066	.996	.04
	2	.133	.994	.06		1	.878	.737	2.63		3	.238	.978	.22
	1	.577	.931	.69	.130	4	.022	.999	.01		2	.614	.890	1.10
.070	3	.022	.999	.01		3	.110	.994	.06		1	.975	.549	4.51
	2	.140	.994	.06		2	.392	.958	.42	.190	5	.016	.999	.01
	1	.591	.926	.74		1	.893	.718	2.82		4	.078	.995	.05
.075	3	.026	.999	.01	.135	4	.025	.999	.01		3	.269	.973	.27
	2	.158	.992	.08		3	.121	.993	.07		2	.654	.873	1.27
	1	.627	.913	.87		2	.415	.952	.48		1	.982	.522	4.78
.080	3	.031	.999	.01		1	.907	.699	3.01	.200	5	.020	.999	.01
	2	.177	.990	.10	.140	4	.028	.999	.01		4	.092	.994	.06
	1	.660	.899	1.01		3	.132	.991	.09		3	.300	.968	.32

TABLE 4 Continued
FINITE QUEUING TABLES

POPULATION 10

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
	2	.692	.854	1.46		1	.999	.370	6.30		2	.968	.584	4.16
	1	.987	.497	5.03	.280	6	.018	.999	.01	.360	7	.014	.999	.01
.210	5	.025	.999	.01		5	.081	.994	.06		6	.064	.995	.05
	4	.108	.992	.08		4	.252	.972	.28		5	.205	.978	.22
	3	.333	.961	.39		3	.571	.896	1.04		4	.474	.923	.77
	2	.728	.835	1.65		2	.903	.692	3.08		3	.798	.787	2.13
	1	.990	.474	5.26		1	.999	.357	6.43		2	.978	.553	4.47
.220	5	.030	.998	.02	.290	6	.022	.999	.01	.380	7	.019	.999	.01
	4	.124	.990	.10		5	.093	.993	.07		6	.083	.993	.07
	3	.366	.954	.46		4	.278	.968	.32		5	.247	.971	.29
	2	.761	.815	1.85		3	.603	.884	1.16		4	.533	.906	.94
	1	.993	.453	5.47		2	.918	.672	3.28		3	.840	.758	2.42
.230	5	.037	.998	.02		1	.999	.345	6.55		2	.986	.525	4.75
	4	.142	.988	.12	.300	6	.026	.998	.02	.400	7	.026	.998	.02
	3	.400	.947	.53		5	.106	.991	.09		6	.105	.991	.09
	2	.791	.794	2.06		4	.304	.963	.37		5	.292	.963	.37
	1	.995	.434	5.66		3	.635	.872	1.28		4	.591	.887	1.13
.240	5	.044	.997	.03		2	.932	.653	3.47		3	.875	.728	2.72
	4	.162	.986	.14		1	.999	.333	6.67		2	.991	.499	5.01
	3	.434	.938	.62	.310	6	.031	.998	.02	.420	7	.034	.993	.07
	2	.819	.774	2.26		5	.120	.990	.10		6	.130	.987	.13
	1	.996	.416	5.84		4	.331	.957	.43		5	.341	.954	.46
.250	6	.010	.999	.01		3	.666	.858	1.42		4	.646	.866	1.34
	5	.052	.997	.03		2	.943	.635	3.65		3	.905	.700	3.00
	4	.183	.983	.17	.320	6	.036	.998	.02		2	.994	.476	5.24
	3	.469	.929	.71		5	.135	.988	.12	.440	7	.045	.997	.03
	2	.844	.753	2.47		4	.359	.952	.48		6	.160	.984	.16
	1	.997	.400	6.00		3	.695	.845	1.55		5	.392	.943	.57
.260	6	.013	.999	.01		2	.952	.617	3.83		4	.698	.845	1.55
	5	.060	.996	.04	.330	6	.042	.997	.03		3	.928	.672	3.28
	4	.205	.980	.20		5	.151	.986	.14		2	.996	.454	5.46
	3	.503	.919	.81		4	.387	.945	.55	.460	8	.011	.999	.01
	2	.866	.732	2.68		3	.723	.831	1.69		7	.058	.995	.05
	1	.998	.384	6.16		2	.961	.600	4.00		6	.193	.979	.21
.270	6	.015	.999	.01	.340	7	.010	.999	.01		5	.445	.930	.70
	5	.070	.995	.05		6	.049	.997	.03		4	.747	.822	1.78
	4	.228	.976	.24		5	.168	.983	.17		3	.947	.646	3.54
	3	.537	.908	.92		4	.416	.938	.62		2	.998	.435	5.65
	2	.886	.712	2.88		3	.750	.816	1.84	.480	8	.015	.999	.01

TABLE 4 Continued

FINITE QUEUING TABLES

POPULATION 10-20

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
	7	.074	.994	.06		8	.072	.994	.06		6	.995	.667	3.33
	6	.230	.973	.27		7	.242	.972	.28	.950	9	.630	.938	.62
	5	.499	.916	.84	.600	6	.518	.915	.85		8	.934	.841	1.59
	4	.791	.799	2.01		5	.795	.809	1.91		7	.994	.737	2.63
	3	.961	.621	3.79		4	.953	.663	3.37					
	2	.998	.417	5.83		3	.996	.500	5.00	POPULATION 20				
.500	8	.020	.999	.01	.650	9	.021	.999	.01					
	7	.093	.992	.08		8	.123	.988	.12	.005	1	.095	.999	.02
	6	.271	.966	.34		7	.353	.954	.46	.009	1	.171	.998	.04
	5	.553	.901	.99		6	.651	.878	1.22	.011	1	.208	.997	.06
	4	.830	.775	2.25		5	.882	.759	2.41	.013	1	.246	.996	.08
	3	.972	.598	4.02		4	.980	.614	3.86	.014	1	.265	.995	.10
	2	.999	.400	6.00		3	.999	.461	5.39	.015	1	.283	.994	.12
.520	8	.026	.998	.02	.700	9	.040	.997	.03	.016	1	.302	.993	.14
	7	.115	.989	.11		8	.200	.979	.21	.017	1	.321	.992	.16
	6	.316	.958	.42		7	.484	.929	.71	.018	2	.048	.999	.02
	5	.606	.884	1.16		6	.772	.836	1.64		1	.339	.991	.18
	4	.864	.752	2.48		5	.940	.711	2.89	.019	2	.053	.999	.02
	3	.980	.575	4.25		4	.992	.571	4.29		1	.358	.990	.20
	2	.999	.385	6.15	.750	9	.075	.994	.06	.020	2	.058	.999	.02
.540	8	.034	.997	.03		8	.307	.965	.35		1	.376	.989	.22
	7	.141	.986	.14		7	.626	.897	1.03	.021	2	.064	.999	.02
	6	.363	.949	.51		6	.870	.792	2.08		1	.394	.987	.26
	5	.658	.867	1.33		5	.975	.666	3.34	.022	2	.070	.999	.02
	4	.893	.729	2.71		4	.998	.533	4.67		1	.412	.986	.28
	3	.986	.555	4.45	.800	9	.134	.988	.12	.023	2	.075	.999	.02
.560	8	.044	.996	.04		8	.446	.944	.56		1	.431	.984	.32
	7	.171	.982	.18		7	.763	.859	1.41	.024	2	.082	.999	.02
	6	.413	.939	.61		6	.939	.747	2.53		1	.449	.982	.36
	5	.707	.848	1.52		5	.991	.625	3.75	.025	2	.088	.999	.02
	4	.917	.706	2.94		4	.999	.500	5.00		1	.466	.980	.40
	3	.991	.535	4.65	.850	9	.232	.979	.21	.026	2	.094	.998	.04
.580	8	.057	.995	.05		8	.611	.916	.84		1	.484	.978	.44
	7	.204	.977	.23		7	.879	.818	1.82	.028	2	.108	.998	.04
	6	.465	.927	.73		6	.978	.705	2.95		1	.519	.973	.54
	5	.753	.829	1.71		5	.998	.588	4.12	.030	2	.122	.998	.04
	4	.937	.684	3.16	.900	9	.387	.963	.37		1	.553	.968	.64
	3	.994	.517	4.83		8	.785	.881	1.19	.032	2	.137	.997	.06
.600	9	.010	.999	.01		7	.957	.777	2.23		1	.587	.962	.76

TABLE 4 Continued
FINITE QUEUING TABLES

POPULATION 20

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
.034	2	.152	.996	.08		2	.392	.978	.44	.095	5	.031	.999	.02
	1	.620	.955	.90		1	.922	.785	4.30		4	.112	.996	.08
.036	2	.168	.996	.08	.062	4	.029	.999	.02		3	.326	.980	.40
	1	.651	.947	1.06		3	.124	.996	.08		2	.733	.896	2.08
.038	3	.036	.999	.02		2	.413	.975	.50		1	.998	.526	9.48
	2	.185	.995	.10		1	.934	.768	4.64	.100	5	.038	.999	.02
	1	.682	.938	1.24	.064	4	.032	.999	.02		4	.131	.995	.10
.040	3	.041	.999	.02		3	.134	.996	.08		3	.363	.975	.50
	2	.202	.994	.12		2	.433	.972	.56		2	.773	.878	2.44
	1	.712	.929	1.42		1	.944	.751	4.98		1	.999	.500	10.00
.042	3	.047	.999	.02	.066	4	.036	.999	.02	.105	5	.046	.999	.02
	2	.219	.993	.14		3	.144	.995	.10		4	.151	.993	.14
	1	.740	.918	1.64		2	.454	.969	.62		3	.400	.970	.60
.044	3	.053	.999	.02		1	.953	.733	5.34		2	.809	.858	2.84
	2	.237	.992	.16	.068	4	.039	.999	.02		1	.999	.476	10.48
	1	.767	.906	1.88		3	.155	.995	.10	.110	5	.055	.998	.04
.046	3	.059	.999	.02		2	.474	.966	.68		4	.172	.992	.16
	2	.255	.991	.18		1	.961	.716	5.68		3	.438	.964	.72
	1	.792	.894	2.12	.070	4	.043	.999	.02		2	.842	.837	3.26
.048	3	.066	.999	.02		3	.165	.994	.12	.115	5	.065	.998	.04
	2	.274	.989	.22		2	.495	.962	.76		4	.195	.990	.20
	1	.815	.881	2.38		1	.967	.699	6.02		3	.476	.958	.84
.050	3	.073	.998	.04	.075	4	.054	.999	.02		2	.870	.816	3.68
	2	.293	.988	.24		3	.194	.992	.16	.120	6	.022	.999	.02
	1	.837	.866	2.68		2	.545	.953	.94		5	.076	.997	.06
.052	3	.080	.998	.04		1	.980	.659	6.82		4	.219	.988	.24
	2	.312	.986	.28	.080	4	.066	.998	.04		3	.514	.950	1.00
	1	.858	.851	2.98		3	.225	.990	.20		2	.895	.793	4.14
.054	3	.088	.998	.04		2	.595	.941	1.18	.125	6	.026	.999	.02
	2	.332	.984	.32		1	.988	.621	7.58		5	.088	.997	.06
	1	.876	.835	3.30	.085	4	.080	.997	.06		4	.245	.986	.28
.056	3	.097	.997	.06		3	.257	.987	.26		3	.552	.942	1.16
	2	.352	.982	.36		2	.643	.928	1.44		2	.916	.770	4.60
	1	.893	.819	3.62		1	.993	.586	8.28	.130	6	.031	.999	.02
.058	3	.105	.997	.06	.090	5	.025	.999	.02		5	.101	.996	.08
	2	.372	.980	.40		4	.095	.997	.06		4	.271	.983	.34
	1	.908	.802	3.96		3	.291	.984	.32		3	.589	.933	1.34
.060	4	.026	.999	.02		2	.689	.913	1.74		2	.934	.748	5.04
	3	.115	.997	.06		1	.996	.554	8.92	.135	6	.037	.999	.02

TABLE 4 Continued

FINITE QUEUING TABLES

POPULATION 20

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
	5	.116	.995	.10		6	.099	.995	.10		8	.054	.998	.04
	4	.299	.980	.04		5	.248	.983	.34		7	.140	.992	.16
	3	.626	.923	1.54		4	.513	.945	1.10		6	.306	.975	.50
	2	.948	.725	5.50		3	.838	.830	3.40		5	.560	.931	1.38
.140	6	.043	.998	.04		2	.993	.587	8.26		4	.834	.828	3.44
	5	.131	.994	.12	.180	7	.044	.998	.04		3	.981	.649	7.02
	4	.328	.976	.48		6	.125	.994	.12	.240	9	.024	.999	.02
	3	.661	.912	1.76		5	.295	.978	.44		8	.068	.997	.06
	2	.960	.703	5.94		4	.575	.930	1.40		7	.168	.989	.22
.145	6	.051	.998	.04		3	.879	.799	4.02		6	.351	.969	.62
	5	.148	.993	.14		2	.996	.555	8.90		5	.613	.917	1.66
	4	.358	.972	.56	.190	8	.018	.999	.02		4	.870	.804	3.92
	3	.695	.900	2.00		7	.058	.998	.04		3	.988	.623	7.54
	2	.969	.682	6.36		6	.154	.991	.18	.250	9	.031	.999	.02
.150	7	.017	.999	.02		5	.345	.971	.58		8	.085	.996	.08
	6	.059	.998	.04		4	.636	.914	1.72		7	.199	.986	.28
	5	.166	.991	.18		3	.913	.768	4.64		6	.398	.961	.78
	4	.388	.968	.64		2	.998	.526	9.48		5	.664	.901	1.98
	3	.728	.887	2.26	.200	8	.025	.999	.02		4	.900	.780	4.40
	2	.976	.661	6.78		7	.074	.997	.06		3	.992	.599	8.02
.155	7	.021	.999	.02		6	.187	.988	.24	.260	9	.039	.998	.04
	6	.068	.997	.06		5	.397	.963	.74		8	.104	.994	.12
	5	.185	.990	.20		4	.693	.895	2.10		7	.233	.983	.34
	4	.419	.963	.74		3	.938	.736	5.28		6	.446	.953	.94
	3	.758	.874	2.52		2	.999	.500	10.00		5	.712	.884	2.32
	2	.982	.641	7.18	.210	8	.033	.999	.02		4	.924	.755	4.90
.160	7	.024	.999	.02		7	.093	.995	.10		3	.995	.576	8.48
	6	.077	.997	.06		6	.223	.985	.30	.270	10	.016	.999	.02
	5	.205	.988	.24		5	.451	.954	.92		9	.049	.998	.04
	4	.450	.957	.86		4	.745	.874	2.52		8	.125	.992	.16
	3	.787	.860	2.80		3	.958	.706	5.88		7	.270	.978	.44
	2	.987	.622	7.56		2	.999	.476	10.48		6	.495	.943	1.14
.165	7	.029	.999	.02	.220	8	.043	.998	.04		5	.757	.867	2.66
	6	.088	.996	.08		7	.115	.994	.12		4	.943	.731	5.38
	5	.226	.986	.28		6	.263	.980	.40		3	.997	.555	8.90
	4	.482	.951	.98		5	.505	.943	1.14	.280	10	.021	.999	.02
	3	.813	.845	3.10		4	.793	.852	2.96		9	.061	.997	.06
	2	.990	.604	7.92		3	.971	.677	6.46		8	.149	.990	.20
.170	7	.033	.999	.02	.230	9	.018	.999	.02		7	.309	.973	.54

TABLE 4 Continued

FINITE QUEUING TABLES

POPULATION 20

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
	6	.544	.932	1.36		8	.309	.973	.54		5	.992	.624	7.52
	5	.797	.848	3.04		7	.529	.935	1.30	.420	13	.019	.999	.02
	4	.958	.708	5.84		6	.766	.862	2.76		12	.055	.997	.06
	3	.998	.536	9.28		5	.933	.748	5.04		11	.131	.991	.18
.290	10	.027	.999	.02		4	.993	.605	7.90		10	.265	.977	.46
	9	.075	.996	.08	.340	11	.029	.999	.02		9	.458	.949	1.02
	8	.176	.988	.24		10	.079	.996	.08		8	.678	.896	2.08
	7	.351	.967	.66		9	.179	.987	.26		7	.863	.815	3.70
	6	.592	.920	1.60		8	.347	.967	.66		6	.965	.711	5.78
	5	.833	.828	3.44		7	.573	.924	1.52		5	.996	.595	8.10
	4	.970	.685	6.30		6	.802	.846	3.08	.440	13	.029	.999	.02
	3	.999	.517	9.66		5	.949	.729	5.42		12	.078	.995	.10
.300	10	.034	.998	.04		4	.995	.588	8.24		11	.175	.987	.26
	9	.091	.995	.10	.360	12	.015	.999	.02		10	.333	.969	.62
	8	.205	.985	.30		11	.045	.998	.04		9	.540	.933	1.34
	7	.394	.961	.78		10	.112	.993	.14		8	.751	.872	2.56
	6	.639	.907	1.86		9	.237	.981	.38		7	.907	.785	4.30
	5	.865	.808	3.84		8	.429	.954	.92		6	.980	.680	6.40
	4	.978	.664	6.72		7	.660	.901	1.98		5	.998	.568	8.64
	3	.999	.500	10.00		6	.863	.812	3.76	.460	14	.014	.999	.02
.310	11	.014	.999	.02		5	.971	.691	6.18		13	.043	.998	.04
	10	.043	.998	.04		4	.998	.555	8.90		12	.109	.993	.14
	9	.110	.993	.14	.380	12	.024	.999	.02		11	.228	.982	.36
	8	.237	.981	.38		11	.067	.996	.08		10	.407	.958	.84
	7	.438	.953	.94		10	.154	.989	.22		9	.620	.914	1.72
	6	.684	.893	2.14		9	.305	.973	.54		8	.815	.846	3.08
	5	.892	.788	4.24		8	.513	.938	1.24		7	.939	.755	4.90
	4	.985	.643	7.14		7	.739	.874	2.52		6	.989	.651	6.98
.320	11	.018	.999	.02		6	.909	.777	4.46		5	.999	.543	9.14
	10	.053	.997	.06		5	.984	.656	6.88	.480	14	.022	.999	.02
	9	.130	.992	.06		4	.999	.526	9.48		13	.063	.996	.08
	8	.272	.977	.46	.400	13	.012	.999	.02		12	.147	.990	.20
	7	.483	.944	1.12		12	.037	.998	.04		11	.289	.974	.52
	6	.727	.878	2.44		11	.095	.994	.12		10	.484	.944	1.12
	5	.915	.768	4.64		10	.205	.984	.32		9	.695	.893	2.14
	4	.989	.624	7.52		9	.379	.962	.76		8	.867	.819	3.62
.330	11	.023	.999	.02		8	.598	.918	1.64		7	.962	.726	5.48
	10	.065	.997	.06		7	.807	.845	3.10		6	.994	.625	7.50
	9	.154	.990	.20		6	.942	.744	5.12	.500	14	.033	.998	.04

TABLE 4 Continued
FINITE QUEUING TABLES

POPULATION 20–30

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
	13	.088	.995	.10		14	.129	.991	.18		16	.272	.980	.40
	12	.194	.985	.30		13	.266	.978	.44		15	.487	.954	.92
	11	.358	.965	.70		12	.455	.952	.96		14	.703	.913	1.74
	10	.563	.929	1.42		11	.662	.908	1.84		13	.864	.859	2.82
	9	.764	.870	2.60		10	.835	.847	3.06		12	.952	.798	4.04
	8	.908	.791	4.18		9	.941	.772	4.56		11	.988	.733	5.34
	7	.977	.698	6.04		8	.986	.689	6.22		10	.998	.667	6.66
	6	.997	.600	8.00		7	.998	.603	7.94	.800	19	.014	.999	.02
.520	15	.015	.999	.02	.600	16	.023	.999	.02		18	.084	.996	.08
	14	.048	.997	.06		15	.072	.996	.08		17	.242	.984	.32
	13	.120	.992	.16		14	.171	.988	.24		16	.470	.959	.82
	12	.248	.979	.42		13	.331	.970	.60		15	.700	.920	1.60
	11	.432	.954	.92		12	.532	.938	1.24		14	.867	.869	2.62
	10	.641	.911	1.78		11	.732	.889	2.22		13	.955	.811	3.78
	9	.824	.846	3.08		10	.882	.824	3.52		12	.989	.750	5.00
	8	.939	.764	4.72		9	.962	.748	5.04		11	.998	.687	6.26
	7	.987	.672	6.56		8	.992	.666	6.68	.850	19	.046	.998	.04
	6	.998	.577	8.46		7	.999	.583	8.34		18	.201	.988	.24
.540	15	.023	.999	.02	.650	17	.017	.999	.02		17	.451	.965	.70
	14	.069	.996	.08		16	.061	.997	.06		16	.703	.927	1.46
	13	.161	.988	.24		15	.156	.989	.22		15	.877	.878	2.44
	12	.311	.972	.56		14	.314	.973	.54		14	.962	.823	3.54
	11	.509	.941	1.18		13	.518	.943	1.14		13	.991	.765	4.70
	10	.713	.891	2.18		12	.720	.898	2.04		12	.998	.706	5.88
	9	.873	.821	3.58		11	.872	.837	3.26	.900	19	.135	.994	.12
	8	.961	.738	5.24		10	.957	.767	4.66		18	.425	.972	.52
	7	.993	.648	7.04		9	.990	.692	6.16		17	.717	.935	1.30
	6	.999	.556	8.88		8	.998	.615	7.70		16	.898	.886	2.28
.560	15	.035	.998	.04	.700	17	.047	.998	.04		15	.973	.833	3.34
	14	.095	.994	.12		16	.137	.991	.18		14	.995	.778	4.44
	13	.209	.984	.32		15	.295	.976	.48		13	.999	.722	5.56
	12	.381	.963	.74		14	.503	.948	1.04	.950	19	.377	.981	.38
	11	.586	.926	1.48		13	.710	.905	1.90		18	.760	.943	1.14
	10	.778	.869	2.62		12	.866	.849	3.02		17	.939	.894	2.12
	9	.912	.796	4.08		11	.953	.783	4.34		16	.989	.842	3.16
	8	.976	.713	5.74		10	.988	.714	5.72		15	.999	.789	4.22
	7	.996	.625	7.50		9	.998	.643	7.14	POPULATION 30				
.580	16	.015	.999	.02	.750	18	.031	.999	.02					
	15	.051	.997	.06		17	.113	.993	.14	.004	1	.116	.999	.03

TABLE 4 Continued
FINITE QUEUING TABLES

POPULATION 30

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
.007	1	.203	.998	.06		2	.286	.992	.24		2	.634	.951	1.47
.009	1	.260	.997	.09		1	.843	.899	3.03		1	.997	.616	11.52
.010	1	.289	.996	.12	.034	3	.083	.999	.03	.056	4	.086	.998	.06
.011	1	.317	.995	.15		2	.316	.990	.30		3	.267	.991	.27
.012	1	.346	.994	.18		1	.876	.877	3.69		2	.665	.944	1.68
.013	1	.374	.993	.21	.036	3	.095	.998	.06		1	.998	.595	12.15
.014	2	.067	.999	.03		2	.347	.988	.36	.058	4	.096	.998	.06
	1	.403	.991	.27		1	.905	.853	4.41		3	.288	.989	.33
.015	2	.076	.999	.03	.038	3	.109	.998	.06		2	.695	.936	1.92
	1	.431	.989	.33		2	.378	.986	.42		1	.999	.574	12.78
.016	2	.085	.999	.03		1	.929	.827	5.19	.060	5	.030	.999	.03
	1	.458	.987	.39	.040	3	.123	.997	.09		4	.106	.997	.09
.017	2	.095	.999	.03		2	.410	.983	.51		3	.310	.987	.39
	1	.486	.985	.45		1	.948	.800	6.00		2	.723	.927	2.19
.018	2	.105	.999	.03	.042	3	.138	.997	.09		1	.999	.555	13.35
	1	.513	.983	.51		2	.442	.980	.60	.062	5	.034	.999	.03
.019	2	.116	.999	.03		1	.963	.772	6.84		4	.117	.997	.09
	1	.541	.980	.60	.044	4	.040	.999	.03		3	.332	.986	.42
.020	2	.127	.998	.06		3	.154	.996	.12		2	.751	.918	2.46
	1	.567	.976	.72		2	.474	.977	.69	.064	5	.038	.999	.03
.021	2	.139	.998	.06		1	.974	.744	7.68		4	.128	.997	.09
	1	.594	.973	.81	.046	4	.046	.999	.03		3	.355	.984	.48
.022	2	.151	.998	.06		3	.171	.996	.12		2	.777	.908	2.76
	1	.620	.969	.93		2	.506	.972	.84	.066	5	.043	.999	.03
.023	2	.163	.997	.09		1	.982	.716	8.52		4	.140	.996	.12
	1	.645	.965	1.05	.048	4	.053	.999	.03		3	.378	.982	.54
.024	2	.175	.997	.09		3	.189	.995	.15		2	.802	.897	3.09
	1	.670	.960	1.20		2	.539	.968	.96	.068	5	.048	.999	.03
.025	2	.188	.996	.12		1	.988	.689	9.33		4	.153	.995	.15
	1	.694	.954	1.38	.050	4	.060	.999	.03		3	.402	.979	.63
.026	2	.201	.996	.12		3	.208	.994	.18		2	.825	.885	3.45
	1	.718	.948	1.56		2	.571	.963	1.11	.070	5	.054	.999	.03
.028	3	.051	.999	.03		1	.992	.663	10.11		4	.166	.995	.15
	2	.229	.995	.15	.052	4	.068	.999	.03		3	.426	.976	.72
	1	.763	.935	1.95		3	.227	.993	.21		2	.847	.873	3.81
.030	3	.060	.999	.03		2	.603	.957	1.29	.075	5	.069	.998	.06
	2	.257	.994	.18		1	.995	.639	10.83		4	.201	.993	.21
	1	.805	.918	2.46	.054	4	.077	.998	.06		3	.486	.969	.93
.032	3	.071	.999	.03		3	.247	.992	.24		2	.893	.840	4.80

TABLE 4 Continued
FINITE QUEUING TABLES

POPULATION 30

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
.080	6	.027	.999	.03		5	.289	.985	.45		3	.986	.687	9.39
	5	.088	.998	.06		4	.570	.950	1.50	.150	9	.024	.999	.03
	4	.240	.990	.30		3	.890	.833	5.01		8	.065	.998	.06
	3	.547	.959	1.23		2	.998	.579	12.63		7	.155	.993	.21
	2	.929	.805	5.85	.120	7	.057	.998	.06		6	.322	.980	.60
.085	6	.036	.999	.03		6	.147	.994	.18		5	.580	.944	1.68
	5	.108	.997	.09		5	.327	.981	.57		4	.860	.849	4.53
	4	.282	.987	.39		4	.619	.939	1.83		3	.991	.665	10.05
	3	.607	.948	1.56		3	.918	.808	5.76	.155	9	.029	.999	.03
	2	.955	.768	6.96		2	.999	.555	13.35		8	.077	.997	.09
.090	6	.046	.999	.03	.125	8	.024	.999	.03		7	.177	.992	.24
	5	.132	.996	.12		7	.069	.998	.06		6	.357	.976	.72
	4	.326	.984	.48		6	.171	.993	.21		5	.622	.935	1.95
	3	.665	.934	1.98		5	.367	.977	.69		4	.887	.830	5.10
	2	.972	.732	8.04		4	.666	.927	2.19		3	.994	.644	10.68
.095	6	.057	.999	.03		3	.940	.783	6.51	.160	9	.036	.999	.03
	5	.158	.994	.18	.130	8	.030	.999	.03		8	.090	.997	.09
	4	.372	.979	.63		7	.083	.997	.09		7	.201	.990	.30
	3	.720	.918	2.46		6	.197	.991	.27		6	.394	.972	.84
	2	.984	.697	9.09		5	.409	.972	.84		5	.663	.924	2.28
.100	6	.071	.998	.06		4	.712	.914	2.58		4	.910	.811	5.67
	5	.187	.993	.21		3	.957	.758	7.26		3	.996	.624	11.28
	4	.421	.973	.81	.135	8	.037	.999	.03	.165	9	.043	.999	.03
	3	.771	.899	3.03		7	.098	.997	.09		8	.105	.996	.12
	2	.991	.664	10.08		6	.226	.989	.33		7	.227	.988	.36
.105	7	.030	.999	.03		5	.451	.966	1.02		6	.431	.967	.99
	6	.087	.997	.09		4	.754	.899	3.03		5	.702	.913	2.61
	5	.219	.991	.27		3	.970	.734	7.98		4	.930	.792	6.24
	4	.470	.967	.99	.140	8	.045	.999	.03		3	.997	.606	11.82
	3	.816	.879	3.63		7	.115	.996	.12	.170	10	.019	.999	.03
	2	.995	.634	10.98		6	.256	.987	.39		9	.051	.998	.06
.110	7	.038	.999	.03		5	.494	.960	1.20		8	.121	.995	.15
	6	.105	.997	.09		4	.793	.884	3.48		7	.254	.986	.42
	5	.253	.988	.36		3	.979	.710	8.70		6	.469	.961	1.17
	4	.520	.959	1.23	.145	8	.055	.998	.06		5	.739	.901	2.97
	3	.856	.857	4.29		7	.134	.995	.15		4	.946	.773	6.81
	2	.997	.605	11.85		6	.288	.984	.48		3	.998	.588	12.36
.115	7	.047	.999	.03		5	.537	.952	1.44	.180	10	.028	.999	.03
	6	.125	.996	.12		4	.828	.867	3.99		9	.070	.997	.09

TABLE 4 Continued
FINITE QUEUING TABLES

POPULATION 30

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
	8	.158	.993	.21		10	.123	.994	.18		8	.676	.915	2.55
	7	.313	.980	.60		9	.242	.985	.45		7	.866	.841	4.77
	6	.546	.948	1.56		8	.423	.965	1.05		6	.970	.737	7.89
	5	.806	.874	3.78		7	.652	.923	2.31		5	.997	.617	11.49
	4	.969	.735	7.95		6	.864	.842	4.74	.280	14	.017	.999	.03
	3	.999	.555	13.35		5	.976	.721	8.37		13	.042	.998	.06
.190	10	.039	.999	.03		4	.999	.580	12.60		12	.093	.996	.12
	9	.094	.996	.12	.240	12	.031	.999	.03		11	.185	.989	.33
	8	.200	.990	.30		11	.074	.997	.09		10	.329	.976	.72
	7	.378	.973	.81		10	.155	.992	.24		9	.522	.949	1.53
	6	.621	.932	2.04		9	.291	.981	.57		8	.733	.898	3.06
	5	.862	.845	4.65		8	.487	.955	1.35		7	.901	.818	5.46
	4	.983	.699	9.03		7	.715	.905	2.85		6	.981	.712	8.64
.200	11	.021	.999	.03		6	.902	.816	5.52		5	.999	.595	12.15
	10	.054	.998	.06		5	.986	.693	9.21	.290	14	.023	.999	.03
	9	.123	.995	.15		4	.999	.556	13.32		13	.055	.998	.06
	8	.249	.985	.45	.250	13	.017	.999	.03		12	.117	.994	.18
	7	.446	.963	1.11		12	.042	.998	.06		11	.223	.986	.42
	6	.693	.913	2.61		11	.095	.996	.12		10	.382	.969	.93
	5	.905	.814	5.58		10	.192	.989	.33		9	.582	.937	1.89
	4	.991	.665	10.05		9	.345	.975	.75		8	.785	.880	3.60
.210	11	.030	.999	.03		8	.552	.944	1.68		7	.929	.795	6.15
	10	.073	.997	.09		7	.773	.885	3.45		6	.988	.688	9.36
	9	.157	.992	.24		6	.932	.789	6.33		5	.999	.575	12.75
	8	.303	.980	.60		5	.992	.666	10.02	.300	14	.031	.999	.03
	7	.515	.952	1.44	.260	13	.023	.999	.03		13	.071	.997	.09
	6	.758	.892	3.24		12	.056	.998	.06		12	.145	.992	.24
	5	.938	.782	6.54		11	.121	.994	.18		11	.266	.982	.54
	4	.995	.634	10.98		10	.233	.986	.42		10	.437	.962	1.14
.220	11	.041	.999	.03		9	.402	.967	.99		9	.641	.924	2.28
	10	.095	.996	.12		8	.616	.930	1.20		8	.830	.861	4.17
	9	.197	.989	.33		7	.823	.864	4.08		7	.950	.771	6.87
	8	.361	.974	.78		6	.954	.763	7.11		6	.993	.666	10.02
	7	.585	.938	1.86		5	.995	.641	10.77	.310	15	.017	.999	.03
	6	.816	.868	3.96	.270	13	.032	.999	.03		14	.041	.998	.06
	5	.961	.751	7.47		12	.073	.997	.09		13	.090	.996	.12
	4	.998	.606	11.82		11	.151	.992	.24		12	.177	.990	.30
.230	12	.023	.999	.03		10	.279	.981	.57		11	.312	.977	.69
	11	.056	.998	.06		9	.462	.959	1.23		10	.494	.953	1.41

TABLE 4 Continued
FINITE QUEUING TABLES

POPULATION 30

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
	9	.697	.909	2.73		12	.392	.967	.99		18	.041	.998	.06
	8	.869	.840	4.80		11	.578	.937	1.89		17	.087	.996	.12
	7	.966	.749	7.53		10	.762	.889	3.33		16	.167	.990	.30
	6	.996	.645	10.65		9	.902	.821	5.37		15	.288	.979	.63
.320	15	.023	.999	.03		8	.974	.738	7.86		14	.446	.960	1.20
	14	.054	.998	.06		7	.996	.648	10.56		13	.623	.929	2.13
	13	.113	.994	.18	.380	17	.020	.999	.03		12	.787	.883	3.51
	12	.213	.987	.39		16	.048	.998	.06		11	.906	.824	5.28
	11	.362	.971	.87		15	.101	.995	.15		10	.970	.755	7.35
	10	.552	.943	1.71		14	.191	.988	.36		9	.994	.681	9.57
	9	.748	.893	3.21		13	.324	.975	.75		8	.999	.606	11.82
	8	.901	.820	5.40		12	.496	.952	1.44	.460	19	.028	.999	.03
	7	.977	.727	8.19		11	.682	.914	2.58		18	.064	.997	.09
	6	.997	.625	11.25		10	.843	.857	4.29		17	.129	.993	.21
.330	15	.030	.999	.03		9	.945	.784	6.48		16	.232	.985	.45
	14	.068	.997	.09		8	.988	.701	8.97		15	.375	.970	.90
	13	.139	.993	.21		7	.999	.614	11.58		14	.545	.944	1.68
	12	.253	.983	.51	.400	17	.035	.999	.03		13	.717	.906	2.82
	11	.414	.965	1.05		16	.076	.996	.12		12	.857	.855	4.35
	10	.608	.931	2.07		15	.150	.992	.24		11	.945	.793	6.21
	9	.795	.876	3.72		14	.264	.982	.54		10	.985	.724	8.28
	8	.927	.799	6.03		13	.420	.964	1.08		9	.997	.652	10.44
	7	.985	.706	8.82		12	.601	.933	2.01	.480	20	.019	.999	.03
	6	.999	.606	11.82		11	.775	.886	3.42		19	.046	.998	.06
.340	16	.016	.999	.03		10	.903	.823	5.31		18	.098	.995	.15
	15	.040	.998	.06		9	.972	.748	7.56		17	.184	.989	.33
	14	.086	.996	.12		8	.995	.666	10.02		16	.310	.977	.69
	13	.169	.990	.30	.420	18	.024	.999	.03		15	.470	.957	1.29
	12	.296	.979	.63		17	.056	.997	.09		14	.643	.926	2.22
	11	.468	.957	1.29		16	.116	.994	.18		13	.799	.881	3.57
	10	.663	.918	2.46		15	.212	.986	.42		12	.910	.826	5.22
	9	.836	.858	4.26		14	.350	.972	.84		11	.970	.762	7.14
	8	.947	.778	6.66		13	.521	.948	1.56		10	.993	.694	9.18
	7	.990	.685	9.45		12	.700	.910	2.70		9	.999	.625	11.25
	6	.999	.588	12.36		11	.850	.856	4.32	.500	20	.032	.999	.03
.360	16	.029	.999	.03		10	.945	.789	6.33		19	.072	.997	.09
	15	.065	.997	.09		9	.986	.713	8.61		18	.143	.992	.24
	14	.132	.993	.21		8	.998	.635	10.95		17	.252	.983	.51
	13	.240	.984	.48	.440	19	.017	.999	.03		16	.398	.967	.99

TABLE 4 Continued
FINITE QUEUING TABLES

Population 30

X	M	D	F	L_q	X	M	D	F	L_q	X	M	D	F	L_q
	15	.568	.941	1.77		13	.972	.772	6.84		13	.999	.667	9.99
	14	.733	.904	2.88		12	.993	.714	8.58	.700	25	.039	.998	.06
	13	.865	.854	4.38		11	.999	.655	10.35		24	.096	.995	.15
	12	.947	.796	6.12	.580	23	.014	.999	.03		23	.196	.989	.33
	11	.985	.732	8.04		22	.038	.998	.06		22	.339	.977	.69
	10	.997	.667	9.99		21	.085	.996	.12		21	.511	.958	1.26
.520	21	.021	.999	.03		20	.167	.990	.30		20	.681	.930	2.10
	20	.051	.998	.06		19	.288	.980	.60		19	.821	.894	3.18
	19	.108	.994	.18		18	.443	.963	1.11		18	.916	.853	4.41
	18	.200	.988	.36		17	.612	.936	1.92		17	.967	.808	5.76
	17	.331	.975	.75		16	.766	.899	3.03		16	.990	.762	7.14
	16	.493	.954	1.38		15	.883	.854	4.38		15	.997	.714	8.58
	15	.663	.923	2.31		14	.953	.802	5.94	.750	26	.046	.998	.06
	14	.811	.880	3.60		13	.985	.746	7.62		25	.118	.994	.18
	13	.915	.827	5.19		12	.997	.690	9.30		24	.240	.986	.42
	12	.971	.767	6.99		11	.999	.632	11.04		23	.405	.972	.84
	11	.993	.705	8.85	.600	23	.024	.999	.03		22	.587	.950	1.50
	10	.999	.641	10.77		22	.059	.997	.09		21	.752	.920	2.40
.540	21	.035	.999	.03		21	.125	.993	.21		20	.873	.883	3.51
	20	.079	.996	.12		20	.230	.986	.42		19	.946	.842	4.74
	19	.155	.991	.27		19	.372	.972	.84		18	.981	.799	6.03
	18	.270	.981	.57		18	.538	.949	1.53		17	.995	.755	7.35
	17	.421	.965	1.05		17	.702	.918	2.46		16	.999	.711	8.67
	16	.590	.938	1.86		16	.837	.877	3.69	.800	27	.053	.998	.06
	15	.750	.901	2.97		15	.927	.829	5.13		26	.143	.993	.21
	14	.874	.854	4.38		14	.974	.776	6.72		25	.292	.984	.48
	13	.949	.799	6.03		13	.993	.722	8.34		24	.481	.966	1.02
	12	.985	.740	7.80		12	.999	.667	9.99		23	.670	.941	1.77
	11	.997	.679	9.63	.650	24	.031	.999	.03		22	.822	.909	2.73
	10	.999	.617	11.49		23	.076	.996	.12		21	.919	.872	3.84
.560	22	.023	.999	.03		22	.158	.991	.27		20	.970	.832	5.04
	21	.056	.997	.09		21	.281	.982	.54		19	.991	.791	6.27
	20	.117	.994	.18		20	.439	.965	1.05		18	.998	.750	7.50
	19	.215	.986	.42		19	.610	.940	1.80	.850	28	.055	.998	.06
	18	.352	.973	.81		18	.764	.906	2.82		27	.171	.993	.21
	17	.516	.952	1.44		17	.879	.865	4.05		26	.356	.981	.57
	16	.683	.920	2.40		16	.949	.818	5.46		25	.571	.960	1.20
	15	.824	.878	3.66		15	.983	.769	6.93		24	.760	.932	2.04
	14	.920	.828	5.16		14	.996	.718	8.46		23	.888	.899	3.03

POPULATION 30

X	M	D	F	L_q	X	M	D	F	L_q	X	D	F	L_q	
	22	.957	.862	4.14		26	.683	.953	1.41		28	.574	.973	.81
	21	.987	.823	5.31		25	.856	.923	2.31		27	.831	.945	1.65
	20	.997	.784	6.48		24	.947	.888	3.36		26	.951	.912	2.64
	19	.999	.745	7.65		23	.985	.852	4.44		25	.989	.877	3.69
.900	29	.047	.999	.03		22	.996	.815	5.55		24	.998	.842	4.74
	28	.200	.992	.24		21	.999	.778	6.66					
	27	.441	.977	.69	.950	29	.226	.993	.21					

TABLE 5
TABLE OF RANDOM DIGITS

78466	83326	96589	88727	72655	49682	82338	28583	01522	11248
78722	47603	03477	29528	63956	01255	29840	32370	18032	82051
06401	87397	72898	32441	88861	71803	55626	77847	29925	76106
04754	14489	39420	94211	58042	43184	60977	74801	05931	73822
97118	06774	87743	60156	38037	16201	35137	54513	68023	34380
71923	49313	59713	95710	05975	64982	79253	93876	33707	84956
78870	77328	09637	67080	49168	75290	50175	34312	82593	76606
61208	17172	33187	92523	69895	28284	77956	45877	08044	58292
05033	24214	74232	33769	06304	54676	70026	41957	40112	66451
95983	13391	30369	51035	17042	11729	88647	70541	36026	23113
19946	55448	75049	24541	43007	11975	31797	05373	45893	25665
03580	67206	09635	84612	62611	86724	77411	99415	58901	86160
56823	49819	20283	22272	00114	92007	24369	00543	05417	92251
87633	31761	99865	31488	49947	06060	32083	47944	00449	06550
95152	10133	52693	22480	50336	49502	06296	76414	18358	05313
05639	24175	79438	92151	57602	03590	25465	54780	79098	73594
65927	55525	67270	22907	55097	63177	34119	94216	84861	10457
59005	29000	38395	80367	34112	41866	30170	84658	84441	03926
06626	42682	91522	45955	23263	09764	26824	82936	16813	13878
11306	02732	34189	04228	58541	72573	89071	58066	67159	29633
45143	56545	94617	42752	31209	14380	81477	36952	44934	97435
97612	87175	22613	84175	96413	83336	12408	89318	41713	90669
97035	62442	06940	45719	39918	60274	54353	54497	29789	82928
62498	00257	19179	06313	07900	46733	21413	63627	48734	92174
80306	19257	18690	54653	07263	19894	89909	76415	57246	02621
84114	84884	50129	68942	93264	72344	98794	16791	83861	32007
58437	88807	92141	88677	02864	02052	62843	21692	21373	29408
15702	53457	54258	47485	23399	71692	56806	70801	41548	94809
59966	41287	87001	26462	94000	28457	09469	80416	05897	87970
43641	05920	81346	02507	25349	93370	02064	62719	45740	62080
25501	50113	44600	87433	00683	79107	22315	42162	25516	98434
98294	08491	25251	26737	00071	45090	68628	64390	42684	94956
52582	89985	37863	60788	27412	47502	71577	13542	31077	13353
26510	83622	12546	00489	89304	15550	09482	07504	64588	92562
24755	71543	31667	83624	27085	65905	32386	30775	19689	41437
38399	88796	58856	18220	51016	04976	54062	49109	95563	48244
18889	87814	52232	58244	95206	05947	26622	01381	28744	38374
51774	89694	02654	63161	54622	31113	51160	29015	64730	07750
88375	37710	61619	69820	13131	90406	45206	06386	06398	68652
10416	70345	93307	87360	53452	61179	46845	91521	32430	74795
99258	03778	54674	51499	13659	36434	84760	76446	64026	97534
58923	18319	95092	11840	87646	85330	58143	42023	28972	30657
39407	41126	44469	78889	54462	38609	58555	69793	27258	11296
29372	70781	19554	95559	63088	35845	60162	21228	48296	05006
07287	76846	92658	21985	00872	11513	24443	44320	37737	97360
07089	02948	03699	71255	13944	86597	89052	88899	03553	42145
35757	37447	29860	04546	28742	27773	10215	09774	43426	22961
58797	70878	78167	91942	15108	37441	99254	27121	92358	94254
32281	97860	23029	61409	81887	02050	63060	45246	46312	30378
93531	08514	30244	34641	29820	72126	62419	93233	26537	21179

SOURCE: The Rand Corporation, *A Million Random Digits with 100,000 Normal Deviates* (Glencoe, Ill.: The Free Press, 1955), pp. 180–183. (Reproduced with permission.)

TABLE 6
FACTORS USEFUL IN THE CONSTRUCTION OF CONTROL CHARTS*

Number of Observations in Sample	Chart for Averages			Chart for Standard Deviations						Chart for Ranges						
	Factors for Control Limits			Factors for Central Line		Factors for Control Limits				Factors for Central Line		Factors for Control Limits				
n	A	A_1	A_2	c_2	$1/c_2$	B_1	B_2	B_3	B_4	d_2	$1/d_2$	d_3	D_1	D_2	D_3	D_4
2	2.121	3.760	1.880	0.5642	1.7725	0	1.843	0	3.267	1.128	0.8865	0.853	0	3.686	0	3.267
3	1.732	2.394	1.023	0.7236	1.3820	0	1.858	0	2.568	1.693	0.5907	0.888	0	4.358	0	2.575
4	1.500	1.880	0.729	0.7979	1.2533	0	1.808	0	2.266	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	1.596	0.577	0.8407	1.1894	0	1.756	0	2.089	2.326	0.4299	0.864	0	4.918	0	2.115
6	1.225	1.410	0.483	0.8686	1.1512	0.026	1.711	0.030	1.970	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	1.277	0.419	0.8882	1.1259	0.105	1.672	0.118	1.882	2.704	0.3698	0.833	0.205	5.203	0.076	1.294
8	1.061	1.175	0.373	0.9027	1.1078	0.167	1.638	0.185	1.815	2.847	0.3512	0.820	0.387	5.307	0.136	1.864
9	1.000	1.094	0.337	0.9139	1.0942	0.219	1.609	0.239	1.761	2.970	0.3367	0.808	0.546	5.394	0.184	1.816
10	0.949	1.028	0.308	0.9227	1.0837	0.262	1.584	0.284	1.716	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.973	0.285	0.9300	1.0753	0.299	1.561	0.321	1.679	3.173	0.3152	0.787	0.812	5.534	0.256	1.744
12	0.866	0.925	0.266	0.9359	1.0684	0.331	1.541	0.354	1.646	3.258	0.3069	0.778	0.924	5.592	0.284	1.716
13	0.832	0.884	0.249	0.9410	1.0627	0.359	1.523	0.382	1.618	3.336	0.2998	0.770	1.026	5.646	0.308	1.692
14	0.802	0.848	0.235	0.9453	1.0579	0.384	1.507	0.406	1.594	3.407	0.2935	0.762	1.121	5.693	0.329	1.671
15	0.775	0.816	0.223	0.9490	1.0537	0.406	1.492	0.428	1.572	3.472	0.2880	0.755	1.207	5.737	0.348	1.652
16	0.750	0.788	0.212	0.9523	1.0501	0.427	1.478	0.448	1.552	3.532	0.2831	0.749	1.285	5.779	0.364	1.636
17	0.728	0.762	0.203	0.9551	1.0470	0.445	1.465	0.466	1.534	3.588	0.2787	0.743	1.359	5.817	0.379	1.621
18	0.707	0.738	0.194	0.9576	1.0442	0.461	1.454	0.482	1.518	3.640	0.2747	0.738	1.426	5.854	0.392	1.608
19	0.688	0.717	0.187	0.9599	1.0418	0.477	1.443	0.497	1.503	3.689	0.2711	0.733	1.490	5.888	0.404	1.596
20	0.671	0.697	0.180	0.9619	1.0396	0.491	1.433	0.510	1.490	3.735	0.2677	0.729	1.548	5.922	0.414	1.586
21	0.655	0.679	0.173	0.9638	1.0376	0.504	1.424	0.523	1.477	3.778	0.2647	0.724	1.605	5.950	0.425	1.575
22	0.640	0.662	0.167	0.9655	1.0358	0.516	1.415	0.534	1.466	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.647	0.162	0.9670	1.0342	0.527	1.407	0.545	1.455	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.632	0.157	0.9684	1.0327	0.538	1.399	0.555	1.445	3.895	0.2567	0.712	1.759	6.031	0.452	1.548
25	0.600	0.619	0.153	0.9696	1.0313	0.548	1.392	0.565	1.435	3.931	0.2544	0.709	1.804	6.058	0.459	1.541
Over 25	$\frac{3}{\sqrt{n}}$	$\frac{3}{\sqrt{n}}$	—	—	—	†	‡	†	‡	—	—	—	—	—	—	—

† $1 - \frac{3}{\sqrt{2n}}$ ‡ $1 + \frac{3}{\sqrt{2n}}$

Chart	Central Line	3σ Control Limits
$\bar{X}$	$\bar{\bar{x}}$	$\bar{\bar{x}} \pm A_1\bar{s}_x$ or
		$\bar{\bar{x}} \pm A_2\bar{R}$
	μ_x	$\mu_x \pm A\sigma_x$
R	$\bar{R}$	$D_3\bar{R}$ and $D_4\bar{R}$
	$d_2\sigma_x$	$D_1\sigma_x$ and $D_2\sigma_x$
σ_x	$\bar{s}_x$	$B_3\bar{s}_x$ and $B_4\bar{s}_x$
	σ_x	$B_1\sigma_x$ and $B_2\sigma_x$

Definitions: $A = 3/\sqrt{n}$, $A_1 = \frac{3}{c_2\sqrt{n}}$, $A_2 = \frac{3}{d_2\sqrt{n}}$, $B_1 = c_2 - K$, $B_2 = c_2 + K$, $B_3 = 1 - \frac{K}{c_2}$, $B_4 = 1 + \frac{K}{c_2}$, $D_1 = d_2 - 3d_3$,

$$D_2 = d_2 + 3d_3,\ D_3 = 1 - 3\frac{d_3}{d_2},\ \text{and}\ D_4 = 1 + 3\frac{d_3}{d_2}\ \text{where}\ K = 3\sqrt{\frac{(n-1)}{n} - c_2^2}$$

Warning: The fourth significant figures for D_1, D_2, D_3, and D_4 are in doubt for n greater than 5.

*Reproduced from Table B2 of the A.S.T.M. *Manual on Quality Control of Materials*, p. 115. Reprinted by permission of the American Society of Testing and Materials.

Index