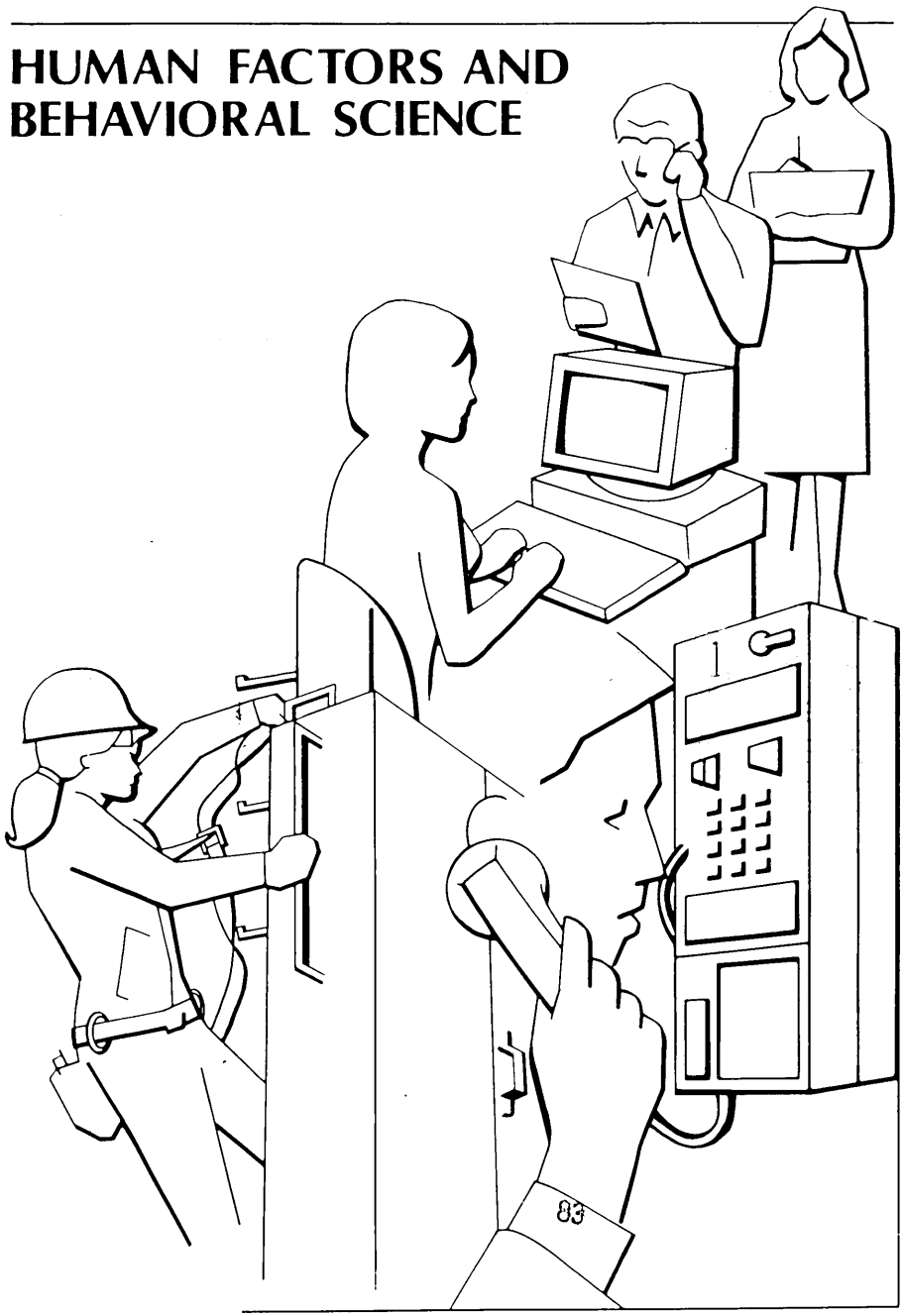
JULY-AUGUST 1983
VOL. 62, NO. 6, PART 3

THE BELL SYSTEM TECHNICAL JOURNAL

HUMAN FACTORS AND BEHAVIORAL SCIENCE



THE BELL SYSTEM TECHNICAL JOURNAL (ISSN0005-8580) is published by the American Telephone and Telegraph Company, 195 Broadway, N. Y., N. Y. 10007; C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; T. O. Davis, Secretary.

The Journal is published in three parts. Part 1, general subjects, is published ten times each year. Part 2, Computing Science and Systems, and Part 3, single-subject issues, are published with Part 1 as the papers become available.

The subscription price includes all three parts. Subscriptions: United States—1 year $35; 2 years $63; 3 years $84; foreign—1 year $45; 2 years $73; 3 years $94. Subscriptions to Part 2 only are $10 ($12 foreign). Single copies of the Journal are available at $5 ($6 foreign). Payment for foreign subscriptions or single copies must be made in United States funds, or by check drawn on a United States bank and made payable to The Bell System Technical Journal and sent to Bell Laboratories, Circulation Dept., Room 1E-335, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room 1J-319, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

Printed in U.S.A. Second-class postage paid at Short Hills, N. J. 07078 and additional mailing offices. Postmaster: Send address changes to The Bell System Technical Journal, Room 1E-335, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 62 July–August 1983 Number 6, Part 3

HUMAN FACTORS AND BEHAVIORAL SCIENCE

P. A. Turner, Guest Editor

Introduction 1561
E. E. Sumner

History and Methods 1569

A Brief History of Applied Behavioral Science at Bell Laboratories 1571
B. L. Hanson

Methods for Field Testing New Telephone Services 1591
D. J. Eigen

Characteristics of Human Performance 1617

Textons, The Fundamental Elements in Preattentive Vision and Perception of Textures 1619
B. Julesz and J. R. Bergen

Central Control of Movement Timing 1647
D. A. Rosenbaum

Experiments on Quantitative Judgments of Graphs and Maps 1659
W. S. Cleveland, C. S. Harris, and R. McGill

Retrospective Reports Reveal Differences in People's Reasoning 1675
D. E. Egan

New Technological Demands 1699

Human Factors Engineering for the Loop Plant 1701
J. Donegan and D. N. Koppes

Effects of Shape and Size of Knobs on Maximal Hand-Turning Forces Applied by Females 1705
G. A. Kohl

Human Factors Comparison of Two Fiber-Optic Continuous-Groove Field-Repair Splicing Techniques 1713
L. M. Paul

Performance in Locating Terminals on a High-Density Connector 1723
L. E. Flamm

Membrane Keyboards and Human Performance 1733
K. M. Cohen Loeb

Interface Design 1751

Statistical Semantics: Analysis of the Potential Performance of Key-Word Information Systems 1753
G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais

On Abbreviating Command Names 1807
L. A. Streeter, J. M. Ackroff, and G. A. Taylor

Designing and Evaluating Standard Instructions for Public Telephones 1827
C. J. Karhan, C. A. Riley, and M. S. Schoeffler

A Study of the Match Between the Stylistic Difficulty of Technical Documents and the Reading Skills of Technical Personnel 1849
E. U. Coke and M. E. Koether

Toward Bell System Applications of Automatic Speech Recognition 1865
J. E. Holmgren

New Functions for Technology 1881

The *UNIX*™ Writer's Workbench Software: Philosophy 1883
L. T. Frase

***The UNIX*™ Writer's Workbench Software: Rationale and Design** 1891
N. H. Macdonald

The *UNIX*™ Writer's Workbench Software: Results of a Field Study 1909
P. S. Gingrich

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Introduction

By E. E. SUMNER*

(Manuscript received March 15, 1983)

Human factors and its parent discipline, behavioral science, help us understand human capacities, interests, and needs. This understanding is especially important now, because modern technology makes unusual demands on its users. Knowledge and techniques derived from the behavioral sciences can help identify and design new products. This paper discusses the background and roles of the behavioral sciences, and it summarizes the contents of this issue.

This special issue of *The Bell System Technical Journal* presents a sample of recent activities of the behavioral science and human factors community at Bell Laboratories. Articles were drawn from several areas to give the reader an idea of the ways behavioral science and human factors contribute to the work at Bell Laboratories and, now, American Bell, a new subsidiary of AT&T.

This issue is intended for three groups of readers. First, behavioral scientists should gain a perspective of how their discipline is used in our Company. Second, members of the engineering and technical community who read this journal will see how behavioral science and human factors can contribute to the best use of technical innovation. The third audience this volume serves is the growing population of students who might find work in the field to be as interesting and challenging as we have.

* Bell Laboratories.

I. HUMAN USES FOR NEW TECHNOLOGY

We are now in a qualitatively new era of invention. Throughout history, imaginative people have known what they would like to do, but they have not had the means to do it. They wanted to fly, to communicate with each other from afar, to be entertained easily, and to have knowledge at their fingertips. Until now, reality has lagged behind imagination. Today, inventions such as the transistor, the laser, and large-scale circuit integration have reduced the gap between what we can imagine and what we can do. The future seems unlimited; but we face a new problem—how to make wise choices among unlimited opportunities.

Choices between alternative designs of a product have always reflected trade-offs among technical feasibility, cost, and utility to the user. But new technology has increased engineering flexibility and lowered the cost of components until the balance among these trade-offs has shifted dramatically toward the needs and desires of the user. Ways of designing a product to best fit users' interests can now be addressed as never before.

Word processing is a vivid example of this shift. When Christopher Sholes and his associates designed the typewriter a little over 100 years ago, even a slow operator would tie up the keys. Since faster machine action wasn't possible, Sholes' answer was to spread the high-use keys far apart in the key basket, which in turn forced a keyboard arrangement that is less than optimal for the user.[1] In contrast, it is not surprising to anyone now that a computer-based word processing system can keep up with a dextrous typist entering text from a keyboard. The machine can even operate on the text between key strokes. Such processing systems can instruct new users, provide page- or line-oriented editors, and format output according to users' needs. Computer-based text handling can thus be "user friendly"; it can allow the novice typist to substitute rudimentary knowledge of a set of computer programs for the psychomotor skills of an advanced typist. The balance is seen to have shifted even further toward the user when we consider the computer language analysis techniques described later in this issue (see the articles by Frase,[2] Macdonald,[3] and Gingrich[4]).

As human issues become more central to design, pressure increases to provide a technology of invention, design, and evaluation for human use. This technology is the focus of human factors or human performance engineering. It is an application of behavioral science concerned with the design of machines and procedures for human use. Bell Laboratories established the first industrial human factors laboratory in the late 1940s under the direction of John E. Karlin. Examples of early work include all-number-dialing studies, design of better dial layout, and development of objective preference methods. (Hanson[5]

provides a history of human factors work at Bell Laboratories in this issue.) The need for this work has increased, and participation of human factors specialists has grown explosively in recent years. For example, the number of people with primary work assignments in human factors or behavioral science has tripled in the last decade, from a starting point of about 100 workers. Many of these people recently joined American Bell, a new subsidiary of AT&T, which designs and markets enhanced services and equipment such as telephones and PBXs. Human factors specialists are now involved in virtually every Bell Laboratories and American Bell development having an important human performance component.

Even so, the potential for the applied behavioral sciences has only begun to be realized, partially because technology has just begun to challenge our ability to imagine new applications, and partially because the contribution of behavioral and related sciences has been difficult to foresee. This volume shows, by example, the contribution of behavioral science to an understanding of human perceptual, cognitive, and motor behavior, and how human factors applies this understanding to the design of new products.

Today, we are not only faced with choices between alternative designs, we must also decide what *should* be designed. This is the hardest question. Currently, we have technologies searching for applications, and new computer applications increasingly resemble intelligent, almost human, actions. Hence, an active role for behavioral science in planning these applications is crucial.

II. BEHAVIORAL SCIENCE RESOURCES

Most of the people doing applied and basic behavioral science research at Bell Laboratories and American Bell have been trained as experimental psychologists. Not many have degrees in human factors, because until recently there have been few programs in the field. In the past fifteen years, psychological research on learning and perception has broadened to include cognitive psychology, which studies thinking, language, and other aspects of information processing. Psychologists are well acquainted with sensory and central nervous system limits on human information handling, through studies of thought processes, linguistic and problem-solving skills, learning and forgetting, social communication, perceptual motor skills, and strength and endurance. Psychologists are also trained in experimental design, statistics, and the measurement of human performance, attitudes, preferences, and motivation.

Behavioral science work in Bell Laboratories has been differentiated mainly by its intent and the clientele that uses it. Basic research has been directed more at understanding the processes underlying behav-

ior, while applied research, emanating from human factors groups, has been associated with designing specific products to meet the needs, interests, and performance capabilities of its users. If research and applications specialists were to collaborate on a study, the measures taken of a person's performance—for instance, reaction time to red or green lights—would be of equal interest to both, but would have been collected for different reasons. For research, we would interpret these measures as a reflection of some underlying process, while for applications we would see the behavior as a fact to be reckoned with in product design. For example, Pierce and Karlin[6] found that reading rate is independent of the number of alternative words on a list of fixed length. The researcher might interpret this result as a reflection of channel capacity of humans as information processors. The applications designer might find the invariance of reading rates important for determining transmission requirements for a speech communication system.

In basic research, we attempt to structure work to provide generality beyond the experimental conditions under which data were collected. Researchers routinely publish work, which is read by other researchers, academics, students, and their colleagues. Applied research, on the other hand, often requires team cooperation, and people are responsible to others for providing information needed to complete a design. Although information gathered is often private and is generally collected to answer questions about the product being designed, data collected in one application often find good use in other applications (see McCormick[7] and Woodson[8]). Applied work has sometimes been seen by some as just the use, refinement, and dissemination of existing methods, rather than the invention and development of tools that generalize across tasks. This is wrong. Many behavioral concepts and techniques, important to theory as well as application, have arisen from practical problems. Aptitude testing and the measurement of human channel capacity are only two examples.

The role of basic research in Bell Laboratories in the study of psychological processes is well known to the outside world through research published over the last 20 years. Hanson[5] briefly summarizes highlights of that work in this volume. An example of applied work completes this section.

The human factors specialist contributes to product design at each stage in product development. One trade-off of particular concern is the extent to which the operation of a product can be made self-evident and require only minimal user training. Training can represent a substantial investment for a new electronic telephone system being installed in a large business having hundreds or thousands of users. The number and range of features available to each user mean that even a well-designed system will have to be carefully explained.

Dooling and Klemmer[9] have described work conducted on training procedures for business telephone users. The problem was that, although customers were happy with their systems, they wanted more training. But telephone companies found that two-hour training sessions, run with small groups of users, were too expensive. The results of several studies led to a more efficient training package that required only one hour of training time. The highlights of these studies are as follows.

Originally, participants in a session were trained in the detailed operation of as many as a dozen features in the two-hour session. More information was being presented than many participants could absorb. Testing showed that a card summarizing the operation of the features was a powerful performance aid. Consequently, training people to use the card replaced much of the detailed explanation and contributed to a reduction in training time. Learning to use telephone features is a cognitive task with only a minor motor component. If it is not a difficult manual task is it necessary to give hands-on experience? A set of laboratory and field studies showed that working the features on the telephone in the session did not significantly aid user performance. Here was a way to further shorten the training session, by eliminating unnecessary activities. Finally, self-instruction was found to be as effective as the group sessions both in laboratory experiments and in field studies. Thus, it was not crucial for all users to attend the training sessions.

Human factors specialists contributed here by proposing alternative training schemes based on what they knew about training and the ways people learn efficiently. They then devised studies and gathered data on the efficiency of those alternatives. Providing recommendations from theory alone would not have been enough to convince either the human factors people or telephone company management to use new procedures.

The skills of behavioral science personnel are critical now. Throughout the computer industry, much is said about "user-friendly" systems, yet in few places are the tools of behavioral science well understood, much less effectively applied. Computers are no longer just science and engineering tools; they are used every day at work, at school, and at home. Computers are for everyone, and for the Bell System this means finding new ways to relate technology to human needs and abilities.

III. STRUCTURE AND THEMES OF THIS SPECIAL ISSUE

This issue is divided into five sections. The "History and Methods" section describes the growth of human factors work at Bell Laboratories and techniques that have been used to study human response

to communications products. Hanson's paper reviews human factors history, while Eigen's paper describes field test methods. Many techniques are used by behavioral scientists, of which field testing is only one. However, Eigen's paper gives an overview of one traditional use for human factors work—i.e., to evaluate products created by communications engineers.

The second section, "Characteristics of Human Performance," describes fundamental properties of human perception and performance. Julesz' and Bergen's paper describes mechanisms of visual perception, concentrating on the perception of texture. The paper by Rosenbaum explains how people control the timing of finger movements, and this may, in the future, help us understand how better to design keyboards and other control devices and tools. The final two papers in this section address higher-level cognitive processes. The paper by Cleveland, Harris, and McGill describes studies of how people derive information from graphic displays, while the paper by Egan explores how people reason when solving verbal problems. In particular, Egan shows that people can be trained to think in new ways, using strategies they normally do not use. These papers provide context for understanding possible limits and potentials for human adaptation to new technology.

The third section, "New Technological Demands," includes papers that deal with the unusual requirements that modern technology places on us. The paper by Donegan and Koppes introduces the three papers that follow it; their introduction emphasizes the wide variety of skills and characteristics that are needed by the craft personnel who install, maintain, and repair telecommunications equipment. The paper by Kohl describes work on the design of shapes to improve handling physical objects, the paper by Paul describes improved designs for splicing connectors, and the final paper in this trilogy, by Flamm, describes connecting-block design changes that reduce errors often made with complex apparatus. Following these papers, the article by Cohen moves on to the design of keyboards and people's ability to use new technology, such as membrane keyboards. Thus, papers in this section progress from simple to more complex performance.

The fourth section, "Interface Design," expands the theme of verbal interactions demanded by new technology. The paper by Furnas, Landauer, Gomez, and Dumais explores alternate ways to design key words to ease people's interactions with computers. The paper by Streeter, Ackroff, and Taylor extends this concern to the design of abbreviations for command names. The theme of human interactions involving verbal information is carried forward in the paper by Karhan, Riley, and Schoeffler, as they describe the design of telephone cards to provide easy access to dialing information people need when phoning. This theme is extended in the paper by Coke and Koether, which

explores the match between people's ability to read and the reading demands placed on them by the language of telecommunications documents. The final paper, by Holmgren, shows that technology, too, has limits, and that proper adaptation of humans to the limited speech of machines can make it possible for machines to understand what humans say.

The fifth and final section of this issue, "New Functions for Technology," is an example of bringing behavioral science to bear on the design of new products. The papers in this section, by Frase, Macdonald, and Gingrich, describe the *UNIX* Writer's Workbench software, a set of programs that do many things editors do when they proofread and comment on various features of written material. This section exemplifies the issue raised at the start of this introduction; we have the technology, we understand the physical principles needed to design new electronic tools; we have the engineering skills to build those tools, and we have the computing techniques to program them. But we need a behavioral perspective to know where human minds falter and need help, and where they excel and are best left alone. Human factors—behavioral science, generally—is an essential resource in this new era of communication design.

IV. ACKNOWLEDGMENTS

This special issue was created through the efforts of many people. Paul A. Turner proposed this special issue to the Editorial Committee of *The Bell System Technical Journal*, with the encouragement of Thomas H. Crowley, a member of that committee. V. G. Stetter, followed by Eric Wolman, served as administrative coordinator for the issue. A working-level group, chaired by Barry L. Lively (now of American Bell) and including Lawrence T. Frase, Thomas K. Landauer, and Christine A. Riley (American Bell) solicited, edited, and organized the papers in the volume. Special recognition should be given to Lawrence T. Frase and Barry L. Lively for their contributions. A panel of more than 30 behavioral scientists and human factors specialists within Bell Laboratories and American Bell read and commented on the papers. The ground rules for that procedure discourage acknowledgment of those reviewers by name, but the reader should know that their contributions were vital to the issue.

REFERENCES

1. C. Leckberg, "The Tyranny of Qwerty," Saturday Review, *55*, No. 40 (September 1972), pp. 37–40.
2. L. T. Frase, "*UNIX*™ Writer's Workbench Software: Philosophy," B.S.T.J., this issue.
3. N. H. Macdonald, "The *UNIX*™ Writer's Workbench Software: Rationale and Design," B.S.T.J., this issue.

4. P. S. Gingrich, "The *UNIX*™ Writer's Workbench Software: Results of a Field Study," B.S.T.J., this issue.
5. B. L. Hanson, "A Brief History of Applied Behavioral Science at Bell Laboratories," B.S.T.J., this issue.
6. J. R. Pierce and J. E. Karlin, "Reading Rates and the Information Rate of a Human Channel," B.S.T.J., *36* (March 1957), pp. 497–516.
7. E. J. McCormick, *Human Factors in Engineering and Design*, 4th ed., New York: McGraw-Hill, 1976.
8. W. E. Woodson, *Human Factors Design Handbook*, New York: McGraw-Hill, 1981.
9. D. J. Dooling and E. T. Klemmer, "New Technology for Business Telephone Users: Some Findings from Human Factors Studies," in *Information Technology and Psychology: Prospects for the Future* (edited by R. A. Kasschau, et al.), New York: Praeger Publishers, 1982, pp. 148–65.

AUTHOR

Eric E. Sumner, B.M.E., Cooper Union, 1948; M.S. (Physics), 1953 and professional degree (Electrical Engineering), 1960, Columbia University; Bell Laboratories, 1948—. Upon joining Bell Laboratories, Mr. Sumner worked on wire spring relays, trouble recording apparatus, and ESS circuits. He was responsible for the first commercial PCM system for exchange trunks. He was promoted to Director and served in positions responsible for various transmissions systems. He was Executive Director of the Loop Transmission Division from 1967 to 1979, when he was appointed Executive Director of the Customer Network Operations Division. He assumed his present position as Vice President, Software and Processor Technologies, in 1981. Mr. Sumner is a member of the Advisory Board to Georgia Institute of Technology's School of Electrical Engineering; Chairman of the Advisory Council to Cooper Union's School of Engineering; President of IEEE Communications Society; Chairman of IEEE Policy Board; and recipient of the Alexander Graham Bell Medal of the IEEE for pioneering contributions to the field of digital communications. Mr. Sumner is author of a number of technical articles and talks and has been granted 11 patents. Fellow, IEEE; member, Tau Beta Pi and Pi Tau Sigma.

History and Methods

This special issue begins with a brief history of human factors and behavioral science at Bell Laboratories. Bruce Hanson traces the expansion of applied and basic behavioral science at Bell Laboratories over more than three decades. His account shows the important role of laboratory studies and field tests in helping us understand how people interact with new telecommunication systems. The second paper in this section, by Daryl Eigen, discusses field test methods. Laboratory studies are often conducted early in a product or system development, while field tests occur later and involve realistic simulations. Eigen shows how field tests help to ensure that the human interface to complex telephone services meets the user's needs.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

A Brief History of Applied Behavioral Science at Bell Laboratories

By B. L. HANSON*

(Manuscript received March 16, 1982)

The history of applied behavioral sciences at Bell Laboratories follows two main paths. The first, the primarily customer-oriented "human factors" tradition, began in the late 1940s and has been characterized by an empirical approach, relying heavily on laboratory and field simulation. The second, the employee-oriented "human performance technology" tradition, had its roots in Bell Laboratories behavioral research organization formed in the late 1950s. It has been characterized by a more rule-oriented approach to the integration of human users and operators into large, computerized systems. This paper traces the evolution of these two applied traditions and the behavioral research organization and examines the people and the events that influenced their growth and success.

I. INTRODUCTION

From the earliest days of telephony when Alexander Graham Bell and Thomas Watson worked to perfect Bell's invention into a useful and convenient communication device, designing for users has been an important goal of the Bell System. As the number of Bell System

* American Bell.

employees has grown to more than a million, there has been an equal concern for using people efficiently in satisfying assignments. But it was not until the 1940s that specialists began to be hired to address customers' and employees' needs, and today Bell Laboratories employs hundreds of such specialists working in the behavioral sciences.

This article traces the major lines of evolution of applied behavioral science at Bell Laboratories and examines some of the forces that have shaped and continue to shape its growth, particularly the behavioral research organization. While there were some early applications elsewhere in the Bell System (the Western Electric Hawthorne studies conducted in the late 1920s and early 1930s, for example[1]), these are not treated here, except as they influenced events at Bell Laboratories.

II. SETTING THE STAGE

Human factors, human performance technology, and engineering psychology are a few of the many names used to describe the application of the behavioral sciences (psychology, sociology, anthropology, etc.) to the design of systems that involve people. While each of these names has its own connotations, human factors will be used here for simplicity to denote the entire discipline. Human factors traces back to World War II when American and British psychologists worked to match complex new weapons systems to the people who would employ them. Most human factors specialists, then and now, have been trained in psychology as specialists in learning, human performance, visual and auditory perception, motivation, social behavior, or decision making. Many have received training in systems analysis, industrial engineering or specifically in human factors engineering.

Human factors at Bell Laboratories has evolved along two paths which have only recently begun to come together. The first arose from the needs of telephone customers. The second arose in response to the needs and skills of employees, and led to a "pure" behavioral research organization as well as applied activities. This article describes the two applied paths and the path of behavioral research. To prevent getting lost in digressions, a chronological road map is provided in Fig. 1. References point to sources of additional information.

III. TELEPHONES AND CUSTOMERS

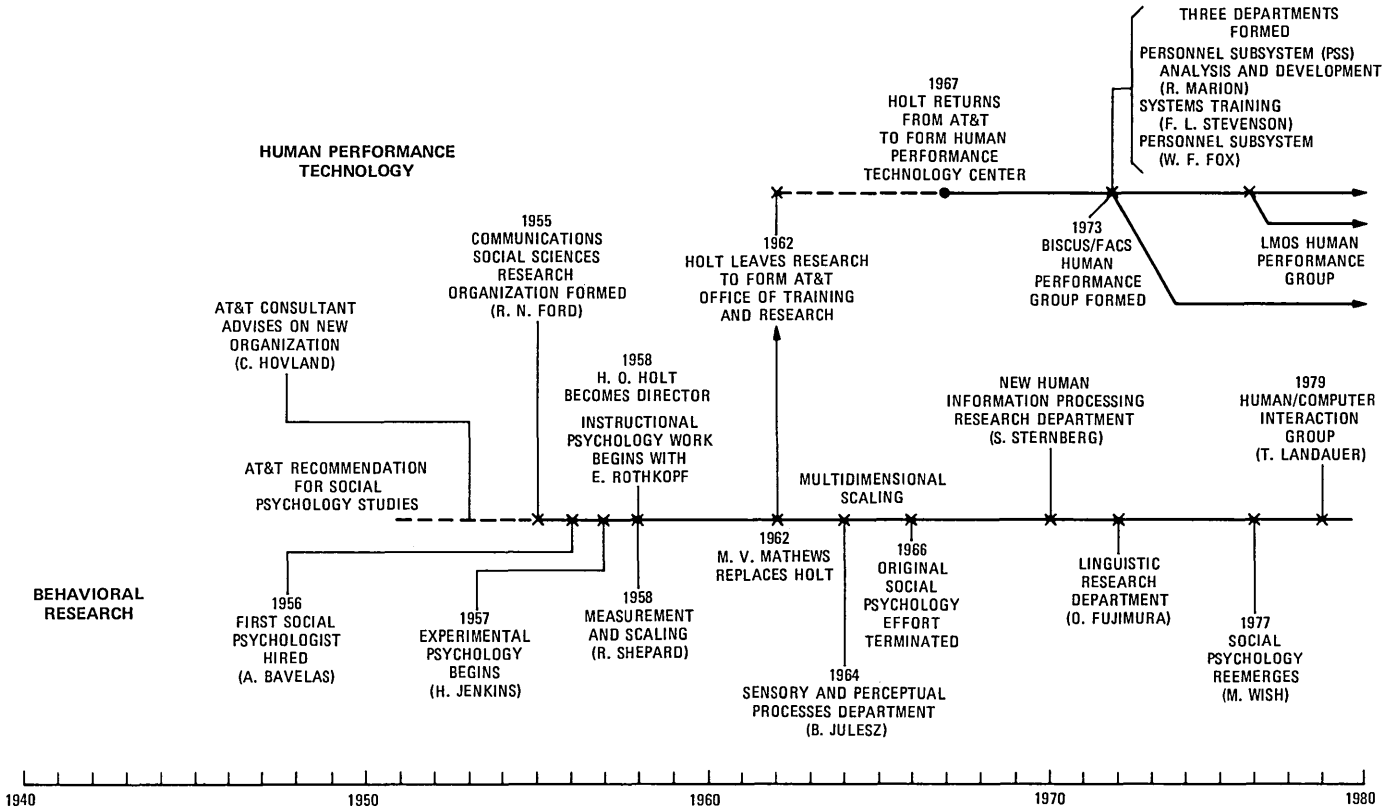
From the beginning of telephony, human factors decisions were being made about telephones. The first concern was to improve the quality and intelligibility of transmitted speech. Signalling was also important, especially as the number of users started to grow. Other, more subtle innovations also helped to make the telephone more useful

and convenient. The invention of the switchhook meant users no longer had to remember to throw a switch after each call. More calls went through and many batteries were saved because when the receiver was "hung up" on the switchhook (it had to go somewhere), the job was done automatically. The one-piece telephone handset added portability and convenience, but only because designers calculated the dimensions of the heads of potential users to ensure that when the receiver was placed to the ear the microphone would be the right distance from the lips. At Bell Laboratories in the 1920s and 1930s, work was focused on designing better telephone sets by considering the physical dimensions of customers' heads and hands[2] and on understanding the properties of the human ear and voice so that electrical transducers and circuits could be improved.[3] It was in this second area, known as psychoacoustics, that behavioral science was formally instituted at Bell Laboratories.

The pioneering work in this field was conducted at Bell Laboratories under the leadership of Harvey Fletcher, Wilden A. Munson, and others,[4] but, by the 1940s, laboratories at other institutions were doing similar work for communications systems for the World War II effort. One such laboratory was managed by S. Smith Stevens, an experimental psychologist at Harvard. Fletcher was aware of the work in Stevens' lab and, in 1945, hired John E. Karlin from the list of notable psychologists who worked there.

Karlin worked only briefly in psychoacoustics before realizing that there were many other opportunities for valuable behavioral sciences work. On the basis of his observations, he made a proposal to his management for a broad program of customer studies. Shortly thereafter, the User Preference Research Department was formed, headed by Walter A. Shewhart (widely known for his work in quality control), and staffed by Karlin, Robert R. Riesz, and others. In 1951, Karlin succeeded Shewhart as head of the department.

Karlin's department continued to expand, performing a growing range of studies on user preference, telephone design, and network applications. Notable among these was a series of studies confirming people's inability to make reliable preference judgments about things they have never experienced. In one of these studies, customers were asked to judge their preference for handsets lighter than those already on their telephones. They all preferred the existing 18-ounce weight. But when they had the opportunity to handle handsets of various weights, they preferred ones that were much lighter. No one showed a preference for handsets as heavy as the standard one. Even 12-ounce handsets were heavier than people preferred. The customers could not predict their own preferences without having actual experience with the alternative choices. This rejection of armchair opinions set the


HUMAN PERFORMANCE TECHNOLOGY
BEHAVIORAL RESEARCH
THREE DEPARTMENTS FORMED
PERSONNEL SUBSYSTEM (PSS)
ANALYSIS AND DEVELOPMENT
(R. MARION)
SYSTEMS TRAINING
(F. L. STEVENSON)
PERSONNEL SUBSYSTEM
(W. F. FOX)
1967
HOLT RETURNS FROM AT&T TO FORM HUMAN PERFORMANCE TECHNOLOGY CENTER
1973
BISCUS/FACS HUMAN PERFORMANCE GROUP FORMED
LMOS HUMAN PERFORMANCE GROUP
1962
HOLT LEAVES RESEARCH TO FORM AT&T OFFICE OF TRAINING AND RESEARCH
1955
COMMUNICATIONS SOCIAL SCIENCES RESEARCH ORGANIZATION FORMED
(R. N. FORD)
AT&T CONSULTANT ADVISES ON NEW ORGANIZATION
(C. HOVLAND)
1958
H. O. HOLT BECOMES DIRECTOR
INSTRUCTIONAL PSYCHOLOGY WORK BEGINS WITH
E. ROTHKOPF
AT&T RECOMMENDATION FOR SOCIAL PSYCHOLOGY STUDIES
MULTIDIMENSIONAL SCALING
NEW HUMAN INFORMATION PROCESSING RESEARCH DEPARTMENT
(S. STERNBERG)
1979
HUMAN/COMPUTER INTERACTION GROUP
(T. LANDAUER)
1962
M. V. MATHEWS REPLACES HOLT
1966
ORIGINAL SOCIAL PSYCHOLOGY EFFORT TERMINATED
LINGUISTIC RESEARCH DEPARTMENT
(O. FUJIMURA)
1956
FIRST SOCIAL PSYCHOLOGIST HIRED
(A. BAVELAS)
1957
EXPERIMENTAL PSYCHOLOGY BEGINS
(H. JENKINS)
1958
MEASUREMENT AND SCALING
(R. SHEPARD)
1964
SENSORY AND PERCEPTUAL PROCESSES DEPARTMENT
(B. JULESZ)
1977
SOCIAL PSYCHOLOGY REEMERGES
(M. WISH)
1940
1950
1960
1970
1980

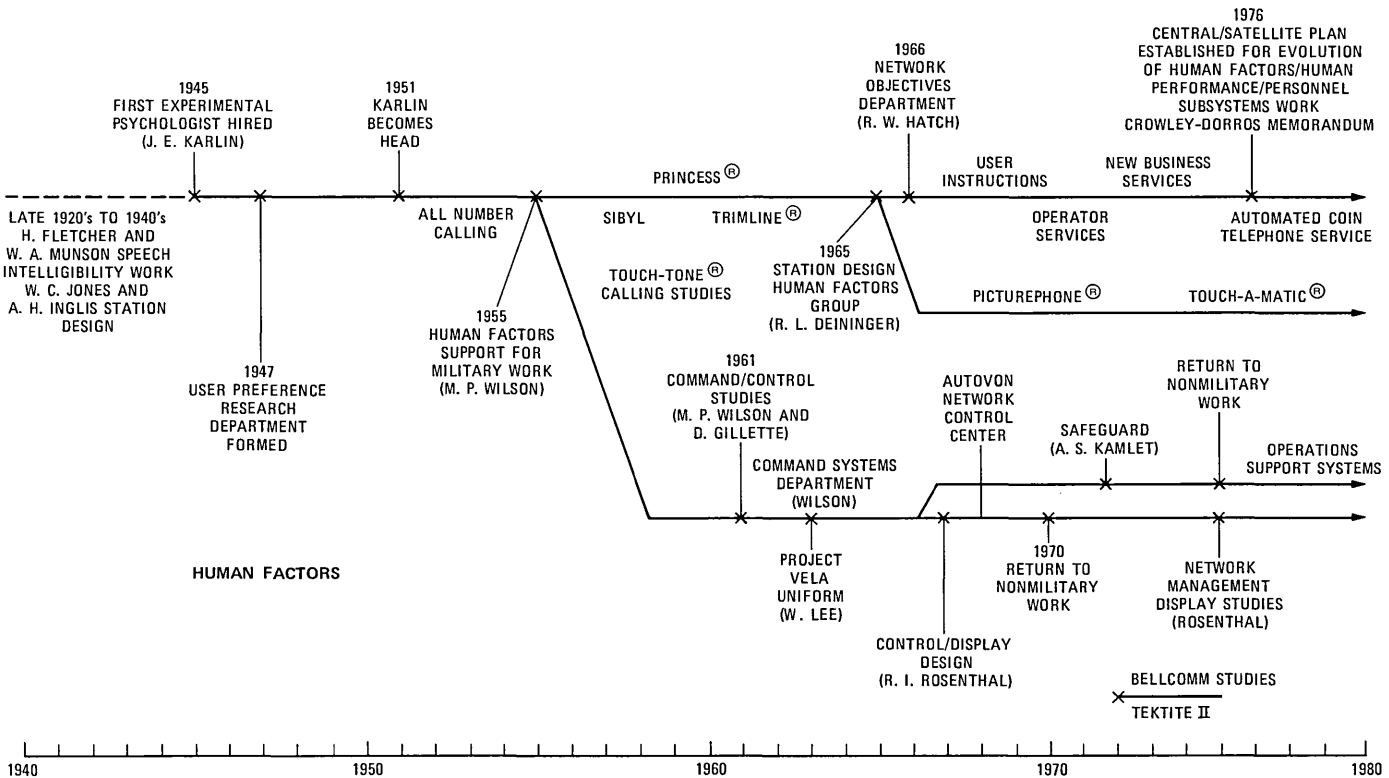


Fig. 1—Development of applied behavioral sciences organizations at Bell Laboratories.

tone for the activities in Karlin's department. The work was empirically based and it depended on providing people with product experience through laboratory simulation of alternatives.[5] At the root of this approach were the research methods and empirical orientation of experimental psychology.

The activities of Karlin's department in the early 1950s included studies comparing the use of seven-digit telephone numbers, termed "All Number Calling," with the two-letter, five-digit numbers in use at the time.[6] Both the telephone companies and customers were concerned that this change, designed to increase the number of available telephone numbers, would make dialing more of a problem. Laboratory studies showed that dialing performance was slightly faster with all digits, although long-term memory for numbers appeared to be slightly reduced. The small size of these effects led to the overall prediction, later confirmed, that All Number Calling would not impair customer dialing.[7] Other projects included design of dials for the 500-type telephone set (the little dots in the dial make dialing significantly faster, because they make it easier to see when the dial has stopped rotating from the previous number dialed). There were some futuristic studies demonstrating the feasibility, from the customers' viewpoint, of machine recognition of spoken digits (people with speech difficulties actually did better when they thought they were talking to a machine), and effects on customer perceptions of transmission quality of Time-Assigned Speech Interpolation (TASI)—a procedure for using the natural pauses in a telephone conversation to send bits of another conversation over the same line.

3.1 Military work

By the mid-1950s, word of the success of Karlin's enterprise had reached Bell Laboratories military development organization at Whippany, New Jersey, which was dealing with some complex control systems with significant human factors implications. At that organization's request, M. Paul Wilson was transferred from Karlin's department to organize a military human factors effort.

In complex systems, the division of responsibility between human operators and hardware called for some creative new analytic techniques. In a series of projects, each characterized by complex command-control systems—such as the XM3-H tactical radar: the Nike-Ajax, Nike-Hercules, and Nike Zeus missile systems; and the SAGE air defense system—Wilson and his colleagues developed this new approach to command/control systems. Two major characteristics were:

1. The emphasis on the human's role as decision maker in the system; and

2. The importance of information displays in conveying critical information required for correct and timely decisions.

Shortly thereafter, Wilson became head of the newly formed Command Systems Department, with a single human factors group supervised by W. L. Lee, another former Karlin employee. In the mid-1960s, this group was supervised by Robert I. Rosenthal, who extended the Wilson formulations to the design of control panels and display systems.

One of the interesting developments from this organization was the design of a network control center for the AUTOVON military communications system.[8] It was significant for two reasons. First, the human factors specialists conceived and designed the system rather than serving as consultants. Second, it introduced the concept of "exception reporting"—limiting displayed information to unusual or abnormal conditions that the system operator needs to know. As a result, system operators could focus on analyzing and solving problems, the most critical job and the one they could do best, while the computer's ability could be used to sort through large quantities of data to find abnormal conditions. This work exemplified the kinds of things the military human factors people could do for the Bell System. In fact, Rosenthal's group later designed the display system for the No. 4 Electronic Switching System (ESS) electronic toll switching machine using the principles established in military work.

Human factors studies had a significant part in notable government-sponsored projects that included the Safeguard Antiballistic Missile System and the Bellcomm company that was organized to provide technical support for NASA's Apollo project. Work on the antiballistic missile system began in 1967 with a human factors system evaluation at White-Sands, New Mexico, and continued with Arthur S. Kamlet's Command and Control Test Facility.[9] Human factors at Bellcomm was led by A. N. Kontaratos and focused on basic studies of problems of extended isolation and the dynamics of small technical groups. Some of this work, led by Nicolas Zill, used Tektite II, a small undersea laboratory where groups of scientists worked for extended periods. The Bellcomm effort was quite independent of other Bell System human factors activities.

By the mid-1970s most of the military work had been phased out at Bell Laboratories and the people had been reassigned to telecommunications projects, bringing with them the command/control system approach and experience in broader aspects of system design.

3.2 Sibyl, station equipment, and network applications

About the time Wilson left in the mid-1950s, the Karlin organization began development of a new, complex system for conducting simula-

tion studies. This new system, called Sibyl, could be programmed to insert a wide range of network transmission and switching impairments into real telephone calls, record user-calling behavior, and record user-preference judgments.[10] Up to 100 Bell Laboratories volunteers could be connected to the system.

Sibyl's first use was in the evaluation of push-button vs. rotary dialing. It recorded dialing times, intervals between digits, and errors. In later applications, Sibyl varied factors such as dial tone delay, network blocking, or delay after dialing. Following each test call, Sibyl would call the user back to get a subjective evaluation. Users would dial one of several digits to express their opinions. Knowledge of customer opinions of various levels of service enabled designers to engineer the telephone network to satisfy customers without spending money on improvements that customers did not care about.

In 1966, Sibyl and other transmission quality activities were consolidated in a separate department under Richard W. Hatch. In 1970, psychologist Herman R. Silbiger was chosen to supervise a new human factors group in Hatch's department, specializing in subjective evaluation of network performance. Although it has undergone much modification Sibyl is still in use.[11]

Karlin's department also continued its work in customer instructions (design of flowchart dialing instructions for centrex users), transmission (defining the optimal parameters for trading off echo suppression with the listener's ability to interrupt, especially on long-delay satellite connections) and telephone set design. Particularly notable was Richard L. Deininger's work on the design of the *Touch-Tone** telephone. Most of the telephone set design work was carried out in cooperation with the Station Instrumentation Department (headed up by Alfred H. Inglis, and then Harris F. Hopkins), and the *Touch-Tone* telephone design was no exception. Deininger's responsibility was to determine the optimal parameters of the push-button dial, particularly the arrangement of the 10 buttons.

Deininger's studies led to the selection of the now ubiquitous *Touch-Tone* telephone dialing arrangement. His studies showed that it was strongly preferred over the now equally familiar calculator arrangement.[12]

Karlin's department also worked jointly with the Station Instrumentation people to design the *Trimline** telephone.[13] Early efforts to design a dial-in-handset telephone had been unsuccessful because the sets were uncomfortably bulky and unattractive. Invention of the space-saver dial by Charles F. Mattke enabled Lionel W. Mosing of Karlin's department to fashion a comfortable and attractive telephone

* Trademark of AT&T.

of angular design, which he called the *Trimline* telephone. The final version of this set is now installed in millions of homes and appears on display in the design collection of New York's Museum of Modern Art. Much additional customer work was conducted primarily by engineers in the Station Instrumentation department. In 1964 this work was inherited by Leo Schenker's Telephone Station Studies Department, and Deininger was named to supervise the human factors work.

3.3 Station studies

For several years, Deininger's group was heavily involved in design and evaluation of the *Picturephone** visual telephone service, particularly aspects of image and quality and camera placement.[14] In 1970, after Deininger's untimely death, the group was managed by Gaber P. Torok, who worked to bring human factors work closer to the product-development activity, speeding up the development cycle by reducing the need for lengthy field testing and subsequent modifications.

In 1973, Murray J. Katz replaced Torok as supervisor of this group. Advancing technology, increasing customer demand for new products and the Bell System's entry into a more competitive environment all led to steadily increasing demand for the services of Katz's group. As a result of Katz's human factors evaluations, modifications were made to many new station products including the *Touch-a-matic*† repertory dialer (changes in labels, button characteristics, handset location), *Design Line*† decorator telephones (modifications to improve transmission characteristics, balance, and comfort), and *Dataphone*† II data communications terminal (design of maintenance panels and instructions). They also conducted extensive studies of the effects of mobile telephone usage on driving behavior[15] and worked jointly with Karlin's department on design of calling procedures for the Advanced Mobile Phone Service (AMPS).

Throughout the 1970s, station design work was changing from largely physical design ("knobs and dials" as it is sometimes called) to more procedural or feature-oriented work. Human factors specialists were addressing questions of how to implement new and complex features for residence and business customers. There was also a growing awareness of the role of human factors in identifying and satisfying customer needs. As a result, human factors was playing a larger role in the product design process. In the late 1970s, increasing demand for human factors support led to the formation of new groups in both station and business terminal development.

---

* Registered service mark of AT&T.
† Trademark of AT&T.

3.4 Repositioning of human factors

In the mid-1960s, Karlin's department addressed a wide variety of customer issues such as the potentially disruptive effects of satellite delay on telephone conversations,[16] design of flowchart instructions, and diagnosis of network impairments. The problem with diagnosing network troubles was to correlate customer descriptions of problems with the problems, themselves. The solution was to create the same network impairments in the laboratory and ask test customers to describe the problems. Once the customer descriptions were tied to the difficulties, customer complaints could be used to pinpoint and repair the troubles.

Feasibility studies were also conducted to see whether business customers, if given computerized switching systems, could successfully program their own rearrangements and changes, e.g. reassignment of telephone numbers and features when employees moved offices. Such changes tend to occur frequently, and can be costly to both the customer and the telephone company if an installer must be dispatched to make them. It was found that, with the right user interface, customers could make their own changes with little difficulty, so long as they had a way to trace and correct their errors. An added benefit of this capability, which now exists in a number of Bell System customer switching systems, is that customers can make changes instantly, without having to wait for a telephone company employee to be dispatched.

The Human Factors Department had started in the research area and still viewed itself largely as an applied research department, but by 1970 it found itself in a development organization which was interested in increasing the direct, near-term payoff of human factors. Where the department had previously been organized around psychological topics (visual studies, interpersonal communication, etc.) it was asked to realign along Bell System lines (e.g., operator services, loop and outside plant). Along with this realignment, groups were given the responsibility for identifying and addressing existing human factors opportunities in their assigned areas and for contacting the appropriate organizations to discuss them.

During this same period there was a growing concern throughout Bell Laboratories about the impact of technology on customers and employees. While there is question as to cause and effect, the growing sensitivity to the impact of new products on customers and employees and the refocusing of human factors research were followed by remarkable growth in the human factors area. Within two years, Bell Laboratories "consumers" of human factors work were negotiating for long-term commitments of human factors support and were exploring the option of starting their own human factors efforts to guarantee

adequate staffing of their work. By 1976, new human factors efforts had begun in both operator services and loop operations, and a director-level committee, conceived and led by Herbert M. Zydney, was assembled to plan an orderly expansion strategy.

The resultant plan, known as the Crowley-Dorros plan after the two executive directors who commissioned it—Thomas H. Crowley and Irwin Dorros—outlined a strategy for all human factors work at Bell Laboratories, including both customer and employee issues. The plan called for Karlin's department to focus on customer issues and to serve as a centralized resource to a growing number of satellite human factors groups that would be closely tied to specific development activities. The plan also viewed Karlin's department as a training ground for individuals to start such satellite groups.

As a case in point, an entire group, led by Edmund T. Klemmer, was transferred from Karlin's department to continue its work on business customer telephone systems. The group started with five people and an emphasis on procedures and instructions, and has since grown to two groups working on all aspects of business services, from customer needs studies to field introduction.[17]

In 1977, Karlin retired from Bell Laboratories after 32 years of service. (See Ref. 18 for Karlin's entertaining and insightful parting comments on the state of human factors at Bell Laboratories.) He was replaced by Charles B. Rubinstein, an electrical engineer from the Research Area who had done work in the psychophysics of color perception and visual thresholds. Shortly after Rubinstein arrived, Robert I. Rosenthal, who had supervised military control-display work under M. Paul Wilson and later worked on network management display systems, joined the department. In the late 1970s, the department's work included design of customer dialing procedures for AMPS,[19] and other new services that relied on recorded instructions to customers, such as the Voice Storage System,[20] the Automated Credit Card Service,[21] and Automated Coin Telephone Service (ACTS).[22] Each of these services depended on recorded instructions that could explain to customers what they were to do next. Computer-controlled voice recording and playback equipment enabled human factors people to conduct extensive laboratory simulations of these services, identifying deficiencies in procedures or instructions that led users to make mistakes. ACTS, for example, uses recorded messages to tell coin telephone customers how much money to deposit in the telephone. Laboratory and field studies were required to determine how long to wait for customers to deposit coins, and how often to prompt them for the remainder. Time limits had to be established to ensure that a human operator could come on the line to assist customers having difficulty. Customer satisfaction was also monitored to

ensure that people would be comfortable with the less personal approach to coin service. As a result of the attention to human factors, ACTS has met with complete success. Customers have few difficulties with the service, and customer acceptance has been outstanding.

Recently a new focus has emerged. Bell Laboratories is developing new, complex products whose success will depend on the quality of the user interface. As a result, Rubinstein's department and other customer-oriented human factors groups are directing their efforts to human/computer interface issues.

IV. THE RESEARCH CONNECTION

4.1 Behavioral research beginnings

The second path in the evolution of applied behavioral sciences at Bell Laboratories began in the early 1950s at AT&T. The Personnel Relations department had been doing applied field work for some time, but had become frustrated by lack of knowledge of the social processes which influenced organizational success. Bolstered by a 1953 commitment by the AT&T Board of Directors to attracting and developing capable employees and first-rate leaders, Robert K. Greenleaf, Director of Personnel Research at AT&T, asked Bell Laboratories to start a new research group, the Communications Social Science Research Department.

The new department was led by Robert N. Ford who transferred from AT&T Personnel Relations, and its charter was to focus on problems of communication, organization, and leadership in small groups. The department was established at Bell Laboratories because BTL had experience in managing basic research and because of the stimulation other Bell Laboratories research activities could provide. Ford was to report directly to William O. Baker, then Vice President and Director of Research. To help get the effort under way, and to help recruit good research people, psychologist Carl I. Hovland was brought in from Yale as a consultant.[23]

The first person hired was Alex Bavelas, whose specialty was human motivation. Bavelas left after only a short time to join the Stanford faculty. Morton Deutsch, another social psychologist, was the second to arrive, and stayed for several years. Both Bavelas and Deutsch went on to establish major reputations in the academic world.

In addition to this venture in social psychology, efforts were begun in experimental and educational psychology. The experimental psychology activity began in 1957 with Herbert Jenkins (a disciple of Harvard behaviorist B. F. Skinner), who established a lab to study learning in pigeons. He was followed by Ernest Z. Rothkopf, an educational psychologist specializing in learning and instruction, and

Roger Shepard, another Harvard experimentalist recruited by Hovland and specializing in information processing.

In 1958, H. O. (Ollie) Holt, an educational psychologist who had been Deputy Director of George Washington University's Human Resources Research Office (HumRRO), took over operation of the center from Ford, who returned to AT&T to pursue his interest in job enrichment. Holt immediately began to develop a plan for entering the new field of programmed instruction. He worked primarily with consultants such as Hovland, Tom Gilbert from the University of Tennessee, and Skinner, himself. Given the massive sums of money spent by the Bell System in training its employees, the payoff from such work was potentially enormous.

This work showed such promise that Holt was transferred to AT&T in 1962 to head the Office of Training Research which had the job of implementing individualized instructional technology throughout the Bell System.[24] It is Holt who provides the ultimate connection to applied behavioral sciences, so the path of applied work follows him. But the Behavioral Research organization, with its worldwide reputation for excellence, deserves a short history of its own.

4.2 Basic research in psychology

Research on the social psychology of group interaction was the initial focus of the Behavioral Research Center. The problems ranged over issues of cooperation, influence, and social perception and the communicative behavior entailed. The amount of work in these areas gradually diminished, and the original social psychologists had all left by 1966. However, social psychology reemerged from 1978 to 1982 with a new central concern in interpersonal communications. The issues now involved the reasons why people choose one communication modality over another (e.g., voice-only versus face-to-face) and the consequences. This work was headed by Myron Wish, whose involvement in a study of the use of *Picturephone* visual telephone service sparked renewed interest in such matters.

There were other topics that, like group psychology, were pursued only for limited periods. Most notable of these were efforts in auditory neurophysiology between 1964 and 1968, and the psycholinguistics of grammar from 1963 to 1971.

By contrast, the research in human learning and instructional technology, which was under way by 1958, has continued, expanded, and diversified up to the present. For example, Ernst Z. Rothkopf and Lawrence T. Frase conducted a large number of experiments on the effect of adjunct questions on learning from text. These not only showed practical means for greatly improving instruction, but also deepened understanding of the active role of learners in studying.

Pioneering studies of the role of organization (e.g., the placement of repetitions of information) and surface structure (e.g., word choice) in learning from written prose led, among other things, to the development of computer methods for determining text difficulty (see Ref. 25). The fundamental work on instructional materials contributed to the development of widely used editorial aids such as the *UNIX** Writer's Workbench software (see Refs. 26, 27, and 28), and AT&T's Training Development Standards, a guide used in course development throughout the Bell System.

Two other areas that have continued since their establishment in the 1950s are human information processing and psychological measurement. Human information processing research at Bell Laboratories goes at least as far back as the work of John R. Pierce and John Karlin on estimating human channel capacities[29]; but as a continuous institutional activity, it can be dated from George Sperling's discovery and description of a visual memory that stores, for about a second, considerably more information than can be read out intact.[30] Saul Sternberg's elegant work on rapid scanning of active symbolic memories followed shortly.[31] He and his collaborators also made important advances in the study of temporal order judgments, and most recently, the control of rapid action sequences in speech and typing.

The Human Information-Processing Research Department, headed by Sternberg since 1970, has also been the home of pioneering work and major contributions in many other parts of cognitive psychology: verbal learning, picture memory, semantic memory, word recognition, visual perception, motor control (see article by Rosenbaum,[32] this issue) and adaptation, attention, and reasoning. In 1979 a new group was formed under Tom K. Landauer's leadership to explore cognitive problems in human use of computers. Its work has centered on issues in interactive language design and the proper representation of human knowledge to assure mutual understanding of the messages passed between machines and users (see article by Furnas et al., this issue).[33]

Measurement and scaling were among Roger Shepard's many interests when he first joined the Labs in 1958. Along with Joseph B. Kruskal, he developed the first effective methods for nonmetric multidimensional scaling of human similarity judgments.[34] There has been an almost continuous development of new theory, computational methods, and applications ever since. For example, in recent years J. Douglas Carroll and his collaborators have developed the method of individual differences multidimensional scaling, a method that identifies psychological dimensions by their orderly variations in perceptual importance as reflected in judgments of object similarity by

* Trademark of Bell Laboratories.

different people. These methods have found extensive use in such diverse areas as marketing research and telephone tone-ringer design.[35]

Over the last 20 years, the Research division has also sponsored a variety of other efforts with behavioral science content. There has been much noteworthy work in vision, color vision, and visual perception under the leadership of Bela Julesz. One example is his own invention and use of random-dot stereograms to demonstrate the independence of depth perception from object recognition and his explorations of fundamental mechanisms of perceptual processing. (See Ref. 36.) Speech analysis and evaluation research was conducted in a department headed by Peter B. Denes, and elsewhere. Linguistic research has been under way since 1971 in a department led by Osamu Fujimura, who himself has studied the production of motor action sequences, an interest shared recently by several other investigators. Finally, mention should be made of the important fundamental analyses of speech acoustics and perception carried out by James F. Flanagan's Acoustics Research Department and Manfred R. Schroeder's Hearing and Speech Synthesis Research Department.

Over the years there has been fairly regular communication between the behavioral research groups and the applied behavioral scientists. There have been many seminar series attended by both, frequent consultations and visits, and since 1977 a Bell Laboratories-wide convention of all behavioral scientists every 18 months. There are occasional cross-assignments of months or a year's duration, and a steady trickle—averaging perhaps one per year—of people moving from one kind of work to the other; most movement has been from the smaller population of research to the larger applied areas. There has been a great deal of useful stimulation in both directions.

5. HUMAN PERFORMANCE TECHNOLOGY

5.1 Human performance and employee systems

The Bell operating telephone companies, with their huge billing and accounting systems, were among the first businesses to make widespread use of computers in their operations. During the late 1950s and early 1960s, computers became an integral part of their accounting operations. While the development of computerized systems had resulted in substantial savings and better service, there had also been some difficulties, largely because of a mismatch between such systems and the employees who worked with them. The initial solution to these problems was thought to be in better training, so Ollie Holt and Harry A. Shoemaker of AT&T's Office of Training Research were asked to get involved.

At about the same time, it became clear that centralized planning

and development of such systems would be more efficient and would ensure system-wide standards. Finally, in 1967, largely through the efforts of AT&T Assistant Vice President Bruce Warner, the Business Information Systems Programs (BISP) organization was created from parts of the AT&T Planning Department. Holt was asked to join the new organization as Director of the Human Performance Technology Center. After several months Holt moved, along with the whole BISP organization, to Bell Laboratories.

Holt brought with him several of his former staff members including Fred L. Stevenson, a training specialist originally with Pacific Telephone. The group's early experience showed that training was not going to be a universal cure for design deficiencies, so they set about to develop an organization that could influence system design.

One of the first additions to Holt's staff was Bill F. Fox, a psychologist with human factors training and experience. Fox had worked with HumRRo and Lockheed Aircraft. Fox's assignments were to adapt military experience with large hardware systems to the development of large software systems and to build a team of psychologists to do the work.

Holt also recruited Ralph Marion, a psychologist and personnel specialist from Southern Bell. Marion's assignment was to develop documentation that would enable system designers to do as much of the job as possible, themselves. Together, they continued to recruit both operations people from the Telephone Companies and behavioral scientists. By the end of 1970 they numbered over 40.

They called their work personnel subsystems (PSS), a phrase with military roots, rather than human factors roots, a term which they felt referred narrowly to control-display design rather than systems concerns. In later years, the name was changed to human performance engineering to more clearly characterize the major objective. In 1970 the Human Performance Center was organized into three departments. Fox's Personnel Subsystems Department provided direct consultation and support to projects and worked to develop PSS technology; Marion's Personnel Subsystem Analysis and Development Department refined the technology and documented it; and Stevenson's Systems Training Department trained others in the use of the technology.[37] This organization worked well, and the technology, documentation, and training it developed are in use in information systems organizations throughout the Bell System.[38]

In many ways Fox's department was the BISP homologue of John Karlin's Human Factors Department. Both were staffed primarily by behavioral science professionals who provided consultation to system designers and developers, and both experienced significant successes in the 1970s. The primary difference between them was in their way

of approaching the job. The Karlin approach was to solve each design problem empirically; the Fox approach was to assist the project in utilizing the PSS technology to solve its own problems. In each case, the approach was appropriate to the problem—Karlin dealt with a number of specific problems, each with significant implications for many telephone users, while Fox was faced with vast needs for job requirements, training, etc., which could not have been efficiently done by behavioral scientists alone.

Typically, Fox's people would be assigned to projects as technical advisors to the project people who had design responsibility. For the most part, these project people were telephone company people on rotational assignment who had no prior experience with PSS. These people used the documentation produced by Marion's department and were trained in Stevenson's. The success of this work and the desire of the project organizations to have complete responsibility and guaranteed support led to the spin-off of separate PSS groups. In 1973, George A. Schweickert and Mort H. Kahn left Holt's center to form groups in Business Information System Customer Service Facilities Assignment and Control System (BISCUS/FACS).[39] Shortly after, the organization involved in the Trunks Integrated Record Keeping System (TIRKS) formed its own group under Roy J. Porterfield.[40] Barry K. Schwartz also left to work on systems engineering for network administration systems and later formed a group to work on the Total Network Data System (TNDS). By 1976, most of the projects which had started as part of BISP had their own groups of specialists, and few of Fox's people were assigned to specific BISP projects. Instead, many of them were working on user aspects of Operations Support Systems (OSSs) developed elsewhere in Bell Laboratories to aid in the operation and maintenance of various aspects of the Bell System network. The first of these to come to Fox's attention was the Loop Maintenance Operations System (LMOS).[41] By 1976, one of Fox's groups, led by Grace H. Leonard, had transferred to the LMOS organization.

In late 1976, the Crowley-Dorros plan—which had called for Karlin's department to serve as a centralized customer human factors resource—named Fox's department as the resource for employee human performance activities, serving as both consultants and the source of trained people to start satellite groups. Since that time, a number of new groups have been created in LMOS and other OSS projects, employing the methods and procedures developed in the late 1960s by Holt and Fox and their colleagues.

The growth of groups and activities has continued unabated through the late 1970s and early 1980s. Again the increasing complexity resulting from new technology has placed new burdens on employees

and provided new opportunities and challenges for behavioral scientists.

VI. SUCCESSES. . .AND CHALLENGES

Measured by its impact and its growth, human factors has had notable success at Bell Laboratories, particularly since the early 1970s. Both the empirical orientation of Karlin and the technological approach of Holt and Fox have had significant impact. Why has human factors been a success at Bell Laboratories when external efforts in both industry and the military have fared less well? There are several possible explanations.

First, the philosophy and structure of the Bell System have been important. Its strong tradition of good customer service implies a need to give explicit consideration in the design process to the customers. The vertical integration of the system encouraged Bell Laboratories to design products and systems which are not only attractive at time of purchase, but which also continue to perform cost-effectively. In such an environment, design goals such as minimizing customer errors or increasing employee efficiency and decreasing turnover become more important. The numbers of employees and customers potentially affected by new designs also multiply the benefits that result from human factors.

The growing complexity of technology has also played a significant role. While technological advances have provided many new products and services, these products and services have become more complicated for the customers who use them and the employees who install and maintain them. Computerization of internal operations has led to substantial complications in some aspects of employee jobs. Human factors specialists have the responsibility to minimize the problems imposed by this complexity and to make new products and services "friendlier," more useful, and more attractive.

In addition to the technical challenges of the information age, the major challenge to human factors in the 1980s is to continue to expand its role in the design/development process. This implies both applying existing human factors skills in new areas and developing new skills which complement existing ones. For example, extension of human factors methods to the definition of customer needs for new products will lead to earlier and more influential involvement in product design. Development of systems engineering skills will result in improved communication with engineers and greater influence over design decisions. Specialized knowledge in electrical engineering and computer science will enable evaluation of the trade-offs that must often be made between human factors and hardware and software constraints. With a broader view of its responsibilities, human factors should continue to grow and contribute at Bell Laboratories.

VII. ACKNOWLEDGMENTS

The author would like to thank the many people who provided information and insights for this history. I am particularly indebted to H. O. Holt for his encouragement and perspective and to W. F. Fox and T. K. Landauer for their generous assistance. Thanks are also due to J. E. Karlin, M. J. Katz, and R. I. Rosenthal for the significant background material they provided. Special thanks go to B. L. Lively for resolving the many thorny issues generated by a project of this nature.

REFERENCES

1. G. Homans, "Group Factors in Worker Productivity," in E. Maccoby, T. Newcomb, and E. Hartley, eds., *Readings in Social Psychology*, 3rd ed., New York: Holt, Reinhart and Winston, 1958, pp. 583–95.
2. W. C. Jones and A. H. Inglis, "The Development of a Handset for Telephone Stations," B.S.T.J., *11,* No. 2 (April 1932), pp. 245–63.
3. H. Fletcher, "The Nature of Speech and Its Interpretation," B.S.T.J. *1,* No. 1 (July 1922), pp. 129–44.
4. H. Fletcher and W. A. Munson, "Loudness, Its Definition, Measurement, and Calculation," B.S.T.J., *12,* No. 5 (October 1933), pp. 377–430.
5. R. R. Riesz and H. D. Irvin, "Simulation in Engineering," Bell Lab. Rec., *36,* No. 7 (July 1958), pp. 238–41.
6. J. E. Karlin and R. K. Potter, "Preference Research," Bell Lab. Rec., *32,* No. 5 (May 1954), pp. 161–6.
7. J. E. Karlin, "All Numeral Dialing, Would Users Like It?" Bell Lab. Rec., *36,* No. 8 (August 1958), pp. 284–8.
8. J. W. Gorgas, "AUTOVON, Switching Network for Global Defense," Bell Lab. Rec., *46,* No. 4 (April 1968), pp. 106–11.
9. A. S. Kamlet and J. D. Musa, "Human Factors Testing of Computerized Systems," Bell Lab. Rec., *52,* No. 8 (September 1974), pp. 245–51.
10. "Simulating Speech Through Space," Bell Lab. Rec., *38,* No. 8 (August 1960), pp. 296–8.
11. J. L. Sullivan, "Is Transmission Satisfactory? Telephone Customers Help Us Decide," Bell Lab. Rec., *52,* No. 3 (March 1974), pp. 90–8.
12. R. L. Deininger, "Human Factors Engineering Studies of the Design and Use of Pushbutton Telephone Sets," B.S.T.J., *34,* No. 4 (July 1960), pp. 995–1012.
13. C. L. Krumreich and L. W. Mosing, "The Evolution of a Telephone," Bell Lab. Rec., *44,* No. 1 (January 1966), pp. 8–14.
14. R. R. Stokes, "Human Factors and Appearance Design Considerations of the Mod II PICTUREPHONE Station Set," IEEE Trans. Communic. Tech., *COM-17,* No. 2 (April 1969), pp. 318–23.
15. A. J. Kames, "A Study of the Effects of Mobile Telephone Use and Control Unit Design on Driving Performance," IEEE Trans. Vehicular Tech., *VT-27,* No. 4 (November 1978), pp. 282–7.
16. E. T. Klemmer, "Human Factor Problems in Satellite Telephoning," Human Factors, *8* (December 1966), pp. 475–80.
17. S. H. Ellis and R. A. Coskren, "A New Approach to Customer Training," Bell Lab. Rec., *57,* No. 2 (February 1979), pp. 60–6.
18. J. E. Karlin, "The Changing and Expanding Role of Human Factors in Telecommunications Engineering at Bell Laboratories," Eighth International Symposium on Human Factors in Telecommunications, Churchill College, Cambridge, England, September 1977. Reprinted in J. P. Duncanson, ed., *Getting It Together: Research and Applications in Human Factors,* Proceedings of a Symposium sponsored by the Metropolitan Chapter of the Human Factors Society, Stevens Institute of Technology, Hoboken, New Jersey, November 17, 1977.
19. B. L. Hanson and C. E. Bronell, "Human Factors Evaluation of Calling Procedures for the Advanced Mobile Phone System (AMPS)," IEEE Trans. Vehicular Tech., *VT-28,* No. 2 (May 1979), pp. 126–31.
20. B. L. Hanson, R. J. Nacon, and D. P. Worrall, "New Custom Calling Services," Bell Lab. Rec., *58,* No. 6 (June 1980), pp. 174–80.

21. J. O. Hardy, D. G. Raj-karne, and K. A. Raschke, "Handling Coin Toll Calls—Automatically," Bell Lab. Rec., *58,* No. 8 (September 1980), pp. 256–62.
22. M. R. Allyn, T. M. Bauer, and D. J. Eigen, "Human factors in the design of a new service," Bell Lab. Rec., *58,* No. 5 (May 1980), pp. 155–61.
23. C. I. Hovland, "Two New Social Science Research Units in Industrial Settings," American Psychologist, *16* No. 2 (February 1961), pp. 87–91.
24. H. O. Holt, "Programmed Self Instruction," Bell Telephone Magazine, *42* (Spring 1963), pp. 24–7.
25. E. U. Coke and M. E. Koether, "A Study of the Match Between the Stylistic Difficulty of Technical Documents and the Reading Skills of Technical Personnel," B.S.T.J., this issue.
26. L. T. Frase, "The *UNIX*™ Writer's Workbench Software: Philosophy," B.S.T.J., this issue.
27. N. H. Macdonald, "The *UNIX*™ Writer's Workbench Software: Rationale and Design," B.S.T.J., this issue.
28. P. S. Gingrich, "The *UNIX*™ Writer's Workbench Software: Results of a Field Study," B.S.T.J., this issue.
29. J. R. Pierce and J. E. Karlin, "Reading Rates and the Information Rate of a Human Channel," B.S.T.J., *36,* No. 2 (March 1957), pp. 497–516.
30. G. Sperling, "Successive Approximations to a Model for Short Term Memory," Acta Psychologica, *27* (1967), pp. 285–92.
31. S. Sternberg, "Memory Scanning: Mental Processes Revealed by Reaction-Time Experiments," American Scientist, *53,* No. 4 (1969), pp. 421–57.
32. D. A. Rosenbaum, "Central Control of Movement Timing," B.S.T.J., this issue.
33. G. W. Furnas et al., "Statistical Semantics: Analysis of the Potential Performance of Key-Word Information Systems," B.S.T.J., this issue.
34. R. N. Shepard, "The Analysis of Proximities: Multidimensional Scaling With an Unknown Distance Function," Psychometrica, *27* (1962), pp. 219–46.
35. J. D. Carroll and M. Wish, "Measuring Preference and Perception with Mathematical Models," Bell Lab. Rec. *49,* No. 5 (May 1971), pp. 146–54.
36. B. Julesz and J. R. Bergen, "Textons, the Fundamental Elements in Preattentive Vision and Perception of Textures," B.S.T.J., this issue.
37. F. L. Stevenson, "An Individualized Approach to BISP Training," Bell Lab. Rec. *52,* No. 9 (October 1974), 271–78.
38. H. O. Holt and F. L. Stevenson, "Human Performance Considerations in Complex Systems," Science, *195* (March 18, 1977), pp. 1205–209.
39. J. J. Yostpille, "BISCUS/FACS Processes Service Orders Automatically," Bell Lab. Rec. *55,* No. 4 (April 1977), pp. 96–102.
40. J. W. Timko, "TIRKS Takes the Paperwork Out of Trunk Recordkeeping," Bell Lab. Rec. *53,* No. 9 (October 1975), pp. 368–75.
41. J. J. Appel and J. L. Woodruff, "Streamlining Loop Operations," Bell Lab. Rec., *56,* No. 6 (June 1978), pp. 147–54.

AUTHOR

Bruce L. Hanson, B.A. (Psychology), Johns Hopkins University; M.S., Ph.D. (Psychology), Pennsylvania State University; Bell Laboratories, 1972–1982; American Bell, 1983—. After joining Bell Laboratories in 1972, Mr. Hanson worked on the human factors aspects of central office craft jobs. He also has helped evaluate hardware, instructions, and procedures for outside plant craftspeople, engineers, and managers, and he has studied human factors in new services such as the Advanced Mobile Phone Service. In 1977, he became Supervisor of the Network Systems group, designing user interfaces for voice storage services. Before he assumed his present position, he also supervised studies of customer aspects of billing, PhoneCenter® and Business Office service. Mr. Hanson is currently Supervisor of the Customer Interfaces Group in the Product Family Services Planning Department and ABI. This group is engaged in the design and planning of user-computer interfaces for ABI products. Member, Human Factors Society, the American Psychological Association, and Sigma Xi.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Methods for Field Testing New Telephone Services

By D. J. EIGEN*

(Manuscript received December 29, 1981)

Telephone services are increasingly complex and diverse, and they require more human-machine interaction than ever before. A field test can help improve a new service by ensuring that it is easy to use with little chance for error. This paper discusses the methodology of field testing. Specially tailored telephone service evaluation methods, based on field test experience with the Calling Card Service, are presented in detail.

I. INTRODUCTION

New telephone services involve more customer-system interaction than ever before, and making the use of these services easy and error-free is a major goal of their development. Properly designed dialing plans, announcements, timings, tones, and instructions increase customer acceptance, minimize customer errors, and promote use of the service. The design of new services can be evaluated by coordinated studies that include:

- Analysis of present services,
- Interviews with customers,

* Bell Laboratories.

- Laboratory studies of proposed protocols for customer-system interactions,
- Field tests of services, and
- Product follow-up studies.

Field testing is the largest and most costly step in this coordinated set of studies. This paper discusses the methodology of field testing, using a field test of the Calling Card Service as an example.

II. CALLING CARD SERVICE FIELD TEST OVERVIEW

An analysis of operator-handled credit card service indicated that a reasonable proportion of credit card calls could be automated. Interviews with credit-card, bill-to-third-number, and collect customers verified their interest in and need for an automated Calling Card Service, and laboratory studies provided evidence that customers could use the Calling Card Service successfully.

The Calling Card Service field test was conducted in Milwaukee from November 1977, to June 1978.[1,2] Regular telephone credit card numbers could be dialed in order to place Calling Card calls from about 3000 noncoin phones in the Milwaukee area and from 70 coin phones at Milwaukee's airport, two downtown hotels, and a few local restaurants. Bright orange placards on the trial coin phones instructed customers on how to use their telephone credit card number. In addition, operators were trained to assist callers and answer questions. (Calls from unequipped stations were handled as usual.)

To use the trial Calling Card Service, customers first dialed zero plus the number they wished to call. Special programs in the Traffic Service Position System (TSPS) routed incoming "0+" calls from trial stations to a small team of specially trained operators who helped simulate Calling Card Service—in actual service no operators are used. Besides the TSPS console, these operators had a video display terminal linked to a minicomputer (see Fig. 1).

When a call arrived from a specially equipped station, the trial operator notified the minicomputer, which then delivered a tone to prompt the customer to dial a Calling Card number. Detectors received the dialed digits and sent them to the minicomputer over a data link for verification. Calls with valid Calling Card or credit card numbers were allowed to proceed and were billed appropriately.

Depending on the protocol being tested, the minicomputer displayed step-by-step instructions on a terminal screen to guide the operator in handling each call. For example, to encourage customers to redial after making errors, the minicomputer might display the instruction, "Please hang up and dial zero plus the number you are calling," which was to be read to the customer. By making simple changes in the minicomputer program, the operator's treatment of calls could be

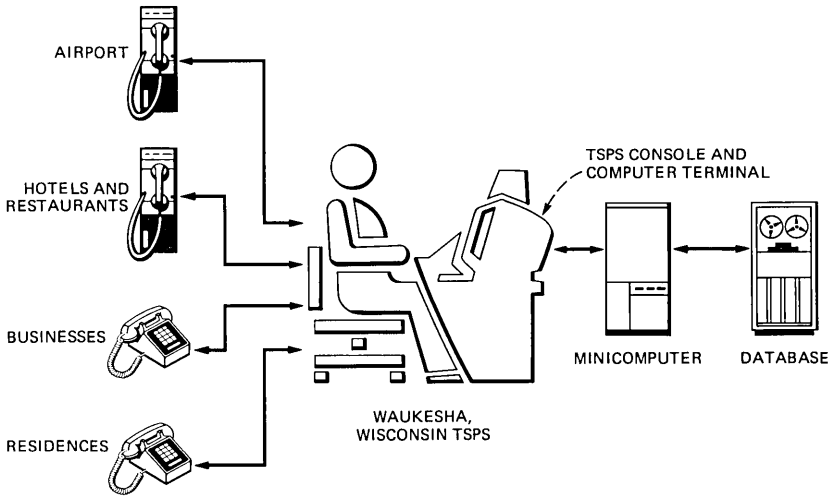


Fig. 1—Test setup.

altered, often without additional training. This flexible arrangement allowed easy testing of many different protocols and rapid changes among them.

The minicomputer recorded the time and details of each call. These records were analyzed rapidly to determine how the protocol could be improved. Throughout the trial, protocols were varied by changing announcements, timings, access to operators, error-correction procedures, and other aspects of the caller interface. In all, 24 variations of the protocol were tested at equipped coin phones, 14 at noncoin phones. Over 10,000 customers dialed more than 30,000 automated Calling Card calls during the trial and more than 5,000 customer interviews were obtained.

III. FIELD TEST ACTIVITIES

3.1 Field tests

In a field test such as that for Calling Card Service, a simulation of the proposed service is actually offered to a limited, but representative, set of customers on a trial basis. In some circumstances, when there is sufficient confidence in the form of the service, the actual product can be used as a test vehicle.

The field test can be used to adjust the technical and operational aspects of the service. It can be used to determine whether the service fits a customer need, and it can also be used to improve estimates of willingness to pay. And finally, it can be used to evaluate customer performance, satisfaction, and usage in the effort to provide a service

that optimizes the customer-system interaction on which new services rely.

The greater the fidelity of the test to the real service situation, the greater one's confidence in the final success of the service being tested. Field tests provide increased fidelity over more indirect techniques for evaluating services.

3.2 Study plan

Developing a study plan is the first task in planning a field test. A study plan must include the following five steps:

1. State the objectives of the study and delineate issues to be resolved by it. The primary objectives of the evaluation of the human-machine aspects of telephone services are to:
   a. Determine if usage, satisfaction, and performance are at acceptable levels.
   b. Predict the levels of usage, satisfaction, and performance in the final service.
   c. Refine the service to improve usage, satisfaction, and performance.

Many other detailed issues for a particular service may require resolution.

2. Determine constraints on study service and resources necessary to accomplish the test. Decisions on test methodology necessarily involve practical choices. For example, the marketplace often imposes serious time constraints on the development, deployment, and evaluation of a telephone service.

3. Design and refine the service and the human-machine interfaces involved. Some initial human-machine telephone service interface must be defined before the evaluation process can be initiated. Continued analyses, interviews, and laboratory studies are best used to generate and refine the service alternatives prior to the field test itself.

4. Delineate variables that may influence the results and hypothesize their interactions. Three categories of variables need to be specified for the field test:
   a. Independent variables—those variables that are to be deliberately manipulated or held constant. Among the possible variables of this type (with some examples) are the following:
      (1) Service protocol—announcements, tones, timings, error-handling strategies, and digit strings.
      (2) Capabilities of the service—billing, routing, and screening.
      (3) Availability—geographic constraints, time-of-day constraints, and station-type constraints.
      (4) Type of instruction for customers—media and format.
      (5) Marketing effort—promotion.
      (6) Rate—price structure.

b. Dependent variables—those variables whose values are affected by changes in the independent variables. Some examples are given below.
   (1) Subscription—initial interest and sign-up rate.
   (2) Usage—rate of first and repeated use.
   (3) Acceptance—judged worth and satisfaction.
   (4) Customer performance—speed, error, and abandonment rate.
c. Parameters—those identifiable variables that are free to vary. Among these are:
   (1) A priori condition—predisposition toward service, demographic mix of customer population, etc.
   (2) Internal characteristics—the test design or method used, intrinsic characteristics of the customer in the test, etc.
   (3) External characteristics—the geographic, environmental, and temporal setting of the test.

These variables are not necessarily mutually exclusive. Some parameters might be fixed or systematically varied, and, thus, be made independent variables.

It helps to have hypotheses about how these variables will interact. Figure 2 demonstrates one model of several possible models of the relationships among variables for telephone services. A different model might assume that instruction may affect acceptance and subscription more directly than is shown in Fig. 2.

5. Define methods for increasing confidence in the results of the study. Some of the most important aspects of a good study plan are motivated by the need to counteract rival explanations of the results

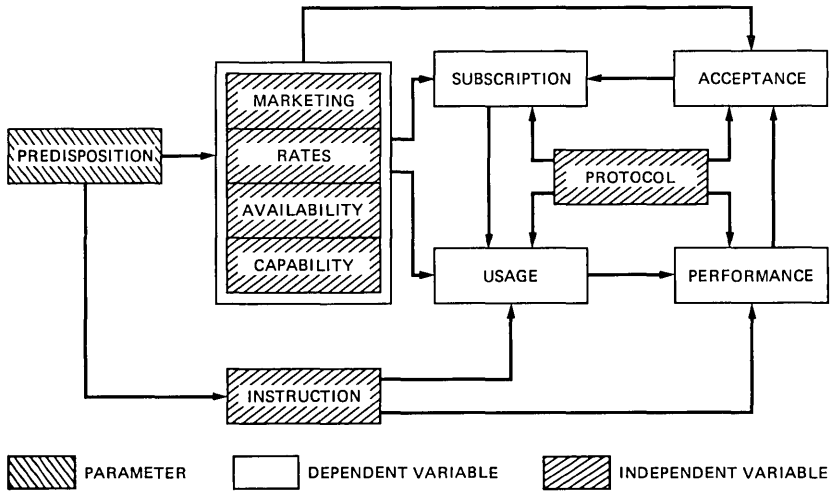


Fig. 2—Variable block model.

and, thereby, increase confidence in the validity of the field test as an indicator of the success of the service.[3–5] Section IV is devoted to this component of the study plan.

3.3 Test development

After the study is designed, the following tasks must be done.

1. Define data gathering tools.
2. Define data analysis techniques.
3. Define test service delivery vehicle.
   a. Define test service requirements.
   b. Design test service delivery vehicle.
   c. Implement service delivery vehicle (hardware/software development).
   d. Integrate trial system and test.
4. Acquire and prepare customers.
5. Acquire and prepare site(s).
6. Define test operations and procedures for utilizing results.

Figure 3 is an illustration of the interrelationships among these tasks.

3.3.1 Data collection

The following data collection methods were used in the Calling Card Service field trial.

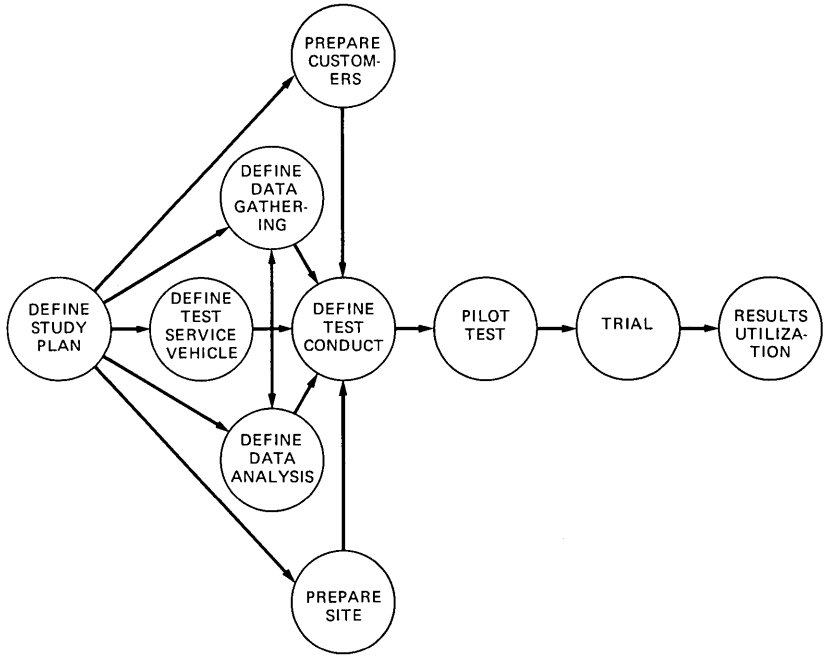


Fig. 3—Evaluation development paradigm.

3.3.1.1 Interviews/questionnaires. Interviews and questionnaires provide measures of attitude, as well as corroborative measures of usage, and can be given before, during, and after the field test. In-person, face-to-face interviews with customers who had used the Calling Card Service provided some of the more compelling evidence of the success of the service. The following spontaneous comments illustrate the range of such data.

Example of positive response:

1. "It's excellent. I compliment the phone company for coming through with this. Makes it a lot easier."

Example of negative response:

1. "In the past, you dialed Helen (operator) and said, 'I want to talk to Joe,' and right off Joe was on the line—it seems you are trying a lot of new services to get rid of Helen."

3.3.1.2 On-line measurements. These measurements are the machine recording of customer and system events, such as picking up a receiver or placing it back on the switchhook (hanging up), dialing digits, tones or announcements, error detection, and call progress states. By recording, time stamping, and storing these events, the sequence of customer-machine interactions can be reconstructed.*

3.3.1.3 Billing. Billing records routinely provide various data on a call, such as calling and called number, call duration, date, time of day, class of charge (e.g., collect, DDD, and so forth).

3.3.1.4 Observation. Observation provides a crude but valuable technique for gathering data that cannot be obtained by other means.

3.3.1.5 Customer feedback. Feedback mechanisms for customer complaints and suggestions are maintained by the telephone companies offering the service.

3.3.1.6 Records. Records and logs of equipment malfunction and events recorded in newspapers are important for detecting or counteracting erroneous or misleading results. Data in such records can be used to measure dependent variables.

These methods of collecting data are all relatively insensitive to one another in that a specific error in one measurement is very unlikely to affect another measurement. Thus, the methods can be used to corroborate each other.

---

* The Bell System as a common carrier is entitled to certain privileges to monitor the quality of its services. Bell Laboratories as an agent of the Bell System is extended these privileges. Only data necessary to ensure good service are gathered, and they are kept secure, in strict confidence, and statistically anonymous. To guarantee customer privacy, data are never gathered after a call is placed, i.e., after called party answers.

3.3.2 Analytic tools and techniques

The practice of feeding back the results as input to the experimental design process, while providing an efficient data gathering technique, places an additional burden on analysis. Results must be provided in a timely manner to be of use in formulating the next service improvements to be tested. Figure 4 illustrates a data processing stream used to field test the Calling Card Service.

Standard statistical routines must be used with caution since the assumptions on which they are based may be violated. For example, sampling is often nonrandom.

One heuristic is to leave the service unperturbed for some reasonable period of time (two or more weeks depending on the call volume). The extent of variations or noise in the data is measured and later used as a benchmark to assess meaningful variations potentially due to service manipulations.

3.3.3 Test service provisioning

A system to deliver a voice-prompted test service should include the ability to deliver tones and announcements, to route to an operator, and to receive dialed digits. Connection to the network, billing, and data processing and storage also need to be considered.

3.3.4 Acquire and prepare sample

Obtaining sufficient numbers of representative customers for the field test is important. The field test approach usually requires thou-

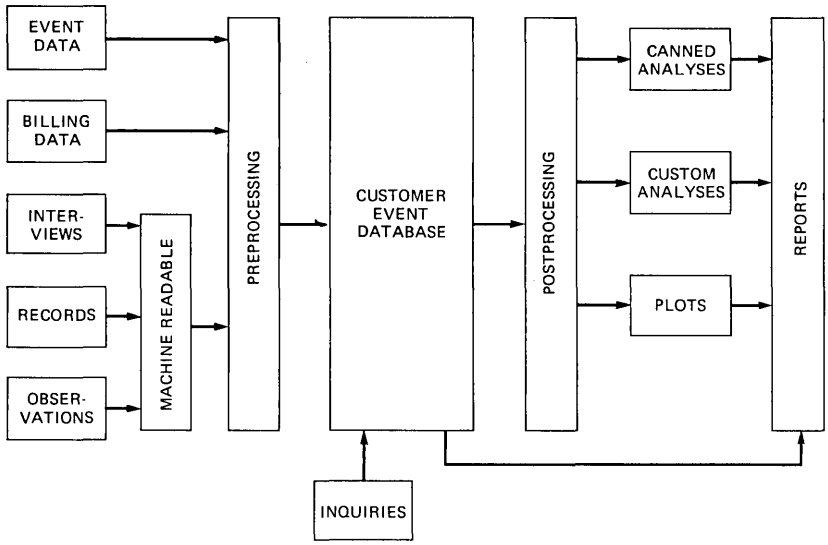


Fig. 4—Data processing stream.

sands of calls to provide adequate data. Inability to obtain sufficient customers, if not attributable to test limitations alone, may be a key indicator of the potential failure of a new service. The means for selecting and acquiring customers has to be designed with test validity in mind.

The process of obtaining participants and assigning them to groups can be the source of some of the strongest rival hypotheses. Random selection is the principal counteraction to ensure the representativeness of the population sample and group equivalence. But, randomization is not always possible. Moreover, even when it can be used, aspects of the customer acquisition and assignment process can still introduce biases that will limit the ability to generalize.

Selection processes are outlined below:

1. Determine target population using interviews and market studies.
2. Characterize the history, predisposition, environmental influences, etc., of target population.
3. Draw customers randomly from within the target population and from a similar group outside the target population. Sampling customers outside the target population will help to validate the use of the target groups for tests. If random selection is not possible, select customers who are matched to characterization of the target population.
4. Solicit participation for the test in a representative way. Preferably, contact customers in the same way they would be contacted for the final service.
5. Check participants for degree of similarity to target population characterization.
6. Interview random samples of those who agree and disagree to participate to determine the reasons for participation or nonparticipation. For example, the fact that it is a test rather than an actual service may have influenced either decision. Also, use a priori factors to account for any differences between those who agreed and those who refused.
7. Randomly assign participants to control groups and treatment groups. If random assignment is not possible, use self-selection of treatments.
8. Use measurements taken before the test to help determine the nature of group differences.

3.3.5 Results utilization

One of the largest pitfalls in the practice of implementing an evaluation of telephone service can be the lack of assurances and mechanisms for integrating evaluation results into the final product. Results from the test must be timed to allow development of changes

in the final product. A responsible organization must be named for coordinating the integration of results and the final product itself must be designed to allow changes. Appropriate flexibility in the initial design of the product can reduce delays in product introduction.

IV. DESIGNING TO COUNTERACT OTHER EXPLANATIONS

Steps taken in designing the test—which decrease confidence in rival hypotheses and increase confidence in the working hypotheses (the one we wish to prove)—are called counteractions. In laboratory research, some typical counteractions are:

- Orthogonality and counterbalancing of independent variables
- Control and preclusion of extraneous variables
- Random collection of subjects in the attempt to eliminate selection bias.

Field settings require adapting such measures to real-world constraints.

The counteractions useful in field tests can be divided into those required for evaluating a static service (i.e., determining if a service meets a priori acceptability standards) and those required for evaluating attempts to improve a service. Counteractions can be adapted to preclude, disconfirm, or control rival hypotheses.

4.1 Service evaluation

One of the simplest counteractions is to take multiple measurements of the same service, such as measurements of usage, satisfaction, and performance. Taking these measurements in more than one way (multiple methods) is also useful. Confidence in a particular result is increased if different sources point to the same conclusion.

One fear (rival hypothesis) that occasionally strikes is that the data are somehow mutilated to spuriously inflate results. This fear became real in the Calling Card Service field trial when the service began performing better than expected. Interviews, observation, and billing data corroborated the computer-collected data so strongly as to explode this rival hypothesis. Test calls removed any remaining shards of doubt.

Many rival hypotheses are based on test time. For example, the results may be due to some unusual circumstance or event, or the test may give rise to a trend in usage or performance that may be transitory (e.g., novelty effects or start-up effects) or the test may give rise to cyclic or delayed effects. One simple counteraction used in the Calling Card Service field trial to account for these problems was to repeat a test or take continued measurement of the service over time. If a unique circumstantial or historical event affected the outcome of a

test, then subsequent measures could provide evidence to support or negate this rival hypothesis.

Before, during, and after usage measurements of all coin stations (trial and nontrial) at the airport were inspected to determine if any events, trends, or cyclic phenomena confounded the data. Specifically, the Christmas holiday season and an airline strike had to be taken into account.

While multiple measurements and time-spaced testing counteract many problems, they give rise to problems of their own. One measurement may affect another. For example, repeated interviewing may cause customers to change their behavior more as a function of the interviewing process than the service itself.[6] Fortunately, the bulk of the measurements typically taken in a telephone service field trial are made on-line and are unknown to the telephone caller.* The customer is not usually aware in these tests that digits dialed, on-hooks and off-hooks are recorded and time stamped. These measures are presumably nonreactive, and thus there is no plausible explanation of measurement affecting customer behavior and attitude.

One could avoid the problem of reactivity by using only nonreactive measures. But interviews and questionnaires provide invaluable data. We handled this problem in the Calling Card Service field trial by interviewing some customers once, others twice, and still others not at all. Comparisons of these groups in terms of nonreactive measures and subsequent interviews, however, did not substantiate a reactivity effect.

Establishing a control group is another strategy that can be used to help determine if the results were due to historical events or arbitrary circumstances. The assumption is that those extraneous events or circumstances that affect the treatment group also affect the control group. To the extent that the precision of the measures allow, differences between the control group and the treatment group can, therefore, be attributed to the presence of the treatment, that is, the service.

People at nontrial coin stations at the airport were interviewed as a control for interviews taken at trial coin stations. This procedure allowed us to determine if there were any changes in customer attitudes that might have been attributable to events alone and not to service. The interview control group thus served as a counteraction to the rival hypothesis of history. For example, those people interviewed some days were angry because of plane delays. If this irritation had an effect, it would show up in both treatment and control group scores.

---

* The customer generally is aware that a test is being conducted and that service is being measured; thus, the problem of test reactivity cannot be completely discounted.

Randomization in the selection process assures equivalence between the treatment and control groups. If randomization is not possible, a plausible rival hypothesis is that the people assigned to one group differ, on the average, from those in the others and that these intrinsic differences are the real cause of any observed difference among the groups. A method for determining whether such rival hypotheses are correct is to make measurements of the groups before the treatment (service) is administered. Comparing the pretreatment and post-treatment observations of the group given the trial service with simultaneous observations of the control group (not given the trial service) tests the rival hypotheses that these two groups are inherently different, or in other words, that results are due to the history of the customers or their maturation or learning during testing.

Customers were interviewed and billing data were tracked before and after the Calling Card Service field trial. These data were used to determine the differences among customers and between trial and nontrial stations.

4.2 Improving service

In the preceding discussion, counteractions for a static service evaluation were discussed. A new service can be said to consist of a set of variables that may be manipulated to possibly improve service.

The key question is: How can we be sure that changes in service outcome are due to our service manipulations? Establishing causal connections for what improves or degrades service increases understanding. Increased understanding will tend to increase confidence in predictions of how the final service will fare.

4.2.1 Varying the service

A simple model of service variation is depicted in Fig. 5. The knobs in the figure represent the independent variables; the meters, the dependent variables. The object of varying the service is to manipulate the knobs in such a way as to cause the meters to register a beneficial change in service. If the meters are considered to show customer satisfaction, usage, and performance, the object then is to find the combination of knob positions that maximizes the readings on each meter (metaphorically speaking).

With every knob manipulation, knowledge and understanding is increased, whether the meters change positively, negatively, or not at all. Knowledge gained from the previous adjustment permits efficient and effective design of subsequent service adjustments. However, such a test-adjust-test method requires a system that can provide very quick analysis of results.

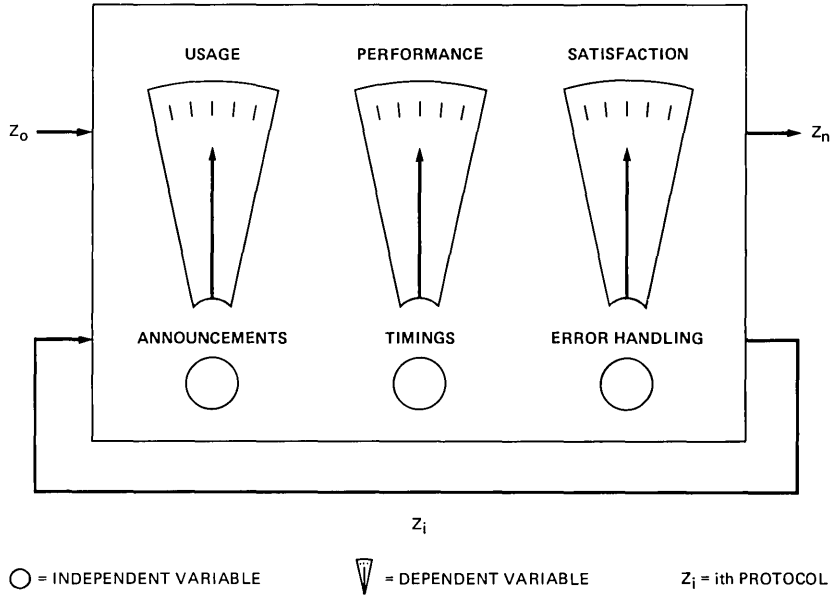


Fig. 5—Field trial metaphor—with results fed back as input into the field trial design.

4.2.2 Increasing confidence in service manipulations

Improving service can be thought of as repetitions of static service evaluation. The same counteractions discussed earlier can be used and augmented for service improvement. To account for the possibility of other events interposing effects on the dependent variables coincident with the adjustment of a service, the concept of the control group is again appropriate.

One new problem that is introduced in improving service is that multiple treatments (changes) may interact with each other. That is to say, customers' responses to a subsequent service may not be the same as those to the first service they experienced. Further, the very act of changing the service may impact attitudes and customer performance.

Adding a new group of customers for every new protocol iteration counteracts this threat to validity. (Appendix A contains notations for representing this design consideration and the others discussed here.)

The rudiments of these design principles were used in the Calling Card Service field trial. Service was manipulated in the trial in an attempt to improve service. A particular service configuration is here called a service protocol. Because many of the trial phones were in the airport, there were always a great number of new customers trying the service, more than enough to support the manipulations.

When customers dialed 0+NPA NXX XXXX at trial stations a tone prompt was given (sometimes followed by an announcement prompt) and the following things (in most protocols) could happen:

1. The customer could dial a credit card number.
2. The customer could time out without dialing and be connected to an operator.
3. The customer could dial 0 after the prompt and be connected to an operator.
4. The customer could abandon the call.

Ideally, all customers with credit card numbers would dial them. All others would dial 0 after the tone prompt to reach an operator. Practically speaking, this was not possible. Rather, different protocols were tested to attempt to increase the proportion of dialed credit card calls, increase the proportion of customers who needed an operator to dial 0, or decrease abandons.

Figure 6 summarizes the service manipulations for the placarded coin stations. Figure 7 shows the proportion of the four classes of call dispositions: (1) dial credit cards, (2) dial 0, (3) abandons, and (4) time-outs for each protocol for the placarded coin stations. The area under each curve represents the incremental proportion of 0+ calls.

While these curves are revealing, they cannot be used alone to determine the best protocol, i.e., the one with the highest satisfaction, performance, and usage. However, unacceptable protocols can readily be identified: they are those with abandonment rates in excess of 25 percent and those with dialed credit card call rates less than 25 percent.

Fortunately, the first coin placarded service protocol had low abandonment rates and high credit card dialing rates. The subsequent

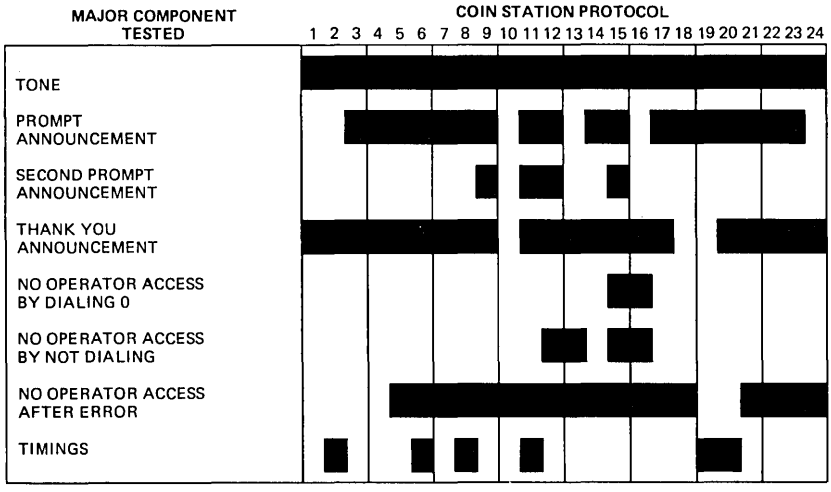
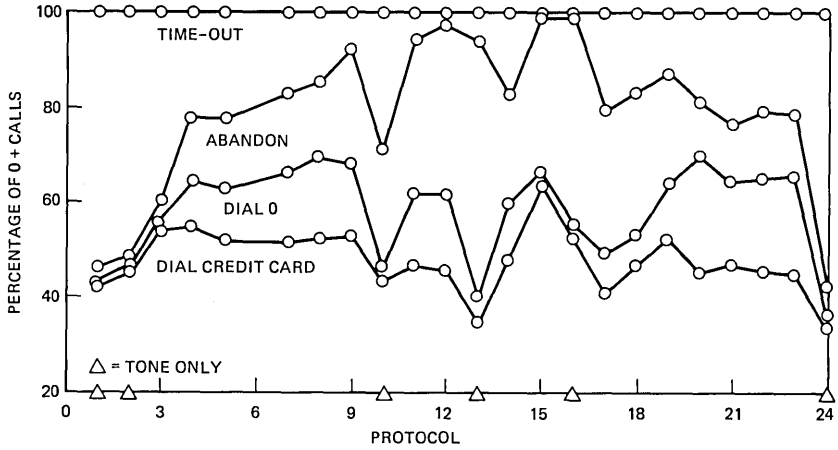


Fig. 6—Service manipulations.



NOTE: PROTOCOLS WITHOUT A TONE-ONLY INDICATOR HAVE A TONE AND AN ANNOUNCEMENT. PROTOCOL 17 INCLUDES A SPECIAL ANNOUNCEMENT DESIGNED TO DISCOURAGE CALLS FROM UNENABLED ROTARY STATIONS.

NOTE: THESE CURVES REPRESENT INCREMENTAL PERCENTAGES.

Fig. 7—Dialing performance at placarded coin stations by protocol.

manipulations were primarily aimed at improving service, although clearly the opposite effect was sometimes achieved.

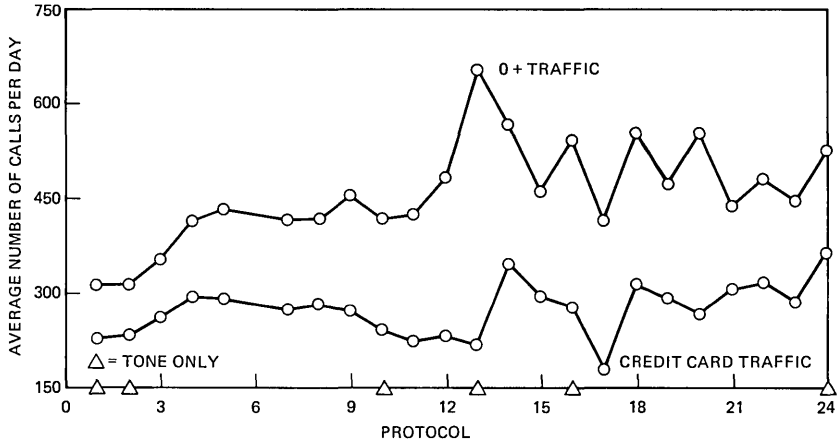
A changing volume of 0+ calls and a changing call mix were plausible explanations for one service protocol doing marginally better than another. As Figure 8 shows this was sometimes a consideration.

Another rival hypothesis was related to assumptions about those who abandoned. If all or a large number of those who abandoned were credit card customers, estimates of the proportion of credit card dialers could vary significantly. Figure 9 shows the percentage of dialed credit card calls (of all credit card calls) with and without abandons counted as credit card calls. Including and excluding abandons in this way provides lower and upper bounds, respectively, of the percent of credit card calls dialed.

Still another rival hypothesis was that usage (or attempts) may be higher, but performance lower when comparing one protocol to another. Figure 10 illustrates success rates by protocol.

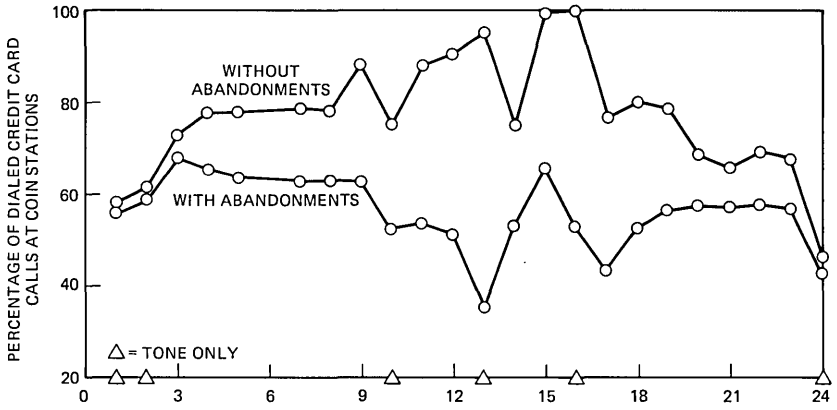
Rival hypotheses due to repeated usage (familiarity and learning) were accounted for by sorting out and examining results from repeat users.

Service protocols were compared to determine the effect of a particular manipulation. Figures 11 and 12 compare two protocols in terms of abandonments and percent of 0+ dialing. These figures lead to the conclusion that repeating a prompt announcement in this service does not increase dialing but does increase abandons. Thus, repeating announcements is a less desirable alternative.



NOTE: PROTOCOLS WITHOUT A TONE-ONLY INDICATOR HAVE A TONE AND AN ANNOUNCEMENT. PROTOCAL 17 INCLUDES A SPECIAL ANNOUNCEMENT DESIGNED TO DISCOURAGE CALLS FROM UNENABLED ROTARY STATIONS.

Fig. 8—Volume of 0+ traffic at placarded coin stations.



NOTE: PROTOCOLS 12, 13, 15 AND 16 MANIPULATIONS INVOLVED LIMITING OPERATOR ACCESS WHICH REDUCES 0 VOLUME AND CAUSED 0 + CUSTOMERS TO EITHER ABANDON OR DIAL A CREDIT CARD NUMBER.

NOTE: PROTOCOLS WITHOUT A TONE-ONLY INDICATOR HAVE A TONE AND AN ANNOUNCEMENT. PROTOCOL 17 INCLUDES A SPECIAL ANNOUNCEMENT DESIGNED TO DISCOURAGE CALLS FROM UNENABLED ROTARY STATIONS.

Fig. 9—Dialed credit card calls with and without abandonments by protocol.

4.2.3 Comparison groups

There are two ways of making service comparisons in the test design just discussed, which consisted of staggered subjects and measurements of responses to service refinements. First, customer comparisons can be made across service changes. This consists of multiple measurements or time-spaced measurements of the same customer. The

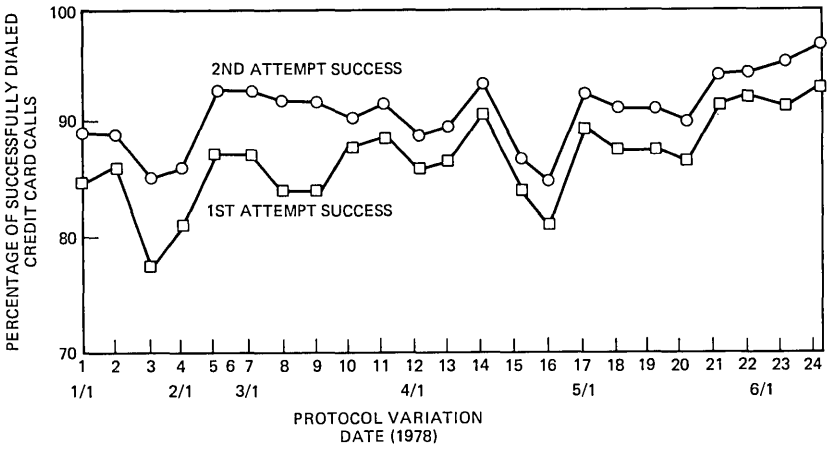


Fig. 10—Customer dialing success rate.

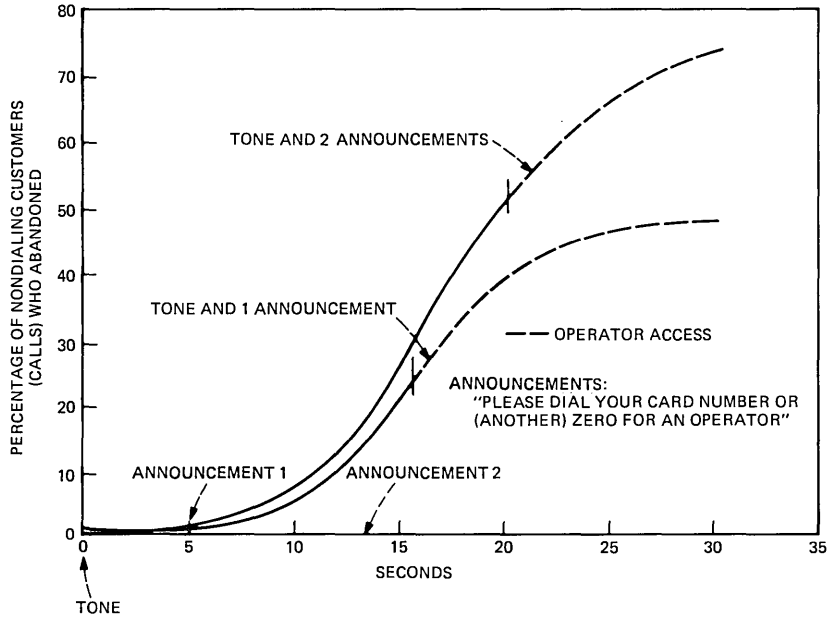


Fig. 11—Abandonments as a function of repeated announcements at coin placarded stations.

error variance is small because there are no intersubject differences, but there is multiple treatment interference. Second, service comparisons can be made across new subject groups, but treatment effects are confounded with history and by any bias introduced by the selection of groups.

Comparison groups can be arranged to preclude history as a con-

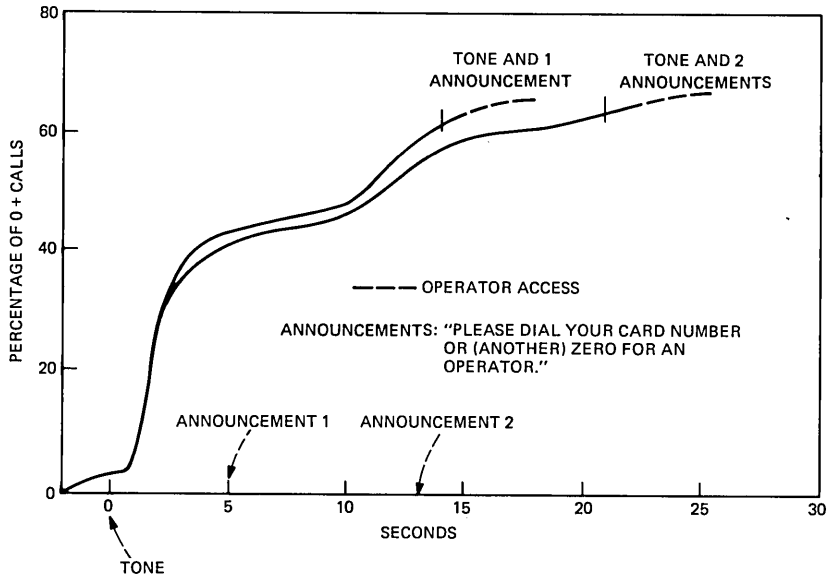


Fig. 12—Dialing as a function of repeated announcements at coin placarded stations.

founding variable by offering different service arrangements simultaneously (at different telephones, for example). For best results, each of these comparison groups should be as similar as possible except for any planned differences in treatment (service).

In the Calling Card Service field trial, we devised three comparison groups: noncoin stations, placarded coin stations (with special bright orange instruction cards), and placardless coin stations. These comparison groups were not mutually exclusive. Crossovers occurred, especially between placarded coin and placardless coin groups. But for every comparison group, and, as discussed, for every manipulation, there were new customers. Service changes were replicated across these comparison groups further increasing confidence and providing information about the effects of the station type.

4.2.4 Refinement

Because the number of variables and variable states in a field test of this kind is usually large, only a small subset of the number of possible treatment combinations can be used. Moreover, the protocol is evolving and a delay is often necessary in the testing of different combinations. Unfortunately, this delay means the manipulation is confounded with the history of the customer population. However, prudent selection of changes and repetitions within a treatment group and across comparison groups can yield a more complete understanding of the effects of the service components and their interactions.

In the Calling Card Service, the primary objective of the service manipulations was to make reasonable changes which were hypothesized to increase usage, satisfaction, and performance. This objective was sometimes subjugated to the purpose of increasing confidence in the belief that the changes in the independent variables (and not the parameters) caused the observed changes in the dependent variables. One such manipulation was to deliberately make it progressively more difficult for the customer to get to an operator, i.e., protocols 12, 13, 15, and 16. It was hypothesized that more customers would use their credit card if it were made relatively more difficult to obtain an operator. However, this hypothesis was not borne out; the proportion of credit card calls dialed was either unchanging or decreasing, and abandons increased as one would suspect. Clearly, other manipulations were more successful.

As the treatment change chart shows, changes in the state of more than one independent variable were made more often than not. The large number of variables prohibited single sequential changes in each variable. For example, the error-leg protocol was manipulated independently and simultaneously with the access-leg protocol many times. Evidence gathered from protocols prior to this practice showed little or no interaction from changes in the access leg and the error leg, except that the number of people making errors fluctuates.

Inferences made about the protocol change were made in terms of the multiple change because the independent variables were confounded. Inferences based on effect separation and independence assumptions were made cautiously and verified, when possible, in later manipulations.

4.2.5 Replication

Any changes in the dependent variables caused by a change in the protocol (or service) and not by history should be present in successive identical tests (within the bounds of measurement error). Replication of effects increases confidence in the hypothesis that a change in protocol actually causes a change in the dependent variable being measured. As a practical matter, the utility of repeated replications of a single treatment is sharply reduced by the need to manipulate as many variables as possible in the time allowed. However, intermittent replications can serve to benchmark drifting or trending data.

On several occasions, we repeated protocols within and across comparison groups in the Calling Card Service field trial. Replications afforded an increase in confidence by allowing an assessment of changes in dependent variables due to confounding of external and internal parameters with the independent variables. The replications, in effect, provided a basis for comparison. And the replications in-

creased confidence in the assertion that the observed changes in the dependent variable were caused by the deliberate change in the independent variables. For example, protocol 14 amidst protocols 12, 13, 15, and 16 (coin placarded stations) was a replication of an earlier protocol and supported the hypothesis that the poor performance of the other protocols was real and not an artifact of internal or external conditions.

Some replications were designed to provide more information about the contributing effects of certain service components. Most notably, major service changes were tested with both the tone-only and tone-and-announcement conditions. For example, protocol 13 (coin placard station) was the same as protocol 12, but without the prompt announcement. This is similarly true for protocols 15 and 16.

4.2.6 Partial counterbalancing

Arranging the treatment order in special ways—e.g., the Latin Square design—across comparison groups provides for pairwise comparisons of different treatment orderings. Such arrangements, however, are often not practical in field trials of telephone services because of the large number of variables and variable states (treatment conditions) that should be tested. The large number of treatment combinations required of typical counterbalancing techniques preclude their use. Manipulation of protocols on the basis of data gathered from prior manipulations also precludes planned counterbalancing.

But it may be possible partially to counterbalance treatment pairs across comparison groups in order to assess some multiple treatment order effects. If the manipulations evolve, the application of counterbalancing schemes have to be delayed by at least one time frame to have knowledge of the treatment pair.

4.2.7 Post-manipulation static study

After service manipulations are finished, it is useful to study the final service at rest. In this period in the Calling Card Service field trial (protocol 24), transient effects due to changes in the service were not present, thus improving predictions of the final service.

V. CONCLUSION

Many service issues were resolved by the trial of Calling Card Service. Most important, the customers who tried the service continued to use it because they felt it was faster and more convenient than operator-assisted credit card calling. Moreover, the design of the service was critically dependent on the results of the field trial manipulations. Both announcements and instruction placards were found to be very effective in stimulating customers to dial. The field trial data

provided objective measurements of optimum customer usage and acceptance and lower rates of error and abandonment in the service offered. Table I lists the criteria found most useful in selecting the attributes of the final service.

When Calling Card Service was first offered in Buffalo in July 1980, a product follow-up evaluation study was conducted to monitor customer use, performance, and acceptance levels. The measurements of actual service conformed closely to estimates made from field trial data and, thus, support the utility of test methodology in guiding service design choices. The methodology developed for evaluating Calling Card Service is now being applied to the next generation of services, such as teleconferencing.

VI. METHODOLOGY SUMMARY

The pre-field trial activities of analysis, interviews, and laboratory studies are undertaken to refine the service definition and assess the potential of its success. A study plan is constructed which defines objectives, resources and constraints, study variables, acceptance criteria, and study design. Field test requirements are developed that specify the data collection, analysis, and customer acquisition and preparation procedures as well as specify the hardware, software, and network aspects of a telephone service test vehicle.

Given adequate requirements the test is developed and implemented. During the test, the service is manipulated and data are collected and analyzed to optimize criteria of usage, performance, and satisfaction. Trade-offs are made relative to available resources, data validity, confidence in results, and time constraints. If success is predicted for the service being tested, results are then integrated into the design and the service is offered and monitored. Figure 13 summarizes the total process of conducting a field test of a telephone service.

Table I—Criteria for selecting service attributes

Independent Variables	Dependent Variables
1. Announcement Presence	a. Percentage Dialed b. Abandons
2. Announcement Wording	a. Percentage Dialed b. Confusion (interviews) c. Abandons
3. Number of Attempts Allowed	a. Success Rate b. Percentage of Subsequent Attempts
4. Operator Access	a. Percentage Dialed b. Abandons c. Satisfaction (interviews)
5. Interevent Timing	a. Distribution of Times to First Digit
6. Interdigit Timing	a. Distribution of Pauses Between Digits
7. Protocol (service in general)	a. Percentage Dialed b. Satisfaction (interviews) c. Success Rate
8. Delays	a. Abandons

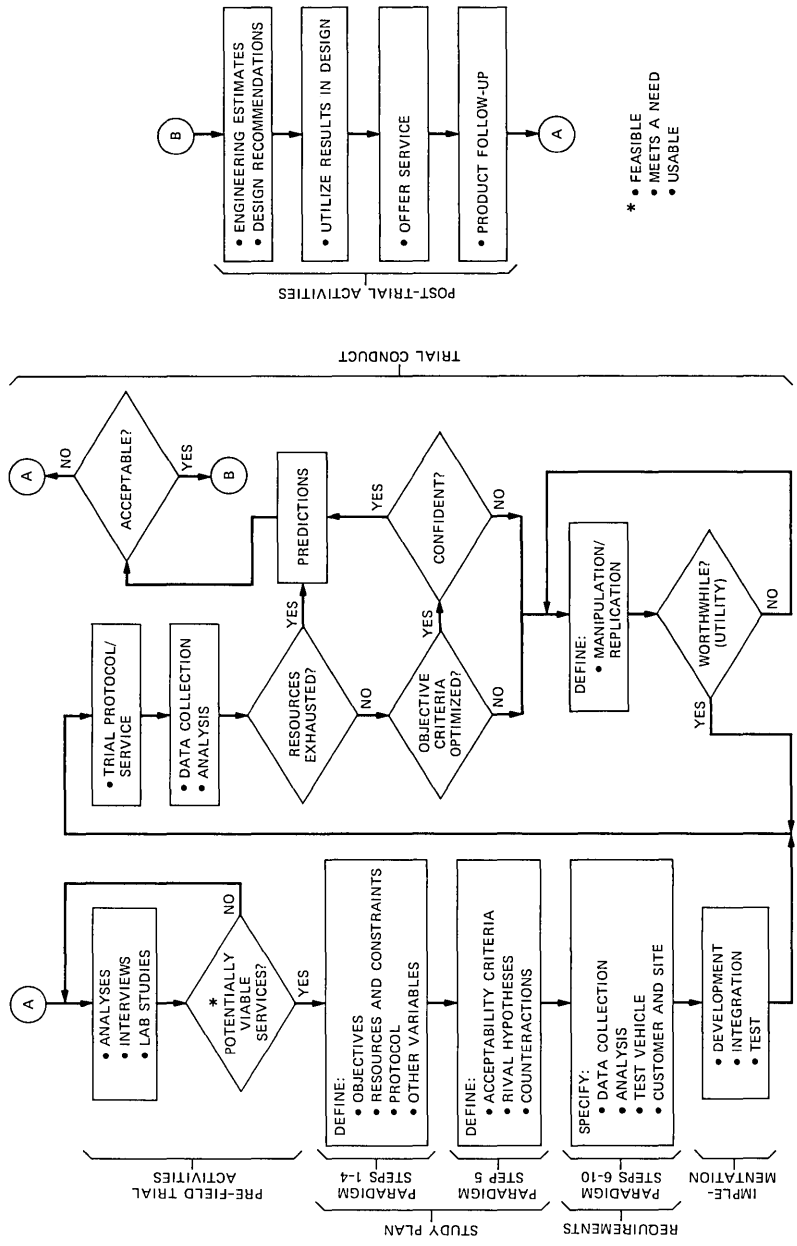


Fig. 13—Design and evaluation process for human-machine telephone service protocols.

VII. ACKNOWLEDGMENTS

I would like to thank E. C. T. Walker and C. A. Riley for their valuable comments on this manuscript. K. R. Hickey, R. J. Jaeger, N. S. Pearson, and J. L. Santee are thanked for their input as well. The Calling Card Service test depended heavily on the contributions of T. M. Bauer and E. A. Youngs.

REFERENCES

1. M. R. Allyn, T. M. Bauer, and D. J. Eigen, "Planning for People: Human Factors in the Design of a New Service," Bell Lab. Rec., *58,* No. 5 (May 1980), pp. 55–161.
2. D. J. Eigen and E. A. Youngs, "Calling Card Service—Human Factors Studies," B.S.T.J., *61,* No. 7 (September 1982), pp. 1715–35.
3. D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Design for Research,* Chicago: Rand McNally, 1966.
4. D. T. Cook and D. T. Campbell, *Quasi-Experimentation Design and Analysis Issues for Field Settings,* Chicago: Rand McNally, 1979.
5. H. M. Parsons, *Man-Machine System Experiments,* Baltimore: Johns Hopkins, 1972.
6. E. J. Webb et al., *Unobtrusive Measures: Nonreactive Research in the Social Sciences,* Chicago: Rand McNally, 1966.

APPENDIX A

Field Evaluation Design Notation

To provide another perspective on the field design methods, the following notation is provided.

A.1 Service

A service (or treatment) can be defined by set of dependent variables with specific values, represented here as a vector $\dot{Z}$.

$$\dot{Z} = \begin{vmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{vmatrix}$$

A.2 Measurements

Let 0 stand for observation. A matrix of n kinds of measurement taken in m different ways is represented by $\ddot{0}$.

$$\ddot{0} = \begin{vmatrix} 0_{11} & 0_{12} & \cdots & 0_{1m} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ 0_{n1} & 0_{n2} & \cdots & 0_{nm} \end{vmatrix}.$$

When multiple measures (taken multiple ways) are repeated, the notation is:

$$\ddot{O}^m = \ddot{O}_1 \cdots \ddot{O}_m .$$

A.3 Improving service

Two aspects of the field study, static service evaluation and improving service, could be represented as:*

Static Service Evaluation		Improving Service
$\dot{O}_0^m \dot{Z}_0 \ddot{O}_1^m$	$\dot{Z}_1$	$\ddot{O}_2^m \cdots \ddot{O}_n^m \dot{Z}_n \ddot{O}_{n+1}^m$

This design effectively constitutes repetitions of the static service evaluation.

A.4 Control group

${}_{A'}\ddot{O}_0^m \dot{Z}_0 {}_A\ddot{O}_1^m \dot{Z}_1 {}_A\ddot{O}_2^m \cdots {}_A\ddot{O}_n^m \dot{Z}_n {}_A\ddot{O}_{n+1}^m$	Treatment Group
${}_{A'}\ddot{O}_0^m {}_{A'}\ddot{O}_1^m {}_{A'}\ddot{O}_2^m \cdots {}_{A'}\ddot{O}_n^m {}_{A'}\ddot{O}_{n+1}^m$	Control Group

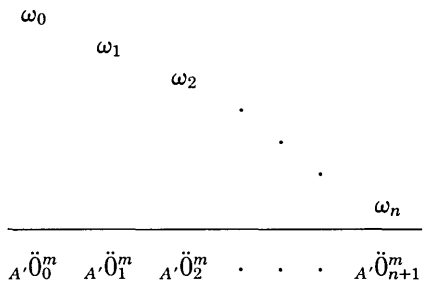
A.5 Staggered treatment groups

New subjects are added (or are available) at each treatment change. Limited resources usually necessitate the continued manipulation of the variables within treatment groups. The line offsets the control group from the treatment groups and denotes unequivalence of the groups.

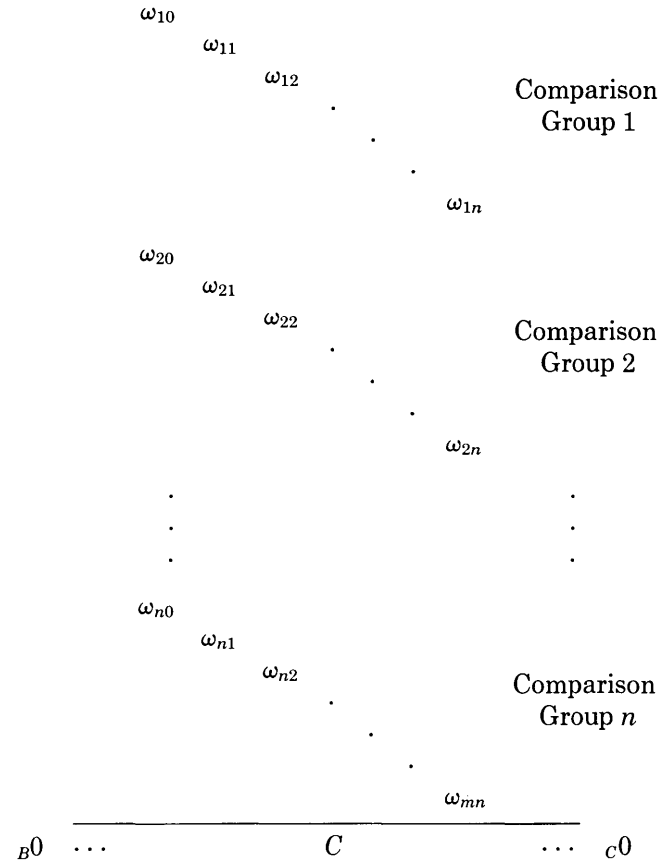
${}_{A'}\ddot{O}_{10}^m \dot{Z}_0$	${}_{A'}\ddot{O}_{11}^m \dot{Z}_1$	${}_A\ddot{O}_{12}^m$	$\cdots$	${}_A\ddot{O}_{1n}^m$	$\dot{Z}_n$	${}_A\ddot{O}_{1n+1}^m$	Treatment Group 1
	${}_{A'}\ddot{O}_{21}^m \dot{Z}_1$	${}_A\ddot{O}_{22}^m$	$\cdots$	${}_A\ddot{O}_{2n}^m$	$\dot{Z}_n$	${}_A\ddot{O}_{2n+1}^m$	Treatment Group 2
				${}_{A'}\ddot{O}_{nn}^m$	$\dot{Z}_n$	${}_A\ddot{O}_{nn+1}^m$	Treatment Group n
${}_{A'}\ddot{O}_{n+10}^m$	${}_{A'}\ddot{O}_{n+11}^m$	${}_{A'}\ddot{O}_{n+12}$	$\cdots$			${}_A\ddot{O}_{n+1n+1}^m$	Control Group

* $\dot{Z}_i$ denotes the vector of service-independent values for the ith service change (treatment). ${}_A\ddot{O}_i^m$ denotes the matrix of multiple measurements by multiple methods measured m times for the ith service (treatment) change. The left-hand A' subscript (versus A subscript) denotes that some measurements present in the trial may not be available in the pretrial measures (or in the control group).

To reduce the notational burden, let ω_i denote the ith staggered treatment group.



A.6 Staggered treatment groups with comparison groups



Note that the Control Group, C, is augmented with historical and follow-up measurements.

A.7 Replications

Replications are denoted by the appropriate repetition of protocol time frame subscripts on the independent variable vector:

$$ {}_{A'}\ddot{O}_0^m \boxed{\dot{Z}_0} {}_{A'}\ddot{O}_1^m \dot{Z}_1 {}_{A}\ddot{O}_2^m \boxed{\dot{Z}_0} {}_{A}\ddot{O}_3^m \dot{Z}_2 {}_{A}\ddot{O}_4^m \cdots \dot{Z}_n {}_{A}\ddot{O}_p^m. $$

A.8 Partial counterbalancing

$$ \begin{array}{llll} \boxed{{}_{A'}\ddot{O}_{101}^m \; \dot{Z}_0 \; {}_{A}\ddot{O}_{111}^m \; \dot{Z}_1} & & \ddot{O}_{121}^m \; \dot{Z}_2 \; \cdots & \left.\begin{array}{l}\\ \\ \\ \end{array}\right\} \text{Comparison Group 1} \\ & {}_{A'}\ddot{O}_{211}^m \; \dot{Z}_1 & {}_{A}\ddot{O}_{221}^m \; \dot{Z}_2 \; \cdots & \\ & & {}_{A'}\ddot{O}_{321}^m \; \dot{Z}_2 \; \cdots & \\ {}_{A'}\ddot{O}_{102}^m \; \dot{Z}_0 & {}_{A'}\ddot{O}_{112}^m \; \dot{Z}_1 & {}_{A}\ddot{O}_{122}^m \; \dot{Z}_0 \; \cdots & \left.\begin{array}{l}\\ \\ \\ \end{array}\right\} \text{Comparison Group 2} \\ & \boxed{{}_{A'}\ddot{O}_{212}^m \; \dot{Z}_1} & \boxed{{}_{A}\ddot{O}_{222}^m \; \dot{Z}_0} \; \cdots & \\ & & {}_{A'}\ddot{O}_{322}^m \; \dot{Z}_0 \; \cdots & \end{array} $$

The two boxed groups are the time-delayed counterbalanced treatment pairs $\dot{Z}_0$ and $\dot{Z}_1$.

AUTHOR

Daryl J. Eigen, B.A. (Psychology), 1972, M.S. (Electrical Engineering), 1973, University of Wisconsin; Ph.D. (Industrial Engineering), 1981, Northwestern University; Bell Laboratories, 1973—. Mr. Eigen initially worked in the Human Performance Technology Center. He then was involved in feature and service planning for the Traffic Service Position System and, later, the No. 4 ESS. He is currently Supervisor of the System Analysis and Human Factors Group for No. 4 ESS. Member, IEEE, APA, Human Factors Society, and Tau Beta Pi.

Characteristics of Human Performance

Vigorous efforts to improve understanding of underlying processes in human behavior have been made at Bell Laboratories. Current interests include most major aspects of human information processing and performance, including perception, thinking, movement control, and learning. This section includes samples of recent work on fundamental mechanisms in visual pattern recognition (Julesz and Bergen), the timing of repetitive actions (Rosenbaum), the perceptual interpretation of graphical displays (Cleveland, Harris, and McGill), and individual variations in modes of mental problem solving (Egan). While these four papers by no means exhaust the problem areas under investigation, they do give a fair picture of the wide range of issues, methods, theory, and data involved, and the diverse domains of knowledge being enriched.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July-August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Textons, The Fundamental Elements in Preattentive Vision and Perception of Textures

By B. JULESZ* and J. R. BERGEN*

(Manuscript received September 23, 1981)

Recent research in texture discrimination has revealed the existence of a separate "preattentive visual system" that cannot process complex forms, yet can, almost instantaneously, without effort or scrutiny, detect differences in a few local conspicuous features, regardless of where they occur. These features, called "textons", are elongated blobs (e.g., rectangles, ellipses, or line segments) with specific properties, including color, angular orientation, width, length, binocular and movement disparity, and flicker rate. The ends-of-lines (terminators) and crossings of line segments are also textons. Only differences in the textons or in their density (or number) can be preattentively detected while the positional relationship between neighboring textons passes unnoticed. This kind of positional information is the essence of form perception, and can be extracted only by a time-consuming and spatially restricted process that we call "focal attention". The aperture of focal attention can be very narrow, even restricted to a minute portion of the fovea, and shifting its locus requires about 50 ms. Thus preattentive vision serves as an "early warning system" by pointing out those loci of texton differences that should be attended to. According to this theory, at any given instant the visual information intake is relatively modest.

* Bell Laboratories.

I. INTRODUCTION

In this article we give an overview of some insights into the workings of the human visual system gained during two decades of research at Bell Laboratories, and culminating in the discovery of a few local conspicuous features that we call textons. Textons appear to be the basic units of preattentive texture perception,[1] when textures are viewed in a quick glance with no further effort or analysis. Although this article goes beyond texture perception into preattentive vision in general, studies of texture discrimination led to the basic insights presented here and provide excellent demonstrations of the main findings. Based on our findings we propose a novel theory of vision in which the preattentive visual system inspects a large portion of the visual field in parallel and detects only density differences in textons. It then directs focal attention to these loci of texton differences for detailed scrutiny.

Now, after 20 years of research, when we know what textons are and their role in vision is clarified, we can save the reader from following the rather difficult steps that led to their discovery. [The reader interested in the history of these developments, and in the sophisticated mathematics necessary to generate textures with certain stochastic constraints, should turn to the original articles referred to in a recent review by one of us[1] and to the Appendix.] Here we follow an axiomatic treatment. The main findings are presented in Section II as heuristics (similar to axioms, but not necessarily totally independent), immediately followed by many demonstrations. The reader can test the power of these newly acquired heuristics by being able to predict and then verify which texture pairs will be perceived to be different, and which will appear as a single texture. The reader can thus understand the new theory of vision without mathematical knowledge.

Section III emphasizes the essentially local nature of texture perception. Section IV relates the psychologically identified textons to some neurophysiological results concerning local feature analyzers in primate cortex. Section V extends the texton theory from texture perception to the discrimination of briefly presented patterns. In Section VI a model of human vision is proposed that postulates two different modes of visual system function. Section VII discusses some implications of this model.

II. HEURISTICS: DEFINITION OF TEXTONS AND THEIR INTERACTIONS IN PREATTENTIVE VISION

Visual textures are defined as aggregates of many small elements. The elements can be either dots of certain colors (e.g., black, white, grey, red) or simple patterns. For purposes of this article, we consider

only elements that do not overlap, and are placed at either regular or random positions, in identical or in random angular orientations.

Usually in our demonstrations two textures (composed of two different elements) are placed side-by-side, or one is embedded in the other, as shown in Fig. 1. When the reader cursorily inspects Fig. 1, an area made up of +'s will appear to stand out from the surrounding texture composed of L's. Indeed, without scrutiny, that is without detailed element-by-element inspection, the reader might not notice that a third area composed of T-shaped elements is also embedded in the texture of L's. We call this effortless perceptual segregation of the texture composed of +'s from the surrounding texture of L's *preattentive texture perception.* On the other hand, if texture discrimination requires element-by-element scrutiny, as is the case of finding the T's in the L's, we call this way of looking with scrutiny *focal attention.* We will show many other preattentively indiscriminable texture pairs (e.g., Figs. 3c and 6b), which, because they do not segregate, often are not even perceived as containing different elements until this is pointed out.

Although in all texture perception the preattentive system is dominant, the role of focal attention can be even further reduced by brief presentation. The reader who is not convinced by the qualitative difference between preattentive and attentive texture discrimination

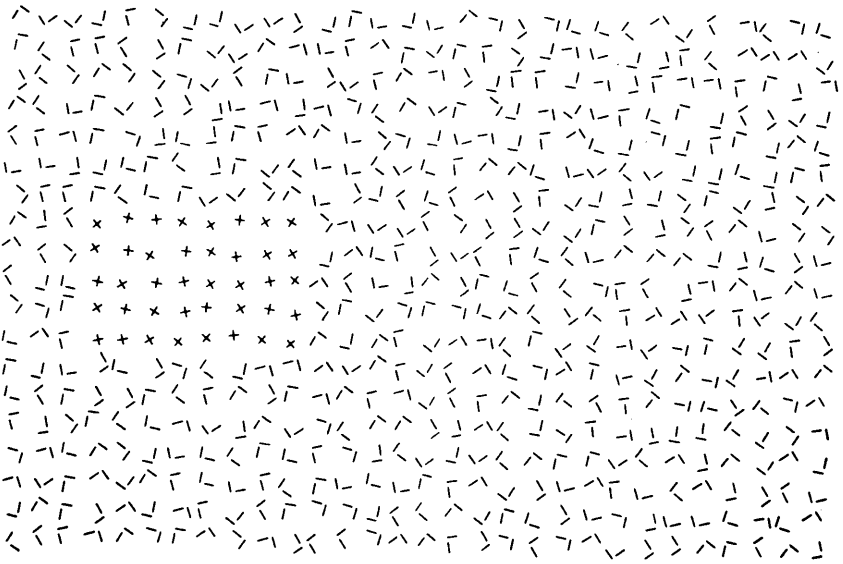


Fig. 1—"Preattentive texture discrimination" is shown between areas composed of +'s and L's, while element-by-element scrutiny, called "focal attention" is required to find the T's embedded in the L's.

might inspect Fig. 1 through a camera shutter set at 1/10 second exposure time.

Heuristic 1: Human vision operates in two distinct modes

1. Preattentive vision—parallel, instantaneous, without scrutiny, independent of the number of patterns, covering a large visual field, as in texture discrimination.
2. Attentive vision—serial search by focal attention in 50-ms steps limited to a small aperture, as in form recognition.

Heuristic 2: Textons

1. Elongated blobs—e.g., rectangles, ellipses, line segments with specific colors, angular orientations, widths, and lengths.
2. Terminators—ends-of-line segments
3. Crossings of line segments

Heuristic 3: Preattentive vision directs attentive vision to the locations where differences in the density (number) of textons occur, but ignores the positional relationships between textons.

Before we discuss the implications of these heuristics, let us apply them to a few pairs of elements and predict whether the texture pairs formed from these elements will yield preattentive texture discrimination or not. This application of the rules also helps to clarify them. For instance, elongated blobs of different widths or lengths are different textons, as Fig. 2a demonstrates. The larger sized R's containing longer and wider line segments form a texture that segregates (i.e., is preattentively discriminable) from its surround, which is composed of smaller R's with shorter and narrower line segments.

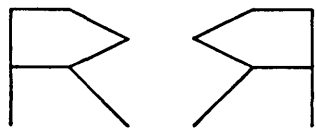
Similarly, elongated blobs of different orientations are different

(a) (b)

Fig. 2—Preattentive texture discrimination based on texton differences between line segments of (a) length and width and (b) angular orientation. (*Nature*, March 12, 1981[1])

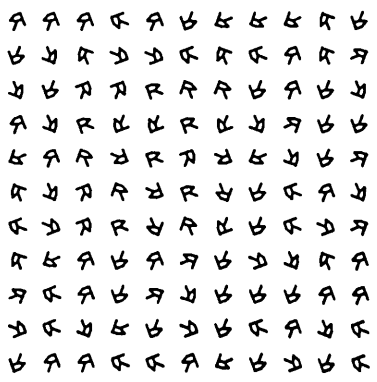
textons. Indeed, in Fig. 2b the texture pair composed of the same sized R's having two different orientations in the two textures, yields preattentive discrimination. Obviously, the same elongated blob shape with the same orientation yields different textons if the colors (e.g., black, gray, white, red, green, etc.) are different.

Now, let us predict what would happen if we took an R and a mirror-image R, as shown in Fig. 3a, and formed a texture pair by throwing them in random orientations. Obviously, without randomizing the orientations, the two textures would yield texture discrimination, since even though their widths and lengths agree, some of the line segment textons have different orientations in the R and in its mirror image, though the widths and lengths agree, as shown in Fig. 3b. However, if the two elements are thrown at random orientations, then the two textures formed have the same average density of textons (i.e., in some area of integration the number of line segments with the same color, width, length, and orientation is identical). Therefore, the preattentive visual system should not be able to direct focal attention to loci of texton differences that form the boundary between the two regions. Indeed, an inspection of Fig. 3c yields a single, uniform texture. It requires laborious element-by-element inspection for several seconds



(a)

(b)



(c)

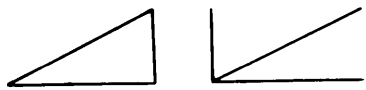
Fig. 3—Demonstration of how the heuristics given in text predict why (a) R and its mirror image in aggregates yield texture discrimination (b), or are indistinguishable (c). (*Perception*, 1973[5])

to find the boundary between the array of R's and mirror-image R's. Obviously, in a 100-ms presentation discrimination of these textures is impossible.

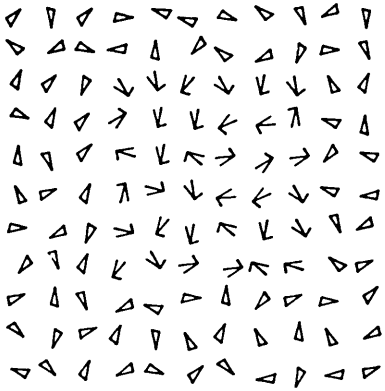
Let us note that if one were to select a pair of elements without knowing the rules given above, most probably the resulting texture pair would be discriminable. Only through the joint effort of our colleagues (D. Slepian, M. Rosenblatt, E. Gilbert, L. Shepp, H. Frisch, T. Caelli, and J. Victor) from 1962 to 1978 were some elegant methods found that yielded indistinguishable textures, even though their elements appeared very different.

In the next examples we stress the importance of terminator textons. For instance, in Fig. 4a the two elements are composed of three identical line segments (i.e., same orientation, width, and length). The only difference is in the number of their ends-of-lines (terminators). The triangle-shaped element has no open ends, while the "dual" element has three ends-of-lines. One should expect texture segregation, given such a large difference in terminator number, and as Fig. 4b demonstrates, this is the case.

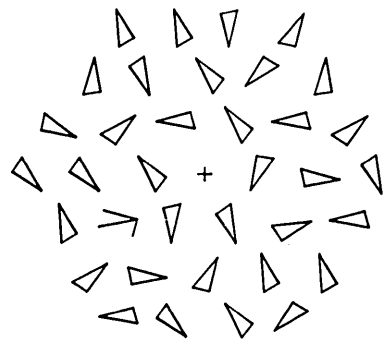
As a matter of fact, discrimination is so strong that a single element can be preattentively detected among 35 dual elements, as shown in Fig. 4c. This arrangement is now routinely used by us in studying



(a)



(b)

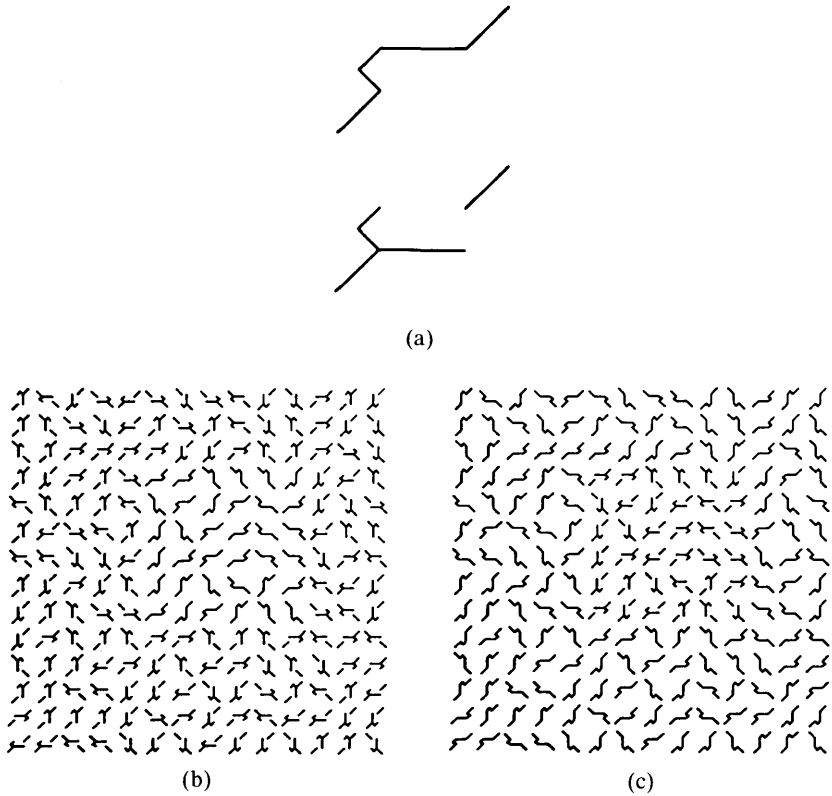


(c)

Fig. 4—Demonstration of how the heuristics given in text predict preattentive texture discrimination (b) and even discrimination of a single element among many (c), based on terminator number difference (zero versus three) between elements (a). (*Nature*, March 12, 1981[1])

pattern discrimination in preattentive vision, as discussed in Section V. Here we note only that when there is a texton difference (as in Fig. 4c) detecting one element in the midst of 35 other elements is almost as easy as detecting the difference between two elements (shown in Fig. 4a) for presentation times as brief as 100 ms.

In the next example, both members of the element pair of Fig. 5a are again composed of the same five line segments (each corresponding line segment in the two elements has identical width, length, and orientation, respectively) but one element contains only two ends-of-lines, whereas the other contains five. This large difference in terminator numbers should yield texture segregation, and inspection of Fig. 5b demonstrates that it does. Figure 5c consists of the same texture pair as Fig. 5b, except that the texture containing the five terminators is now the surround. Although, as predicted, the large difference in terminator numbers again yields texture segregation, the appearance of the boundary between the two regions is different for Fig. 5b and 5c.



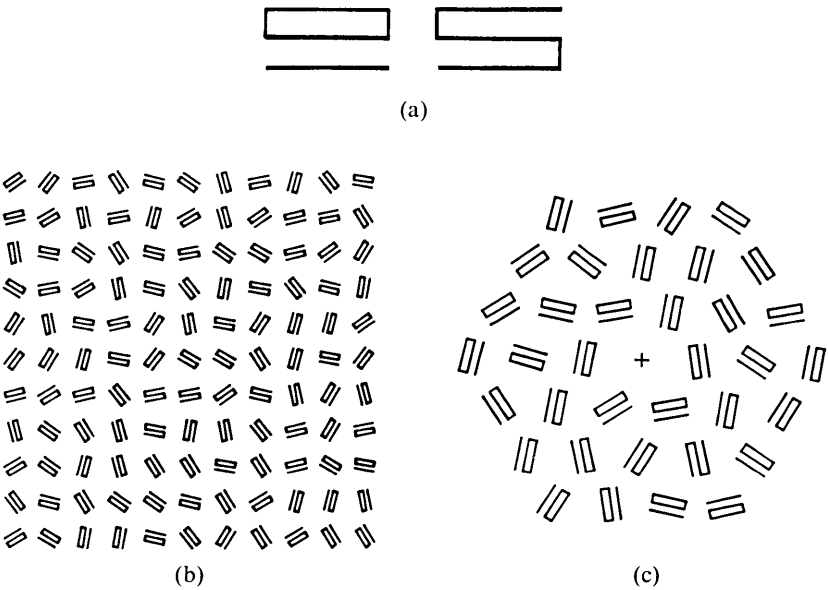
(a)

(b) (c)

Fig. 5—Similar to Fig. 4 except the terminator number between elements is two versus five.

The next example, shown in Fig. 6a, consists of the "S"- and "10"-shaped elements, that in isolation appear quite different. However, the two contain the same number of line segment textons (three identical horizontal and two identical vertical line segments) and both contain two ends-of-lines. The fact that the positional relationship between these textons is different (as it is in Fig. 3b) can be perceived only by the attentive visual system (yielding the percept of an S versus a 10). However, according to Heuristic 3 the preattentive system can count only the density (number) of textons and ignores their relative positions. So, according to our rules, a texture pair composed of these elements contains the same average density (number) of textons, and thus should be indistinguishable. Surprising as it may seem, the texture pair is indeed preattentively indistinguishable as demonstrated by Fig. 6b. [Readers who find this demonstration of the distinction between preattentive and focal vision not adequately convincing without brief presentation should note the contrast between the attentively different percepts of Fig. 6a, and the texture pair in Fig. 6b, which remains difficult to distinguish even with element-by-element scrutiny.]

Finally, let us demonstrate the third texton, the crossing of elongated blobs (line segments). Figure 7a shows the conspicuous difference between a texture pair that segregates based on the presence or absence of elements having crossing versus not-crossing line segments.



(a) (b) (c)

Fig. 6—Demonstration of how the heuristics given in text predict why (a) the differently appearing S- and 10- shaped elements in aggregates (b) and one S in 10's (c) are indistinguishable. (*Nature*, March 12, 1981[1])

If the elements have identical textons, including crossing (or not-crossing line segments) the texture pairs become preattentively indistinguishable. The positional relationship between the line segment textons is unnoticed by the preattentive system. The difference in gap size between the L-shaped elements in Fig. 7b yields a preattentively indistinguishable texture pair. Particularly interesting is the demonstration in Fig. 7c where T- versus L-shaped elements yield an indistinguishable texture pair. Although we have kept a small gap between the perpendicular line segments that make up the L's and T's, preattentive discrimination of texture pairs composed of these elements is impossible even when the gaps are not resolvable. Apparently, the difference of a single end-of-line terminator is not adequate to yield texture segregation. Finally, Fig. 7d depicts a preattentively indistinguishable texture pair, where, with scrutiny, it is obvious that the

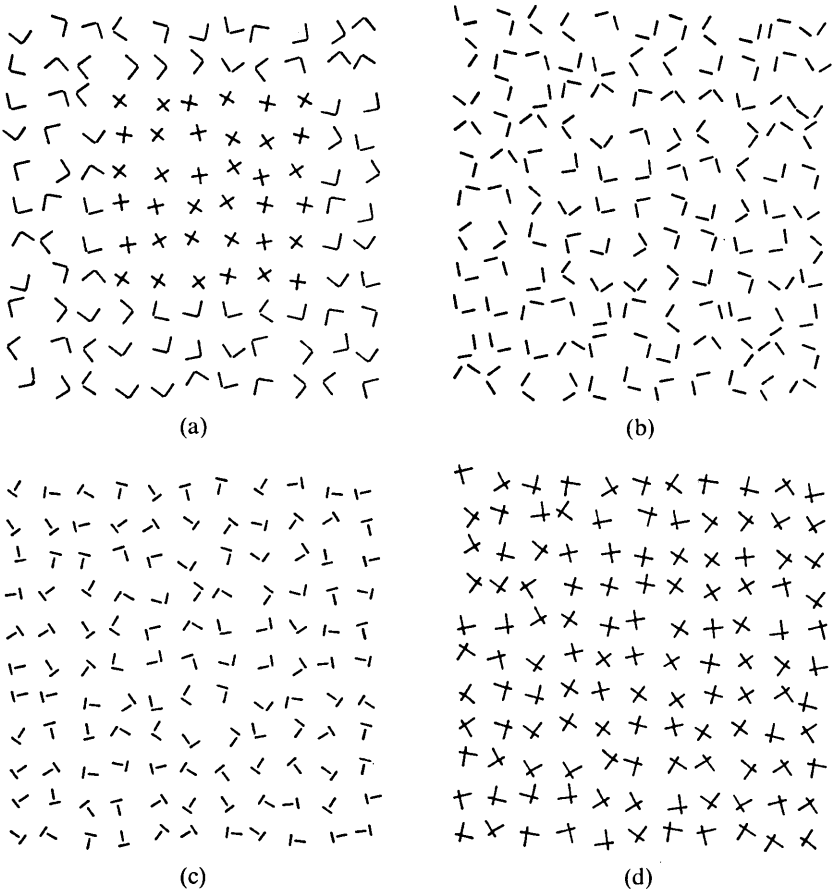


Fig. 7—Demonstration that crossing of line segments is a texton.

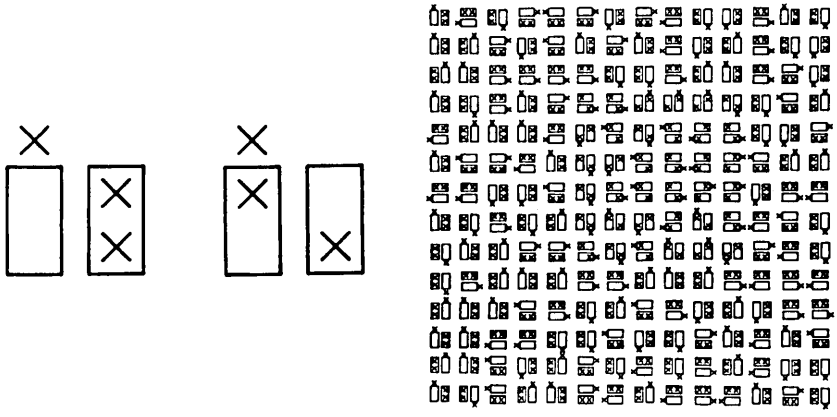
elements contain line segments that either cross at midpoint or cross far from the midpoint.

The last two examples are given in Figs. 8a and b and Figs. 9a, b, and c. From the element pairs containing the same textons, the reader can predict that although their elements in isolation appear very different, the resulting texture pairs will be indistinguishable.

In all these demonstrations the texture elements consisted of line segments. For line segments the definition of terminators (ends-of-lines) and their crossings are straightforward. For elongated bars with substantial width these definitions are less direct. Particularly difficult is the notion of terminators, because instead of terminators some combination of white elongated bars in a black surround with black elongated bars in white surround might suffice. So, we are not certain whether terminators are independent textons. Nevertheless, as a first approximation these three heuristics work remarkably well.

III. PREATTENTIVE TEXTURE PERCEPTION IS ESSENTIALLY A LOCAL PROCESS

The essence of all the findings reported in the previous section can be summed up as follows: In texture perception the preattentive visual system utilizes only local conspicuous features, textons, and these textons are not coupled to each other (i.e., a vertical and horizontal line segment do not cohere to form an L or T). The preattentive system utilizes globally only the textons in the simplest possible way by counting their numbers (densities). This might surprise many of our readers who assume that texture perception utilizes complex global statistical interactions between textural elements.



(a) (b)

Fig. 8—Since the element pair in (a) is composed of the same textons, the texture pair (b) composed of these elements is preattentively indistinguishable. (*Philosophical Transactions,* 1980[25])

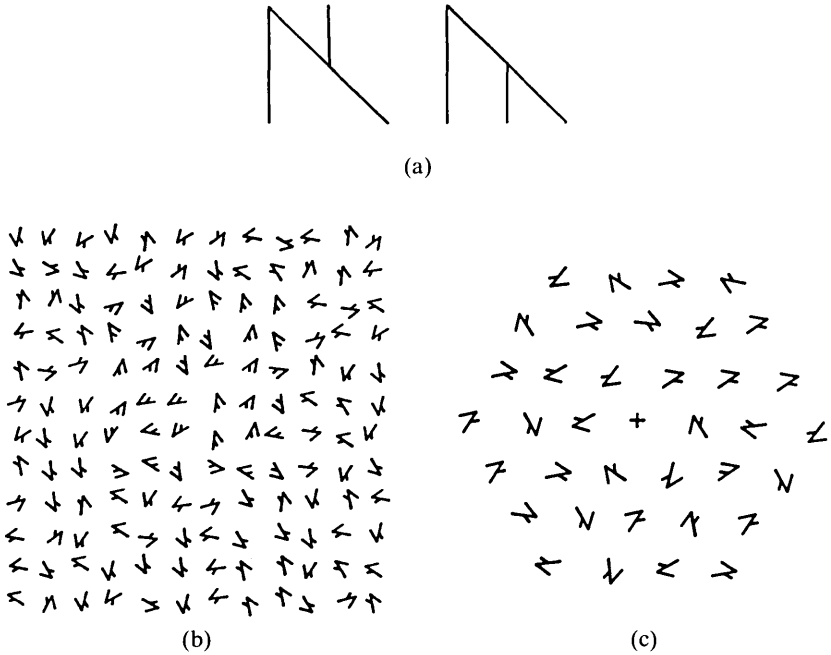
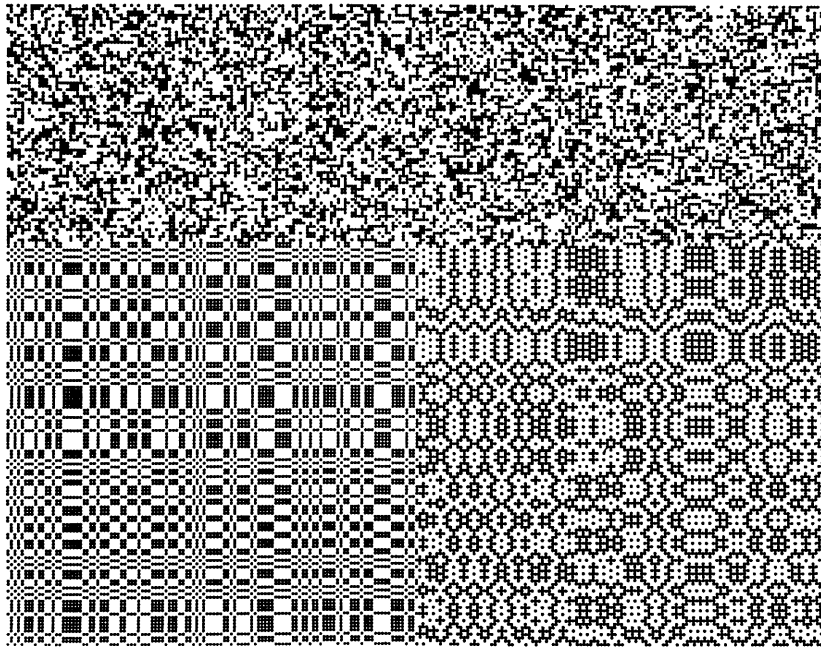


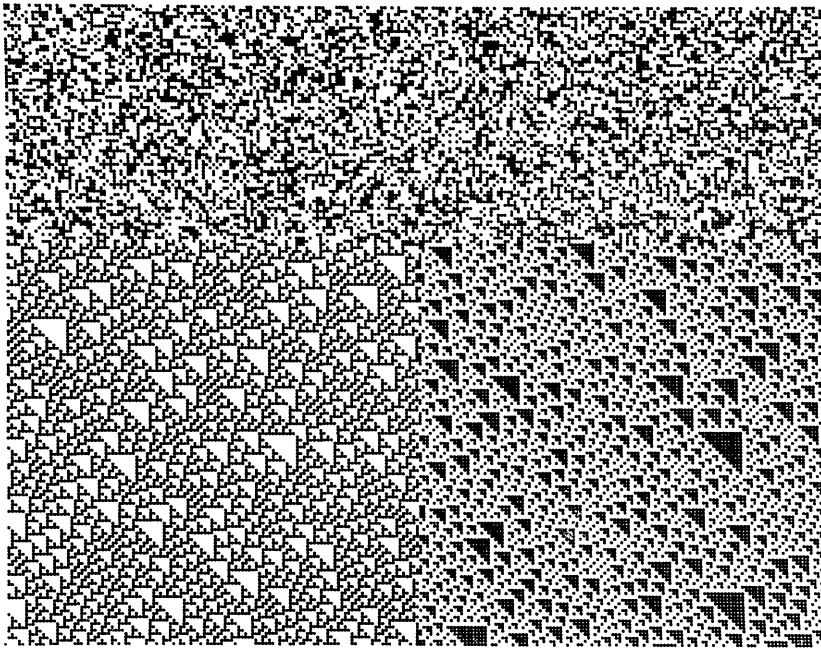
Fig. 9—Similar to Fig. 8, showing that aggregates of elements composed of the same textons cannot be preattentively discriminated. (*Philosophical Transactions*, 1980[25])

One of the simplest *global* computations routinely performed on images by engineers, and recently by psychologists in vision research, is to determine the images' Fourier power spectra. This process involves the decomposition of the images into one-dimensional sinusoidal luminance gratings whose specific amplitudes, spatial frequencies, phases, and angular orientations depend on the spatial characteristics of luminance distributions across the entire image. The amplitude of the spectral components ignoring phase determine the power spectra. When Fourier power spectra of textures are taken, it is a common misconception that differences in these will reveal differences in texture granularity. That the preattentive visual system does not perform Fourier analysis is demonstrated next.

Figure 10a consists of three areas that have identical Fourier power spectra (invented by Julesz, Gilbert and Victor[2]) and yet appear as very distinct textures. [The mathematically sophisticated reader might appreciate that the three areas have identical third-order statistics, and differ only in their fourth-order statistics. Those interested in the definition of *n*th-order statistics should consult Refs. 3 and 4 and the Appendix.] Figure 10b also consists of three areas with identical Fourier power spectra, and again these areas appear conspicuously different. Conversely, in Fig. 11a the lower left quadrant of the bottom



(a)



(b)

Fig. 10—Discriminable texture pairs with identical Fourier power spectra (*a* has even identical third-order statistics) based on local granularity (texton) differences. (*Biological Cybernetics,* 1981[2])

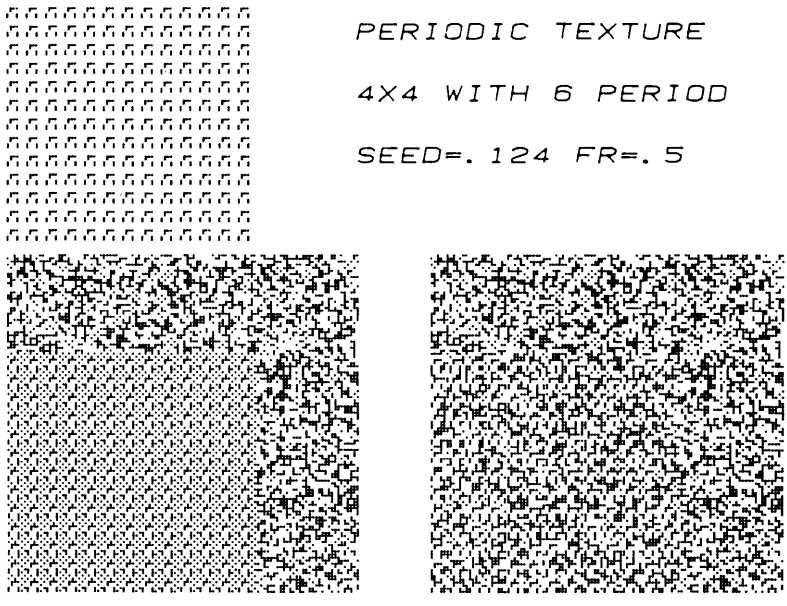
right array has a very different power spectrum from the remainder of the array, yet no preattentive texture discrimination results.[5] The derivation of this texture pair is presented in three steps. The top left array, in Fig. 11a, consists of 4x4 dot elements (8 black and 8 white) with 6-dot periodicity. The bottom left array contains this periodic array in one quadrant, but the 2-dot-wide gaps are filled by a checkerboard screen, while the rest is covered with uniformly random black and white dots. The bottom right array is similar to the bottom left array, except the 2-dot-wide gaps between the periodic patterns are now randomly speckled with dots. Obviously, the periodic patterns in the lower quadrant of the bottom right array in Fig. 11a yield a very different Fourier power spectrum from the rest, which has a flat (white noise) spectrum. The reason that this texture pair is indistinguishable can be easily understood in the light of the texton theory. The periodic patterns are not different from the surrounding random-dot array in the density of elongated blob textons, and therefore are indistinguishable. Indeed, if the 4x4 dot micropattern consists of vertical stripes, which contain textons different from the surrounding random-dot array, as shown in Fig. 11b, the periodic quadrant embedded in randomness is easily perceived.

In all these densely packed dot textures, discrimination is based on local granularity differences that correspond to differences in the density (number) of elongated blobs of certain sizes and orientations. Global statistical descriptors of textures, including the Fourier power spectrum, apparently are ignored in preattentive vision.

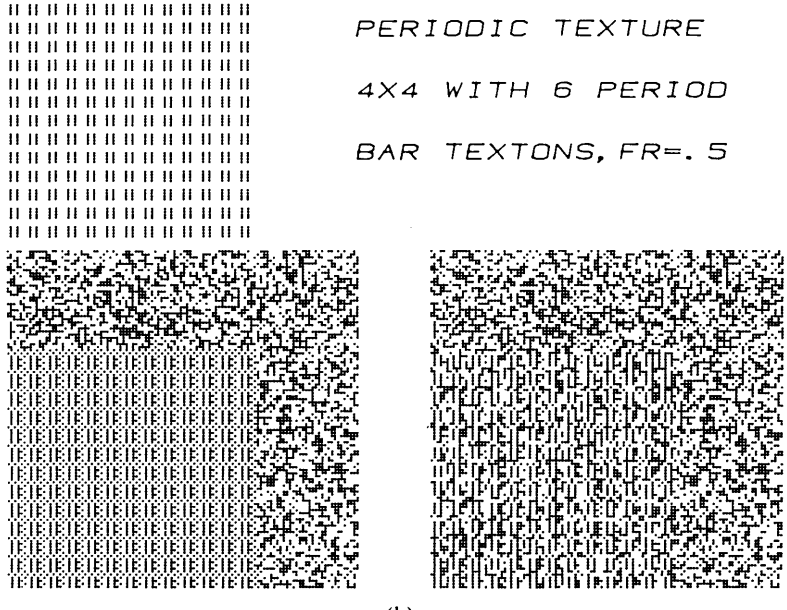
IV. TEXTURES AND NEUROPHYSIOLOGICAL FEATURE ANALYZERS

We have seen how elongated blob textons are crucial in preattentive texture perception. These human psychological findings have a parallel in primate neurophysiology. Neural units have been found by Hubel and Wiesel[6] in the visual cortex of monkeys that fire optimally for elongated blobs of specific width, length, and orientation. These neural units in the cortex have *retinal receptive fields* consisting of elongated, blob-shaped, excitatory regions, which are surrounded by inhibitory regions. Some of these elongated blob detecting units—which fire optimally for black elongated blobs surrounded by white flanking areas—are called simple "off" detectors. Other neural units are excited optimally by white elongated blobs surrounded by black. These are called simple "on" detectors. The exact shape of the receptive fields of these simple neural units varies a great deal, and is of secondary importance. The important property of these cortical units is that the weighting of the excitatory and inhibitory areas of their receptive fields is about equal, so that for homogeneous stimuli they do not fire.

It should be stressed that the textons reported here were found by



(a)



(b)

Fig. 11—Demonstration that the preattentive system cannot perform even such a simple global computation as Fourier power spectra, as described in text. (*Biological Cybernetics*, 1978[5])

psychological methods, and imply that simple neural units found as early as the striate cortex of the monkey might be used in texture perception. However, the relationship between a texton—for example, a perceived line segment—and a Hubel and Wiesel type of neural feature analyzer with a receptive field whose excitatory center matches the shape of the line segment is not a simple isomorphism. As we pointed out years ago (Ref. 7, p. 3), a single simple neural unit might respond equally for a broad line of high contrast or a narrow line of low contrast, while perceptually one can preattentively perceive both the width and contrast of a line segment. Thus, obviously a perceived line segment is encoded by many neural units of similar orientations but tuned to different widths, and having different firing thresholds. It is some combination of these units that would correspond to a perceived line segment. Until more is known about the relationship between perception and neurophysiology, the textons must be defined as perceptual entities, that is conspicuous local features as we actually perceive them. Nevertheless, even though textons and neural units are not simply related, one can easily conceptualize how a "perceptual analyzer" could be built from known neural analyzers that could extract, say, a line segment texton. The question of whether terminators and crossings of line segments—which have been regarded as textons—could be related to the complex and hypercomplex neural analyzers found by the neurophysiologists remains to be seen.[6]

David Marr, in his primal-sketch model of machine vision, also incorporated such elongated blob detectors, by assuming that the neurophysiological findings had direct relevance to vision.[8] The work reported here followed an opposite trend. It took almost two decades to find evidence for the utilization of simple cortical units in texture perception. Caelli and Julesz found the first elongated blob textons that could account for texture discrimination locally, when all global statistical properties of the texture pairs were kept identical.[9] Later demonstrations such as Figs. 10a and b illustrate even more strikingly the importance of local blob textons.

To demonstrate the possible role of the Hubel and Wiesel type of neural units in preattentive texture perception, we developed a computer program called TEXTONS that filters any image with a pool of elongated bar-shaped receptive fields. Each pool of filters consists of "on" and "off" types having the same width, length, orientation, and firing threshold and placed at each point of the array. Figure 12 (bottom) shows the three largest response levels of a pool of 3x3 dot square-shaped receptive fields as this pool processes the texture pair of Fig. 10a, shown also in Fig. 12 (top). These filters have 2x2 dot excitatory centers flanked by one-dot-wide inhibitory margins, as shown in Fig. 12 (right). Of course, there are several pools consisting

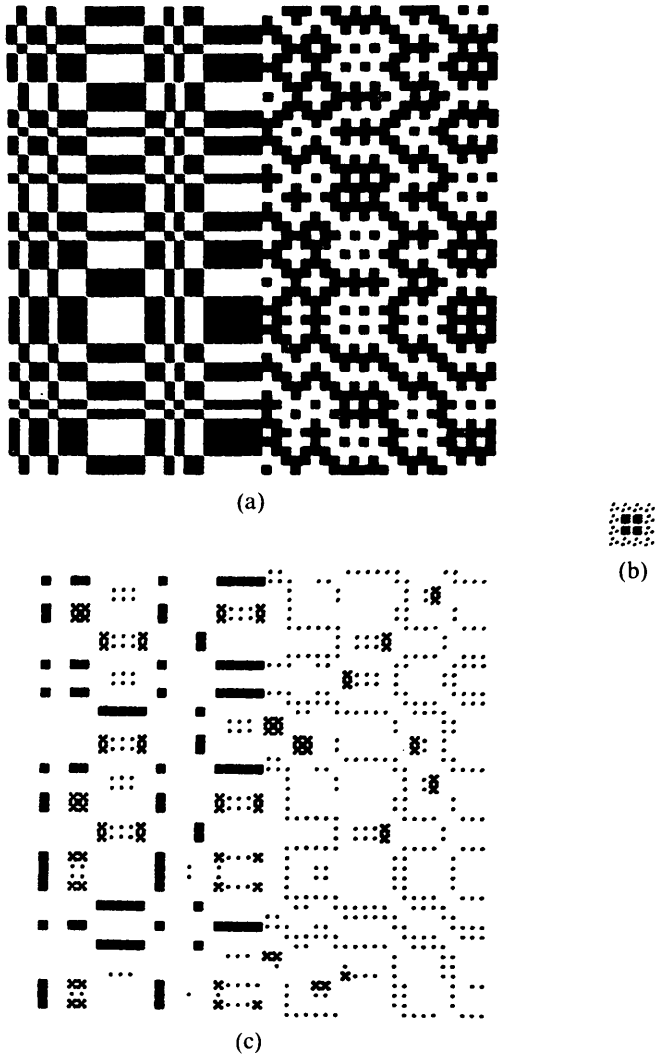


Fig. 12—Automatic texture segregation, shown in (c), by applying a texton filter (b) to texture pair (a) (also shown in Fig. 10a).

of filters having receptive fields with some other dimensions and orientations that would be even more effective in segregating the two textures of Fig. 12 (top). Here we stress again that the combination of several filters would be required to yield the best texture segregation, corresponding to human texture discrimination. This combination of filters would correspond to a texton detector.

What our psychological findings show, however, could not have been

guessed by physiologists and theoreticians of artificial intelligence. In preattentive texture perception the various textons are not coupled, that is their relative positions are ignored. T- and L-shaped pairs of line segments cannot be discriminated preattentively in textures. Marr thought that elongated blobs and terminators would form some higher molar unit, which he called place tokens.[8] However, in preattentive vision no such higher interactions are found; the textons appear to be independent of each other.

V. EXTENSION OF THE TEXTON THEORY TO RAPID PATTERN DISCRIMINATION

The success of the texton theory in predicting phenomena of texture perception is the result of the spatial complexity of the patterns. This complexity over a large area exceeds the capacity of focal attention and thus allows the preattentive system to dominate. This same deemphasis of focal attention can be achieved in simpler patterns by very brief presentation. We will show that under these conditions the same texton theory can be applied.[10]

Because brief temporal presentation is required, the stimuli used in these experiments can be produced only in the laboratory. Consequently, they cannot be demonstrated as the texture discrimination results have been. Thus, in this section we present the main findings as curves describing observers' performance.

The stimuli used in these experiments are shown in Fig. 13. In Fig. 13a there are 35 T's and one L arranged on a hexagonal grid with slight random positional jitter added. In Fig. 13b the T's have been replaced by + elements, and in Fig. 13c only two of the 36 possible positions actually contain an element. In all cases, a disk surrounding the central fixation marker is kept empty. Stimuli of this type are presented for 40 ms, followed by a blank interval of variable duration and a 40-ms erasing field. This erasing field consists of elements, which are the union of the two being discriminated, arranged in the same way as the test field. Use of this erasing technique allows restriction of the inspection interval to times shorter than the duration of the retinal afterimage. The times used are all too short to allow eye movements to be initiated during the presentation. In half of the presentations the test field consists of all identical elements, while in the other half one element is different, as in the examples shown. The task of the observer is to discriminate between these two conditions.

Results obtained using the three stimuli of Fig. 13 are shown in Fig. 14. On the abscissa is the time in milliseconds between the onsets of the test and erasing fields, or the stimulus onset asynchrony (SOA). On the ordinate is the percentage of correct discrimination. The results

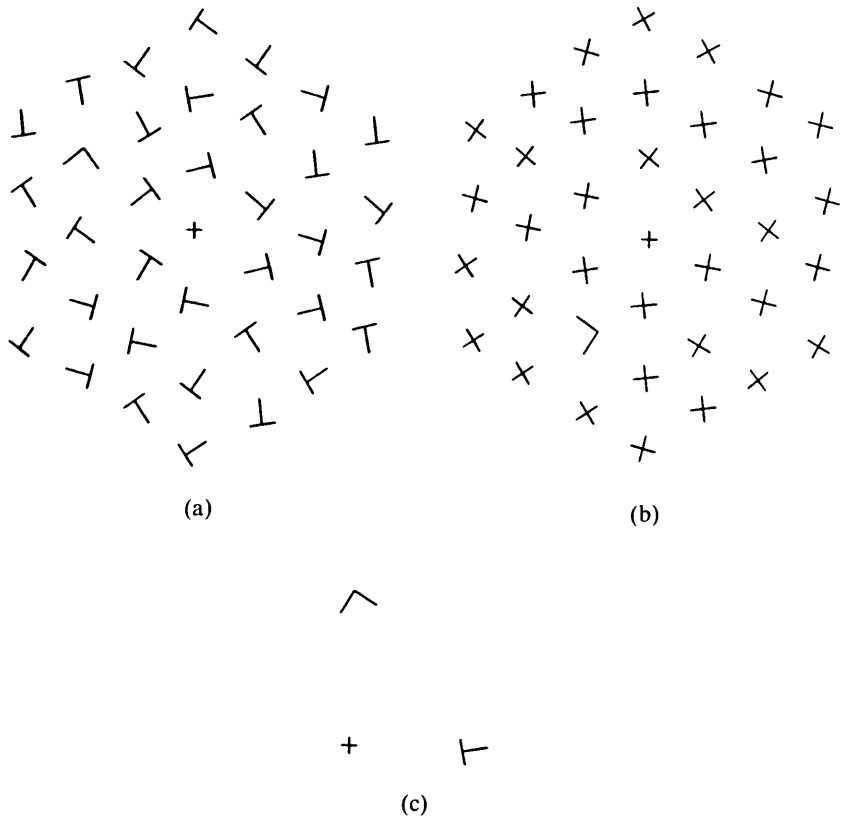


Fig. 13—The three types of stimuli used to obtain the results of Fig. 14. In all cases the task is to discriminate these from a control stimulus in which all elements are identical.

are very different for the case in which the elements share the same textons (T vs L. solid circles) from that in which they contain different textons (+ vs L, open circles). Note that in the T vs L case, not only does performance never exceed 65 percent correct, but it takes over 300 ms to reach this asymptote, while in the + vs L case the asymptote is reached within 200 ms. In fact, by the time the asymptote in the T vs L case is reached, the afterimage resulting from the test flash has largely disappeared. In the case in which only one T and one L are presented (filled squares), the results closely resemble those for 35 +'s and one L. Perceptually, the difference between the same-textons and different-textons cases is simply that the L in the field of T's stands out almost as if presented alone on a blank field, while the same L in a field of T's must be sought out. Attention is rapidly shifted to the L in the former case, while in the latter the search process is apparently still going on after 300 ms have passed. When there are only two elements to choose from, this search time is very brief.

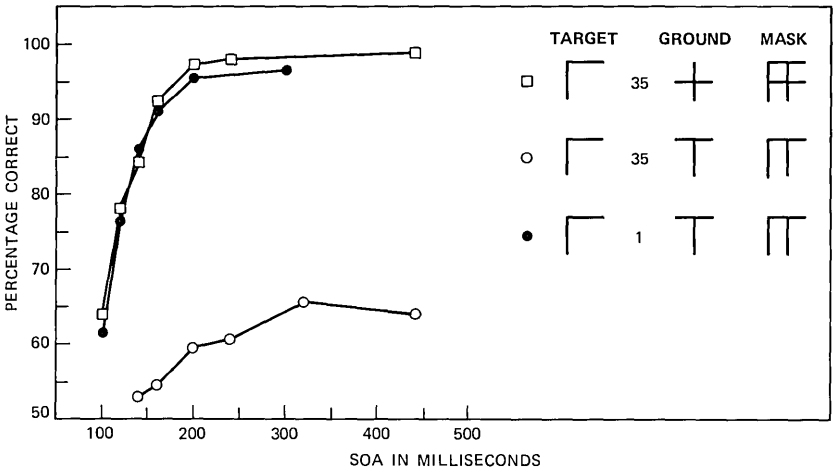


Fig. 14—Results of discrimination experiments using Fig. 13. The open circles, filled circles, and squares correspond to stimuli of Fig. 13a,c,b, respectively. SOA is the abscissa, and percent correct discrimination is the ordinate.

It is interesting to note that the observed asymptotic level of about 65 percent correct is what would be expected if seven or eight of the possible positions could be searched in the time available. Combining this number with the afterimage persistence time of 300–400 ms[11,12] gives a figure of about 50 ms per position inspected.[10]

This process of sequential inspection seems to be essentially independent of the overall angular subtense of the stimulus. Figure 15 shows results from an experiment in which the observer is required to distinguish a stimulus consisting of six T's and one L, or vice versa, from one in which all elements are identical. The stimulus was uniformly contracted so as to fall entirely within the fovea (<3 degrees across), or dilated to extend almost 14 degrees across, with no systematic variation in performance.

Another way of describing this is to say that the measurements are independent of the distance from which the stimulus is viewed, assuming that all of the elements remain resolvable: This independence suggests two important points. First, the fovea is not better than the near periphery in the extraction of this type of visual information. Second, the aperture of attention changes its spatial scale according to the size of the feature being sought. Thus, the same number of sequential fixations of attention are needed when the stimulus is reduced in size uniformly, because the sizes of the features upon which the discrimination is based are proportionally reduced. This extension of the scope of the texton theory from texture perception to rapid pattern discrimination suggests a model of vision in general as described in the following section.

VI. A MODEL OF THE "TWO VISUAL SYSTEMS"

When a visual scene changes suddenly in time or space, and our attention encompasses the entire visual scene, only those areas in which density differences in textons occur are conspicuous. These textons are elongated blobs with specific colors, widths, lengths, orientations, terminators, and crossings between them. Furthermore, because binocular disparity, movement disparity, and flicker are locally conspicuous features that can be detected in a brief presentation,[7,11,13] they, like color, are also properties of elongated blob textons.

Focal attention is directed to areas of spatial or temporal texton changes. The preattentive process appears to work in parallel and extends over a wide area of the visual field, while scrutiny by local or foveal attention is a serial process, which at any given time is restricted to a small patch. Focal attention can be shifted in 50-ms steps, four times faster than the fastest scanning eye movements. Furthermore, the aperture of focal attention can vary in size and can be a minute portion of the fovea, that is, extending to only a few minutes of arc (as shown in Fig. 15). Therefore, if the visual environment is rich in detail even when slowly changing in time, or is rather lacking in spatial detail but changes rapidly, we perform the major portion of our spatio-temporal processing in the preattentive state.

The focus of visual attention seems to be characterized by a texton class as well as a spatial locus. In particular, just as it apparently is impossible to attend simultaneously to two different places, it also seems impossible simultaneously to attend to very different sizes of features. This fact has been noted previously by other psychologists.[14] Stimuli widely separated in space produce cortical responses which are far apart. Similarly, stimuli of differing sizes often generate re-

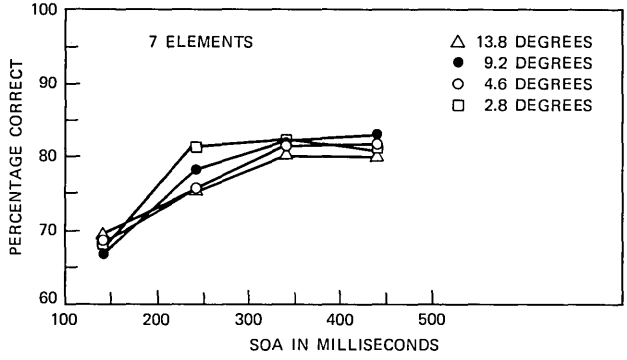


Fig. 15—Size invariance while the angular subtense of seven elements is varied from 2.8 degree of arc diameter to 13.8 degree of arc. Findings imply that the aperture of focal attention can be as small as a few minutes of arc.

sponses in different cortical areas.[6,15] These results seem to imply that the focus of attention is restricted to a very small region of visual cortex, and that stimuli producing responses far apart in the cortex cannot be attended simultaneously.

The essence of our findings is illustrated in Fig. 16. The left array contains a texture composed of "L"-shaped elements (formed by two perpendicular line segments with a gap), except for one "+" shaped and one "T" shaped element (formed by two perpendicular line segments which cross, or have a gap, respectively). The "+" shaped element (target) differs from the many surrounding L's in one texton, namely the "crossing", and perceptually stands out immediately. On the other hand, the T-shaped target can be detected only after some search, by directing the aperture of attention to the target itself.

The right array of Fig. 16 is identical to the left but illustrates our model of vision. The parallel preattentive system instantly detects the location of texton differences and directs the aperture of focal attention to this location, as indicated by the dotted disk around the "+". Since the T contains the same textons as its surround, its detection requires the aperture of attention (symbolized as a "cone" of a searchlight) to scrutinize the texture elements in sequence. Therefore, this

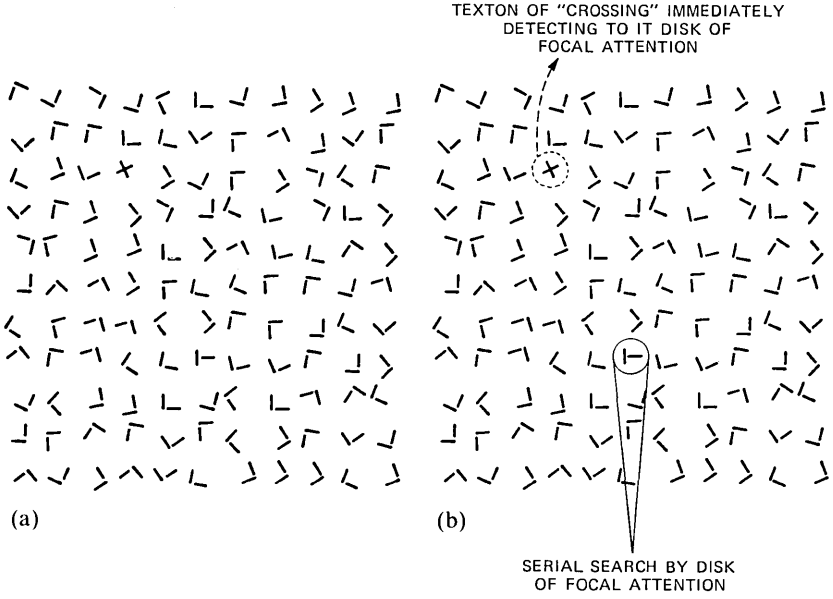


Fig. 16—Model of the two visual systems (b), showing how the preattentive system directs the aperture of focal attention to the loci of texton differences [the + in the L's in (a)], while without such texton differences [the T in the L's in (a)] focal attention requires time-consuming search.

serial search for the T-shaped target depends on the number of texture elements and may take considerable effort and time. However, after the T has been found, and the aperture of focal attention surrounds it, both the + and the T targets are seen with the same clarity. Obviously, form recognition, restricted to the aperture of focal attention, does not depend on the way attention has been directed to the targets. Whether a local difference in textons quickly directed focal attention to the target, or in the absence of texton differences it required time-consuming search to find the target, is immaterial for processing of the target by the attentive visual system.

This mode of behavior of the preattentive and attentive visual systems can also be observed in texture perception, when the reader inspects Fig. 1. The preattentive system immediately detects the texton differences at the boundary of the + and L aggregates, and a quick inspection by focal attention of a few elements on the two sides of the boundary lets the observer conclude that the two areas must contain +'s and L's. Only detailed scrutiny will reveal that the area believed to contain L's only has a region of T's as well.

In summary, the reason that texture discrimination is such a revealing process for showing the workings of the two visual systems is that textures usually cover wide areas of the visual field, while the texture elements are a small portion of the textural area. When the observer is inspecting an extended field, there is an "uncertainty region" in which the relative spatial position of local features is ignored. This is very different from a resolution limit due to visual acuity. In all of the indistinguishable texture pairs, the line segments which make up the texture elements are clearly resolved; nevertheless, if these textons fall within this uncertainty region, it is impossible to tell a T from an L. Many physiologists and psychologists have proposed two visual systems, one ambient and the other focal.[11-20] Yet, without the notion of textons, whose spatial and temporal changes are detected by the preattentive system, which in turn directs focal attention to these loci, the model of the two visual systems is not complete. We hope that the model outlined here gives some useful insights into human vision.

VII. IMPLICATIONS AND CONCLUSIONS

Some conspicuous local features called textons have been identified by psychological means. These textons, particularly the elongated blobs, are quite similar to features found to stimulate the simple neural units in the striate cortex of the monkey, which are selectively tuned to elongated blobs of certain colors, orientations, width, and length.

Our findings, that in preattentive vision objects are distinguished only through their texton decompositions, might be of considerable

importance. Since in preattentive vision these textons are not coupled, and furthermore the resolution of texton properties—i.e., the perceptual threshold for color, width, length, and orientation differences—is rather limited, the number of distinguishable textons is within practically useful bounds. (For example, the width of periodic bars can be judged with an error of 4 to 6 percent,[21] while accuracy of bar orientation is measured to be only 6 degrees of arc.[22]) This limitation makes practical the devices that simulate preattentive vision. This contrasts with attentive vision for which virtually an infinite number of recognizable patterns exist whose biological, social, or intellectual interest to the observer is unknown. Whether additional textons will be discovered remains to be seen. But as long as they remain independent of the previously isolated textons, the model outlined here will not be importantly affected.

The main implication of our findings is as follows: A considerable amount of vision is carried out by the preattentive system whose workings appear to be much simpler than that of the attentive system. This is important in judging the information requirements of the human visual system realistically. Furthermore, it is important to realize that even in the attentive mental state, with all its prodigious processing powers, complex feats of form recognition are restricted to a small spatial aperture, often as small as a few minutes of arc. Also, changing the position or extent of the aperture of focal attention requires considerable time. The shortest time is about 50 ms when eye movements are prevented, and as long as about 200 ms if saccadic eye movements are necessary.

This dichotomy between preattentive and attentive mental states, the first limited in its power of information processing, the latter limited in its spatial extent, gives a model of human vision that could be exploited in visual communication. Here we do not want to invent specific methods, but only indicate some obvious possibilities. With the advent of fast, perhaps parallel computers, the textons that direct the human observer's attention could be simultaneously extracted by hardware. Detailed images need only be presented in such areas.

Also, one could program computers to extract local features other than textons. For instance, a parallel computer might rapidly detect the difference between an L and a T, rather than between a + and an L. If an observer's attention were directed by such a machine, whose capabilities are very different from human preattentive vision, perhaps a new way of inspecting the visual environment could be made available and possibly learned.

The textons reported here help to discriminate textures, mainly surfaces of objects, without the need of complex familiarity cues. Such an early separation of the visual environment into figure and ground,

or objects and their backgrounds, is a fundamental operation of visual perception. Lack of understanding of this process is, as of now, the greatest bottleneck in machine vision, which in turn is necessary in extending the capabilities of robots.

Regardless of the feasibility of such ambitious schemes, the finding that texton differences can be almost instantaneously perceived over large areas of the visual field can be practically exploited in traffic signs and in directing attention to select areas of visual displays. Traditionally, flickering or static colored lights have been used as traffic signs, or in instrument panels. Now we can add other texton classes—for instance, gaps to increase the terminator number—to enhance visibility. For example, in Fig. 17 we show how a single gap introduced in the conventional alphabet draws attention to the word STOP, which otherwise would require a long time to be segmented and detected. Such slight modification of the alphanumeric characters (amounting to a new "font") might be beneficial in improving legibility. For instance, dyslexic children—children who cannot distinguish well between similar characters with different symmetric transformations such as b, d, or p—might greatly benefit if a gap or stroke were added to one of the characters, so that all characters would differ in at least one texton.

It should be stressed that the textons of preattentive vision only draw attention to certain areas, and we do not claim that these same textons are also the building blocks of form vision. If they were, our findings would prove preattentive vision to be the basis of attentive vision. Even if textons are restricted to vision in the preattentive state, we feel that to know those conspicuous features that grab our attention, wherever they appear, is of interest to everyone who wants to communicate through visual means.

VIII. ACKNOWLEDGMENTS

We thank our many colleagues who contributed to this research effort throughout the years, and who are mentioned in Section II and in the references. We also thank Max V. Mathews for his helpful comments while reading the manuscript. We are indebted to Walter

Fig. 17—Demonstration that the introduction of textons into the alphabet (here through increasing the terminator numbers by adding a gap) can help to segment and detect certain areas in a dense letter array.

Kropfl who developed the display hardware, and to Peter Burt who wrote the GENTEX program permitting the rapid generation, display, and manipulation of texture arrays. We thank our summer student, Franklin Schmidt, for developing the TEXTONS program.

REFERENCES

1. B. Julesz, "Textons, the Elements of Texture Perception, and Their Interactions," Nature, *290* (March 12, 1981), pp. 91–7.
2. B. Julesz, E. N. Gilbert, and J. D. Victor, "Visual Discrimination of Textures with Identical Third-Order Statistics," Biol. Cybernetics, *31* (1978), pp. 137–40.
3. B. Julesz et al., "Inability of Humans to Discriminate Between Visual Textures that Agree in Second-Order Statistics—Revisited," Perception, *2* (1973), pp. 391–405.
4. B. Julesz, "Experiments in the Visual Perception of Texture," Sci. Am., *232* (April 1975), pp. 34–43.
5. B. Julesz, "A Theory of Preattentive Texture Discrimination Based on First-Order Statistics of Textons," Biol. Cybernetics, *41* (1981), pp. 131–8.
6. D. H. Hubel and T. N. Wiesel, "Receptive Fields and Functional Architecture of Monkey Striate Cortex," J. Physiol., *195* (1968), pp. 215–43.
7. B. Julesz, *Foundations of Cyclopean Perception*, Chicago: University of Chicago Press, 1971.
8. D. Marr, "Early Processing of Visual Information," Philos. Trans. R. Soc. London Ser. B, *275* (1976) pp. 483, 524.
9. T. Caelli and B. Julesz, "On Perceptual Analyzers Underlying Visual Texture Discrimination: Part I," Biol. Cybernetics, *28* (1978), pp. 167–75.
10. J. R. Bergen and B. Julesz, "Discrimination with Brief Inspection Times," J. Opt. Soc. Am., *71*, No. 12 (December 1981), p. 1570. Also see "Parallel Versus Serial Processing in Rapid Pattern Discrimination," Nature, *303* (June 23–29, 1983), pp. 696–8.
11. B. Julesz, "Binocular Depth Perception of Computer-Generated Patterns," B.S.T.J., *39* No. 5 (September 1960), pp. 1125–62.
12. E. Averbach and G. Sperling, "Short-term Storage of Information in Vision," in C. Cherry (ed.) *Information Theory, Fourth London Symposium*, London: *Butterworth, 1961*, pp. 196–211.
13. B. Julesz and J. J. Chang, "Interaction Between Pools of Binocular Disparity Detectors Tuned to Different Disparities," Biol. Cybernetics, *22* (1976), pp. 107–19.
14. G. Sperling and M. J. Melchner, "Visual Search, Visual Attention and the Attention Operating Characteristics," in J. Requin (ed.) *Attention and Performance VII*, Hillsdale, NJ: Erlbaum, 1978, pp. 675–86.
15. S. M. Zeki, "The Functional Organization of Projections From Striate to Prestriate Cortex in the Rhesus Monkey," Cold Spring Harbor Symposia on Quantitative Biology, *15* (1976), pp. 591–600.
16. R. Held et al., "Locating and Identifying: Two Modes of Visual Processing, *31*, Psychol. Forsch., pp. 44–62; 299–348 (1967–1968).
17. C. B. Trevarthen, "Two Mechanisms of Vision in Primates," Psychol. Forsch. *31* (1968), pp. 299–337.
18. J. E. Hoffmann, "Hierarchical Stages in the Processing of Visual Information," Perception and Psychophysics, *18* (1975), pp. 348–54.
19. A. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," Cognitive Psychol., *12* (1980), pp. 97–136.
20. B. Julesz, "Visual Pattern Discrimination," IRE Trans. Inform. Theory, *IT-8* (February 1962), pp. 84–92.
21. F. W. Campbell, J. Nachmias, and J. Jukes, "Spatial-Frequency Discrimination in Human Vision," J. Opt. Soc. Am., *60* No. 4 (April 1970), pp. 555–9.
22. J. P. Thomas and J. Gille, "Bandwidths of Orientation Channels in Human Vision," J. Opt. Soc. Am., *69*, No. 5 (May 1973), pp. 652–60.
23. M. Rosenblatt and D. Slepian, "Nth Order Markov Chains With Any Set of N Variable Independent," J. Soc. Indust. Appl. Math., *10* (1962), pp. 537–49.
24. T. Caelli, B. Julesz, and E. N. Gilbert, "On Perceptual Analyzers Underlying Visual Texture Discrimination: Part II," Biol. Cybernetics, *29* (1978), pp. 201–14.
25. B. Julesz, "Spatial Nonlinearities in the Instantaneous Perception of Textures with Identical Power Spectra," Phil. Trans. R. Soc. Lond. B, *290* (1980), pp. 83–94.
26. B. Julesz, "Perceptual Limits of Texture Discrimination and Their Implications to

Figure-Ground Separation," in E.L.J. Leeuwenberg and H.F.J.M. Buffart (eds.), *Formal Theories of Visual Perception*, New York: Wiley, 1978, pp. 205–16.
27. H. L. Frisch and F. H. Stillinger, "Contribution to the Statistical Geometric Basis of Radiation Scattering," J. Chem. Phys., *38* (1963), pp. 2200–7.
28. N. Wiener, "Extrapolation, Interpolation and Smoothing of Stationary Time Series, With Engineering Applications," New York: Cambridge University Press.

APPENDIX

It required two decades of research efforts to discover that preattentive texture perception depends on local features alone and that global higher-order statistical parameters can be ignored. In 1962, Julesz asked mathematicians to generate stochastic texture pairs that would be identical in their first (n-1)th order statistics, but different in the nth- and higher than nth-order statistics.[20] The nth-order statistics are similar to the well-known nth-order joint probability distribution of n samples. The n samples are n points of a texture selected at random. However, in random geometry the shape of the n samples is of importance.

These n points can be regarded as the vertices of an n-gon. The n-gon (or nth-order) statistics are obtained when these n points (having the same n-gon shape) are selected at random, and statistics indicate that these n points have certain color values. For instance, the second-order statistics can be obtained if a 2-gon (dipole, or needle) is randomly thrown at the texture and the probability is determined that the two end-points of the dipole—of given lengths and orientations—fall on certain color combinations: e.g., black and black; or black and white; or black and gray, etc.

In the intervening years many such stochastic textures were discovered, particularly with identical first- and second-order statistics, but different third- and higher-order statistics.[3,23–25] As a matter of fact, the texture pairs in Figs. 3–6, and 8–10 have this property. The finding that many of these *iso-second-order* texture pairs differing only in third- and higher-orders are indistinguishable suggests that the preattentive visual system cannot compute statistical difference beyond the second order. The recent finding by Julesz, demonstrated in Fig. 11a, suggests that the preattentive visual system cannot even process second-order statistical parameters.[4] From the second-order statistics the autocorrelation function can be uniquely determined—as a matter of fact, for two-tone textures composed of black and white dots, the second-order (dipole) statistic *is* the autocorrelation function[26,27]—and the Fourier transform of the autocorrelation is the Fourier power spectrum.[28] Therefore, all the texture pairs with identical second-order statistics also have identical power spectra. The finding that texture segregation can be obtained in iso-second-order textures, after it was established that the preattentive system cannot process third-order statistics (and, as Fig. 11a demonstrates, not even second-order statis-

tics), implies that this segregation must be based on *local* density differences. Finally, it was proposed that the density changes of certain local conspicuous features, the textons, explain preattentive texture discrimination.[1,25]

AUTHORS

James R. Bergen, A.B. (Mathematics and Psychology), 1975, University of California, Berkeley; Ph.D. (biophysics and theoretical biology), 1981, University of Chicago; Postdoctoral Fellow, Bell Laboratories, 1981–82; RCA, 1983—. Mr. Bergen's work concerns the quantitative analysis of information processing in the human visual system. At the University of Chicago he was involved in the development of a model of the spatial and temporal processing which occurs in the early stages of the system. At Bell Laboratories, his work has concentrated on the effect of visual system structure on the extraction of information from a visual image. Mr. Bergen recently joined RCA Laboratories in Princeton, NJ.

Bela Julesz, Diploma, 1950 (Electrical Engineering), Technical University, Budapest; Ph.D., 1956, Hungarian Academy of Sciences; Bell Laboratories, 1956—. Mr. Julesz taught and did research in communications systems for several years prior to 1956. Since joining Bell Laboratories, he has devoted himself to visual research, particularly depth perception and pattern recognition. He is the originator of the random-dot stereoimage technique and of the method of studying texture discrimination by constraining second-order statistics. He has written extensively in the area of visual and auditory perception and is the author of *Foundations of Cyclopean Perception.* Mr. Julesz was Head of the Sensory and Perceptual Processes Department from 1964 to 1982, and in 1983 was made Head of the Visual Perception Research Department. He has been visiting professor of experimental psychology at M.I.T. and other universities. In February 1983 he received the MacArthur Prize Fellow Award in Experimental Psychology and Artificial Intelligence. He was a Fairchild Distinguished Scholar at the California Institute of Technology from 1977 to 1979. Fellow, AAAS, OSA, and the American Academy of Arts and Sciences; Corresponding Member of the Goettingen Academy of Sciences.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Central Control of Movement Timing

By D. A. ROSENBAUM*

(Manuscript received October 14, 1981)

How do people control the delay between successive finger movements? A reaction-time experiment suggests that the delay is controlled by setting an internal "alarm clock."

I. INTRODUCTION

A fundamental concern of behavioral science is to understand how we control the movements of our bodies. This paper examines the timing of body movements: As we walk, talk, or type, how are the delays between the phases of these various activities controlled?

A popular means of approaching this question has been to consider two extreme views of how movements might be timed. One view holds that delays between the phases of a movement sequence are controlled by using feedback from each phase to trigger the next. Since the feedback can arise from receptors in the peripheral nervous system—within the muscles, joints, and skin—this view is often referred to as the "peripheralist" theory.[1] The other view holds that movement timing is controlled by plans or "programs" that allow sequences to be performed without the aid of feedback. This view is often called the "centralist" theory.[2]

* Work done at Bell Laboratories. Now at School of Language and Communication, Hampshire College, Amherst, Massachusetts.

Two main arguments have been leveled against the peripheralist position. One argument is that times between successive movements may often be shorter than the time needed for feedback from one movement to be used to trigger the next.[3] The other argument is that animals and people in whom peripheral feedback is physiologically disrupted can often perform movements effectively.[4,5] Both of these observations indicate than when feedback cannot be relied on, central programs may be used. What the observations do not resolve, however, is whether central programs are also used when feedback is available. The experiment reported here was meant to answer this question. The experiment was also meant to provide detailed information about the nature of the programming of movement sequences when it appears that programming is used.

II. THE EXPERIMENT

The experiment, which has been reported in detail elsewhere,[6,7] used a simple procedure (see Fig. 1). The subject's task was to perform a sequence of two responses, made with the left and right index fingers, in such a way that the produced interresponse interval approximated a target interval. Feedback given at the end of each trial indicated whether the produced interval was longer or shorter than the target interval and by how much. In addition to producing the specified

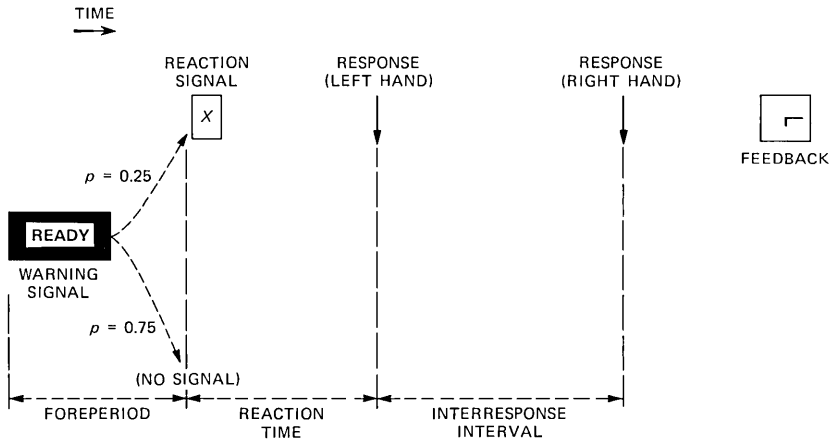


Fig. 1—Overview of the experimental procedure. "Catch trials," in which no signal was presented and no response was allowed, were used to discourage anticipation responses. The warning, reaction, and feedback signals were displayed on a CRT screen. The feedback signal consisted of a vertical line that pointed up if the produced interval was too long and down if the produced interval was too short, where the length of the line indicated the size of the proportional error; the length of a horizontal line indicated how much the Reaction Time (RT) exceeded a target of 130 ms. The foreperiod was fixed at 0.5 second, the reaction signal remained on the screen until the last required response was made, and the feedback signal came on 0.5 second after the offset of the reaction signal and remained on the screen for 2.5 seconds.

interresponse interval, the subject was also required to make the first response as quickly as possible after the onset of a visual reaction signal.

The rationale for the experiment can be understood by noting that several investigators have found that RTs are longer for the first of two responses, performed approximately simultaneously, than for single finger responses.[8–14] If it is assumed that the lengthening of RTs for response doublets results from the demands associated with coordinating two responses, the question that arises is how long the delay between two responses must be before the RT for the first response is as short as the RT for a single, isolated response. According to the peripheralist theory, one would expect the RT for the first of two responses to equal the RT for a single, isolated response whenever the delay between the two responses equals or exceeds the time required to use peripheral feedback from the first response to trigger the second response. Thus, if it takes 200 ms to use peripheral feedback from one response to trigger the next, the peripheralist theory would predict that, whenever the delay between the two responses is greater than or equal to 200 ms, the RT to initiate the two responses in series should be no longer than the RT to initiate the first response alone. By contrast, according to the centralist theory, the RT to initiate the two responses in series should be longer than the RT to initiate the first response alone, even if the delay between the two responses exceeds the feedback loop time; for according to the centralist theory the availability of peripheral feedback does not suffice to control the onset of the second response relative to the first.

Figure 2 shows the results from the main experiment designed to test these predictions. In this experiment, the target interresponse intervals ranged (in a blocked design) from 0 to 1050 ms. In the control condition no second (right-hand) response was required, so the target interval was effectively infinite (∞). The data are averaged over three adult female subjects. In the "stringent" condition, the length of the vertical feedback line was directly proportional to the deviation of the produced interval from the target (nonzero, finite) interval such that the line reached the edge of the CRT screen for deviations exceeding 100 percent of the target interval. In the "relaxed" condition, the scale factor relating the size of the deviation to the length of the vertical feedback line was reduced by a factor of 25 so that generally subjects could only see whether their produced interresponse intervals were too long or too short. About 1300 observations contribute to each point in Fig. 2.

Inspection of Fig. 2a reveals that none of the two-response sequences had mean RTs less than or equal to the mean RT in the control condition, as was confirmed statistically. Since the larger interre-

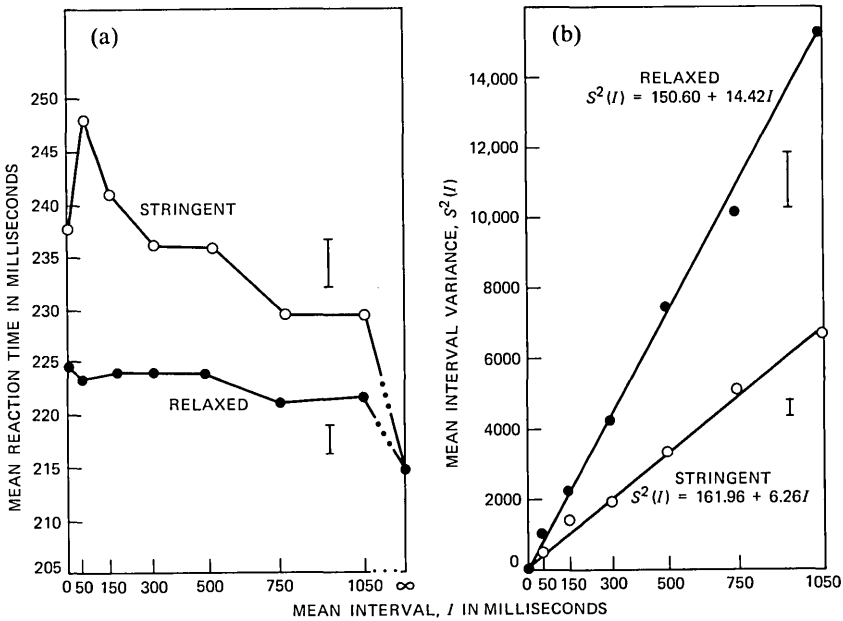


Fig. 2—Results of the experiment. (a) Mean reaction times and estimates of standard error (±SE) for the two accuracy conditions. (b) Mean variances of produced interresponse intervals, fitted linear functions, and estimates of ±SE. The zero intercepts of the fitted linear functions for the two accuracy conditions were not significantly different from one another. (Estimates of ±SE are based on mean squares from fits of mean functions to individual subject data.)

sponse intervals were certainly large enough to allow for the use of peripheral feedback,[15] this result is consistent with the centralist prediction and inconsistent with the peripheralist prediction.

Given that control of the interresponse interval was apparently not based solely on peripheral feedback, what was the nature of the central activity that allowed for control of the intervals? A plausible hypothesis is that the interresponse intervals were controlled by a kind of hydraulic system in which the amount of energy made available to the second response increased as the time until its occurrence decreased. If the total energy available to the two responses was fixed, then the amount of energy available to the first response would decrease as the desired interresponse interval decreased, leading to an inverse relationship between RTs and interresponse intervals. This is a response competition model. (For a more general discussion of response competition, see Ref. 16.)

An alternative hypothesis is that the interresponse interval was controlled by a central clock that was set just before the production of the first response (i.e., during the RT). If the time needed to set the clock depended on the precision of setting, which in turn was reflected

in the variability of the interresponse interval, this alarm clock model predicts that RTs could depend on interval variability rather than interval size.

To test these predictions, we manipulated the analog feedback that subjects received concerning the accuracy of their interresponse intervals. Recall that the analog feedback took the form of a vertical line that pointed up if the produced interval was larger than the target interval and down if the produced interval was smaller than the target interval. By changing the scale factor relating the length of the line to the proportional difference between the produced and target interresponse interval, we could impose stringent or relaxed accuracy requirements on the subject's timing performance. When the scale factor was small, a large error resulted in a small error line ("relaxed" condition), but when the scale factor was large, even a small error resulted in a large error line ("stringent" condition). The relaxed-stringent manipulation was a within-subject variable, which we introduced to get subjects to alter the variability of their interresponse intervals while keeping the means of the intervals constant.

As is seen in Fig. 2a, mean RTs in the relaxed condition were essentially constant over the range of finite intervals. However, mean RTs in the stringent condition were longer than in the relaxed condition and decreased as intervals increased for targets 50 ms or greater. [The discontinuity in the stringent mean RT function at the 0-ms target confirmed subjects' reports that control of the intervals in this condition was qualitatively different from the other conditions (e.g., because the fingers and wrists were kept stiff when simultaneous responses were attempted). Sequences begun with the right hand, which were only permitted in the 0-ms condition, did not have shorter RTs than sequences begun with the left hand.] Thus, Fig. 2a shows that mean RTs were not simply related to mean intervals, contrary to the prediction of the response competition hypothesis. Instead, mean RTs were related to the required precision of the intervals, consistent with the alarm clock model.

The question that now arises is whether the actual precision of the intervals was affected by the differential precision requirements imposed by the stringent-relaxed manipulation. If the actual precision was affected, such that greater precision was achieved in the stringent condition, the evidence would be stronger for the hypothesis that during the RT an internal alarm clock was set to varying degrees of precision. In Fig. 2b variances of produced intervals are plotted against their corresponding means. As is seen in the figure, variances were larger in the relaxed condition than in the stringent condition. That the interval means were approximately equal in the two conditions is demonstrated by the approximately vertical alignment of the filled

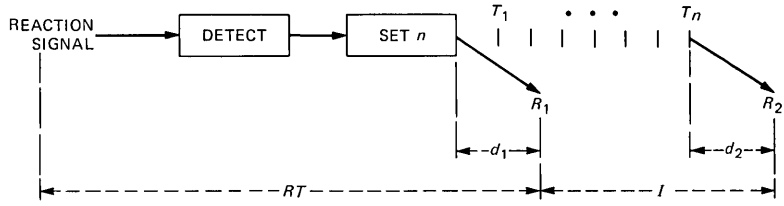


Fig. 3—Overview of an alarm-clock model for performance in the experiment.

and empty points corresponding to each target interval. As is also seen in Fig. 2b, the mean interval variances were well fit by linear functions. Linear regression accounted for 99.7 percent and 99.9 percent of the variance among the mean variances (averaged over subjects) in the stringent and relaxed conditions, respectively. The linearity of the variance functions allows for a simple elaboration of the alarm clock model, which is discussed below.

III. AN ALARM CLOCK MODEL*

The elaborated alarm clock model is shown in Fig. 3. It says that the size of the delay between the first and second response (R_1 and R_2, respectively) depends on the number, n, of pulses of an internal clock, where the value of n is set during the RT interval. The model assumes that the interval, I, between R_1 and R_2 equals the sum of the times, $T_1, T_2, \cdots, T_n$ for the n clock pulses, plus the difference between the motor delays, d_1 and d_2, of R_1 and R_2, respectively. That is,

$$I = T_1 + \cdots + T_n + d_2 - d_1. \tag{1}$$

It is also assumed that the times for the successive clock pulses fluctuate randomly about a mean and are stochastically independent. Hence, as n increases, the variance of the cumulative pulse time increases linearly.[17,18]

The model affords two plausible ways of explaining the difference in slopes for the fitted variance functions in the stringent and relaxed conditions. One possibility is that the variability of interpulse times was greater in the relaxed condition than in the stringent condition. The other possibility, which will be pursued here because of its interesting ramifications for the proposed clock-setting process, is that the variance of n was greater in the relaxed condition than in the stringent condition. It is possible to express precisely how the variance of n in the relaxed condition would be related to the variance of n in the stringent condition, given that the slopes (but not the zero intercepts) of the variance functions differed in the two conditions. Let the

* This is a greatly simplified version of a model presented in Ref. 6.

random variable N be the number of clock pulses between the triggering of response 1 and 2. Let the time between pulses $i-1$ and i, for $1 \leq i \leq n$, be a random variable T, where the random variables T_i are independent of N and identically distributed as a random variable T with finite mean and variance. Allowing the random variable D to represent the difference $d_2 - d_1$, and assuming T, N, and D are mutually independent, it is possible to use the expression for the variance of a random sum[19] to obtain

$$\mathrm{Var}(I) = E(N)\mathrm{Var}(T) + \mathrm{Var}(N)E^2(T) + \mathrm{Var}(D). \tag{2}$$

Assuming that the only terms in (2) that differ in the relaxed and stringent conditions are $\mathrm{Var}(I)$ and $\mathrm{Var}(N)$, and using subscripts R and S for the relaxed and stringent conditions, respectively, we can subtract $\mathrm{Var}_S(I)$ from $\mathrm{Var}_R(I)$ and rearrange terms to obtain

$$\mathrm{Var}_R(N) - \mathrm{Var}_S(N) = [E^2(T)]^{-1} \cdot [\mathrm{Var}_R(I) - \mathrm{Var}_S(I)]. \tag{3}$$

Since $\mathrm{Var}_R(I)$ and $\mathrm{Var}_S(I)$ were found to be linear functions of I with what are assumed to be identical zero intercepts, and since $E^2(T)$ is assumed to be a constant, eq. (3) becomes

$$\mathrm{Var}_R(N) - \mathrm{Var}_S(N) = kI, \tag{4}$$

where k is a constant greater than 0. Equation (4) shows that $\mathrm{Var}_R(N)$ and $\mathrm{Var}_S(N)$ are related in a simple way. As is seen below, however, the relationship can be shown to be even simpler.

Consider how the variance of N changes with I within the stringent or relaxed condition. Relying on the fact that we found $\mathrm{Var}(I)$ and I to be linearly related, and noting that

$$E(N) - [E(I) - E(D)]/E(T), \tag{5}$$

eq. (2) can be rewritten to solve for $\mathrm{Var}(N)$ as

$$\mathrm{Var}(N) = \frac{\alpha + \beta \cdot E(I) - \mathrm{Var}(D) - \left[\dfrac{E(I) - E(D)}{E(T)}\right] \mathrm{Var}(T)}{E^2(T)}. \tag{6}$$

All the terms in eq. (6) are constant except for $\mathrm{Var}(N)$ and $E(I)$. Therefore, eq. (6) can be rewritten

$$\mathrm{Var}(N) = u + vE(I), \tag{7}$$

where u and v are constants. Since the clock presumably would not be used (i.e., N would not be set) when $E(I) = 0$, u can be set to 0, leaving

$$\mathrm{Var}(N) = vE(I). \tag{8}$$

On the basis of the simple relation shown in eq. (8), we can indicate how $\mathrm{Var}(N)$ in the stringent condition, $\mathrm{Var}_S(N)$, is simply related to

$\mathrm{Var}(N)$ in the relaxed condition, $\mathrm{Var}_R(N)$. Let $\mathrm{Var}_R(N) = rE(I)$. As we saw in eq. (4), $\mathrm{Var}_R(N) - \mathrm{Var}_S(N) = kE(I)$. It follows, then, that $\mathrm{Var}_S(N) = (r - k)E(I)$. Thus, $\mathrm{Var}_R(N)$ and $\mathrm{Var}_S(N)$ differ with respect to the size of the proportionality constant relating each of these quantities to $E(I)$.

To summarize, the above discussion reveals that the variance of N is proportional to $E(I)$, and that the slope of this function differs in the relaxed and stringent conditions.

The simplicity of these theoretical results allows for the proposal and test of a simple model of the process that may have given rise to the mean RT effects shown in Fig. 2a. The model is suggested by the observation that shorter RTs were associated with larger (predicted) values of $\mathrm{Var}(N)$. Essentially, the model says that the time to set N increases as the range of possible values of N decreases. (A possible reason will be given in Section IV of this paper.) Let θ denote the time to set N and recall that the range of a distribution is proportional to the standard deviation of that distribution. The model says

$$\theta = \frac{\gamma}{\sqrt{\mathrm{Var}(N)}} = \frac{\gamma}{\sqrt{\delta E(I)}} = c[E(I)]^{-1/2}, \tag{9}$$

where $c = \gamma/\sqrt{\delta}$ is an empirical constant and $0 < E(I) < \infty$. From the previous observation that the proportionality constant relating $\mathrm{Var}(N)$ to $E(I)$ distinguishes the relaxed and stringent conditions, it follows that only the size of c in eq. (9) needs to differ in the relaxed and stringent conditions. Moreover, as is implied in Fig. 3, the RT is assumed to equal θ plus the time to detect the reaction signal and execute the first response. Hence, the model predicts that the RT in the relaxed and stringent conditions can be expressed as

$$\mathrm{RT}_R = k + \theta_R$$

and

$$\mathrm{RT}_S = k + \theta_S, \tag{10}$$

respectively, or

$$\mathrm{RT}_R = k + r[E(I)]^{-1/2}$$

and

$$\mathrm{RT}_S = k + s[E(I)]^{-1/2}, \tag{11}$$

respectively, where k, r, and s are empirical constants. Equation (11) has the appealing property that RTs increase with decreases in $E(I)$ at a rate related to the size of the linear coefficient. Thus, changing the linear coefficient in eq. (11) produces a set of curves two of whose members look like the pair of curves seen in Fig. 2a.

For purposes of finding out how well eq. (11) actually fit the RT data, an iterative procedure was used to find the values of k, r, and s for which it was possible to minimize the sum of squared deviations of the obtained RTs from the predicted RTs. The obtained RT values used were the mean RTs in the conditions where nonzero finite interresponse intervals were required. The best-fitting values of r and s turned out to be 25 and 236, respectively, and the best-fitting value of k turned out to be 221 ms. With these estimates the fitted model accounted for 94.6 percent of the variance of mean obtained RTs. (When k was held at 216 ms, which was the mean RT in the control condition, only 67.5 percent of the variance was accounted for. Because the goodness of fit was so much poorer here, it appears that the time required for detection of the reaction signal and execution of the first response differed in the finite and infinite-interval conditions, or that some extra process, not identified in the model, in fact occurred in the finite-interval conditions.)

IV. CONCLUDING REMARKS

This paper has examined the mechanisms of movement timing. Of necessity, the experiment that has been reported involved a restricted type of movement, namely simple finger sequences. Yet the experiment appears to have fairly broad theoretical importance for our understanding of how movements are timed. In particular, the experiment suggests that *even when peripheral feedback is available, central programs play a controlling role in movement timing.* The basis for this conclusion is that the mean RT for the first of two responses was longer than the mean RT for a single, isolated response, even when the time between the first and second response exceeded 1 second. That the lengthening of RTs at such long delays was in fact attributable to some aspect of the control of movement timing is suggested by the fact that the lengthening of RTs was greater when the interresponse interval was controlled more precisely.

Besides supporting a general, qualitative principle about movement control, the data from the present experiment also allow for a relatively detailed model of the control of movement timing. The model likens the selection and control of an interresponse interval to the setting and running of a conventional alarm clock: First, the desired number, n, of clock pulses is selected (during the RT), and then the clock is allowed to produce the n pulses so that the first response is triggered at the time of the first pulse and the second response is triggered at the time of the nth pulse. The proposal that there is a clock-pulsing mechanism is motivated by the finding that the variance of the interresponse interval increased linearly with the interval mean, which is suggestive of a *stochastic wait* process (see Refs. 17 and 18) where

times between successive pulses fluctuate randomly about a mean and are stochastically independent. Through some algebra, it was seen that another factor contributing to the interval variance is the variance of N, which was seen to be proportional to the interval mean. The size of this proportionality factor was seen to depend on the slope of the linear function relating the interval variance and mean, and this slope was observed to differ in the relaxed and stringent conditions of the experiment. Having established the above relationships, it was observed that the mean RTs in the nonzero, finite-interval conditions were inversely related to the (predicted) variance of N. Then, through a curve-fitting procedure, it was seen that this relationship could account for almost 95 percent of the variance of the mean RTs in the nonzero finite-interval conditions. The inverse relationship between the mean RT and predicted variance of N can be explained by saying that it takes longer to select a particular value of N as the range of allowable values decreases. Such a relationship can be understood in statistical terms as reflecting a decreasing likelihood of selecting an allowable value of N as the range of allowable values of N decreases. If reselection of a value of N is required when an unallowable value has been picked, the average time to select an allowable value would vary in a way consistent with the mean RT results obtained here.

It is interesting to note that the physiological basis for an alarm-clock system has been demonstrated. In the cortex of the cerebellum, rows of regularly spaced neurons known as Purkinje cells are connected to other neurons in such a way that the delay between the entry of a signal into the row and the exit of a signal from the row depends on which Purkinje cell is permitted to conduct the output signal.[20,21] The process of selectively permitting a given Purkinje cell to conduct an output signal on the basis of the delay that would occur prior to that output signal is physiologically analogous to the clock-setting process proposed here.

V. ACKNOWLEDGMENTS

The author wishes to thank J. L. Knight and J. F. Kroll for their helpful comments during the preparation of this paper.

REFERENCES

1. J. A. Adams, "Feedback theory of how joint receptors regulate the timing and positioning of a limb," Psychological Review, *84* (November 1977), pp. 504–23.
2. S. W. Keele, "Movement control in skilled motor performance," Psychological Bulletin, *70*, No. 6, Part 1 (1968), pp. 387–403.
3. K. S. Lashley, "The problem of serial order in behavior," in L. A. Jeffress (Ed.), *Cerebral Mechanism in Behavior*, New York: Wiley & Sons, 1951.
4. E. V. Evarts, E. Bizzi, R. Burke, M. DeLong, and W. T. Thach, Jr., *Central Control of Movement*, Brookline, MA: Neurosciences Research Program, 1971.
5. E. Taub and A. J. Berman, "Movement and learning in the absence of sensory

feedback," in S. J. Freedman (Ed.), *The Neuropsychology of Spatially Oriented Behavior*, Homewood, IL: Dorsey Press, 1968.
6. D. A. Rosenbaum and O. Patashnik, "A mental clock setting process revealed by reaction times," in G. E. Stelmach and J. Requin (Eds.), *Tutorials in Motor Behavior*, Amsterdam: North-Holland Publishing Co., 1980.
7. D. A. Rosenbaum and O. Patashnik, "Time to time in the human motor system," in R. S. Nickerson (Ed.), *Attention and Performance VIII*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.
8. M. DiStefano, M. Morelli, C. A. Marzi, and G. Berlucchi, "Hemispheric control of unilateral and bilateral movements of proximal and distal parts of the arm as inferred from simple reaction time to lateralized light stimuli in man," *Experimental Brain Research, 38* (1980), pp. 197–204.
9. N. Hammond and P. Barber, "Evidence for abstract response codes: Ear-hand correspondence effects in a three-choice reaction-time task," Quart. J. Experimental Psych., *30* (February 1978), pp. 71–82.
10. M. A. Jeeves, "A comparison of interhemispheric transmission times in acollosals and normals," Psychonomic Science, *16*, No. 5 (1969), pp. 245–6.
11. J. A. S. Kelso, D. L. Southard, and D. Goodman, "On the coordination of two-handed movements," Science, *203* (1979), pp. 1029–31.
12. J. R. Peterson, "Response-response compatibility effects in a two-hand pointing task," Human Factors, *7* (1965), pp. 231–6.
13. P. M. A. Rabbitt, S. M. Vyas, and S. Fearnley, "Programming sequences of complex responses," in P. M. A. Rabbitt and S. Dornic (Eds.), *Attention and Performance V*, London: Academic Press, 1975.
14. H. C. Ratz and D. Ritchie, "Operator performance on a chord keyboard," J. Appl. Psych., *45* (1961), pp. 303–8.
15. J. A. Adams, "Issues for a closed-loop theory of motor learning," in G. E. Stelmach (Ed.), *Motor control: Issues and trends*, New York: Academic Press, 1976.
16. L. M. Herman and B. H. Kantowitz, "The psychological refractory period effect: Only half the double-stimulation story?," Psychological Bulletin, *73*, No. 1 (1970), pp. 74–88.
17. A. M. Wing and A. B. Kristofferson, "The timing of interresponse intervals," Perception & Psychophysics, *13* (June 1973), pp. 455–60.
18. A. M. Wing and A. B. Kristofferson, "Response delays and the timing of discrete motor responses," Perception & Psychophysics, *14* (August 1973), pp. 5–12.
19. E. Parzen, *Stochastic Processes*, San Francisco: Holden-Day, 1962, p. 56.
20. V. Braitenberg, "Functional interpretation of cerebellar histology," Nature, *190* (1961), pp. 539–40.
21. H. H. Kornhuber, "Cerebral cortex, cerebellum, and basal ganglia: An introduction to their motor functions," in E. V. Evarts (Ed.), *Central Processing of Sensory Input Leading to Motor Output*, Cambridge, MA: The MIT Press, 1975.

AUTHOR

David A. Rosenbaum, B.A., 1973, Swarthmore College; Ph.D., 1977, Stanford; Bell Laboratories, 1977–1981; Hampshire College, 1981—. Mr. Rosenbaum worked in the Human Information-Processing Research Department at Bell Laboratories, Murray Hill, from 1977 to 1981. Currently, he is Assistant Professor of Cognitive Science at Hampshire College, Amherst, Massachusetts, where he teaches and continues research on the cognitive control of human movement.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Experiments on Quantitative Judgments of Graphs and Maps

By W. S. CLEVELAND,* C. S. HARRIS,* and R. McGILL*

(Manuscript received August 18, 1982)

Behavioral studies are essential for devising guidelines for effective communication of quantitative information from graphs. Three experiments in which subjects made quantitative judgments from three different kinds of graphs lead to several recommendations: use pastel rather than highly saturated colors on statistical maps; standardize the point cloud size relative to the frame on a scatterplot; scale circles by making the circle area proportional to the variable represented, but expect widely varying judgments of the areas.

I. INTRODUCTION

With the proliferation of computer graphics, there is an increasing reliance on visual displays to convey quantitative information. Maps, graphs, and diagrams have been in use for a long time, but in recent years the variety, complexity, and ease of preparing such visual displays have increased greatly. It is often assumed that visual displays allow people to quickly and accurately appreciate quantitative information and relationships that might be much harder to grasp from other representations, such as tables of numbers, equations, or verbal

* Bell Laboratories.

descriptions. Graphs, for example, are regarded as powerful tools for analyzing data[1] and for presenting data to others.[2]

In preparing a graph, the usual assumption is that if the mathematical relationships are represented accurately in visual form, they will convey the correct quantitative impressions. But this may not be right. Numerous experiments with very simple displays have shown that people's perceptual judgments often differ from physical measurements of such attributes as length, area, orientation, separation, brightness, and color (for reference lists, see Refs. 3 through 5). A much smaller number of studies have used displays that are similar to graphs or maps (see Refs. 6 through 9). Considering the enormous usage and variety of graphs, the number of directly applicable experiments is quite small. More information about human factors in graphical judgment is essential for the design of better visual displays.[10,11]

We summarize here three sets of experiments on quantitative judgments of three kinds of graphs. (More detailed accounts are given in Refs. 6, 7, and 12). The experiments differ in the information being conveyed, the coding of the information, the experimental procedures, the subject populations, and the methods of statistical analysis of the results. In passing, we will mention some useful statistical techniques that may be new to many readers.

II. A COLOR-CAUSED ILLUSION OF AREA

A clear example of erroneous perception of a relatively simple display is shown in a study of the perception of areas of colored regions within a map.[12] Figure 1 is a map of the counties in Nevada. Maps like this, with subsets of the counties colored red or green, were shown to 24 subjects (12 scientists and 12 secretaries, clerks, and craftspeople, all from Bell Laboratories). In each map the total red area differed from the total green area by no more than 0.01 percent. Each subject saw ten variations of the map, with different subsets of counties colored red or green. The maps for one group of 12 subjects had the same partitioning of counties as for the other group, but the counties that were red for one group were green for the other, and vice versa. The maps were paper prints produced by a Bell Laboratories Printing System-Multicolor (PRISM) printer (a computer-driven modified Xerox 6500 copier), using standard options, which include highly saturated colors. Matching the colors with Munsell color chips,[13] under illumination similar to that in which subjects viewed the maps, gave Munsell values of hue = 7.5 red, value (brightness) = 4, chroma (saturation) = 14 and hue = 2.5 green, value = 5, chroma = 12. Thus, the red and green were quite similar in brightness and saturation.

An instruction sheet told the subjects that the colors indicated various geological features in each county. The subjects' task was to

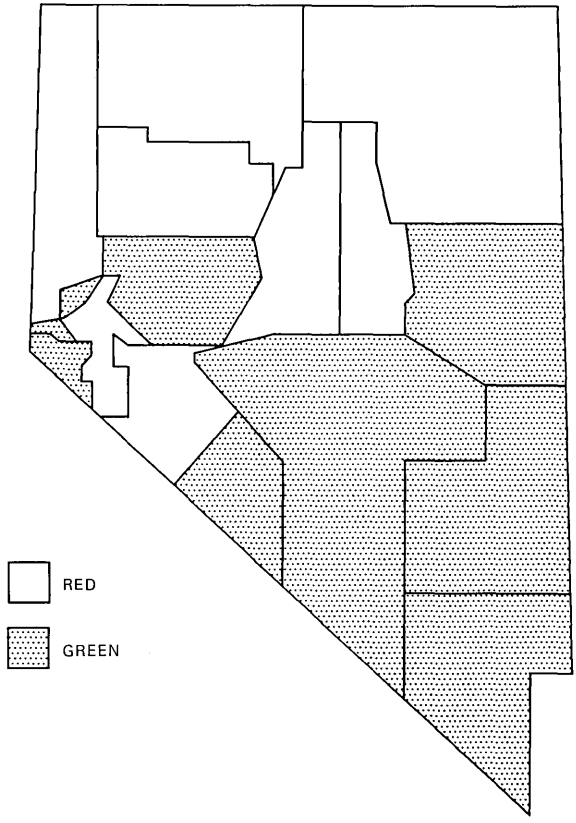


Fig. 1—Map of Nevada, where light and dark areas represent red and green, respectively.

decide which colored area within each map was larger and to mark "more red," "more green," or "same" on a checklist. The outcome was that most of the 24 subjects marked "more red" more frequently than "more green." That is, although the red and green areas were actually equal, they didn't look equal.

Figure 2 shows the data on a trilinear plot,[2] which accommodates within a single graph three variables that sum to a constant (100 percent in this case). Each open circle represents a single subject's data. The vertical axis indicates the percentage of maps for which a subject said the areas of the two colors looked the same. The two diagonal axes measure the percentage of maps in which the red or green area was called larger. Just as in the more familiar Cartesian xy plot, the three percentages for each data point are given by its perpendicular projection onto each of the three axes.

Because each subject made ten judgments, all data points represent multiples of 10 percent. Therefore, to avoid overlapping the values

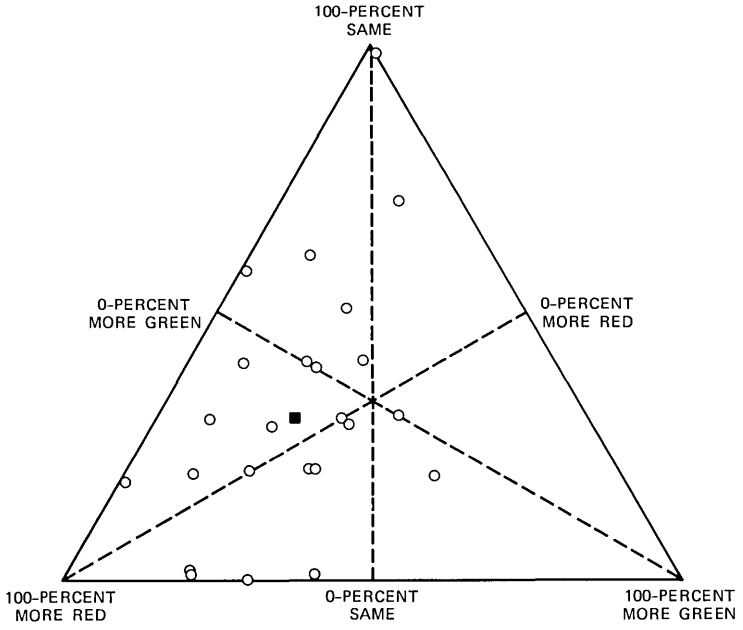


Fig. 2—Area judgments of highly saturated red and green regions.

were jittered: a randomly generated number from the interval −1 percent to +1 percent was added to each value.

First note that most of the 24 subjects' points fall to the left of the vertical axis. This indicates that most subjects judged the red area to be larger more often than they judged the green area to be larger. The filled square is the "robust center of gravity" for the data points. This statistic is calculated by an iterative procedure called "bisquare", using the distances of the data points from the current estimate as input for the next estimate. Using the bisquare[14] produces a statistic that is robust;[15] that is, unlike an arithmetic mean, it is insensitive to outliers, and is highly efficient for a broad class of distributions. The coordinates of this filled square give us another way to summarize the red/green illusion: it falls at about 50 percent "more red" judgments, 20 percent "more green," and 30 percent "same."

To estimate standard errors for the robust center of gravity, the bootstrap method[16] was used. Bisquare estimates were computed on 1000 24-point samples, drawn with replacement. The difference between the green and red coordinates is 49 percent − 22 percent = 26 percent, with a bootstrap standard error of 5.3 percent. Since the bootstrap distribution of the difference was well fit by a normal distribution, the percentage of red-larger judgments is very significantly higher than the percentage of green-larger.

Thus, the likely result of accepting the saturated colors that the PRISM printers normally produce is to make the red areas look too big or the green areas too small. Is there a way around this perceptual distortion? The printer was modified to fill the areas with two unsaturated pastel colors: a half-tone screen composed of red and pale yellow dots and a half-tone screen composed of green and pale yellow dots. The filled square for the robust center of gravity in Fig. 3 shows that for the group as a whole, the percentages of judgments of "more red" and "more green" are nearly equal. Using pastel instead of saturated colors eliminated the overall tendency to call the red area larger than the green.

Additional research could determine whether these findings hold for other pairs of hues, other combinations of brightness and saturation, and other display media (e.g., video displays). For the PRISM printer, it is clear that for accurate judgments of relative area, pastel colors are preferable to the standard saturated red and green.

III. JUDGMENTS OF CIRCLE SIZES

In most of the perceptions that we label "illusions," some extraneous aspect of the stimulus seems to be distorting perception of another aspect. For example, in the familiar Müller-Lyer illusion the arrowheads added to the ends of two horizontal lines distort perception of

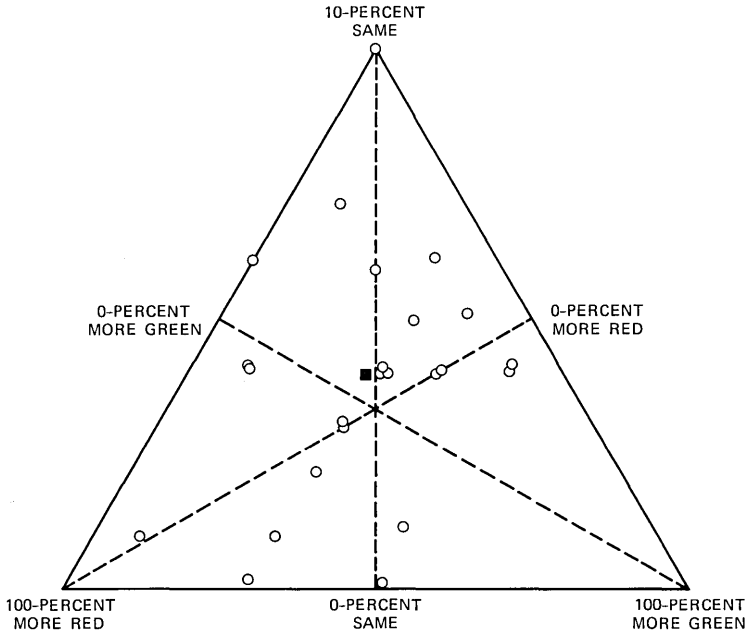


Fig. 3—Area judgments of unsaturated red and green regions.

the lengths of those horizontal lines. Similarly, in the experiment with maps of Nevada, the colors distorted perception of areas. But, in fact, many of our perceptions—perhaps even most of them—are illusions, in the sense that what we see does not match the usual measurements obtained with instruments other than the human observer. For a wide variety of attributes (such as loudness, brightness, tactual roughness, and weight), laboratory studies have found that people's judgments of quantities or intensities seldom vary in direct proportion to these measurements of the physical stimulus. In most cases the relation between subjective judgments and physical measurements can be described by a power function of the form $s = kp^e + c$, where s is the subjective magnitude, p is the physical magnitude, k and c are scaling constants, and e is an exponent that depends on what is being judged, ranging from 0.3 for the brightness of an isolated disk of light to 3.5 for the intensity of an electric shock.[5]

What does this imply for the communication of quantitative information by means of graphs or maps? Consider a common kind of statistical map in which circles of different sizes are used to represent, for example, the number of toll calls from various cities. For laboratory studies of the judgment of area, Stevens (see Ref. 5) gives 0.7 as a typical value for the exponent in the psychophysical function. Such a low exponent would mean that a circle that is double the area of another would be called only 1.6 times as large, and one with five times the area would be called only three times as large, whereas one with 25 percent of the area would be called almost 40 percent as large.

Since the purpose of a map or graph is to convey quantitative relations quickly and accurately, shouldn't areas be scaled to give the correct subjective impressions rather than to be physically correct? Some writers[17,18] have suggested just such a procedure. They recommended scaling the plotted areas to compensate for the low-exponent function found in psychophysical experiments.

However, the exponent found in studies of area judgments might not apply to an actual statistical map. We therefore prepared 24 maplike pages that depicted the average daily telephone toll charges for businesses at different locations in a city.[7] An example is shown in Fig. 4. Fourteen scientists from Bell Laboratories were told that the circle marked with an X represented toll charges of $100. They were to write down, for each of the three circles marked with a dot, their estimate of the dollar amount represented. The word "area" was not used.

The results were quite different from the usual laboratory findings on judgments of area: Most of the exponents in the fitted power functions for individual subjects were closer to 1.0 than to 0.7. There was considerable variation from person to person, with exponents

Figure 467

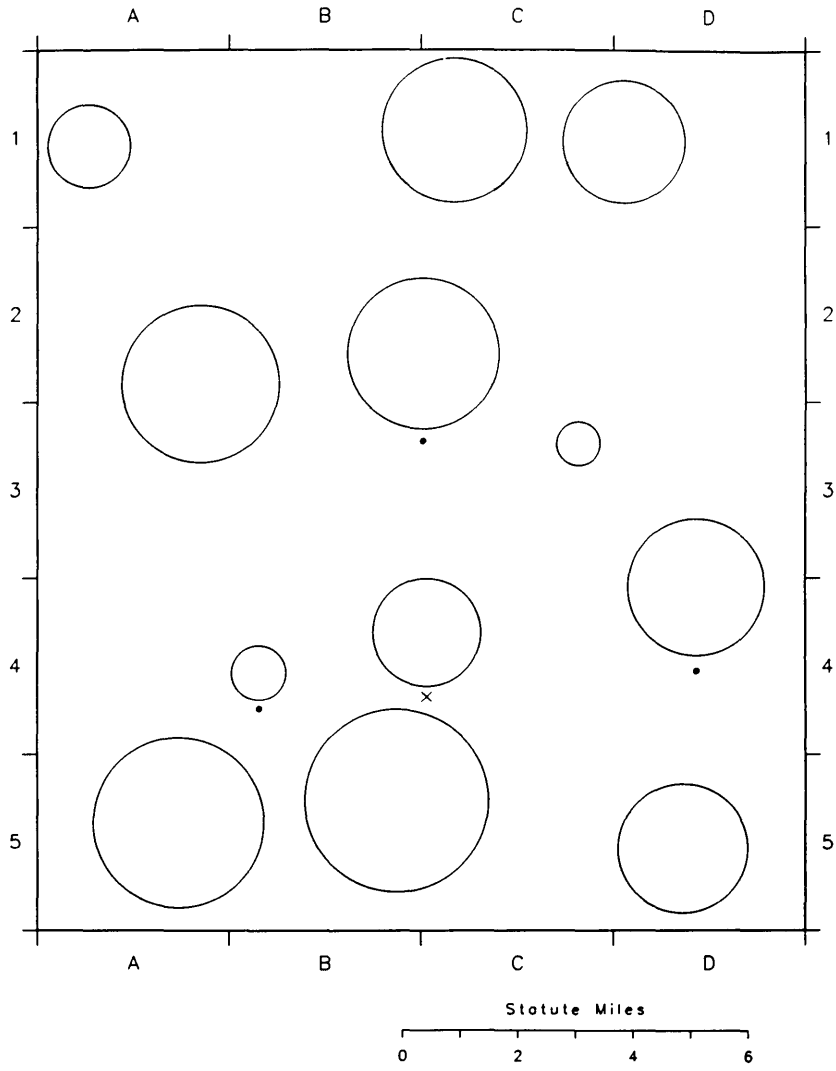


Fig. 4—Maplike display of circles.

ranging from 0.6 to 1.3. This implies that for an actual chart-reading situation, no single form of compensatory scaling of area would give everyone correct impressions. In fact, since the median exponent (0.94) is close to 1.0, these data suggest that the best scale is the simple one with circle areas directly proportional to the represented quantities. (Note that judging diameter or circumference instead of area would

yield an exponent of 0.5 for judgments as a function of area. None of our subjects had an exponent that low.)

Why do our exponents differ from those typically found in psychophysical studies? Is it because we asked for a different kind of judgment—dollars represented, rather than area—and displayed a maplike frame and tick marks? We can answer this question because we had the same subjects make area judgments of the same 24 sets of circles on plain pages, without maplike markings (Fig. 5). All of them did this perceptual task after judging the maplike stimuli. The instruction was to judge the areas of the circles, calling the one marked X 100 units of area. We found no difference between the two types of displays and tasks.

Figure 179

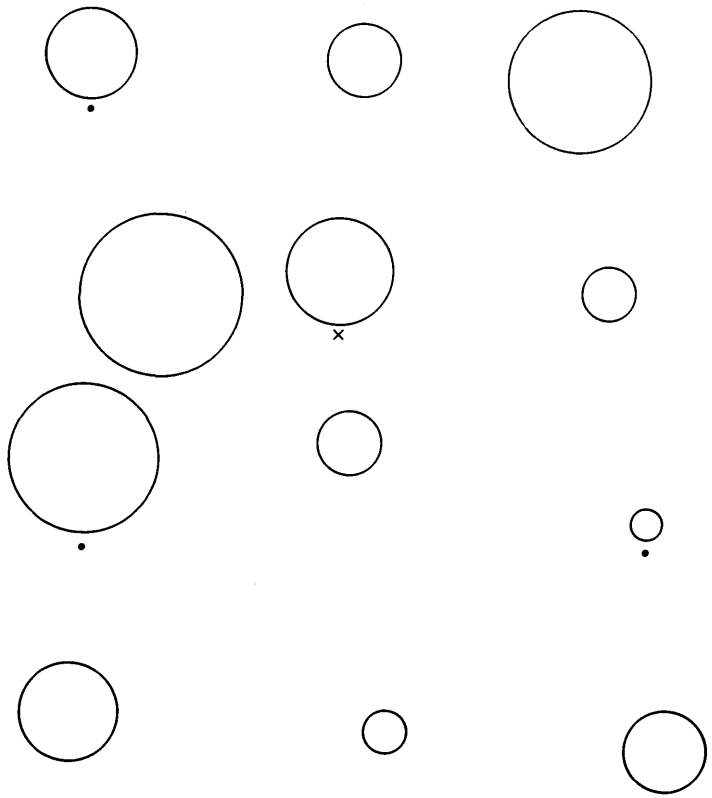


Fig. 5—Display without maplike markings.

Our next thought was that our subjects' scientific training might enable them to judge areas more accurately, on the average, than subjects in previous laboratory experiments. We therefore asked 10 high school students to carry out the same two tasks. Figure 6 presents a comparison of the exponents estimated for the scientists (on the left) and 10 high school students (on the right). Functions of the form $s = kp^e + c$, relating subjective judgment (s) to physical area (p), were fitted to each subject's data. The digits to the left of the colons are the first digits of the estimated exponents (e). Each digit to the right of the colon is the next digit in the exponent for a single subject. The distributions are similar for the scientists and the high school students, with medians at 0.94 and 0.97, respectively. These stem-and-leaf diagrams[19] can be interpreted as enriched histograms. The digits to the left of the colons are the first one or two digits of the estimated exponent, while each single digit to the right of the colon is the next digit in the exponent for a single subject. For example, the stem-and-leaf diagram tells us not only that four scientists had exponents in the range from 1.00 to 1.09, but also that the specific values were 1.00, 1.04, 1.05, and 1.07. There is not much difference between the distributions for the scientists (median = 0.94) and for the students (median = 0.97). Scientific training doesn't have much influence.

The students' data also rule out another possible explanation of the difference from previous psychophysical studies. The scientists judged all of the maplike displays before seeing the ones that contained only circles. When estimating the toll charges, they were free to adopt any strategy they cared to, unlike subjects in psychophysical studies who are told to judge area. Even though the scientists were then asked to judge areas in the second half of the experiment, they might have simply persisted in using the same judgmental strategy as in the first half. This choice of strategy could explain why judgments were the same for maplike displays and plain circles, and also why the results differ from earlier experiments on area estimation. However, half of the students made their area judgments *before* making dollar estimates, and their estimates were quite similar to those made by the students who completed the tasks in the opposite order. The agreement between

EXPONENTS FOR 14 SCIENTISTS	EXPONENTS FOR 10 HIGH SCHOOL STUDENTS
0.5 :	0.5 : 8
0.6 : 3	0.6 : 9
0.7 : 29	0.7 : 4
0.8 : 8	0.8 : 8
0.9 : 01349	0.9 : 7799
1.0 : 0457	1.0 : 3
1.1 :	1.1 : 8
1.2 : 7	1.2 :

Fig. 6—Power-function exponents for judgment of circle sizes.

area and dollar estimates shows that neither the type of judgments, nor the order in which they are made, nor the presence of maplike frames is responsible for the high exponents.

One obvious difference between our experiments and many previous ones is that our subjects always judged circles within a simultaneously visible set, whereas in most psychophysical studies the stimuli are presented one at a time. It may be that higher exponents are obtained when the standard circle is visible along with the circles that are compared with it, instead of being displayed and then removed before the judgments are made. Examination of previous studies offers some support for this hypothesis; for several comparable studies (mostly by psychologists) in which the standard was not visible during the area judgments, the median exponent was 0.7, whereas in several studies (largely by cartographers) in which the standard was always present, the median exponent was 0.9.[7]

These findings suggest that when one is preparing statistical maps, it is probably better simply to make circle areas directly proportional to the quantities represented (exponent = 1) rather than to scale with a very different exponent, as a direct application of psychophysical studies of area judgments might suggest. Future research may enable us to specify what variables contribute to higher or lower exponents, but certainly our multiple-circle displays resemble statistical maps more closely than do circles viewed one at a time.

IV. JUDGED ASSOCIATION IN SCATTERPLOTS

The judgments discussed so far—areas of circles or colored regions on maps—do not require technical training; indeed high school students' judgments proved to be quite similar to those made by scientists. The experiments to be described now, on the other hand, called for subjects with some statistical training. In these experiments[6] all subjects had at least a basic knowledge of statistics (university statistics students and faculty, and practicing statisticians). They were asked to assess the degree of linear association between two variables portrayed by a scatterplot. The most frequently encountered measure of linear association is the correlation coefficient, r. The value of r ranges from 0, when there is no linear association, to +1 or −1, when the linear association is perfect and the plotted points fall on a straight line. In many basic statistics courses, r is the only measure of correlation that is taught.

In the study with maps of Nevada, discussed earlier, perceptions of one aspect of the displays—area—was found to be strongly influenced by another aspect—color—which should have been irrelevant. A similar influence was found with judgments of association in scatterplots. The two scatterplots in Fig. 7, projected alternately on a screen, were

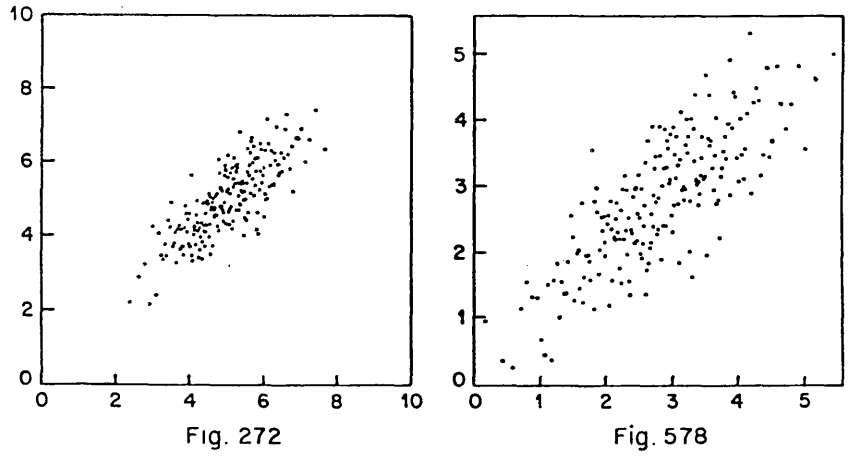


Fig. 7—(Left panel) Scatterplot with correlation $r = 0.8$. (Right panel) Scatterplot with same correlation as in left panel, but with x and y scales expanded.

shown to 109 subjects. They indicated the degree of association of the two variables by assigning a number from 0 to 100 to each plot. In fact, the correlation is identical for the two plots ($r = 0.8$); only the scale of the axes, and hence the size of the point cloud, is different. Judgments of the amount of association on the two plots should therefore be identical also. But they are not; judgments of the association portrayed by the left panel of Fig. 7 were generally higher than for the right panel.

To quantify this difference, each subject's estimate for the right panel of Fig. 7 was subtracted from the estimate for the left panel, and the 10-percent trimmed mean was calculated by dropping the largest 10 percent of the differences and the smallest 10 percent and taking the arithmetic mean of the remaining values. (Unlike means, 10-percent trimmed means are robust measures that are not influenced by extreme outliers. Ten-percent trimmed means can be thought of as a compromise between medians, which are nearly 50-percent trimmed means, and means, which are 0-percent trimmed.) The result (after dividing by 100 to bring the judgments into the range 0 to 1) was a difference of 0.068 between the panels of Fig. 7, with a standard error of 0.011. Thus, the estimated association was significantly higher for the smaller point cloud than for the larger one.

A second experiment corroborated this finding. The scatterplots in Fig. 7 were shown to 32 subjects who were told that the correlation coefficients were identical. The subjects were asked whether one plot *looked* more correlated than the other and if so, which one. Sixty-six percent of the subjects reported that the left panel looked more correlated than the right, 22 percent that they looked the same, and

only 13 percent that the right panel appeared more correlated. So even explicit information that the two correlations are identical does not dispel the illusory impression that the smaller point cloud depicts a higher correlation.

To obtain more detailed information on judgments of association another study was carried out. In this study 74 subjects each judged 19 scatterplots, assigning a number from 0 to 100 to denote the linear association. The 19 stimuli included 10 with the same axis scales, representing 10 different levels of association, plus an additional 9 with 3 different axis scales (that is, different point-cloud sizes) at 3 of those levels of association. A number of other attributes (such as number of points, standard deviations of the variables, sign of the correlation, and size of the square frame) were identical for all of the scatterplots.

The 10-percent trimmed means across subjects are plotted against actual correlation in Fig. 8. The circle radii portray the standard errors of the trimmed means. Thus the circle areas are proportional to the estimated variances of the trimmed means. The numbers to the left of the circles indicate the point-cloud sizes (1 is the smallest). When

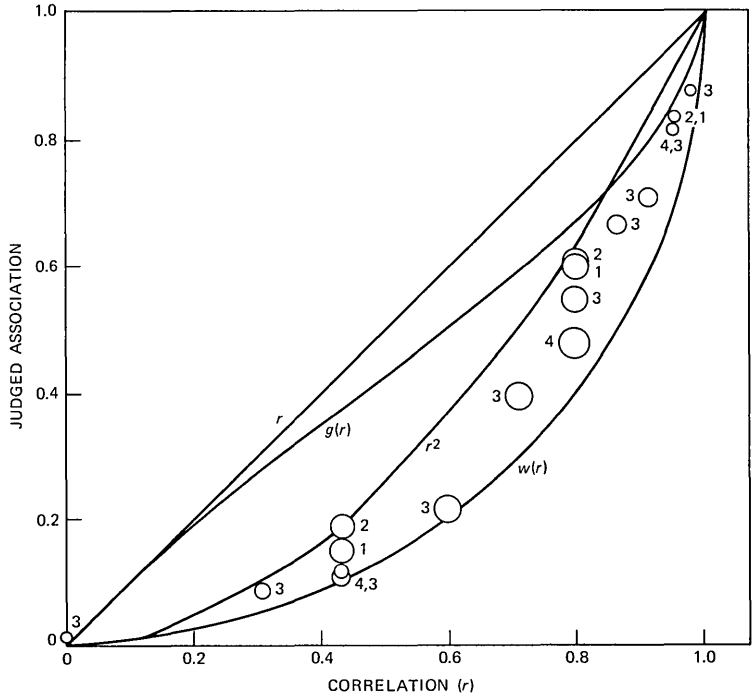


Fig. 8—The 10-percent trimmed means across subjects of judged association divided by 100 for 19 scatterplots are plotted (by the circle centers) against the values of r, the correlation coefficient, of the scatterplots.

two numbers are shown, separated by a comma, two circles are nearly coincident and the first number refers to the circle with the smaller trimmed mean. The information on the plot leads to two conclusions: judged association tends to be higher for smaller point-cloud sizes; judged association is not proportional to any of four standard numerical measures of association [the points do not fall on any of the curves, which plot r, $g(r)$, r^2, and $w(r)$ as a function of r]. Thus, the circles for r near 0.4 and 0.8 corroborate the finding that smaller point clouds tend to elicit higher judgments of association.

The new information is that subjects' judgments are far from proportional to the usual measure of linear association, r. Since the actual correlation coefficient r is given by the horizontal axis, if the subjects had been judging r accurately, the data would have fallen on the 45-degree line. Instead, the points fall well below it. On the average a correlation of 0.4 was judged to be less than 0.2, a difference of more than a factor of 2.

Two other measures of linear association come closer to fitting the data: $g(r) = 1-\sqrt{(1-r)/(1+r)}$ and $w(r) = 1-\sqrt{1-r^2}$ (see Ref. 20). Unlike r, both $g(r)$ and $w(r)$ offer intuitively plausible geometric bases for visual judgments. If we draw the ellipse of the bivariate normal distribution that generated a scatterplot, as in Fig. 9, the ratio of the minor axis to the major axis is $(1-r)/(1+r)$. The smaller this ratio is, the higher the association; so $g(r)$ tells how narrow the ellipse is relative to a zero-correlation circle. The ratio of the area of the ellipse to the area of the rectangle circumscribed around it is $\pi/4\sqrt{1-r^2}$; so $w(r)$ tells how far the ellipse is from filling the rectangle. Unfortunately, neither $g(r)$ nor $w(r)$ fits the data very well, as can be seen from the departure of the data points from both curves in Fig. 8. Another measure, r^2, does not fit the data particularly well either. However, there is a two-parameter family of curves, $1-(1-r)^{\alpha}(1+r)^{\beta}$, that includes all four measures of association that we have mentioned (and many others as well). If such a curve is fitted to the data in Fig. 8, the estimates of α and β and their standard errors are 0.71 ± 0.04 and 0.66 ± 0.11, respectively. These estimated parameter values fall between (and are significantly different from) those for $w(r)$ ($\alpha = 0.5$, $\beta = 0.5$) and for r^2 ($\alpha = 1$, $\beta = 1$). [For r, $\alpha = 1$ and $\beta = 0$; for $g(r)$, $\alpha = 0.5$ and $\beta = -0.5$.] This middle position of the parameter estimates is what we would expect since the circles in Fig. 8 lie between r^2 and $w(r)$.

Thus the parameters of the best-fitting curve have values that are significantly different from those of any of the four measures of association that we have considered.

On the face of it, the poor fit of the subjects' judgments to r, r^2, $g(r)$, and $w(r)$ seems to imply that these highly trained subjects do not base

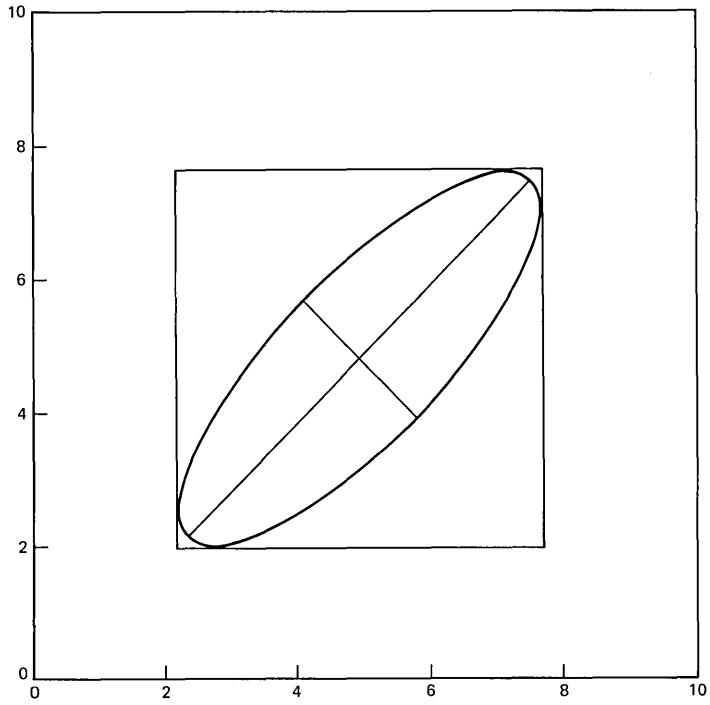


Fig. 9—Geometrical bases for $g(r)$ and $w(r)$.

their estimates of linear association on the visual counterparts of any of these standard measures. However, there is an alternative that remains to be explored. Perhaps subjects do make judgments (for instance, of the ratio of minor to major axis) that are appropriate for one of these measures of association, but misperceive the basic visual attributes on which the judgments are based. For example, if judgments of length were not directly proportional to actual length, then judgments of the ratio of minor to major axis of an ellipse would not be proportional to the actual ratio.

The finding that the size of the point-cloud can have a big effect on judged association means that the axis scaling that many statistical plotting programs apply automatically is not optimal for all purposes. To facilitate comparisons of the degree of association in different plots, it would be wise to make the point-cloud sizes similar; Cleveland and McGill (see Ref. 21) propose a way to do so. The size effect also suggests that when estimates of degree of association are required, the numerical value of a measure of association should also be computed. Even experienced statisticians can have judgments of association affected by extraneous factors.

V. CONCLUSIONS

Behavioral studies of the kind summarized here are essential for devising guidelines for effective communication of quantitative information. These studies confirm that people can make consistent judgments of a wide variety of visual displays. However, those judgments may not match the usual physical measurements of stimulus attributes: People overestimated bright red areas on maps, relative to bright green areas, and judgments of the degrees of linear association in a scatterplot did not agree closely with any of four standard statistical measures of association [r, $w(r)$, $g(r)$, and r^2]. Other findings lead to some recommendations about how to convey quantitative relations more accurately. For example, for output from PRISM plotters, substituting pastel colors for saturated red and green reduced the biasing of area judgments. The finding that smaller point-clouds in a scatterplot are judged to portray higher linear association implies that to permit accurate comparisons of association, axes should be scaled to produce similar point-cloud sizes. And finally, our subjects' judgments of arrays of circles suggest that in statistical maps one should not scale to compensate for the distortions in area judgments that are found when stimuli are viewed one at a time; instead, symbol sizes should be directly proportional to the quantities represented.

With reliance on both old and new forms of visual display becoming increasingly widespread, we will need more behavioral studies like the ones summarized here to guide effective communication of quantitative information.

VI. ACKNOWLEDGMENTS

We acnowledge helpful comments on earlier drafts by Francine Frome, Ram Gnanadesikan, Tom Landauer, Saul Sternberg, Paul Tukey, and several anonymous reviewers.

REFERENCES

1. J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey, *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth, 1983 (hardcover); and Boston: Duxbury Press, 1983 (softcover).
2. C. F. Schmid and S. E. Schmid, *Handbook of Graphical Presentation*, New York: Wiley & Sons, 1979.
3. S. Coren and J. S. Girgus, *Seeing is Deceiving: The Psychology of Visual Illusions*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1978.
4. C. S. Harris (Ed.), *Visual Coding and Adaptability*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.
5. S. S. Stevens, *Psychophysics—Introduction to its Perceptual, Neural and Social Prospects*, New York: Wiley, 1975.
6. W. S. Cleveland, P. Diaconis, and R. McGill, "Variables on Scatterplots Look More Highly Correlated When the Scales are Increased," Science, *216* (June 1982), pp. 1138–41.
7. W. S. Cleveland, C. S. Harris, and R. McGill, "Judgments of Circle Sizes on Statistical Maps," J. Amer. Stat. Assn., *77* (September 1982), pp. 541–7.

8. H. Wainer and D. Thissen, "Graphical Data Analysis," Ann. Rev. Psychol., *32* (1981), pp. 191–241.
9. L. Wilkinson, "An Experimental Evaluation of Multivariate Graphical Point Representations," Proc. Conf. on Human Factors in Computer Systems, Gaithersburg, Maryland, March 15–17, 1982, pp. 202–9.
10. W. Kruskal, "Visions of Maps and Graphs," Proc. Int. Symp. on Computer-Assisted Cartography, September 21–25, 1975, Amer. Congress on Survey and Mapping and the U.S. Bureau of the Census (1977), pp. 27–36.
11. W. Kruskal, "Criteria for Judging Statistical Graphics," Utilitas Mathematica, *21B* (May 1982), pp. 283–310.
12. W. S. Cleveland and R. McGill, "A Color-Caused Optical Illusion on a Statistical Graph," *The Amer. Statistician*, (1983a), in press.
13. A. H. Munsell, *The Munsell Book of Color—Glossy Finish Collection*, Baltimore: Munsell Color Co., 1966.
14. F. Mosteller and J. W. Tukey, *Data Analysis and Regression*, Reading, Massachusetts: Addison-Wesley, 1977.
15. P. J. Huber, "Robust Estimation of a Location Parameter," Ann. Math. Stat., *35* (February 1964), pp. 73–101.
16. B. Efron, "Bootstrap Methods: Another Look at the Jackknife," Ann. of Stat., *7* (January 1979), pp. 1–26.
17. J. J. Flannery, "The Relative Effectiveness of Some Common Graduated Point Symbols in the Presentation of Quantitative Data," Canadian Cartographer, *8* (December 1971), pp. 96–109.
18. G. F. McCleary, Jr., "In Pursuit of the Map User," Proc. Int. Symp. on Computer-Assisted Cartography, Auto-Carto II, (1975), pp. 238–50.
19. J. W. Tukey, *Exploratory Data Analysis*, Reading, Massachusetts: Addison-Wesley, 1977.
20. M. B. Wilk, unpublished work.
21. W. S. Cleveland and R. McGill, unpublished work.

AUTHORS

William S. Cleveland, A.B., 1965 (Mathematics), Princeton; M.S. and Ph.D., 1969 (Statistics), Yale; Bell Laboratories, 1972—. Mr. Cleveland works on the development of statistical methodology. His current research areas include time series analysis, seasonal adjustment, economic forecasting, atmospheric chemistry, graphical methods for data analysis, software systems for presentation graphs, and graphical perception. Mr. Cleveland is co-author of a book *Graphical Methods for Data Analysis* that was published in 1983. Member, AAAS, National Computer Graphics Association, American Statistical Association (Fellow, Chairman of Statistical Graphics Committee); elected member, International Statistical Institute.

Robert McGill, Bell Laboratories, 1969—. Mr. McGill is a Member of Technical Staff in the Statistics and Data Analysis Research Department. Currently, he is working on graphical techniques and perception of statistical graphics. Member, American Statistical Association, Association for Computing Machinery, National Computer Graphics Association, International Association for Statistical Computing, American Association for Advancement of Science.

Charles S. Harris, B.A., 1959, Swarthmore College; M.S., 1961, Yale University; Ph.D. (Experimental Psychology), 1963, Harvard University; Assistant Professor, University of Pennsylvania, 1964–1966; Bell Laboratories, 1966—. Mr. Harris has carried out research on various phenomena of visual and kinesthetic perception, including motion illusions, adaptation to visual distortions, selective attention, color and spatial-frequency aftereffects, and perceptual facilitation by object-like context. He has edited a book entitled *Visual Coding and Adaptability.* He is currently a member of the Human Information-Processing Research Department. Member, Sigma Xi, Phi Beta Kappa, Psychonomic Society, Eastern Psychological Association.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Retrospective Reports Reveal Differences in People's Reasoning

By D. E. EGAN*

(Manuscript received January 14, 1982)

Reasoning for a class of transitive inference problems was studied and the following questions were experimentally investigated: (1) Can people give reliable retrospective reports about their reasoning processes? (2) Do people who report different reasoning processes actually reason in different ways? (3) Can people be trained to use different reasoning processes? In the situations studied, subjects' retrospective reports about reasoning contained sufficient information to classify the subjects reliably. Subjects classified as using different reasoning strategies made different amounts and different kinds of reasoning errors. As a result of training, subjects could use reasoning processes that they would not have used spontaneously. These results have implications for developing theories of reasoning and for assessing and modifying reasoning-like processes in practical situations.

I. INTRODUCTION

Reasoning, the ability to draw conclusions or inferences from given information, is a prized intellectual talent. Reasoning is known to be associated with successful learning of mathematics.[1] A "reasoning

* Bell Laboratories.

factor" plays a prominent role in reading comprehension.[2] Practical problem-solving tasks such as balancing a checkbook or troubleshooting equipment undoubtedly involve reasoning. Because of its pervasive importance, tests of reasoning have been included in virtually every battery of aptitude tests, and experimental psychologists have studied reasoning extensively in various forms.

Reasoning may play an especially prominent role in the future as more occupations and everyday tasks involve computers. For example, reasoning test scores to some extent predict people's success at computer programming,[3] an occupation that is growing in importance. Reasoning probably is required to use the host of new computer-like devices ranging from automated banking machines and appliance timers to systems for editing and data retrieval. The small bit of "programming" required to activate many new telephone services appears to require mental processes akin to reasoning.

1.1 Three basic questions about reasoning

This paper presents results of basic psychological research concerning three questions about reasoning. The first of these questions is, "Can people give reliable retrospective reports about their reasoning processes?" This question is important, because it asks whether a potentially useful kind of data about the reasoning process meets the scientific standard of repeatability. Using subjective reports to analyze reasoning processes poses much more severe problems of reliability than, say, using meter readings to analyze physical processes.

To be useful data, retrospective reports about reasoning processes must, as a minimum, admit to consistent classification. Reports should contain enough information so that two raters would classify a given report in the same category. Different methods of reporting (e.g., verbal reports and nonverbal reports such as drawings or checklists) should agree with each other so that a reasoner would be classified in the same category no matter which particular method of reporting is used. The classification of reports given by the same person on different occasions should be consistent, as long as other indicators show that the person is using the same reasoning process on those occasions. Retrospective reports must meet each of these requirements to be considered reliable indicators of people's reasoning processes.

A second fundamental question considered here is, "Do people who report different reasoning processes actually reason in different ways?" If different people actually reason in different ways consistent with their reports, two conclusions follow. One conclusion would be that retrospective reports about reasoning are valid evidence about the reasoning process. Reports could be valid simply if people who give a particular kind of report tend to use a particular reasoning process,

even if the content of the report does not accurately describe the process. Formally, such reports would have validity because their content would be correlated with performance. A more powerful and interesting kind of validity would require that the content of people's reports accurately reflects how they reason.

Another conclusion would be that data (e.g., reasoning errors) from people giving different kinds of reports do not result from a common underlying reasoning process. To understand reasoning, people using a given reasoning process first may have to be identified and their reasoning data grouped together. Models for the reasoning process appropriate to each group then could be developed.

A very practical aspect of this question is that retrospective reports are often the most easily obtainable data and sometimes they are the only data available in practical situations involving mental processes like reasoning. For example, people's reports about how they do a task or how they use a complex device often are used to evaluate new services or products requiring human interaction. It would be advantageous to have at least one assessment of the validity of these kinds of reports, and to know whether different people are likely to use different reasoning processes spontaneously. The extent to which laboratory studies of reasoning generalize to specific practical situations involving reasoning is not known. However, one benchmark for the validity and variety of retrospective reports about reasoning can be obtained in laboratory studies where it is possible to validate reports.

The third question about reasoning addressed in these studies is, "Can people be trained to use different reasoning processes?" This question introduces a distinction between the reasoning processes people might use spontaneously, and those they could use if trained to do so. In particular, for a specific kind of problem people may be able to follow directions to use a certain reasoning process, even if they would not use that process if left to themselves.

This question has theoretical and practical interest. To the extent that training can induce people to use different reasoning processes, reasoning cannot be conceptualized simply as a stable ability that some people have more of or are better at doing. Instead, a theory of reasoning must explain what factors cause people to discover and use a particular reasoning process when they are competent to use others. Moreover, if people can be trained to reason in different ways, then they probably can be trained to think in various ways about practical tasks involving reasoning. Directions on "how to think about" a task may be an effective aid for learning new procedures (perhaps activating telephone services or interacting with computers) that require reasoning.

1.2 Reasoning problems used in these studies

In all of the studies presented here, people solved reasoning problems involving a simple transitive inference. The problems, known as "three-term series problems," have a common basic form. Examples of the problems are given in Table I.

Three-term series problems were selected for these studies over many other possible kinds of reasoning problems for the following reasons. First, these problems require no specialized content knowledge and therefore seem to be good tools for examining "pure reasoning" uncontaminated by whatever specific information people might know about a problem. Second, the simple verbal structure of problems like those in Table I makes it easy to manipulate certain problem characteristics while controlling others (see below). Finally, three-term series problems have been studied extensively. They have been used on standard tests as markers for deductive reasoning ability[4,5] and much is already known about factors that influence the difficulty of these problems.[6]

Each three-term series problem consists of two premises followed by a question. The premises state a relationship between two of the three terms in the problems. The question concerns the inferred relationship between the remaining pair of terms.

In describing three-term series problems, an important distinction must be made between *relations* and *inverses*. This distinction is based on well-substantiated findings demonstrating that one member of a pair of opposite relational words typically produces better performance than the other member. For example, using the word "above" results in more rapid and accurate performance than the word "below" in a large range of tasks. Similarly, the word "fatter" leads to better performance than "thinner," etc. The psycholinguistic theory of lexical marking[7] may account for some of these differences. For present purposes, an inverse is defined as that member of a pair of opposing relational words that is known to cause the greater difficulty. Conse-

Table I—Sample of the reasoning problems used

Positional Relations	Visual Comparative Relations
Triangle is above circle. Square is below circle. Is triangle above square?	Square is smoother than triangle. Circle is smoother than square. Is triangle smoother than circle?
Circle is left of triangle. Square is left of circle. Is triangle left of square?	Circle is darker than square. Circle is lighter than triangle. Is triangle darker than square?
Square is in back of triangle. Square is in front of circle. Is triangle in front of circle?	Triangle is fatter than circle. Triangle is thinner than square. Is circle thinner than square?

quently, "above" will be referred to as a relation, but "below" will be referred to as an inverse. Similarly, "fatter" is called a relation, but "thinner" is called an inverse.

Having thus defined relations and inverses, the terms of the problems will be distinguished as follows: the "A term" is the initial term in the linear order established by the premises (e.g., when using the relation "rougher" or its inverse "smoother," the A term would be the roughest); the "B term" is the pivot, or middle term; the "C term" is the end term of the linear order (e.g., the smoothest when using rougher/smoother).

Several other details of the problem set are noteworthy. The terms and relationships of each problem were selected to encourage the use of spatial mental representations. The terms were the geometric figures circle, square, and triangle. The relationships involved either positional comparisons (relation/inverse pairs above/below, right of/left of, or in front of/in back of), or nonpositional visual comparisons (rougher/smoother, darker/lighter, or fatter/thinner). For each relation/inverse pair, 16 different problem types were generated. These different problem types resulted from a 2^4 factorial combination of the following factors: (1) the use of a relation or inverse in the premise relating the A term and B term; (2) the use of a relation or inverse in the premise relating the B term and C term; (3) the order of the premises; and (4) the use of a relation or inverse in the problem question. The pattern of reasoning errors made on these different types of problems will be analyzed to test whether various reasoning processes are being used.

II. CAN PEOPLE GIVE RELIABLE RETROSPECTIVE REPORTS ABOUT THEIR REASONING PROCESSES?

2.1 Approach

Regarding the question of the reliability of reports about reasoning, we will consider data from two experiments.[8] In both studies, high school students solved a large number of three-term series problems, and then described their reasoning processes. In the first study, 12 subjects each solved 384 problems. In the second study, 100 subjects each solved 192 problems.

The problems were presented on an audio-tape-playback machine. Each problem began with a speaker saying "Next," then reading the problem, and then pausing five seconds to allow the subject to answer before beginning the next problem. Subjects answered problems by crossing out "Yes" or "No" or "?" next to the number of the problem on an answer sheet. They were not allowed to write anything else. Answers were scored as errors if either the wrong response or a "?" was used.

Several methods of retrospective reporting were used. In the first experiment, subjects were interviewed by the experimenter, and these interviews were tape-recorded and analyzed. In the second experiment, subjects gave written reports—essentially the equivalent of the oral reports in the first study. Following the written reports, subjects then were asked to draw pictures representing how they thought about the problems. As a final method of reporting, subjects were shown two written descriptions of reasoning processes that attempted to capture the two most common kinds of reports found in the first study. Subjects had to choose which of the two descriptions more closely matched their own way of reasoning.

2.2 What people reported

The most striking aspect of subjects' reports was that different subjects claimed to use quite different sets of processes or *strategies* to deal with simple three-term series problems. The difference was most apparent for reports about the visual comparative relations rougher/smoother, darker/lighter, and fatter/thinner. Some subjects claimed to establish an order for the three geometric figures no matter what relation was used. For these subjects, "rougher," for example, would be identified with one end of a vertical or horizontal scale, as would "darker" and "fatter". Subjects described making the transitive inference by mentally arranging the geometric figures roughest to smoothest, darkest to lightest, etc. Other subjects claimed to attribute physical properties to the geometric figures in the case of the visual comparative relations. For example, these subjects described their representation of a rougher triangle as an image of a triangle having a roughly textured surface, or a lighter circle as a picture of a very bright round object. These subjects described making the transitive inference by scanning the images. Distinctions among reports about positional relations were more subtle (see below). Actual examples of written reports are given in Table II.

Reasoning data from subjects who reported similar reasoning strategies were grouped together. The rule for assigning subjects to groups was that if a subject described a representation that clearly involved physical properties for any relation, then the subject was placed in a group labeled "Concrete Properties Thinkers." A subject who claimed to use an ordered mental array for every relation was placed in another group labeled "Abstract Directional Thinkers."

Using this rule, all subjects in the first experiment could be classified into either the Concrete Properties Group ($N = 5$) or the Abstract Directional Group ($N = 7$). In the second experiment, the consensus rating of written reports by two judges identified 18 subjects as

Table II—Examples of written retrospective reports

Abstract Directional Thinkers		
Rougher–Smoother	S#005:	"Rather than imagining a rough/smooth figure, I put the figures in a horizontal line, in my mind, in the order of left/right rather than rough/smooth."
	S#049:	"I pictured the objects in my mind in a line of sequence."
Darker–Lighter	S#003:	"I set up a scale with the lightest on the far right and darkest on the far left and placed the figures on their appropriate spots."
	S#051:	"Placed them in a line up and down, darkest being on top."
Fatter–Thinner	S#086:	"I also used a mental horizontal grid for this relation with the left side of the grid being the 'thin end' and the right side the 'fat end.'"
	S#099:	"Put shapes in order from thinner to fatter."
Concrete Properties Thinkers		
Rougher–Smoother	S#008:	"I also drew a picture, and if something was rough—I would put craters in it in my mind—smooth was just plain white."
	S#098:	"The picture came to mind of corners and smooth edges, then the question was solved."
Darker–Lighter	S#062:	"In my mind, I 'colored in' the object that was darkest."
	S#080:	"I listened to the problem and tried to solve it mentally, at times picturing the objects colored in or not."
Fatter–Thinner	S#022:	"This (fatter/thinner problem) was hard. I had to think of the shapes as squeezed or pulled."
	S#100:	"Made them (the figures) fatter and thinner in my head."

Concrete Properties Thinkers, and 42 as Abstract Directional Thinkers.

The classification rule identifying a subject as a Concrete Properties Thinker on the basis of a single concrete report was motivated by simplicity. Subsequent analyses suggest that the rule, while admittedly crude, did manage to separate people into two groups that used a particular set of reasoning processes fairly consistently. First, subjects who reported a concrete representation for one relation were very likely to report using such a representation for other relations. For example, 17 of 18 subjects identified as Concrete Properties Thinkers in the second study gave reports having a concrete representation for two or more relations. Second, in the statistical analysis of reasoning errors, reasoning groups did not interact with the different relation/inverse pairs, suggesting that each group consistently used one reasoning process. Third, a key-word analysis of the written retrospective reports suggests that the two groups also handled positional relations differently. People classified as Concrete Properties Thinkers used more words in their reports that suggest the use of a visual image (variants of the words "picture" and "draw") for positional problems

(e.g, "I pictured the objects in a row"). People classified as Abstract Directional Thinkers used more words suggesting the use of an order-preserving scale (variants of the words "put", "order", "line," and "horizontal/vertical") for positional problems (e.g., "I put the objects in order on a horizontal line"). This interaction of key-word types and reasoning groups was statistically reliable.

To summarize, the majority of people reported one of two sets of reasoning processes or strategies for the three-term series problems. Some people claimed to preserve the information in the premises of at least some problems by means of an image capturing the visual features or stated position of the geometric terms. Others claimed to preserve the information in all premises by means of an abstract ordering of the geometric terms.

2.3 Reliability of retrospective reports

Several methods were employed to assess various aspects of the reliability of the retrospective reports. The first, alluded to earlier, was to assess the agreement between two different judges who categorized subjects on the basis of their written reports in the second experiment. Each judge classified the 100 subjects as Abstract Directional, Concrete Properties, or Other/Not Clear. Table III shows that the two judges agreed on the classification of 82 percent of the subjects. Almost all cases of disagreement occurred when one judge classified a subject as using one of the two identified strategies, but the other judge classified the subject as using an Other/Not Clear strategy. Compared to several additional studies[9] the consensus shown in Table III is the "worst case." Other estimates of interjudge agreement have ranged up to 95 percent.

A second analysis assessed the agreement among different methods of reporting in the second experiment. The pictures drawn and forced-choice strategy selections made by subjects were compared to the classification of their written reports. Pictures were classified as indicating the Concrete Properties strategy if they depicted geometric objects with altered physical properties (e.g., a pockmarked surface

Table III—Classification of written reports by two judges (Experiment II)

	2nd Judge's Categories		
1st Judge's Categories	Concrete Properties	Abstract Directional	Other/Not Clear
Concrete Properties	18	0	6
Abstract Directional	1	42	5
Other/Not Clear	3	3	22

depicting "rougher," shading depicting "darker," etc.). Pictures showing horizontal or vertical orderings of rather standard geometric figures were classified as indicating the Abstract Directional strategy. The classification of drawings agreed with the classification of written reports for 56 of the 60 subjects (93.3 percent) whose written reports had been classified by consensus as Abstract Directional or Concrete Properties. The analysis of forced-choice strategy selections showed that 51 of the 60 subjects (85 percent) chose the strategy description consistent with the classification of their written report.

To assess the long-term stability of reported reasoning strategies, 38 subjects who participated in the second experiment and who had been classified by consensus as Abstract Directional or Concrete Properties Thinkers were recalled six months later for another study. After solving some warm-up problems, subjects gave written reports describing their reasoning strategies. These reports were classified in the previously described manner and this classification was compared to the classification of the subjects performed six months earlier. The results in Table IV indicate that subjects' reports about reasoning have some, but not perfect, stability over time and across different presentation conditions (the former reports were given after listening to problems, the latter after reading problems). The stability of verbal reports estimated by the four-fold point correlation based on Table IV is $r = 0.59$ ($p < 0.01$). Specifically, 95-percent of the subjects earlier classified as Abstract Directional again reported that strategy, but only 59 percent of the original Concrete Properties Thinkers reported that strategy again six months later. The instability of the latter group may have been due to unreliable reports or classification procedures on one hand, or actual changes in reasoning strategies[9] on the other.

2.4 Summary

The reliability of retrospective reports about reasoning has been established in that: (1) different judges show considerable agreement on how to classify reports, (2) the classification of written reports agrees to a large extent with classifications based on other nonverbal methods of reporting, and (3) the classification of reports given by

Table IV—Number of subjects using strategies initially and six months later

	Strategy Used Six Months Later	
Strategy Used Initially (Experiment II)	Abstract Directional	Concrete Properties
Abstract Directional	20	1
Concrete Properties	7	10

people at different times has some stability. On the other hand, retrospective reports about reasoning or perhaps the present procedures for classifying reports are not perfectly reliable. Compared to paper-and-pencil tests having finely graded scores, the reliability of retrospective reports is somewhat low. In particular, the classification of reports given at different times and under different conditions of presenting problems is not always the same. Despite these difficulties, the great majority of retrospective reports about reasoning in these studies contain sufficient information to be classified consistently. This is a necessary condition for reports to be useful in exploring the process of reasoning.

III. DO PEOPLE WHO REPORT DIFFERENT REASONING PROCESSES ACTUALLY REASON IN DIFFERENT WAYS?

3.1 Approach

Reasoning errors from the two studies previously described will be used to analyze the validity of retrospective reports about reasoning and to gain further understanding of reasoning processes. Error data from the two groups of subjects reporting different reasoning strategies will be compared at successively finer levels of detail. The overall error rates from the two groups will be analyzed first. Then, general patterns of interaction in the error data will be discussed. Next, the effects of specific problem factors hypothesized to affect a particular reasoning process will be tested. Finally, two models of the process of making a transitive inference will be described and tested. The goal of this section is to demonstrate that different models of the reasoning process are required to account for reasoning errors made by the two groups of subjects who reported different reasoning strategies.

3.2 Differences in reasoning errors between groups reporting different strategies

The first attempt at assessing the validity of retrospective reports asked whether the overall reasoning error rate was different for people giving different reports. In the two studies described above, subjects giving Abstract Directional reports made significantly fewer errors than those giving Concrete Properties reports. In the first study, Abstract Directional Thinkers had an error rate of 10.3 percent compared to the Concrete Properties Thinkers' error rate of 38.0 percent. In the second study, the corresponding error rates were 21.0 percent and 27.9 percent. This difference in error rate favoring the Abstract Directional subject now has been found repeatedly.[10] One interpretation of this result is that the Abstract Directional strategy is more efficient for the transitive inference problems used in these studies.

A further study[9] required subjects to describe their reasoning processes at a number of points in a lengthy sequence of three-term series problems. In that study, a change in the strategy reported by a subject was accompanied by a corresponding change in the reasoning error rate. Thus, if a subject reported shifting from the Concrete Properties strategy to the Abstract Directional strategy, the subject's performance improved. For subjects reporting no shift in reasoning strategy, reasoning performance was fairly stable at a low or high level, depending on the strategy reported. This study provides evidence of the validity of different reports given by the same subject. It also suggests why reports by some subjects changed after six months in the previous study: the subjects' reasoning processes may have changed over time.

People reporting different reasoning strategies not only exhibited different overall levels of reasoning errors, but they also exhibited different patterns of reasoning errors. This fact is demonstrated in a general way by a statistically reliable interaction between Report Groups and Problem Types found in the two original experiments.[8] This interaction means that factors causing reasoning problems to be more or less difficult in one group of reasoners were not the same as the factors causing difficulty in the other group. People who gave different reports made different amounts and different kinds of reasoning errors.

Thus far, the validity of reports has been assessed in a formal but rather indirect way. At the next level of detail, we might ask whether the patterns of reasoning errors made by subjects are consistent with the reasoning process subjects claim to be using. Consider reports of Abstract Directional Thinkers who claim to construct an ordered mental array of the geometric terms used in these problems. Previous theories[11,12] have asserted that it should be easier to construct a direct spatial array from the ends toward the middle, rather than from the middle outward. This so-called "end-anchoring principle" leads to a prediction regarding the difficulty of solving various types of three-term series problems. For Abstract Directional Thinkers, reasoning errors in a problem should be directly related to the number of premises that have the middle or pivot term stated first. For people using the Concrete Properties strategy, the end-anchoring principle should be irrelevant.

This prediction was confirmed by patterns of reasoning errors. In both studies, reasoning error rate for Abstract Directional Thinkers was a monotonic function of the number of premises in a problem that began with the middle or pivot term. This factor accounted for highly significant amounts of the variance in the difficulty of different types of problems for Abstract Directional Thinkers (82.2 percent of the variance in Experiment I, 71.8 percent in Experiment II). Data for

Concrete Properties Thinkers were quite different. Reasoning error rate was not a monotonic function of the number of pivot-first premises in either experiment, and this factor accounted for much smaller amounts of the variance in problem difficulty for this group of subjects (17.6 percent in Experiment I, and 20.3 percent in Experiment II). This analysis suggests that Abstract Directional Thinkers are constructing a mental array as they report, and that Concrete Properties Thinkers are reasoning in a different way.

Other analyses of specific problem factors[8] have found that reasoning errors by Concrete Properties Thinkers depend on the number of inverses used in a problem (e.g., using words like "smoother" and "thinner"), as well as the number of times a relation and inverse are alternated in the statement of a problem. Inverses cause extra difficulty for Concrete Properties Thinkers because concrete representations of a property like smoothness may not be as easy to generate as concrete representations of a property like roughness. Abstract Directional Thinkers were less sensitive to such factors. The most difficult kind of problem for Concrete Properties Thinkers was one with premises like, "Circle is rougher than square. Circle is smoother than triangle." In such problems, the Concrete Properties Thinker presumably imagines first a rough circle next to a square, and then imagines a smooth circle next to a triangle. Such problems are difficult to answer when this inconclusive image is scanned.

3.3 Two models of reasoning

Models attempting to capture the reasoning process used by Abstract Directional and Concrete Properties Thinkers are presented in Tables V and VI, respectively. These models try to give a coherent account of the patterns of reasoning errors and the kinds of reports given by subjects. Each model contains parameters representing hy-

Table V—Model for abstract directional thinkers

Process	Problem Factors	Model Parameters
1. Encode Premise 1		
2. Establish abstract scale		
3. Arrange first two terms placing grammatical subject first on scale		
4. Encode Premise 2		
5. Find third term	Is third term grammatical subject or object?	SEARCH
6. Position third term	Does third term fall in "Natural" next position?	POSITION
7. Encode question		
8. Scan the scale		
9. Respond		

Table VI—Model for concrete properties thinkers

Process	Problem Factors	Model Parameters
1. Encode Premise 1		
2. Generate Image Pair 1 by assigning property to grammatical subject	Is difficult (inverse) relation used?	GENERATE
3. Encode Premise 2	Is relation the same as that in #1?	ENCODE
4. Generate Image Pair 2 by assigning property to grammatical subject	Is difficult (inverse) relation used?	GENERATE
5. Encode question	Is relation the same as that in #3?	ENCODE
6. Scan images	Are the images conclusive?	SCAN
7. Respond		

pothetical mental processes that are executed various numbers of times depending on the structure of a specific type of problem.

3.3.1 The abstract directional model

Abstract Directional Thinkers (see Table V) are assumed to encode the first premise and establish a mental scale for a problem. Then, the two terms stated in the first premise are arranged on the scale, the grammatical subject being placed first. The second premise is then encoded, and the subject searches for the third, or missing, term. This search is easier if the third term is the grammatical subject rather than the object of the second premise. The value of the SEARCH parameter (0 or 1, respectively) reflects this difficulty, and accounts for the effect of starting the second premise with the pivot term. Next, the third term is positioned on the mental scale, and it is assumed that there are three distinct cases for this operation. The easiest case (POSITION = 0) occurs when the third term is placed next in the sequence established by the first two terms. For example, if the first two terms are arranged smooth → rough, positioning the third term is easiest if it is roughest. If the first two terms are placed rough → smooth, then positioning the third term is easiest when it is the smoothest. Two more difficult cases exist and correspond to problems beginning with a pivot-first premise. In the easier of these cases (POSITION = 1) the third term does not fall next in sequence, but instead it must be placed at the end of the scale associated with the relation. In the remaining, most difficult case (POSITION = 2), the third term again does not fall in sequence, but must be positioned at the end of the scale associated with the inverse.

3.3.2 The concrete properties model

The model for Concrete Properties Thinkers suggests that these subjects generate and compare images of objects having the stated properties. For each premise, Concrete Properties Thinkers (Table

VI) are assumed to encode the premise and then generate an image pair in which the grammatical subject takes on the property stated in the premise, while the grammatical object remains neutral. After two such pairs have been generated, the question is encoded, and then the two image pairs are scanned for the answer. Differences in difficulty among problems are assumed to arise from three sources, each corresponding to a parameter in the processing model for Concrete Properties subjects. One kind of difficulty has to do with whether the relation or inverse is used in each premise. Using an inverse presumably makes the appropriate image pair more difficult to generate. For a given problem, the parameter GENERATE takes on a value equal to the number of difficult images required (0, 1, or 2). The parameter ENCODE reflects the difficulty of alternately accessing a relation and its inverse. This parameter equals the number of alternations between a relation and inverse as a problem is read (0, 1, or 2). Finally, the parameter SCAN reflects the difficulty of dealing with images that are inconclusive. As noted previously, problems in which the B term takes on a property in one image pair and then takes on the inverse property in the other pair are especially difficult for Concrete Properties Thinkers. Such problems produce inconclusive image pairs in which the A and C terms are both neutral. Confronted with this type of problem, Concrete Properties Thinkers may guess or reformulate one of the premises to arrive at an answer. The SCAN parameter has the value 1 for such problems and 0 otherwise.

3.4 Comparing models to data

The two models were compared to the data of Abstract Directional and Concrete Properties Thinkers from each experiment. The proportion of variance in problem difficulty uniquely associated with each parameter in each model was determined by stepwise multiple regression. The data are shown in Table VII. For both experiments, the Abstract Directional model was the better predictor of performance for Abstract Directional Thinkers, while the Concrete Properties model was the better predictor for Concrete Properties Thinkers. If errors on the various problem types are combined across experiments, the Abstract Directional model accounts for 90.3 percent of the variance in problem difficulty for Abstract Directional Thinkers (the Concrete Properties model accounts for 80.4 percent of the variance for this group). Both the SEARCH and POSITION parameters account for significant and unique portions of variance in problem difficulty for Abstract Directional Thinkers. The Concrete Properties model accounts for 80.1 percent of the variance in problem difficulty for the combined Concrete Properties Thinkers (the Abstract Direc-

Table VII—Proportion of variance* in problem difficulty attributable to parameters of two models

Parameters	Experiment I		Experiment II		Experiment I + II	
	Abstract Group	Concrete Group	Abstract Group	Concrete Group	Abstract Group	Concrete Group
Abstract Directional Model						
1. SEARCH	0.576†	0.021	0.668†	0.105	0.668†	0.076
2. POSITION	0.288†	0.305‡	0.208†	0.201	0.235†	0.311‡
$\sum R^2$	0.864†	0.326	0.876†	0.306	0.903†	0.387‡
Concrete Properties Model						
1. SCAN	0.794†	0.372‡	0.724†	0.374‡	0.768†	0.474†
2. GENERATE	0.001	0.185‡	0.016	0.166‡	0.011	0.221†
3. ENCODE	0.024	0.079	0.024	0.087	0.025	0.106‡
$\sum R^2$	0.819†	0.636†	0.764†	0.627†	0.804†	0.801†

* These proportions are increments in R^2 values due to each parameter. The order in which the parameters are given corresponds to the step at which they entered the regression equation for the group of subjects appropriate to a particular model.

† $p < 0.01$

‡ $p < 0.05$

tional model accounts for 38.7 percent of the variance for this group), and each of the parameters SCAN, GENERATE, and ENCODE accounts for significant and unique variance.

Two aspects of the results of the model-fitting procedure should be clarified. First, the reliability of the error rates on the 16 problem types imposes a theoretical upper limit on the amount of variance for which any model can account. Therefore, it is important to estimate the data's reliability and compare that estimate to the R^2 of the best-fitting model.

Reliability estimates suggest that it would be difficult to improve the fits of the models appropriate to each group of subjects in Table VII. The estimated reliability of the Abstract Directional data combined across experiments was 0.951, so the Abstract Directional model ($R^2 = 0.903$) accounted for 0.903/0.951 or 95.0 percent of the reliable variation in the Problem Type data for those subjects. The fit of the Concrete Properties model ($R^2 = 0.801$) actually slightly exceeded the theoretical upper limit of the reliability of the combined Concrete Properties data (estimated reliability was 0.727).

Second, it is important to note that certain parameters of the two models are correlated in the 16 problem types used. The most important example of this confounding occurs for the parameter SCAN in the Concrete Properties model, which is correlated with both the SEARCH and POSITION parameters in the Abstract Directional model. These correlations account for the contribution of SCAN to variance in problem difficulty for Abstract Directional subjects. This interpretation of the contribution of SCAN is consistent with the fact that it is the only parameter in the Concrete Properties model that correlates with performance for Abstract Directional subjects, and that the two-parameter Abstract Directional model accounts for more variance in that group than the three-parameter Concrete Properties model.

3.5 Summary

Retrospective reporting under the conditions studied here apparently is one case in which reports about reasoning processes contain valid information. The fact that the two groups of people who reported different reasoning strategies actually reasoned in different ways is supported by (1) the different overall levels of reasoning errors made by the two groups, (2) the different general patterns of reasoning errors made by the groups, (3) the differential effects of specific problem factors hypothesized to influence difficulty in one group but not the other, and (4) the fits of different process models of reasoning to the reasoning error data of the two groups.

IV. CAN PEOPLE BE TRAINED TO USE DIFFERENT REASONING PROCESSES?

4.1 Approach

A further study[9] dealt with the question of whether people can be trained to use different reasoning processes. In that study, 65 adult women solved a small number of three-term series problems and reported their reasoning processes. As in the experiments described previously, these reports identified the reasoning strategies that subjects spontaneously used. The subjects were then randomly assigned to two groups. One group received training in applying the Abstract Directional strategy to a new set of three-term series problems. The other group was trained to apply the Concrete Properties strategy to the same problems. Reasoning errors made by the two groups of subjects after receiving training were compared.

4.2 Training in reasoning strategies

The training consisted of short descriptions of the models in Tables V and VI and examples showing how to apply them. The training was tailored specifically to a new set of problems involving the relation/inverse pair happy/sad. The terms for these problems were the names of three imaginary people, "Rich", "Dot", and "Harry". A typical problem was therefore, "Rich is happier than Dot. Harry is sadder than Dot. Is Harry happier than Rich?"

Subjects in the Concrete Properties training group were told to represent premises by vividly imagining faces having different features. Illustrations of the faces were drawn such that the people's names suggested the image of the correct face. Thus, "Rich" was depicted as a man wearing an expensive top hat, "Dot" was drawn with freckles, and "Harry" was pictured with a beard and mustache. Subjects were told to represent each premise by visualizing a pair of faces in which the face of the grammatical subject was smiling or frowning, depending on the wording of the problem. The two pairs of images then were to be scanned to answer the question for each problem.

Subjects given the Abstract Directional training were told to imagine a scale with "Sad" on the left and "Happy" on the right, and to place the names of the people on the scale appropriately as a problem was read. The order of the names on the scale then was to be used to make the transitive inference required to answer the question.

Following training, subjects solved 32 happy/sad problems. Subjects next rated the difficulty of applying the strategy they were trained to use. The rating scale ranged from 1 (extremely easy to use the strategy) to 6 (extremely difficult to use the strategy). After giving these ratings,

subjects described the strategy they would have used if they had not received training.

4.3 Results of strategy training

Reasoning errors made by the two training groups were compared at successively finer levels of detail. The analyses parallel those applied previously to errors from subjects giving retrospective reports of spontaneously adopted strategies.

First, the group trained to use the Concrete Properties strategy had a significantly higher overall error rate (34 percent) than the group trained in the Abstract Directional strategy (17 percent). Second, the two training groups exhibited different general patterns of reasoning errors, as indicated by a statistically reliable interaction of training groups and problem types. Third, the specific problem factors found to distinguish the spontaneous report groups had analogous effects in the training groups. Errors made by the Abstract Directional training group were related strongly to the number of premises in a problem beginning with the pivot term (the "end anchoring effect" described previously), but were not strongly related to the number of inverses in a problem (uses of the word "sadder"). Subjects trained in the Concrete Properties strategy tended to show the complementary pattern. Fourth, when the process models in Tables V and VI were fitted to the error data of each group of subjects, the appropriate process model gave the better fit in each case (see Table VIII).

Two further analyses related the strategy reported by subjects prior to receiving training to their performance after training. In the first of these analyses, subjects were grouped by the strategy they reported prior to training, and the reasoning errors of the different groups were

Table VIII—Proportion of variance* in problem difficulty attributable to two strategy models

Parameters	Abstract Directional Group	Concrete Properties Group
Abstract Directional Model		
1. SEARCH	0.691†	0.330‡
2. POSITION	0.003	0.198
$\sum R^2$ (Percent of Reliable Variance)	0.694† (93.9%)	0.528† (57.4%)
Concrete Properties Model		
1. SCAN	0.372‡	0.743†
2. GENERATE	0.067	0.113‡
3. ENCODE	0.014	0.012
$\sum R^2$ (Percent of Reliable Variance)	0.453 (61.3%)	0.868† (94.4%)

* These proportions are increments in R^2 values attributable to each parameter. The order in which parameters are given corresponds to the step at which they entered the regression equation.

† $p < 0.01$

‡ $p < 0.05$

compared (see Table IX). The general result was that subjects trained to use the Abstract Directional strategy made fewer errors than those trained to use the Concrete Properties strategy no matter what strategy was initially reported. This result suggests that people can be trained to use a particular reasoning strategy even if they have not adopted that strategy spontaneously.

Subjects given Concrete Properties training rated their strategy significantly more difficult to use ($\bar{x} = 3.81$) than those given Abstract Directional training ($\bar{x} = 2.36$). These ratings then were related to the strategy reported by subjects prior to training. All subjects given Abstract Directional training found that strategy relatively easy to use no matter what their initial strategy had been. For people trained to use the Concrete Properties strategy, the results were different. The Concrete Properties strategy was rated more difficult to use by Abstract Directional Thinkers ($\bar{x} = 4.41$) than by Concrete Properties Thinkers ($\bar{x} = 3.20$). Virtually all the subjects who initially reported using the Abstract Directional strategy indicated that they would have applied that strategy to the happy/sad problems if there had been no training. This pattern of results suggests that people can appreciate a good reasoning strategy: subjects rated a less efficient reasoning strategy as "difficult to use," especially if they knew a more efficient strategy.

4.4 Summary

People can be trained to use different mental processes for making transitive inferences in three-term series problems. This fact has been demonstrated by analyses of the reasoning errors made by people after receiving training in different strategies. Results closely parallel the results of previous analyses of errors made by people retrospectively reporting the different reasoning strategies. Further results suggest that training in a particular strategy can be effective even for those people who did not report the strategy in spontaneous reasoning. People also appear to be sensitive to the difficulty of using various reasoning strategies, and rate a less efficient strategy "difficult to use," especially if they know a better one.

Table IX—Reasoning error rates made by subjects after strategy training

	Strategy Reported Prior to Training		
	Abstract Directional	Concrete Properties	Other/Not Clear
Strategy trained to use	$\bar{X}$ (N)	$\bar{X}$ (N)	$\bar{X}$ (N)
Abstract Directional	0.10 (12)	0.09 (6)	0.25 (15)
Concrete Properties	0.33 (11)	0.33 (5)	0.36 (16)

V. GENERAL DISCUSSION

5.1 Retrospective reports about reasoning

These studies provide a direct test of the reliability and validity of retrospective reports about reasoning. Generally, reports by different subjects contained sufficient information to classify the subjects consistently. The content of subjects' reports also was related systematically to the patterns of reasoning errors subjects made.

While retrospective reports proved very useful in these studies, they also had definite limitations. The classification of different people on the basis of their reports was not perfectly reliable. Another limitation was that a sizable minority of people gave reports that were either idiosyncratic or incoherent (Other/Not Clear reports). Subjects also reported metaphors or general descriptions of reasoning strategies (see Table II) rather than complete and detailed models like those in Tables V and VI. Using a crude classification rule, the majority of subjects' reports could be grouped reliably into two categories, and people giving different kinds of reports tended to exhibit different amounts and patterns of reasoning errors.

The fact that different people reported and used different reasoning processes has two practical implications. One is that it may be possible to obtain useful reports of reasoning-like processes in practical situations where a procedure or device is to be evaluated. Reports may lead to redesigning tasks to make them more compatible with people's reasoning processes. Another implication is a methodological suggestion. If reports on reasoning-like processes are used, it may be wise to obtain them from a large number of different people to gauge the range of mental processes people are likely to adopt.

The general conditions under which retrospective reports about thought processes are reliable and valid remain to be established. The method used in these studies was to have subjects carefully describe how they thought through specific kinds of problems immediately after attempting to solve a large number of the problems. Subjects were not asked why they chose a particular strategy, a kind of judgment that people are notoriously poor at making.[13] Subjects also were not asked to "think out loud" while solving problems, a technique that might have yielded more precise descriptions of the reasoning process. The cost of that technique in the present experiments on reasoning is that it would have prevented the collection of unbiased reasoning error data, and therefore would have clouded the test of the validity of reports. For other purposes, the technique of "thinking out loud" may be quite acceptable. What kinds of mental processes are reportable and which techniques are best for reporting them are two questions that must be answered before reports about thought processes generally can be used with confidence.[14]

5.2 Differences in people's reasoning

An important result of these studies is that different people spontaneously adopted different reasoning strategies for solving very simple, highly stereotyped, three-term series problems. Different models of reasoning processes accounted for large amounts of the variance in problem difficulty for people classified as using different reasoning strategies. These models also received support from results of the training study in which people were directed to use one reasoning strategy or the other (Table VIII). While the process models provide a good first approximation to the reasoning processes employed by different people, the more basic and important result is that people spontaneously reason in different ways.

Training has been shown to cause people to adopt different reasoning processes in a rather simple way: people can follow different sets of directions on how to reason. Subsequent studies[9] identified two other factors that influence in a subtle way whether people use one reasoning strategy or another. One factor is aptitude for visualizing spatial transformations of figures. People good at spatial visualization are more likely to adopt the Abstract Directional strategy spontaneously than those who have difficulty with spatial visualization. The use of different reasoning strategies for three-term series problems is not, however, influenced by verbal aptitude. A second factor influencing the adoption of a reasoning strategy is the context in which reasoning problems are posed. The strategy used for a particular problem is influenced by surrounding problems. Some context problems apparently suggest or allow the development of good strategies while others inhibit strategy development.

A new theory of reasoning is required to account for differences in people's reasoning. The emerging picture includes a dynamic and modifiable process in which reasoning strategies are developed. This phase of strategy development may be influenced by people's basic capacities, the context in which the reasoning occurs, and training. The fact that people can rate the difficulty of using different reasoning processes suggests that feedback of this kind also may be involved in strategy development. Certainly this picture is a far cry from the idea that reasoning is a fixed ability in which people differ by the amount they have or how well they can perform. The emerging picture holds the hope that reasoning processes used by different people can be understood, and that the understanding might lead to redesigning some tasks that many people find very difficult.

5.3 Training in reasoning

The training study reviewed here demonstrates that people can learn to use different reasoning processes with consequent effects on

their reasoning performance. A good characterization of the Abstract Directional training would be that it suggested an efficient general approach that most subjects could use, but that some subjects would not have discovered spontaneously. This result tends to confirm informal observation that a suggested strategy for dealing with a complex problem (e.g., thinking of a queueing problem in terms of a "pushdown stack", adopting spatial metaphors for programming problems, etc.) can be very helpful and will lead to certain patterns of performance. Perhaps there are many situations in which directions on "how to think about" a task (e.g., activating a telephone service or interacting with a computer) might reduce errors and lead to more predictable patterns of performance compared to directions in which learners must develop a strategy on their own.

The training study did not address the very important question of whether people can be trained to reason better in general. It is unlikely that the effects of strategy training for a specific simple transitive inference problem would generalize to cause subjects to reason more efficiently in many other situations. People who in general are good reasoners probably have discovered and used a large set of strategies that they can retrieve in given situations. Because of their basic capabilities (e.g., good spatial visualization) and previous experience, good reasoners probably are also very good at generating new alternative strategies for novel problems. The results reviewed here suggest that people can be trained to deal more effectively with some specific situations involving reasoning. This basic fact certainly does not reduce the likelihood of finding ways to train people to reason better in a large range of situations.

VI. ACKNOWLEDGMENTS

Dorothea Grimes-Farrow collaborated on the first two studies reviewed here. Peter Meany and Anthony Rooney assisted in collecting and analyzing data for several of the studies. E. Z. Rothkopf gave many helpful suggestions for revising previous drafts of this paper.

REFERENCES

1. V. A. Krutetskii, *The Psychology of Mathematical Abilities in School Children*, J. Teller, Trans; J. Kilpatrick and I. Wirszup, Eds., Chicago: The University of Chicago Press, 1976.
2. F. Davis, "Psychometric research on comprehension in reading," Reading Res. Quart., *7* (Summer 1972), pp. 628–78.
3. D. M. Irons, "Cognitive correlates of programming tasks in novice programmers," Proc. of Human Factors in Comp. Syst., Gaithersburg, Md., March 15–17, 1982.
4. C. Burt, "The development of reasoning in school children," J. Exp. Pedagogy, *5* (1919), pp. 68–77.
5. L. L. Thurstone, "Primary mental abilities," Psychometric Monogr., *1*, 1938.
6. R. J. Sternberg, "Representation and process in transitive inference," J. Exp. Pscyh.: Gen. *109* (June 1980), pp. 119–59.

7. H. H. Clark, "Influence of language on solving three-term series problems," J. Exp. Psych., *82* (November 1969), pp. 205–15.
8. D. E. Egan and D. D. Grimes-Farrow, "Different mental representations spontaneously adopted for reasoning," Mem. & Cog., *10* (July 1982), pp. 297–307.
9. D. E. Egan, unpublished work.
10. D. E. Egan, unpublished work.
11. C. B. DeSoto, M. London, and S. Handel, "Social reasoning and spatial paralogic," J. Pers. and Soc. Psych., *2* (1965), pp. 513–21.
12. J. Huttenlocher, "Constructing spatial images: a strategy in reasoning," Psych. Rev., *75* (November 1968), pp. 550–60.
13. R. E. Nisbett and T. D. Wilson, "Telling more than we can know: verbal reports on mental processes," Psych. Rev., *84* (May 1977), pp. 231–59.
14. K. A. Ericsson and H. A. Simon, "Verbal reports as data," Psych. Rev., *87* (May 1980), pp. 215–51.

AUTHOR

Dennis E. Egan, A.B., 1969 (Psychology), College of the Holy Cross; M.A., 1973 (Mathematics), Ph.D., 1973 (Experimental Psychology), University of Michigan; Naval Aerospace Medical Research Laboratory, 1973–1976; Bell Laboratories, 1976—. Mr. Egan has worked on individual differences in cognitive abilities, especially spatial ability and reasoning. At Bell Laboratories, he has explored ways of adapting instruction to individual differences in abilities, experience, and other learner characteristics. Mr. Egan is currently a member of the Learning and Instruction Research Department.

New Technological Demands

The articles in this section sample a range of task requirements, from turning a knob to the complex motor activities involved in typing. The common objective in these studies is to find ways to make tasks fit the abilities and prior learning that humans bring to them.

Donegan and Koppes introduce three articles on perceptual and motor factors affecting the design of equipment used by craftspeople who install and maintain telephone equipment. These articles deal with knob characteristics (Kohl), cable splicing techniques (Paul), and the perceptual problem of locating corresponding terminals in two high-density connector fields that are mirror images of one another (Flamm). Finally, Cohen describes a study comparing the perceptual motor effects of typing with membrane keys, which move little when pushed, with conventional travel keys.

No fundamentally new human factors issue is addressed in the articles appearing in this section. They collectively illustrate that old questions will need new answers whenever the human/system interface changes.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Human Factors Engineering for the Loop Plant

By J. DONEGAN* and D. N. KOPPES*

(Manuscript received

Bell System Loop Plant operations employ 150,000 craftspeople. New technology is changing the demanding jobs they perform, and the range of physical characteristics of craftspeople is increasing as women join the traditionally male work force. Both of these factors indicate the need for human factors contributions to designing the tools and apparatus that craftspeople use. This article introduces three other articles describing different aspects of human factors work for the Loop Plant.

I. INTRODUCTION

The installation, testing, operation, and maintenance of the Loop Plant (the customer cable network) requires the full-time employment of some 150,000 outside plant and central office craftspeople—a significant portion of the total Bell System craft work force. An important element of the human factors work in the Loop Transmission Division is knowing the characteristics of the craftspeople and their tasks so that we may supply engineers with human factors criteria for the design of loop apparatus and systems.

The published literature on anthropometrics, biomechanics, human factors, and applied psychology is, of course, a valuable resource.

* Bell Laboratories.

However, the literature is often incomplete with respect to Bell System tasks and the data often describe what can be done by people under ideal conditions, not what must be done by craftspeople under stressful conditions. Traditionally, physically demanding tasks have been performed by these craftspeople in suboptimal conditions: up on poles, down in manholes, and during foul weather conditions. While such work is still done, new technologies such as loop electronics, fiber-optic cable, and computer-controlled equipment have overlaid old jobs with more sophisticated tasks, requiring different physical and mental characteristics. Consider loop plant splicers responding to a major alarm concerning a fiber-optic cable repair. The first set of tasks could include lifting a 300-pound manhole cover, pumping "water" out of the manhole, and setting up a work station—perhaps 10 feet down, but still 10 feet from the bottom of the manhole. The next set of tasks could then involve the precision and fine manipulative skills described in the article entitled "Human Factors Comparison of Two Fiber-Optic Continuous-Groove Field-Repair Splicing Techniques" by Larry Paul, in which he compares two methods of field splicing fiber-optic cable that are currently under development at Bell Laboratories in Atlanta.

In addition to the introduction of new tasks, an increasing number of people of slight stature, particularly females, has altered the anthropometric characteristics of the loop plant work force. Where possible, the physically demanding tasks are redesigned to accommodate this new work force. A case in point is the ubiquitous use of torque knobs in loop plant equipment. In particular they are used to secure splicing equipment, and female craftspeople complain of difficulty in exerting enough torque to hold this heavy equipment in place. The experiment described in the article by George Kohl, entitled "Effects of Shape and Size of Knobs on Maximal Turning Forces Applied by Females," provided design guidelines for matching torque knob shapes and sizes with female capabilities.

The loop plant also includes two main areas for planned manual access to the network: the main distributing frame, and the feeder/distribution interface. At these points a constant "churning" of the loop network takes place in daily installation, removal, and rearrangement of cross-connecting wires. These areas have been very troublesome in the past and any new equipment, hardware, or tasks designed for these crucial access points are closely scrutinized to preclude wiring congestion or poor work practices. Well-motivated, trained craftspeople can always make equipment work but superior designs are required to accommodate the less skilled craftsperson who at one time or another will work on all of the in-place plant. Such attention to design detail and human factors issues is illustrated in the article by Lois

Flamm entitled "Performance in Locating Terminals on a High-Density Connector." This article describes an experiment that identified and helped resolve visual complexities associated with a new and compact protector block designed for use on the main distributing frame.

Many of the more significant loop plant hardware and systems developments are subjected to a full-scale evaluation at the Chester Field Laboratory.[1] Such evaluations take place prior to final production and utilize New Jersey Bell craftspeople to install, operate, and maintain preproduction models in realistic scenarios. However, laboratory experiments play a significant role at a much earlier stage of the design and development process. They may address questions related to the design of specific apparatus or resolve more general questions related to the capabilities of Bell System craftspeople. The articles by Lois Flamm, George Kohl, and Larry Paul are examples of such human factors experiments carried out in Department 54525 as an integral part of the development process of the new Loop Plant.

REFERENCES

1. Brent E. Coy, "Putting the human factor into the outside plant," Bell Lab. Rec., *59*, No. 1 (January 1981), pp. 17–23.

AUTHORS

John Donegan, B.Sc., 1956, Strathclyde University, Scotland; M.S., 1963, New York University; Rolls-Royce in Aero Engine Design, 1956–1957; Union Carbide Company—Mining Equipment Design, 1957–1961; Bell Laboratories, 1961–1976; American Bell, 1976–1979; Bell Laboratories, 1979—. During his initial years with Bell Laboratories, Mr. Donegan worked on microwave radio relay systems and the underground and buried cable systems. With American Bell he was stationed in Tehran as Manager of Outside Plant Standards. Mr. Donegan has been working in the area of human factors engineering of the Loop Plant since 1979 and is currently Supervisor of the Human Factors Engineering Group II.

David N. Koppes, B.C.E., 1959, M.C.E., 1960, Cornell University; Structural Engineer, Aluminum Company of America, 1960–1964; Bell Laboratories, 1964–1971; AT&T, 1971–1973; Bell Laboratories, 1973—. During his initial years at Bell Labs, Mr. Koppes investigated cable plowing techniques, had mechanical design and operations responsibilities for Sea Plow II and Sea Plow III, and worked on maintenance systems for the WT-4 Millimeter Waveguide System. At AT&T, he was an Assistant Manager in Outside Plant Engineering. Mr. Koppes has been doing human factors work in the Loop Plant area since 1975 and currently is Supervisor of the Human Factors Engineering Group I.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Effects of Shape and Size of Knobs on Maximal Hand-Turning Forces Applied by Females

By G. A. KOHL*

(Manuscript received December 23, 1981)

Outside plant craftspeople use knobs to apply turning forces on clamp mechanisms that hold field equipment temporarily in place. A study was performed to develop a set of data that provide guidance for determining knob size and shape characteristics most appropriate for various outside plant working conditions. Forty female participants applied maximal isometric turning force to each member of a set of twenty experimental knobs that systematically varied in shape and size. In half the trials the participants applied force with greased hands and in the other half used nonslip compound. In addition, two arm-wrist positions were observed. In general, triangular knobs allow more hand torque to be generated and require significantly less material than square, pentagonal, hexagonal, or circular knobs of comparable size. However, this effect depends upon the arm-wrist position and grip conditions. A 3.5-inch turning diameter is desirable when both cost and performance are considered.

I. INTRODUCTION

Outside plant craftspeople in the Bell System use knobs to apply turning forces on screw-operated clamp mechanisms used to hold heavy field equipment temporarily in place. Female craft who use these

* Bell Laboratories.

knobs report that the knobs they currently use do not enable them to easily generate sufficient torque to clamp down equipment. The question raised by these reports and addressed here is: What characteristics of a knob allow turning forces to be easily generated? Considering the variety of torque knobs in the world (e.g., door knobs, small valve wheels, faucets), and the considerable body of knowledge the human factors community has acquired about control knobs, it is perhaps surprising to find that the literature sheds little light on this rather straightforward and practical question (see Refs. 1 and 2 for the best existing treatments of torque knobs).

The following experiment was designed to provide a set of data that can assist designers in developing torque knobs, and we hope will be of value to the hand tool industry in general. The approach taken in this study was strictly empirical and consisted of measuring the amount of turning force that participants can generate using knobs of various sizes and shapes under several different conditions.

II. METHOD

2.1 Participants

Forty right-handed females from the Whippany, New Jersey, area responded to newspaper and intracompany bulletin board ads, and served as paid participants. Mean and standard deviation of the group's age, height, and weight were respectively: 37 years, SD = 13.5; 64.7 inches, SD = 3.4, and 132 pounds, SD = 27.2.

2.2 Independent variables

Variables were identified that could affect knob-turning performance, including features of the knobs and the conditions under which the knobs might be used. To render the study manageable in size, only five of the most intuitively and/or practically relevant variables to the Bell System application were included in the study.

2.21 Knob shape

There are a limitless number of possible knob shapes. A shape attribute, "sidedness," was chosen to succinctly capture the spectrum of shapes that could affect knob-turning performance. Five shapes were chosen, as shown in Fig. 1, varying from few sides to infinite sides (i.e., triangle, square, pentagon, hexagon and circle).

2.2.2 Knob size

An operational definition of size was generated so that knobs of different shapes could be compared. The "diameter" of a knob is the diameter of the circle bounding the outermost points of the knob. Thus the triangular knob always has about half the area of a circular

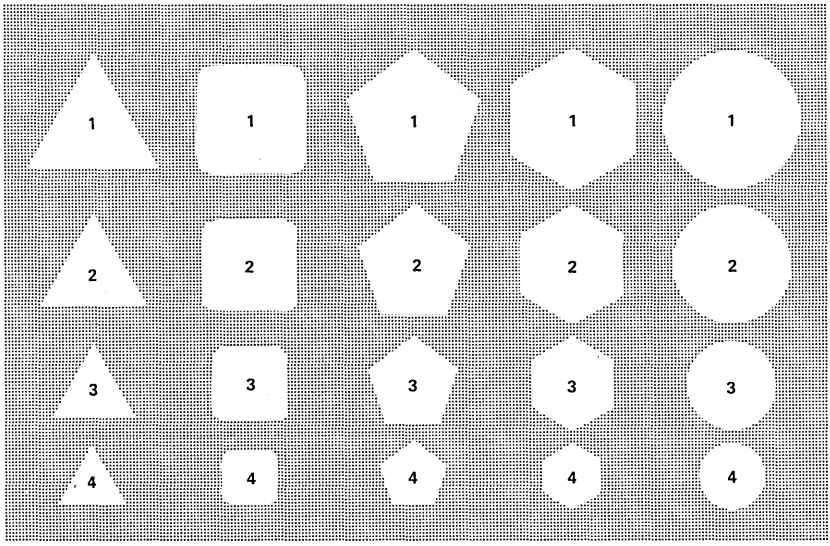


Fig. 1—The set of experimental knobs.

knob with an equivalent diameter. The definition has real-world validity in that the envelope in which a knob turns is likely to be a design constraint. Four diameter sizes were included in the study: 2.5, 3.5, 4.5, and 5.5 inches. Preliminary evidence showed these sizes to represent plausible end- and middle-point values for diameters of knobs to be used to generate turning force.

2.2.3 Grip condition

The types of knob surface conditions that could be tested are numerous, for example, knurled surfaces, serrated edges, slippery surfaces. The surfaces of the knobs in this study were all the same—smooth, anodized aluminum. However, an attempt was made to study the extremes of surface conditions by including two grip friction conditions under which the knob was turned. In the *greased* condition participants greased their hands before turning; in the *nonslip* condition they applied a nonslip compound.

2.2.4 Position

The hand-arm position was varied to emulate various real-world conditions, where knobs must be turned from many different vantage points. Two standing positions were included; in both conditions the height of the knob was adjusted to the level of the elbow joint for each participant. In the *front position* the participant stood straight in front of the knob, forearm parallel to the floor and perpendicular to the upper arm and frontal plane of the body, the axis of knob rotation

coincident with the longitudinal axis of the forearm. The wrist was bent so that the hand pointed up, the palm of the hand against the knob face, fingers spread around the knob edge. In the *side position* the participant stood to the side of the knob with the forearm parallel and perpendicular to the axis of rotation of the knob. The palm of the hand pressed against the face of the knob with the fingers spread around the knob edge.

2.2.5 Participant size

Size was included as a subject variable. Although hand size was originally targeted as the variable that would most affect performance, weight was found to be a better predictor.[3] Four weight groups of ten participants each were formed on a post hoc basis. The mean and standard deviation of the four group's weights in pounds were: 106 pounds, SD = 5.72; 120 pounds, SD = 3.65; 132 pounds, SD = 5.87; 171 pounds, SD = 22.96.

2.3 Experimental design

A complete factorial design was employed (5 knob shapes × 4 knob sizes × 2 grip conditions × 2 positions × 4 participant weight groups). All participants were included in all possible within-participant treatment conditions.

2.4 Procedure

Each participant attended for one day on two consecutive weeks for approximately four hours each day. On each day each participant performed four blocks of twenty trials, each block consisting of one appearance of each of the twenty knobs. For each trial participants were instructed to apply maximal isometric turning force to a knob specified by the experimenter, using the right hand for a period of three one-second beats of a metronome. The order of the knob appearance was randomized for each block of trials for each participant. On each day, the front position was employed on two blocks of trials, the side position on the other two; one of the side position and one of the front position trial blocks were performed with a greased hand. The remaining two blocks were performed using a nonslip compound.

Participants were observed in groups of three or four, every participant performing a single trial before any given participant performed her next trial. Roughly two minutes intervened between any given participant's trials, and breaks of twenty minutes were taken between blocks of trials.

2.5 The knobs

The relative size and shape of the knobs is shown in Fig. 1. They were made of machined aluminum and had an anodized smooth

surface. All corners and edges of the knobs were rounded and the smallest radius on any edge was 0.25 inch. The knobs were 1.25 inches thick.

III. RESULTS

The data were submitted to an analysis of variance,[4] with knob shape, knob size, grip conditions, position, and day of participation treated as within-subjects variables, and participant weight as a between-subjects variable. The most important results are summarized here.

The most informative effect obtained was a four-way interaction of the shape, size, position, and grip variables [$F(12,432) = 5.2$, $p < 0.00001$]. Fig. 2 shows this interaction. The top two panels show performance as a function of knob shape and size for the side position, the left panel for the nonslip blocks and the right panel for the greased blocks of trials. The bottom two panels show performance using the front position, the left panel for the nonslip condition and the right panel for the greased condition.

Several effects are apparent from a visual inspection of Fig. 2. The greased condition performance (right two panels) is lower than the nonslip condition [$F(1,36) = 289.6$, $p < 0.00001$]. Further, over all other variables, performance decreases as the number of knob sides increases [$F(4,144) = 183.2$, $p < 0.00001$]. This main effect is qualified by an interaction with the grip variable [$F(4,144) = 71.0$, $p < 0.00001$], indicating that the main effect of shape is much more pronounced in the greased condition than in the nonslip condition. Figure 3 shows this two-way interaction.

The effect of size is apparent in Fig. 2; across all other variables, the bigger the knob diameter, the more torque developed [$F(3,108) = 246.8$, $p < 0.00001$]. The left-to-right convergence of the curves in Fig. 2 reflects the two-way interaction of the knob size and shape variables: over all other variables, the greater the number of sides, the less advantage size has [$F(12,432) = 53.6$, $p < 0.00001$]. The fact that the convergence of the curves is more pronounced in the greased (right two panels) than the nonslip panels is reflected in a two-way interaction of knob shape, knob size, and grip condition [$F(12,432) = 8.4$, $p < 0.00001$]. The reversal of the sidedness advantage for smaller knobs in the nonslip, top left panel is counter to the patterns in the other three panels and is reflected as the four-way interaction illustrated in Fig. 2.

Participant weight was only a marginally significant main effect [$F(3,36) = 2.6$, $p < 0.07$] but did interact with the knob size variable [$F(9,108) = 4.8$, $p < 0.0001$], indicating that the advantage larger people have over smaller people dwindles as knob size decreases.

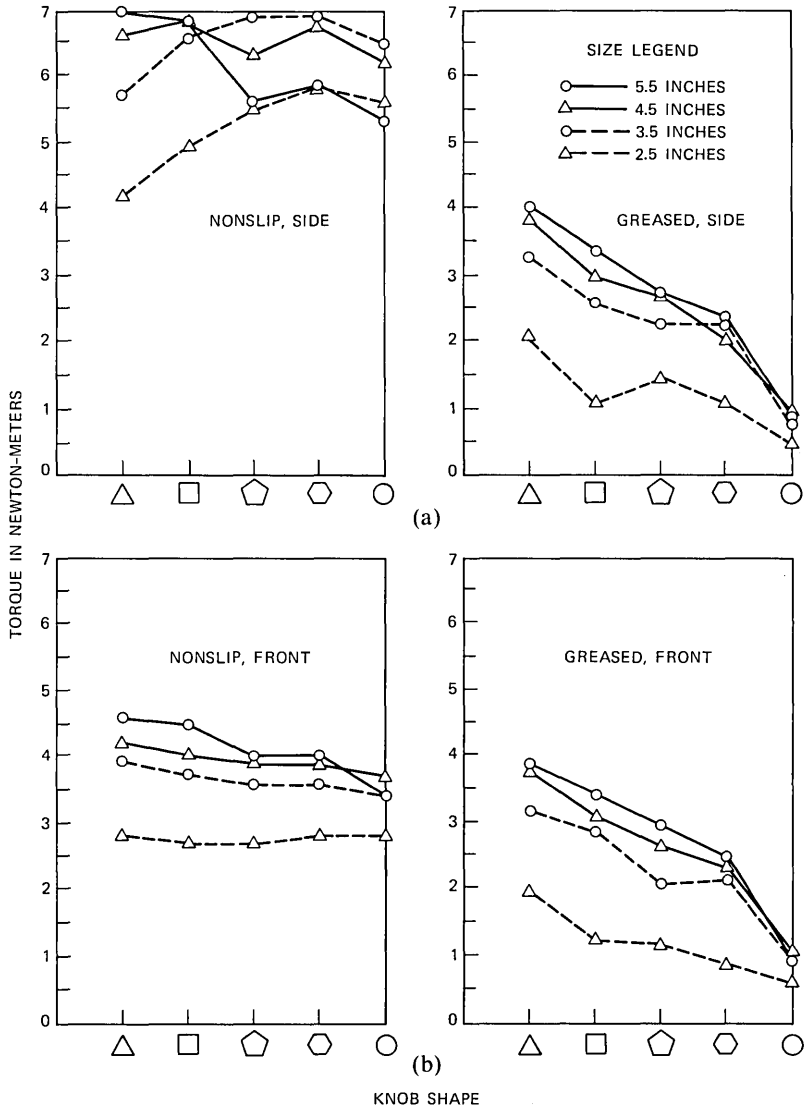


Fig. 2—Four-way interaction of the knob shape, size, position, and grip condition. (a) Performance as a function of the knob shape and size for side position. (b) Performance as a function of knob shape and size for the front position.

Several other main effects and interactions were obtained: position, day, grip × position, grip × knob size, position × knob size, position × shape, day × shape, grip × position × knob size, grip × position × shape, and position × knob size × shape.

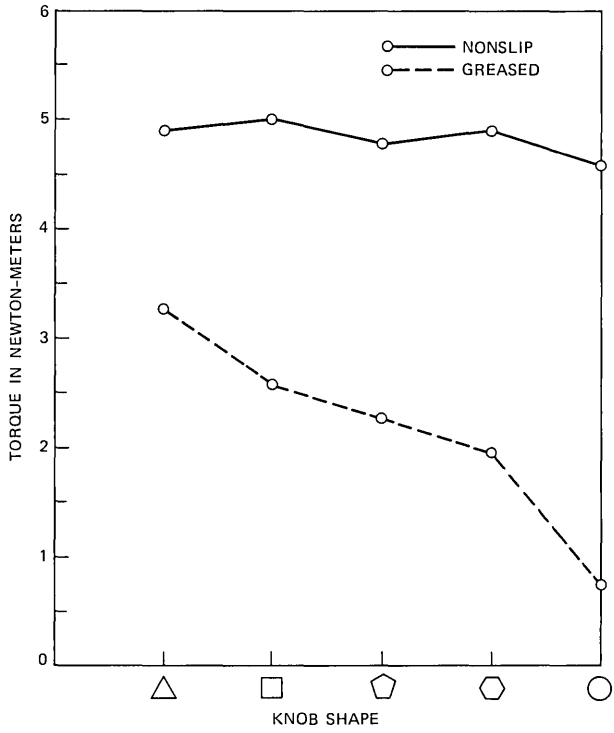


Fig. 3—Knob shape by grip condition.

IV. SUMMARY AND DISCUSSION

The practical difference between means in Fig. 2 will have to be decided by individual designers. Note that if the application involves a nonslip, front approach, the shape of the knob has little effect, and the knob size is very important (Fig. 2, bottom left panel). However, in both the front and side greased conditions, shape is very important; note the smallest triangle results in better performance than the largest circle (Fig. 2, bottom right panel). The savings in materials in this case is roughly 10 to 1.

In general, triangular knobs allow for the generation of as much or more torque than any of the other shapes. However, subjective comfort ratings indicate that the smallest triangular knob causes discomfort. Also, the smallest triangular knob does not fair well in the nonslip, side condition. Therefore, for 2.5-inch diameter applications the square is probably a better choice.

The biggest jump in performance with respect to knob diameters comes between 2.5 and 3.5 inches; thereafter performance increases as diameter increases, but at a smaller rate.

Considering materials, torque and comfort, the triangular and square 3.5-inch knobs are recommended for general application. If more torque is required than can be obtained with these knobs, larger diameter triangular or square-shaped knobs should be used. For these or any other knobs, all corners and points should be rounded for best performance and comfort.

V. ACKNOWLEDGMENTS

The author expresses appreciation to Brian G. Bancroft for assisting in the development of experimental apparatus and conducting the investigation, and Elise Darrow and Ha T. Nguyen for assisting in the data analysis and preparation of this report.

REFERENCES

1. E. O. Sharp, "Maximum Torque Exertable on Knobs of Various Sizes and Rim Surfaces," MRL-TDR-62-17, Behavioral Sciences Laboratory, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, 1962.
2. A. D. Swain, G. C. Shelton, and L. U. Rigby, "Maximal Torque for Small Knobs Operated With and Without Gloves," Ergonomics, *13,* No. 2 (1970), pp. 201–8.
3. G. A. Kohl, unpublished work.
4. B. J. Winer, *Statistical Principles in Experimental Design* (2nd Ed.), New York: McGraw-Hill, 1971.

AUTHOR

George A. Kohl, Jr., B.S., 1968, United States Merchant Marine Academy; M.S., 1974, Experimental Psychology, Purdue University; Ph.D., 1978, Experimental Psychology, Purdue University; U.S. Merchant Marine Engineering Officer, 1968–1974; Assistant Professor of Psychology, Allentown College, 1977–1978; Bell Laboratories, 1978—. Mr. Kohl has been a Human Factors Specialist at Bell Laboratories, performing research, design, and development work on equipment and systems used by central office and outside plant craftspeople. Most recently, he has been a member of the team responsible for the design of a computerized transmission enhancement system to support special services operations. Member, American Psychological Association and the Human Factors Society.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Human Factors Comparison of Two Fiber-Optic Continuous-Groove Field-Repair Splicing Techniques

By L. M. PAUL*

(Manuscript received December 23, 1981)

Two Bell System fiber-optic splicing techniques that are under development were experimentally compared in a human factors study. In addition to splicing technique, the variables explored were the uniformity (evenness) of fiber spacing in the 12-fiber "ribbon," and the hand (preferred or nonpreferred) that was used to splice. Performance was measured by the time to insert fibers, the number of fibers broken, and the time taken to complete a series of insertions for each technique. The preference of the participants between the two techniques was also determined. All of the measures except participant preference showed the vacuum technique to be statistically superior to the hold-down bar technique, with participant preference suggesting the same conclusion, although it was not statistically significant.

I. INTRODUCTION

A human factors study compared two techniques under development for the repair splicing of damaged Bell System fiber-optic cable in the field. The goal of the study was to determine the preferred technique from a human factors viewpoint. This information combined with economic considerations would then allow the technique's designers

* Bell Laboratories.

to choose one of the techniques for further development and release to operating telephone companies.

II. DESCRIPTION

Bell System fiber-optic cable consists of one to twelve multiple-fiber "ribbon" arrays. Each ribbon consists of twelve optical fibers embedded in a plastic tape. Currently, each ribbon of the cable is shipped with a silicon connector on each end to facilitate the splicing (joining) of the ribbons of one cable section to the ribbons of the next section.[1] In joining one ribbon to another, the splicer first places silicon wafers on either side of one ribbon array connector, forming a sandwich, and secures this arrangement with a spring clip. The second ribbon connector is then slipped into this sandwich and a second spring clip is added. The final operation consists of the application and curing of a refraction index-matching gel.

These procedures are reasonably straightforward and have been successfully used by operating telephone companies. On the other hand, in certain situations the splicer must refabricate a ribbon's silicon connector in the field.[2] Refabrication is required when the fragile silicon ribbon connector is received damaged, is damaged in craft handling, when it is necessary to reposition ("swap") fibers within the ribbon, or when an additional splice point becomes necessary. In part, refabrication involves grinding and polishing of the individual fiber ends in the reconstructed silicon connector to minimize losses when the connector is spliced (joined) to another connector. Although field refabrication is successfully being done by operating telephone companies as the field-repair method, the two techniques described in this report are being developed to simplify field-repair splicing basically by eliminating the grinding and polishing operations.

Figure 1 schematically depicts the continuous-groove approach shared by the two splicing techniques. In both techniques, the silicon array connectors are first removed from the ribbons to be spliced. After fiber preparation, which includes removing the plastic tape, removing the individual fiber coatings, and developing square fiber ends, fibers of one ribbon are "combed" into the set of 12 continuous grooves. The fibers of the second ribbon are then similarly inserted into the continuous grooves. Once the fibers of each ribbon are inserted, they are brought into proximity and index-matching gel is applied and cured as in the standard joining procedure. Thus connector refabrication, with its necessary grinding and polishing steps, and the subsequent connector joining have been eliminated and replaced by a single operation.

The two continuous-groove techniques that were evaluated differ in the method for inserting fibers into the continuous grooves. The

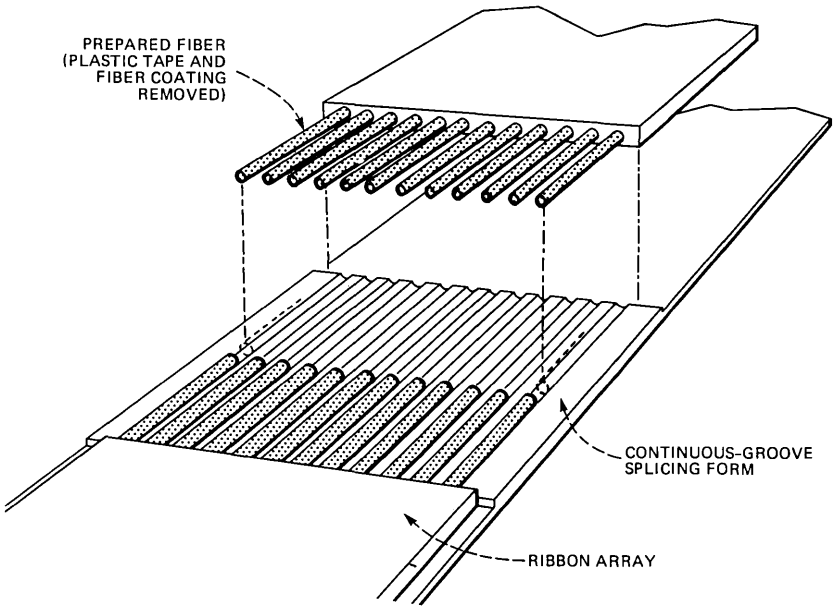


Fig. 1—Basics of a fiber-optic ribbon, continuous-groove splice.

vacuum technique uses a vacuum assist to help secure both the prepared fibers (tape and coating removed) and a portion of the ribbon beyond the prepared fibers. The splicer then uses a lateral motion to "coax" the 12 fibers into their respective grooves. In the hold-down bar technique, the splicer combs the prepared fibers into their respective grooves with the use of an alignment comb without any additional aid. The splicer first registers the twelve prepared fibers in the twelve slots of the alignment comb and then carefully pulls the ribbon back until the ends of the fibers go below a bar in the comb and fall to the continuous grooves below. The splicer then gently pushes the ribbon forward under the bar that holds the ribbon in the continuous grooves (hence the "hold-down bar" name), and pushes the ribbon along the continuous grooves to approximately the middle of the splicing form. The proprietary nature of these techniques, which are still under development, prohibit the use of photographs to illustrate the two techniques.

III. HUMAN FACTORS EXPERIMENTAL COMPARISON OF SPLICING TECHNIQUES

A human factors study was designed to compare the two field-repair splicing techniques on the basis of craft performance. An analysis of the sequence of operations in the techniques suggested that the only difference between fully developed techniques would be the method of

inserting fibers into the grooves. Both techniques would likely share the same or similar methods of fiber preparation, index-matching gel application, etc. Therefore, it was decided to limit the scope of this study to assessing performance differences between techniques for successfully inserting fibers into the continuous grooves.

3.1 Method

3.1.1 Design

The experiment was a within-participants design. Each participant was trained and tested under every combination of experimental variables. The three experimental variables were: splicing technique (vacuum or hold-down bar), fiber spacing uniformity (uniform or nonuniform), and splicing hand (preferred or nonpreferred).

Fiber spacing uniformity refers to the uniformity of spacing of the twelve fibers in the ribbon. Previous experience with handling of fibers showed that nonuniformly spaced fibers are more difficult to guide into grooved structures than uniformly spaced fibers. Nonuniformly spaced fibers can result from manufacturing and from the removal of the coatings on each fiber. Since nonuniform ribbons occur infrequently, they were created for the study by making a hole with a small awl in the plastic tape at the point where the prepared fibers emerge from the tape.

The third variable, splicing hand, was chosen to reflect the fact that *all* splicing techniques require the splicer to use both left and right hands on the ribbons coming from the left and right sides, respectively, and thus the potential interaction between splicing hand and field-repair technique is of interest. Although it is possible to rotate the splicing form so that the splicer uses the preferred hand for both ribbons, concern over building in transmission loss caused by bending the fibers precluded further consideration of this approach.

Four dependent measures selected to contrast the two field-repair techniques were:

1. Average insertion time
2. Number of broken fibers
3. Elapsed (clock) time
4. Participant preference.

Average insertion time, measured with the participants' knowledge, was the time between beginning an insertion and its successful completion. If a fiber was broken the clock was stopped, a new ribbon was provided, and the clock was restarted. Similarly, the clock was stopped and restarted if a question was asked or the experimenter wanted to clarify a procedure.

Elapsed (clock) time was simply the total time required to complete all 14 insertions required in each technique. It included training and

questions in addition to the actual insertion time reflected in the average insertion time.

Participant preference was gauged at the end of the experiment by asking the question: "If you were a Bell System splicer, which of the two splicing techniques you used would you prefer to use on a regular basis?"

3.1.2 Experimental apparatus

The work platform measured 14 × 24 inches and was chosen to closely approximate the current splicing work station. This rather small area is dictated by the limitations of many splicing situations. A stand-mounted 3× magnifier to allow close viewing of the fibers (approximate 0.005″ diameter with coating removed) and splicing form, and two halogen lamps to evenly illuminate the work station were chosen as a result of previous work. A beam splitter plate, television camera, and monitor comprised the observation system that allowed unobtrusive monitoring of participant performance and the demonstration of insertion technique by the experimenter, as discussed below.

3.1.3 Procedure

Participants were individually tested in a session lasting approximately two hours. They were first familiarized with Bell System fiber-optic cable and current procedures for splicing ribbons together. Half of the participants were then trained and tested on the hold-down bar technique first, with the other half receiving the vacuum technique first.

Training consisted of the experimenter demonstrating the correct insertion procedure. Participants first watched "over the experimenter's shoulder" and then observed the detailed aspects on the closed-circuit television monitoring system. After this initial training, all participants performed the sequence of splicing insertions outlined in Table I.

The first insertion was designed as practice with the participant using a uniform ribbon and his or her preferred hand. In addition, the ribbon was free, i.e., it was not anchored at one end. All subsequent

Table I—Sequence of experimental ribbon insertions for both splicing techniques

Step	Fiber Spacing	Task
1	Uniform	Practice with free ribbon using preferred hand
2	Uniform	One practice insertion using preferred hand
3	Uniform	Six insertions alternating hands
4	Nonuniform	One practice insertion using preferred hand
5	Nonuniform	Six insertions alternating hands

insertions were timed with the 20-inch ribbon anchored approximately 16 inches from the splicing form to simulate a ribbon emerging from a cable.

As shown in Table I, nonuniform ribbons were not introduced until later in the splicing sequence, and thus any comparison of the effect of fiber spacing uniformity on splicing performance within a particular splicing technique is confounded by the additional training for nonuniform ribbons relative to uniform ribbons. This point will be discussed further in Section 3.2.

After completing the insertion sequence using one splicing technique, participants were given a five-minute break and then trained and tested in an identical manner using the other technique. After completing both techniques, participants were asked a short series of questions to obtain more detailed information on their performance. Questions were asked concerning any hobbies requiring fine motor control such as knitting, use of glasses (including bifocals) during the experiment or at other times, general comments, and their preference between the two field-repair techniques.

3.1.4 Participants

It was considered useful to select two distinct subpopulations as participants, each with some degree of demonstrated motor skill, in order to be able to generalize the results of the experiment to operating telephone company craftspeople. Nine people from the Bell Labs wiring shop and nine from the Bell Labs clerical pool participated.

3.2 Results

Performance of the two subpopulations, wiring shop and clerical, was virtually identical. The subpopulation variable was not significant ($F < 1$) and did not interact with any of the other independent variables. This similarity of results for the two subpopulations presumably reflects the fact that both groups did indeed possess a similar degree of general motor skill with which to approach the learning of this new motor skill.

3.2.1 Average insertion time

Figure 2 shows average insertion time for the nonpractice insertions (Steps 3 and 5, Table I) as a function of splicing technique and fiber spacing uniformity for all 18 participants. The superiority of the vacuum technique over the hold-down bar technique was significant [$F(1,16) = 19.6, p < 0.001, \mathrm{MS_e} = 0.52$]. This statistical superiority of the vacuum technique was obtained despite the fact that four participants had to be stopped on one or more insertion attempts using the hold-down bar technique and the time to that point used as an

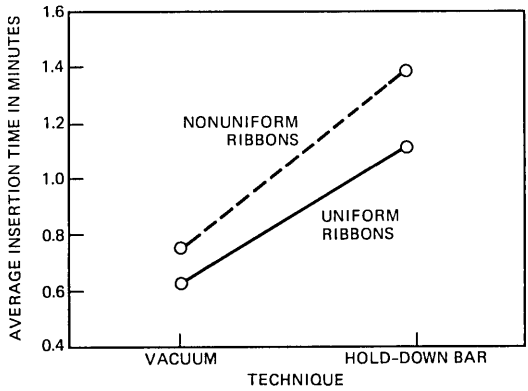


Fig. 2—Average insertion time on uniform and nonuniform ribbons for each splicing technique.

(under) estimate of that insertion attempt. Participants were stopped whenever an attempt exceeded approximately four minutes in order to allow completion of the experiment in a two-hour period.

The smaller average insertion time for uniform ribbons relative to nonuniform ribbons shown in Fig. 2 was statistically significant [$F(1,16) = 5.6, p < 0.05, MS_e = 0.23$]. As previously noted (and shown in Table I), nonuniform ribbons were presented after uniform ribbons, introducing a confounding between fiber spacing uniformity and training. However, since it is reasonable to assume that performance with the nonuniform ribbons benefitted from this additional training relative to the uniform ribbons, the smaller average insertion time for the uniform ribbons may be interpreted as a valid effect. Neither the interaction between splicing technique and fiber spacing uniformity, nor interactions between any of the other independent variables were statistically significant.

Participants performed marginally better with their preferred hand, although this difference was not significant ($p > 0.05$). The lack of an interaction between splicing hand and technique was surprising since a distinct preferred-hand advantage was expected for the hold-down bar technique given the motor skill required for this technique.

3.2.2 Additional measures

Table II contrasts the two field-repair splicing techniques using the other three dependent measures. The number of broken fibers was significantly less in the vacuum technique [$F(1,16) = 14.6, p < 0.005, MS_e = 3.25$]. No other main effects or two-way interactions were found using number of broken fibers as the performance measure. Elapsed (clock) time was also significantly less in the vacuum technique

Table II—Additional ribbon insertion measures

Measure	Technique	
	Vacuum	Hold-Down Bar
* Number of broken fibers (total)	4	24
* Elapsed (clock) time	33.6 min	56.9 min
Participant preference (total)	12	5

* Statistically significant at $p < 0.005$.

(Wilcoxon test, $p < 0.002$). Participant preference (with one abstention) favored the vacuum technique, although it was not significant as assessed by a Sign test.

IV. DISCUSSION

4.1 Vacuum technique superiority

The various measures strongly suggest the superiority of the vacuum technique over the hold-down bar technique. This superiority is basically a time advantage realized both directly in using the technique, and indirectly by reducing the number of repeated fiber preparations required when fibers are broken. However, it should be noted that 14 of the 18 participants were successful in completing all 12 of the nonpractice insertions required of them in the hold-down bar technique. Further, even the four participants who had to have one or more of their insertion attempts terminated because of time limitations completed at least 10 of the 12 attempts. This relative facility is important as it was not clear, a priori, whether inexperienced people could be easily trained to use the hold-down technique, especially with the brief (approximately 10-minute) training.

In interpreting the results of this study, it should be remembered that the study focused on only one of the operations required in field repair, i.e., ribbon insertion. Although ribbon insertion is undoubtedly the most difficult operation for the splicer, the time required by the other operations common to any field-repair technique may somewhat overshadow the vacuum technique time advantage. Thus, other factors such as economics and manufacturability are important in selecting the technique to be developed for release to the operating telephone companies.

4.2 Unanswered questions

Two relevant human factors concerns remain unanswered by this study. One unknown is the amount of additional training that would be required to improve the performance of those people experiencing difficulty with the hold-down bar technique. A second question is the stability of the learned field-repair skills over time, i.e., the memory for the skills. The ability to perform either technique will probably

decrease as a function of the time since a person's last splicing experience. This forgetting is particularly relevant to the scenario in which a period of weeks or months elapses between splicing experiences. It is not clear, for example, that the vacuum technique would remain superior to the hold-down bar technique over time. Further studies will help to answer these questions.

REFERENCES

1. C. M. Miller, "Fiber optic array splicing with etched silicon chips," B.S.T.J., *57*, No. 1 (January 1978), pp. 175–90.
2. H. J. Friedrichsen and P. F. Gagen, "Optical cable field repair using array fabrication," 7th European Conf. on Optical Fiber Commun., September, 1981, Copenhagen, Denmark.

AUTHOR

Lawrence M. Paul, B.S.E.E., 1967, University of Maryland; M.S.E.E., 1969, University of Illinois; Ph.D. (Experimental Psychology), 1975, Purdue University; Design Engineer, Philco-Ford, 1968–69; Digital Project Engineer, Pertec Inc., 1969–71; Assistant Professor of Psychology, Lehigh University, 1975–78; Bell Laboratories, 1978—. Since coming to Bell Laboratories, Mr. Paul has been a human factors specialist consulting on projects relating to operating telephone company outside craftspeople. His principal focus has been on the placing and splicing of lightguide (fiber-optic) cable. Lightguide projects have included the work described in this issue, development of a splicing work station, and the field observation of early lightguide installations to identify human factors concerns. Member, American Psychological Association, Eastern Psychological Association, Human Factors Society, Tau Beta Pi, Eta Kappa Nu, Phi Kappa Phi.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Performance in Locating Terminals on a High-Density Connector

By L. E. FLAMM*

(Manuscript received October 6, 1981)

In two experiments thirty participants located terminals on variations of a new main distributing frame connector. Abutting individual connectors in mirror image led to a higher termination density. Location times and errors were evaluated using deadline procedures. Experiment 1 results showed a high incidence of parallax, counting, and left-right reversal errors. Design modifications aimed at reducing these errors led to improved performance in Experiment 2. Parallax, counting errors, and location times were significantly reduced. The continued occurrence of left-right reversal errors is discussed.

I. INTRODUCTION

Main distributing frame (MDF) connectors provide termination points for outside plant cable and protect central office personnel and equipment from harm due to foreign electrical potentials. Some connectors also have provision for making cross-connections. Over the years connector design has responded to increased space demands on main frames. A new 309-type protected connector achieves a higher density of terminations by putting two independent connectors to-

* Bell Laboratories.

gether in mirror image. Although the physical arrangement of the two abutted connectors differs, functionally they are equivalent.

Figure 1 shows prototypes of the original 309-type connector and a modified version with designation labels to identify terminal numbers. The modified connector incorporates features based on recommendations from the human factors studies described below. The direct spatial correspondence between cross-connect and protector fields and the familiar grouping by fives or tens found on previous connectors is missing in the new composite. Because each connector row consists of two cross-connect pairs and five protector units, it was important that the designation labels reflect users' expectations of direction of count and facilitate their making the appropriate associations.

The aim of a preliminary pencil and paper experiment (unpublished) was to choose the most natural numbering scheme. For both connectors, results showed consistent but independent horizontal left-to-right counting patterns associated with cross-connect and protector locations. Concern was generated from the preliminary study about potentially high error rates within connectors and possible confusions between connectors. The loss of spatial correspondence with the new design and the above results led to the hypothesis that shading connector backgrounds would help users to localize specific cross-connect and protector terminals. In the first experiment, two shading schemes were compared with the unshaded connector version in both slow and fast deadline conditions. It was anticipated that any differences in performance between the shaded and unshaded connector versions would be more evident with the rigorous work pace of the fast deadline.

II. EXPERIMENT 1

2.1 Method

2.1.1 Subjects

The participants in Experiment 1 were 18 male Bell System employees, several of whom had relevant experience as craftspeople.

2.1.2 Design and Materials

The experiment can be characterized as a within-subjects design in which each participant located cross-connect pairs and protector units on each of the three connector versions in both deadline conditions. As schematically illustrated in Fig. 2, the version 1 connector represents the original prototype 309 connector with the preferred labeling; in addition versions 2 and 3 include cross-connect field background patterns grouping cross-connect pairs by fives and tens, respectively. The layout of each connector version is a mirror image, but the

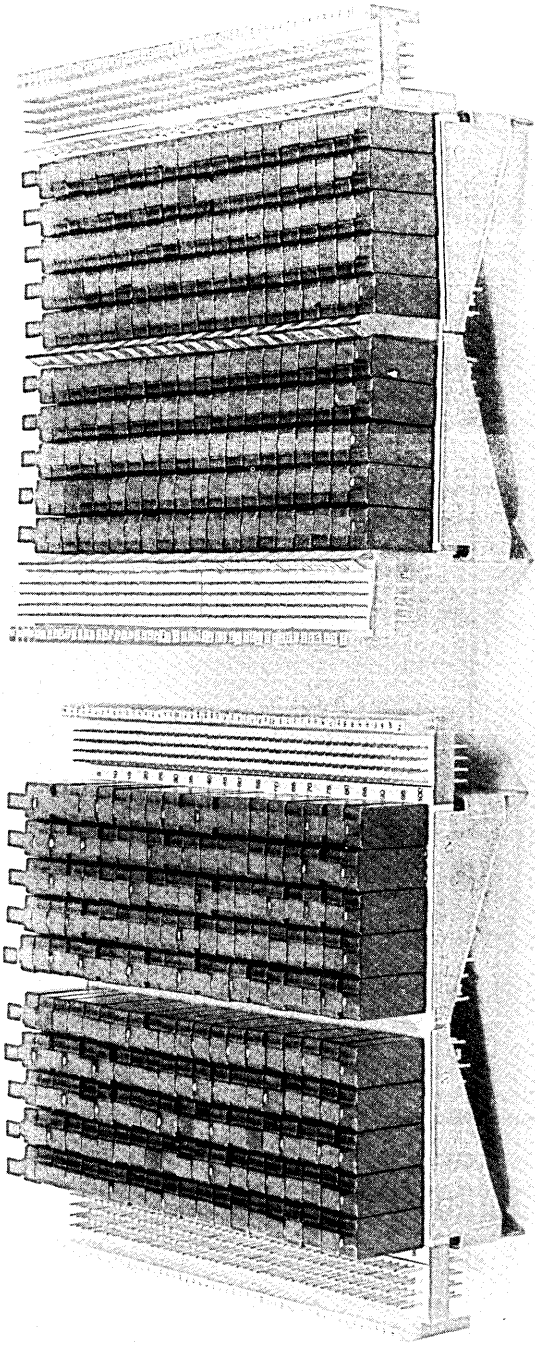


Fig. 1—Originally proposed (left) and modified (right) 309-type protected connector.

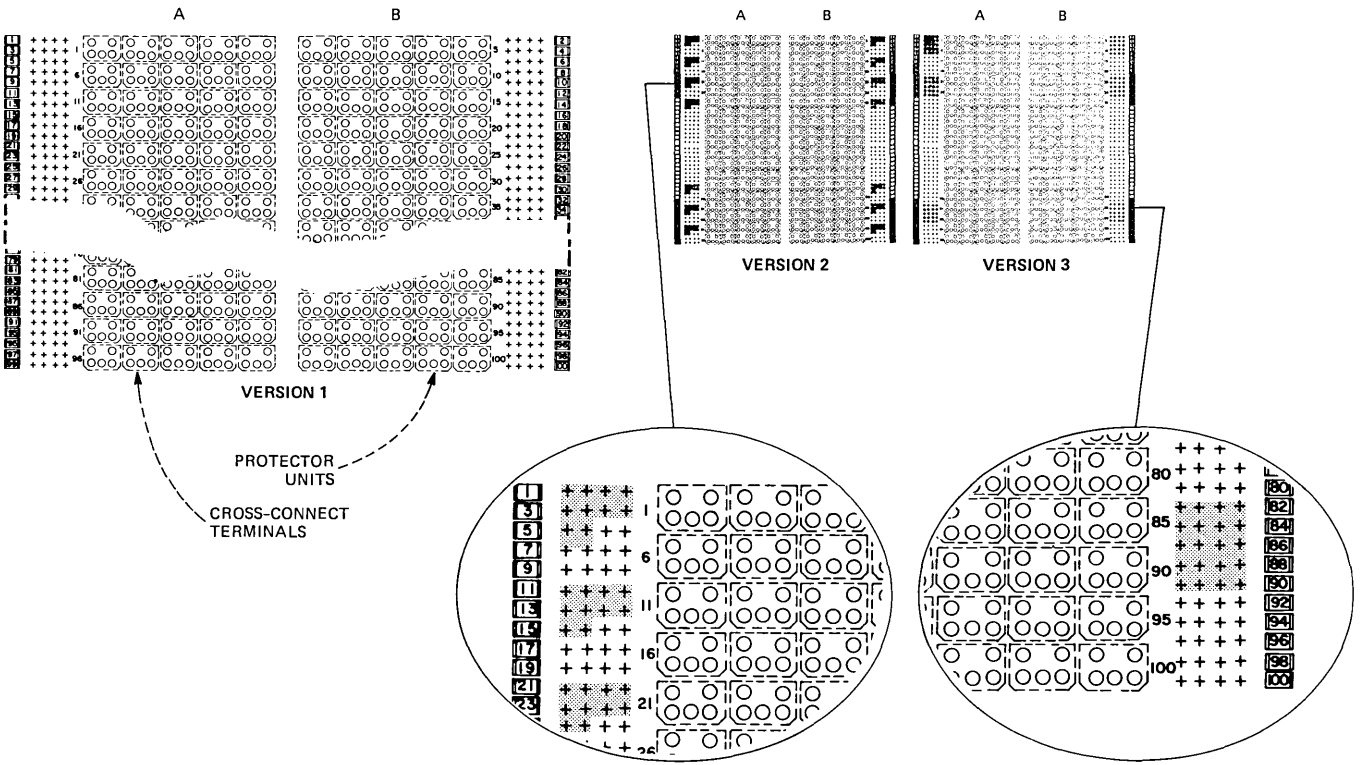


Fig. 2—Schematic illustrations of connectors (Experiment 1).

numbering scheme is not. The option of carrying the cross-connect field background patterns up to the designation strips was rejected as too costly. For purposes of these studies, the independent mirror image connectors of each version are labeled A and B.

A work assignment consisted of twenty trials of locating and marking cross-connect pairs and corresponding protectors. Cardboard trays with 20 plastic cups provided work assignment information. Each cup contained the material for one trial, a cross-connect cap and protector unit each marked with an A or B to designate which connector, and a specific number from 1 to 100 for position within the connector. Time to locate and install each set of cross-connect caps and protector units was recorded on a *Datamyte** Data Collector. If the participant did not complete a trial and press the response button by the imposed deadline, a brief beep occurred. The 34-second slow and 16-second fast deadline times were chosen from the results of a pilot study. In that study, response times measured in an unpaced task showed that 34 seconds was ample time to complete a trial and that a 16-second deadline would be inadequate about 50 percent of the time. Search time was measured as the interval between response button presses. In addition to time data, the position of each cross-connect cap and protector unit was recorded for identification of type as well as number of errors.

2.1.3 Procedure

Each participant completed a practice work assignment of ten trials followed by one work assignment on each of the three connector versions at each of the two deadlines. Order of presentation of connectors within as well as between deadlines was counterbalanced across participants, and the six work assignments were completely rotated through the six connectors every six participants.

Each participant was read instructions describing the physical arrangement of the A and B connectors and the procedure to be followed. For each trial, the procedure was to find the terminal marked on the cross-connect cap and protector unit, place the cap over the cross-connect pair, replace the corresponding protector with the labeled one, and then press the response button to denote completion of the trial. The experimenter emphasized that a clock which was reset by depressing the response button would be running; the goal was to complete each trial before the buzzer sounded.

2.2 Results

Mean time to locate cross-connect and protector unit sets was used

* Trademark of Data Acquisition Equipment, Electro/General Corporation.

in an Analysis of Variance (ANOVA) computed with subjects (18) and connectors (3) with both slow and fast deadline conditions (2) as variables.[1] No significant differences in search times were found for the three connector versions, nor were there deadline x connector interactions.

In determining errors each cross-connect cap member and protector unit was considered separately. The overall error rate was 4.5 percent. Results of the data analysis showed error rates were not affected significantly by the connector background or by the pace of the work (slow versus fast deadline). Percent and types of errors are summarized in Table I.

Three main error categories were found: parallax, left-right reversals, and counting errors. Parallax errors were responses incorrectly displaced vertically by some number of rows. If a terminal was marked on the wrong connector but in the correctly numbered position, e.g., A3 instead of B3, it was called an L-R reversal error. Counting errors were mistakes in locating terminals in any row within a correct connector that did not fit any obvious pattern.

2.3 Discussion

The addition of background patterns to cross-connect fields did not aid in locating connector terminals. Errors were useful in pinpointing areas for design modifications. Parallax, a problem associated with high-density connectors, is likely to be accentuated because cross-connect and protector fields, as well as numbers referring to them, are in different planes. With the 309-type composite connector, additional difficulties may be related to the abutment of two independent connectors. The large number of L-R reversal errors suggested the need

Table I—Errors in locating terminals

Cross-Connect and Protector-Unit Errors	Experiment 1 ($N = 18$)	Experiment 2 ($N = 12$)	
Type	Average for All Connectors	Original Connector	Modified Connector
Left-right reversals	1.4% of all responses (31% of errors)	1.88% of all responses (20.2% of errors)	1.25% of all responses (46% of errors)
Parallax	1.5% of all responses (34% of errors)	5.1% of all responses (55% of errors)	0.5% of all responses (19.2% of errors)
Counting	1.25% of all responses (27% of errors)	1.5% of all responses (15.7% of errors)	0.3% of all responses (11.5% of errors)
Other	0.7% of all responses (8% of errors)	0.8% of all responses (9% of errors)	0.6% of all responses (23.3% of errors)

for a clearer demarcation between the two connectors, and counting errors pointed to difficulties in locating terminal positions.

Based on the results of Experiment 1, the design of the 309-type connector was modified and a second experiment was conducted. A modified 309-type connector included an angled center designation strip with a vertical black line over the ridge for more distinct separation between connectors, and angled designation strips between cross-connect and protector fields for labeling rows on both sides. Angled rather than flat strips were used to minimize any change in connector dimensions and confusion between adjacent labels. To help further, cross-connect and protector fields were placed in more nearly the same plane by raising the cross-connect field and shortening the cross-connect terminal pins slightly so that the ends of the pins were in the same plane as the corresponding designation strip (see Fig. 1).

A similar experimental design was used in a second experiment to compare performance on the modified connector version and the original unpatterned connector of Experiment 1. Neither connector was patterned on the cross-connect field since patterning did not prove to be effective in Experiment 1. Only the fast deadline was used.

III. EXPERIMENT 2

3.1 Method

Twelve new Bell System participants with comparable experience to those of Experiment 1 took part in the study.

Designation strips on the modified connector were angled back 62 degrees from the vertical plane. The original connector was labeled as before. Four of the six work assignment trays and the practice assignment were used. After the practice assignment, the participant completed a 20-set assignment on each of the two connectors, followed by a break and a repeated measure on the two connectors. The order of connector presentation was counterbalanced and work assignments were rotated as before.

3.2 Results

As in Experiment 1, search time and error rate were the dependent measures. Subjects (12), connectors (2), and the repeated measure, trials (2), were the variables for the ANOVAs. Mean times to find and complete a connection on the original connector were 13.6 and 11.9 seconds for the first and second trials, respectively; these were 12.4 and 11.3 seconds on the modified connector. The modified connector version resulted in a modest but significant savings in search time ($F = 9.11$, $df = 1,11$, $p < 0.05$). The improvement in speed from the first

to second trials on both connectors led to a significant practice effect ($F = 34.5$, $df = 1,11$, $p < 0.001$).

The overall mean error rate was 9.25 percent on the original connector and 2.70 percent on the modified version; an ANOVA showed that errors were significantly lower on the modified connector ($F = 10.36$, $df = 1,11$, $p < 0.01$), and improved with practice ($F = 7.31$, $df = 1,11$, $p < 0.05$). Interactions were not significant. Of greatest interest is the marked reduction in errors on the modified 309-type connector. As shown in Table I, using the same error categories, numbers of parallax and counting errors decreased substantially. Unfortunately, L-R reversals did not decrease.

3.3 Discussion

From a human factors standpoint, the lack of a one-to-one correspondence between cross-connect and protector fields and the abutment of two connectors in mirror image are not ideal. On the other hand, there is a real and immediate need for additional space on the MDF.

It is not clear whether the L-R reversal errors were due to difficulties associated with the left-right nature of the task, the mirror image arrangement, or even the proximity of the composite connectors. Although the two sides of each composite connector were called A and B, in all likelihood they were recognized in terms of their relative positions as left and right. People have difficulty discriminating left from right.[2] Despite the random assignment of work order pairs to the A and B connectors, 69 percent of the reversal errors in Experiment 1 and 73 percent in Experiment 2 were preceded by a correct response on the same side, as if participants were not paying attention to which of the connectors they had just worked on, and so continued to work on the same connector.

The terminal location task was chosen as it was assumed it would be sensitive to the spatial arrangement of the connectors. Time-consuming wiring operations were not included in the task as they should be constant. On the main frame, however, these and many other operations would interrupt the left-to-right scan. The search would also necessarily extend over a much wider area since a main frame consists of large arrays of connectors. The identification of main frame connectors at the top of vertical frame modules should also result in vertical top-to-bottom search. Thus, the predominant left-to-right orientation of the task may have been somewhat artificial. It is likely that L-R reversal errors would be reduced on the main frame where the inclusion of other operations should make the craftperson's job more sensitive to a connector's specific position.

While eliminating the one-to-one correspondence between cross-

connect and protector fields results in higher density terminations on the MDF, craftspeople, like the experimental participants, may locate associated elements on the two fields independently. The potentially small difference in time should not affect productivity. In reality, a work assignment would not necessarily require work on both cross-connect and associated protector units. More important for the craftsperson's job may be the advantage of the front-facing work surface. There is some assurance that the craftsperson will keep his/her gaze nearly perpendicular to the angled designation strips on the modified 309-type connector, since numeral positions on the sides of the wedge-shaped strips preclude reading both sides from the same point. Besides providing clear labels and end points for rows, then, the angled designation strips may lead to more concentration that results in greater accuracy in terminal location.

In summary, the number and consistency in errors on the initially proposed 309-type connector were unacceptably high. Performance on a modified version was improved with significant reductions in parallax and counting errors. The resultant 309-type connector represents the integration of design and human factors efforts.

REFERENCES

1. B. J. Winer, *Statistical Principles in Experimental Design*, New York: McGraw-Hill, 1971.
2. W. S. Farrell, Jr., "Coding Left and Right," J. Experimental Psychology: Human Perception and Performance, *5*, No. 1 (February 1979), pp. 42–51.

AUTHOR

Lois E. Flamm, B.A., Skidmore College, 1966; M.A., 1968, Ph.D. (Experimental Psychology) 1971, Northeastern University; Bell Laboratories, 1977—. Ms. Flamm taught at Texas A & M University from 1972–1976. In 1977 she joined the Human Factors Engineering Group, Loop Transmission Division, at Bell Laboratories in Whippany, where she worked on a variety of interface problems associated with the development and evaluation of new apparatus and methods for outside plant craft. In 1982, she was an intern in the Distributed Computer Systems Research Department, Murray Hill. She is presently in the *UNIX*™ Systems Engineering Group in the *UNIX* Development Laboratory, Murray Hill.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Membrane Keyboards and Human Performance

By K. M. COHEN LOEB*

(Manuscript received January 7, 1982)

This paper describes systematic human factors research in which typing performance using a membrane keyboard and using a conventional, full-travel keyboard were compared for subjects representing different levels of typing proficiency. Membrane switch technology has become increasingly popular in many consumer-oriented products because of its low production cost and design flexibility. However, the absence of familiar key travel associated with membrane switches removes an important, direct source of feedback to the user with respect to specific keystrokes. Hence, the conventional wisdom has been that membrane switches without key travel are unacceptable for such keyboard applications as typing tasks. The results of the research discussed here indicate that for nontouch typists there was little difference in performance between keyboards. For touch typists, performance with the conventional keyboard was initially much better than with the membrane keyboard. Rapid learning resulted in improvement in typing performance with the membrane keyboard—both within an experimental session and across sessions—such that the advantage of the conventional keyboard over the membrane one for touch typists was reduced substantially, although not completely. Future work will be aimed at measurement of the additional improvement in performance that may result from extended practice with better-designed membrane keyboards.

*American Bell.

I. INTRODUCTION

Membrane switch technology has become increasingly popular in consumer-oriented products such as pocket calculators and kitchen appliances. Application of this technology to the design of full-size, alphanumeric keyboards has resulted in products that differ markedly from conventional keyboards in terms of keystroke feedback to the user. Membrane switches usually consist of mechanical contacts on two layers of material, separated by a nonconductive third layer of material (see Fig. 1). The upper, membrane layer is usually a thin polyester material with flexible conductors applied to its underside. Graphics may be silk-screened on this surface or on a second surface placed on top of the membrane layer. The substrate may be either a printed circuit board with conductors or a flexible film with printed conductors. This is usually mounted on a rigid, smooth surface. The spacer layer is an insulating material that separates the membrane and substrate layers by 5 to 7 mils. It has holes through which the upper, flexible layer may be depressed, causing contact closure. When pressure on the membrane layer is removed, the resilient, flexible membrane breaks contact with the substrate and returns to its original position.

Membrane switch technology has become popular for several reasons. Its production costs are low because there are no key-plunger mechanisms. It also affords considerable design flexibility in terms of panel layout, "key" size and shape, and graphic labels. Good switch enclosure can also be assured for hostile environments and for protection against dust accumulation, spills, and vermin infestation. Easy cleaning of the upper surface is also an advantage associated with membrane switch technology.

Although there are several technical advantages of membrane switches, there are potential user-related problems when this technology is applied to full-size keyboard design. First, familiar key travel is absent as a source of keystroke feedback to the user. Second, high actuation forces are often used with membrane switches to prevent accidental switch closure that may occur from pressure exerted by hands or fingers in a resting position on the contact area of the switch. Finally, the actual contact area ("sweet spot") of the switch often is indiscriminable for the user, compared with traditional mechanical switches that use key caps on top of key-plunger mechanisms.

Despite these disadvantages, membrane switch technology is currently being introduced into the design of full-size keyboards because of its considerable cost advantages. There is little question that membrane switches may be used successfully for simple on-off function keys. There is, however, considerable doubt about the use of this technology for keyboard tasks such as typing because of the assumed

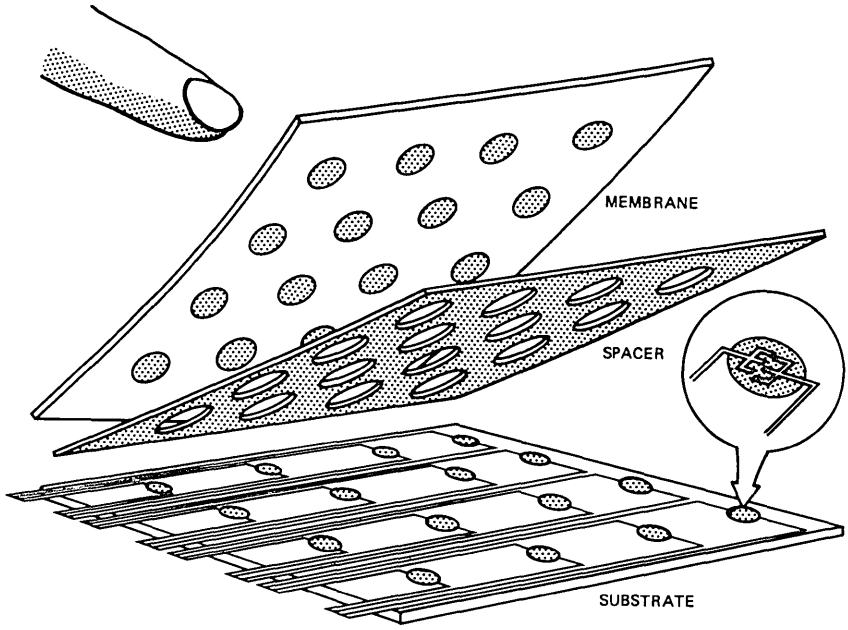


Fig. 1—Membrane switch technology.

importance to the user of key-travel feedback associated with individual keystrokes. Despite an absence of empirical evidence to support this claim, the conventional wisdom has been that membrane keyboards cannot be used as high-speed data- or text-entry devices.

Systematic human factors research was thus undertaken to determine whether performance using membrane keyboards differs from that using a conventional, full-travel keyboard. It was expected that large performance differences between keyboards would be present initially, but that with practice these performance differences might be reduced. Thus, this research was designed to provide subjects with several experimental sessions using each keyboard. It was also hypothesized that typing skill would be directly related to initial performance differences between keyboards and, perhaps, to the speed of learning to use the membrane keyboard. Specifically, it was expected that better touch typists would initially show greater performance differences between the membrane and conventional keyboards than less skilled touch typists because of the greater familiarity of the better touch typists with conventional keyboards. However, it was also possible that skilled touch typists would demonstrate greater transfer of typing skills to a novel keyboard and, hence, would learn to use the membrane keyboard more rapidly than less skilled touch typists.

Unskilled, nontouch typists, who tend to look at the keys as they are struck, were expected to show little difference in performance between membrane and conventional keyboards. A within-subject design with multiple sessions was used so that each subject would serve as his/her own control. This design also facilitated direct comparisons of typing performance on the membrane and conventional keyboards over time.

II. METHOD

2.1 Subjects

Twenty-one subjects (17 females and 4 males) participated in this study. Subjects were either employees of Bell Laboratories or volunteers from the local community. Volunteers were paid nominally for their participation in the study. Subjects varied in their typing ability and their average typing time per week.

2.2 Facilities and equipment

Two commercially available keyboards were used in this study. The presence or absence of familiar key travel associated with each keystroke was the most distinguishable difference between these keyboards. One keyboard was considered a conventional keyboard in terms of key travel (about 100 mils to contact and an additional 50 mils to bottoming out), average actuation force (68 grams or 2.4 oz.), keytop size (450 mils diameter), and interkey spacing (750 mils, center-to-center). The other keyboard was a flat membrane keyboard with virtually no key travel. The average actuation force was 82 grams or 2.89 oz. A 1-mil-thick grid overlay delineated individual "keys." The "keytop" diameter size was about 600 mils with interkey spacing of 750 mils, center-to-center. Both keyboards had a QWERTY layout of alphanumeric keys and also a DELETE key that permitted subjects to make corrections. Audio feedback associated with each keystroke was achieved upon "key" depression with the key plunger mechanism for the conventional keyboard (an audible "click") and with an internal audio tone generator for the membrane keyboard.

Hardware and software interfaces were developed to permit on-line experimentation and data collection using these keyboards. Both keyboards were interfaced with a MINC-11/03 (DEC) minicomputer via an RS-232 port. The output from the keyboard was displayed simultaneously on a 9-inch video monitor placed directly in front of the subject behind the keyboard and on a VT-105 (DEC) display in front of the experimenter (see Fig. 2).

The study was conducted on-line with the MINC computer displaying appropriate prompts (e.g., "type text") to the subject prior to each of the experimental tasks. Three types of data were collected during each task: the actual characters displayed on the monitor, the actual

Fig. 2—Experimental setup.

characters typed by the subject (including deletions and corrections), and each interkeystroke time interval.

2.3 Experimental tasks

Each subject participated in six sessions, three with the membrane keyboard and three with the conventional keyboard. Three types of tasks were used twice during each session: dialing telephone numbers, typing text, and answering questions. The typing-text and dialing-numbers tasks were viewed as experimental tasks, while the answering-questions tasks were viewed as practice tasks. Each experimental task was considered a "trial" so that there were two trials for each type of experimental task per session. In addition, there were two sets of questions for the practice tasks per session. The stimulus materials for each of these tasks were placed on a typing stand to the right of the keyboard and monitor.

Each "dialing" task consisted of 30 seven-digit phone numbers that were to be dialed from the keyboard. Two different lists of numbers were used for each of three sessions with a given keyboard, for a total of six different lists. The two lists used for each of the first, second, and third sessions with the conventional keyboard were identical to those used for the respective sessions with the membrane keyboard. Hence, direct comparison of performance using the two keyboards for a particular session was possible.

Each "typing text" task consisted of a paragraph from a standard

typing textbook that is used to train and assess typing speed.[1] The paragraphs were of comparable typing difficulty, based upon syllabic intensity ratings. These ratings reflect the average number of syllables per word and are traditionally used to assess the typing difficulty of textual material.[1] As with the phone number list, two different paragraphs were used for each session with the given keyboard, for a total of six different paragraphs. The two paragraphs for each of the first, second, and third sessions for the conventional keyboard were identical to those used for the respective sessions with the membrane keyboard. Hence, direct comparison of performance between keyboards for a particular session was possible. These tasks were considered the most critical ones with respect to the major objectives of this study.

Each answering-questions task consisted of a few questions on a general topic (e.g., favorite sport and how it is played, most recent vacation activities). Subjects answered these questions using the conventional or the membrane keyboard. Two different sets of questions were used during each session, for a total of 12 different sets. This task was primarily designed to provide a structured exercise by which the subject would gain experience using each keyboard.

2.4 Procedure

At the beginning of each session, a standard 3-minute typing test was given to each subject. This test was conducted using an IBM Selectric typewriter. Following this test, each subject completed six tasks on either the conventional or the membrane keyboard. The tasks were administered in the following order: dialing numbers, typing text, answering questions, answering questions, typing text, dialing numbers. This order of tasks allowed the subject about one-half to three-quarters of an hour of practice on each keyboard (e.g., the answering-questions trials) in between the two dialing-numbers and typing-text trials. This order of tasks, then, maximized observation of practice effects over trials within a session but also minimized warm-up and fatigue effects on the critical typing-text trials. There were also four general questions that were answered as a warm-up exercise at the beginning of the first session with each type of keyboard.

Each subject completed three sessions with one keyboard and then three sessions with the other keyboard. Sessions took place on consecutive (or nearly consecutive) days, one session per day. At the conclusion of each session, subjects completed a questionnaire that probed their rating of various features of each keyboard. At the conclusion of the third session with each keyboard, another typing test with an IBM Selectric typewriter was administered to assess changes in typing performance over the course of a session. Though these changes in performance on the IBM Selectric are not reported here, they helped

determine the typing skill classification for each subject. Biographical data regarding touch typing ability, weekly amount of time typing, and prior use of membrane switches were also obtained.

2.5 Typing groups

Subjects were categorized according to their typing ability based upon two criteria. First, assessment was made of their use of touch typing techniques in which all ten fingers and "home row" positioning of the fingers are used. Second, their average gross words per minute (WPM) on the eight typing tests given on an IBM selectric typewriter during the six experimental sessions was computed (i.e., the typing tests preceding each session and the test following the third session with each keyboard). These post-hoc criteria resulted in the following five categories of typing proficiency (see Table I): excellent touch typists (more than 60 WPM), good touch typists (50–60 WPM), fair touch typists (40–49 WPM), poor touch typists (26–39 WPM), and non-touch typists (16–25 WPM). There were 3, 5, 5, 4, and 4 subjects in these respective typing groups, for a total of 21 subjects.

The order in which keyboards were used by subjects was counterbalanced within each touch typing group through random assignment of subjects to one of the two possible keyboard orders. This balancing scheme for keyboard order was limited by the fact that the post-hoc categorization of subjects into typing groups resulted in an unequal number of subjects per group. Hence, to the extent possible, half the subjects in each touch typing group experienced the membrane keyboard first (total = 9 subjects) while the remainder of each touch typing group used the conventional keyboard first (total = 8 subjects). Also, by chance, all of the subjects who were categorized post hoc as nontouch typists used the membrane keyboard first. Hence, for the nontouch typing group, there was no variation across subjects in keyboard order.

2.6 Experimental design

Typing Group (5 levels) was the major between-subject factor. Order of keyboards (2 levels) was considered a minor between-subject factor. Analyses using this factor across the touch typing groups were not

Table I—Number of subjects per typing group

Touch Typing Group	Gross Words Per Minute	Total Number	Membrane First	Conventional First
Excellent	>60	3	2	1
Good	50–60	5	3	2
Fair	40–49	5	2	3
Poor	26–39	4	2	2
Non	16–25	4	4	0

reported since the number of subjects per unit of analysis was extremely small, making interpretation of results complex. In addition, there was no variability across the order variable for the nontouch typists, and the basis for estimating the variance attributable to keyboard order for this group was inadequate. Within-subject factors included keyboards (2 levels), sessions (3 levels), and within-session trials (2 levels). Separate analyses were conducted for the different types of experimental tasks (dialing and typing text). The data from the answering-questions trials were not analyzed.

III. RESULTS

Several dependent measures were computed for each subject for the dialing and typing tasks. These measures were based upon the speed and/or accuracy of performance using both the conventional and membrane keyboards. Of particular importance to this study were the performance differences between the two keyboards for the various typing groups. Hence, difference scores and percent difference measures were also computed for each subject, and mixed-factor analyses of variance (ANOVAs) were performed. In addition, for the typing tasks the speed and accuracy data were combined to compute each subject's average words per minute, a throughput measure of performance.

3.1 Dialing results

For each dialing task, the number of phone numbers dialed incorrectly and not subsequently corrected was computed. This measure included both misdialed digits as well as insufficient or additional digits per phone number. However, if a subject corrected a dialing error through use of the DELETE key, the resultant phone number was considered correct. There were 30 phone numbers per dialing task. On a few trials, a subject either omitted a phone number or redialed a phone number, a mistake no doubt attributable to the experimental set-up. These omitted or redialed phone numbers were not scored as dialing errors. Hence, the percent error rate for dialing phone numbers for each subject was computed from these data, and an ANOVA was performed.

As shown in Table II, the error rate for the dialing tasks was low for all groups using either the conventional or the membrane keyboard (less than 3 percent). The ANOVA indicated no significant difference in dialing accuracy as a function of keyboard or typing group ($p = 0.259$ and $p = 0.720$, respectively). There were also no significant effects when difference scores between keyboards were computed for each subject and served as the unit of analysis.

Table II—Percent error rate for dialing tasks*

Keyboard	Touch Typing Group				
	Excellent	Good	Fair	Poor	Non
Membrane	1.11	2.66	2.01	2.50	1.30
Conventional	2.61	2.85	1.88	2.78	1.52
Mean	1.86	2.98	1.95	2.64	1.24

*Error rate is for uncorrected errors.
Each dialing task has 30 phone numbers.

3.2 Dialing summary

Accuracy of dialing performance did not vary as a function of keyboard used or typing proficiency level. Given the simplicity of the task, relative to typing text, this result is not surprising. The data do, however, provide an anchor point for comparison of performance between the two keyboards as task complexity increases.

3.3 Typing results

For each typing task, the number of "words in error" that remained after the subjects' corrections was computed according to traditional methods of scoring typing tests.[2] Hence, an incorrect letter(s), a missing letter, a letter reversal, a punctuation error, and/or a spacing error were all scored as a "word in error." The six passages, while equated in difficulty based upon the traditional syllabic intensity level, varied somewhat in the total number of words per passage (number of words per passage, where 5 character spaces equals 1 word, were: 64, 57, 79, 77, 73, 84).

3.3.1 Typing accuracy

Typing accuracy was measured in terms of both uncorrected errors and corrections. As shown in Table III, the total number of words in error was quite small (range: 1.8 to 4.5 words). The ANOVA did not indicate a reliable difference for this measure among typing groups or between keyboards. There was, however, a significant trial effect [$F(1,16) = 20.01$, $p < 0.0005$], with errors declining as a function of practice (4.2 vs. 2.7 errors).

Unlike standard typing tests, subjects were permitted to correct errors during the typing tests through use of the DELETE key. An ANOVA was performed on the number of corrections made by subjects during each typing task. The results indicated significant effects of keyboard [$F(1,16) = 5.27$, $p = 0.035$] and of session [$F(2,32) = 3.63$, $p = 0.038$]. In general, more corrections were made using the membrane keyboard than using the conventional keyboard (4.61 vs. 3.14), as shown in Table IV. Also, more corrections were made in the second

Table III—Average words typed in error

Keyboard	Touch Typing Group				
	Excellent	Good	Fair	Poor	Non
Membrane	1.99	1.83	3.77	3.79	4.54
Conventional	2.56	4.13	3.90	3.13	3.79
Mean	2.28	2.98	3.83	3.63	4.17

Table IV—Average number of corrections for typing tasks

Keyboard	Touch Typing Group				
	Excellent	Good	Fair	Poor	Non
Membrane	3.22	4.57	6.10	6.13	2.38
Conventional	4.00	2.47	3.83	2.96	2.67
Mean	3.61	3.52	4.97	4.54	2.52

and third sessions than in the first session, even though there was no reliable session difference in typing accuracy (i.e., uncorrected words in error). This session effect for number of corrections may be related to passage length; the average passage length was shorter for the first session (61 words) than for later sessions (78 words), and the probability of making a typing error increases as passage length increases. There was also a significant interaction between keyboard and trial factors [$F(1,16) = 5.48$, $p = 0.033$]. This interaction reflected a more dramatic practice effect (i.e., decline in errors corrected) for the membrane keyboard than for the conventional one.

3.3.2 Typing speed

The total typing time for each trial was also computed and an ANOVA was performed. The results indicated, as expected, a significant effect of typing group [$F(4,16) = 6.94$, $p = 0.002$]. As expected, typing speed varied directly with typing proficiency level, as shown in Table V. There was also a significant keyboard effect [$F(1,16) = 64.7$, $p < 0.005$]. Typing performance was somewhat faster with the conventional keyboard than with the membrane keyboard (130 seconds vs. 171 seconds).

Though there were other reliable effects related to the trial and session factors on speed of typing, it should be noted that, as with the typing accuracy data, variation in passage length (57 to 84 words) for the six test passages could readily have contributed to differences in typing time per trial and/or session.

3.3.3 Differences between keyboards in typing speed

Because of the variation in passage length, ANOVAs were performed on the difference in typing time between the two keyboards for

Table V—Average time for typing tasks (in seconds)

	Touch Typing Group				
Keyboard	Excellent	Good	Fair	Poor	Non
Membrane	134.1	143.5	166.8	193.6	213.1
Conventional	97.0	101.4	106.7	149.0	169.9
Mean	115.5	122.5	136.7	171.3	206.2

corresponding trials. Each difference score was thus based upon the same passage typed on each keyboard.

The ANOVA results indicated significant effects of both the session [$F(2,32) = 3.68$, $p = 0.036$] and the trial [$F(1,16) = 27.0$, $p < 0.0005$] factors. These findings reflect practice effects both across and within sessions. There was also a significant interaction between trial and session factors [$F(2,32) = 3.5$, $p = 0.042$], as shown in Table VI. This interaction was attributable to larger practice effects between trials during the first two sessions, compared with those during the last session. This practice or learning effect within a session is based primarily upon greater improvement in typing speed using the membrane keyboard, compared with that using the conventional keyboard.

The main effect of typing group for this measure approached significance ($p = 0.084$). The largest speed difference between the keyboards occurred for the fair touch typists and the least difference occurred, as expected, for the nontouch typists. Post-hoc analyses (Duncan's Multiple Range Test) indicated that the difference in typing speed between the two keyboards for the fair touch typists was reliably greater than that for all groups other than the poor touch typists ($p < 0.05$), and the difference score for the nontouch typists was reliably less than that of all the touch typing groups ($p < 0.05$).

3.3.4 Words per minute

Perhaps the most relevant measure for the assessment of typing performance is the calculation of each subject's typing ability in terms of words per minute (WPM). This measure reflects a throughput measure of performance that takes into account both speed and accuracy:

$$\frac{\text{number of words typed} - \text{number of words in error}}{\text{total time in minutes}}$$

It was obviously expected that WPM for the typing tasks would vary directly with the typing group categories since this measure for a similar typing task was the basis for classifying individuals into typing groups. It was also likely that the experimental set-up per se might result in lower WPM scores than those calculated from the standard

Table VI—Average typing time difference between membrane and conventional keyboards (in seconds)

	Touch Typing Group				
	Excellent	Good	Fair	Poor	Non
Session 1					
Trial 1	48.5	61.5	89.2	74.6	25.1
Trial 2	24.2	27.0	52.8	40.1	11.6
Session 2					
Trial 1	59.0	67.2	62.9	52.5	32.3
Trial 2	30.4	45.9	50.8	34.0	7.9
Session 3					
Trial 1	39.0	20.6	56.2	42.4	6.1
Trial 2	21.7	32.8	48.7	23.6	16.5
Mean	37.1	42.5	60.1	44.5	14.0

typing tests given to each subject since many touch typists were unfamiliar with an electronic keyboard attached to a visual display.

The ANOVA on WPM scores confirmed these predictions. The main effects of all four experimental factors were significant: typing group [$F(4,16) = 9.45$, $p < 0.0005$]; keyboard [$F(1,6) = 58.7$, $p < 0.0005$]; session [$F(2,32) = 24.6$, $p < 0.0005$]; and trial [$F(1,16) = 23.0$, $p < 0.0005$]. Practice improved performance both across sessions and between trials within a session. The poor and nontouch typing groups differed reliably from the other groups, as shown in Table VII. Performance was also significantly better for the conventional than the membrane keyboard (35.4 vs. 25.9 WPM). However, a reliable interaction between keyboard and typing group [$F(4,16) = 3.74$, $p = 0.025$] indicated that the keyboard effect was attributable to performance differences for the excellent, good, and fair touch typing groups. Post-hoc analyses showed that performance differences between keyboards were not reliably different ($p > 0.05$) for either the poor touch typists or the nontouch typists.

The trial factor interacted significantly with the keyboard factor [$F(1,16) = 1.7$, $p = 0.003$] and with the session factor [$F(2,32) = 4.31$, $p = 0.022$]. The improvement with practice across trials within a session was somewhat greater with the membrane keyboard than with the conventional keyboard. Also, the degree of improvement between trials within a session diminished as practice (i.e., number of sessions) increased, with the largest difference between trials occurring, as expected, during the first session.

3.3.5 Differences between keyboards in words per minute

The relative, rather than the absolute, performance of each subject using the two keyboards was of primary concern in this study; hence,

Table VII—Average words per minute for typing tasks

Keyboard	Touch Typing Group				
	Excellent	Good	Fair	Poor	Non
Membrane	32.1	30.4	25.4	22.1	20.2
Conventional	43.4	43.0	39.0	28.9	21.6
Mean	37.7	36.7	32.2	25.5	20.9

the percent difference in WPM using the two keyboards was computed to evaluate the relative advantage of the conventional keyboard over the membrane keyboard:

$$\frac{\text{WPM on conventional} - \text{WPM on membrane}}{\text{WPM on conventional}} \times 100.$$

This measure was computed for each subject for each typing task such that each difference score was based upon the same paragraph of text using each keyboard.

The ANOVA based upon this measure indicated significant main effects of the session [$F(2,32) = 3.57$, $p = 0.04$] and the trial [$F(1,16) = 17.6$, $p = 0.001$] factors. Both the session and trial effects reflected a decrease with practice in the percent difference in WPM between keyboards. The third session differed reliably from the first and second sessions (19.0% vs. 26.6% and 24.3%, respectively). The performance difference between keyboards was also reduced from the first to the second trial within a session (26.5% to 20.1%). The main effect of typing group approached significance [$F(4,16) = 2.93$, $p = 0.054$]. The conventional keyboard had relatively little advantage over the membrane keyboard for the nontouch typing group (6.5%), while this advantage was much more pronounced for the other touch typing groups (excellent = 26.0%, good = 25.5%, fair = 34.2%, and poor = 21.7%). The pattern of results for this measure for each group is depicted in Fig. 3. The difference between the touch typing groups and the nontouch typing group as well as the effect of practice are readily apparent. For the nontouch typists there is virtually no advantage of the conventional keyboard over the membrane keyboard; for one trial, subjects even performed slightly better with the membrane keyboard than with the conventional keyboard. For the touch typing groups the performance with the conventional keyboard was initially much better than that with the membrane keyboard. However, the percent advantage of the conventional keyboard over the membrane one was reduced substantially, although not completely, as a function of limited practice.

It is noteworthy that, despite this rapid learning function, the excellent touch typists appear to "forget" during the interval between sessions what they learned during a session. This phenomenon is

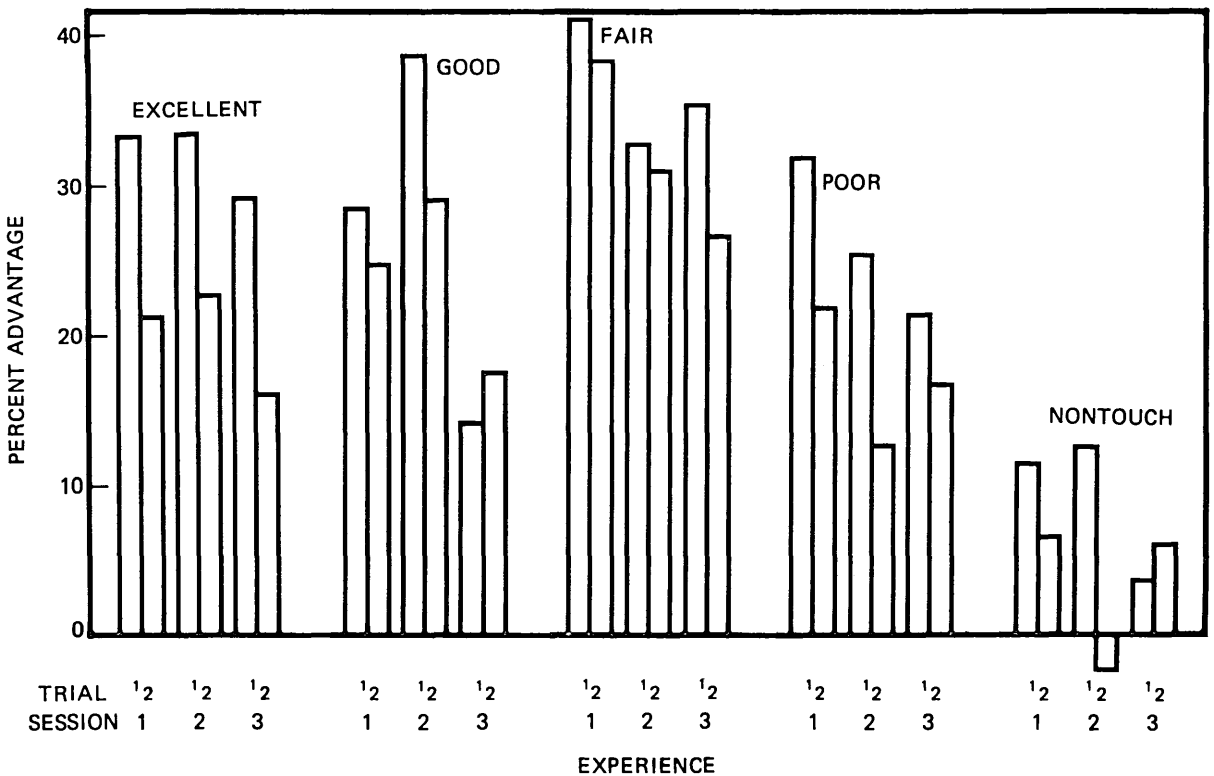


Fig. 3—Average words per minute advantage of conventional keyboard over membrane keyboard as a function of experience and touch typing group.

depicted by the larger advantage of the conventional keyboard over the membrane keyboard for the first trial of a new session, compared with that for the last trial of the prior session. This result, however, is less surprising when biographical data from these subjects are examined; these data indicate that the excellent touch typists spend about 28 hours per week using a conventional keyboard. These subjects were returning between sessions to a keyboard similar to the conventional one used here, and this activity between sessions may have produced retroactive interference, resulting in their forgetting what they had previously learned with respect to the membrane keyboard. Hence, the advantage of the conventional keyboard over the membrane keyboard was most pronounced at the beginning of each new session.

3.4 Typing summary

For the typing tasks, there was evidence of rapid learning that resulted in improved performance on the membrane keyboard, relative to the conventional one, both within and across sessions. Practice effects were found on measures of both speed and accuracy. Keyboard effects were also evident. Performance was faster on the conventional keyboard than on the membrane one. Differences in speed between the two keyboards were most pronounced for the fair touch typists and virtually nonexistent for the nontouch typists. Also, more corrections were made with the membrane keyboard than the conventional one, especially at the beginning of a session.

When speed and accuracy data were combined to produce the WPM measure, practice effects were still evident. Keyboard effects were also evident, but primarily for the excellent, good, and fair touch typists. When the relative difference between keyboards in throughput performance was examined (i.e., percent advantage in terms of WPM), keyboard effects were apparent for the touch typing groups. Performance with the conventional keyboard was initially much better than it was with the membrane keyboard for touch typists. There was, however, rapid learning to use the membrane keyboard such that the advantage of the conventional keyboard, relative to the membrane one, was reduced substantially, although not completely. For nontouch typists there was virtually no advantage of the conventional keyboard over the membrane one. Finally, "forgetting" between sessions for excellent touch typists appears attributable to their customary use of another conventional keyboard during the interval between sessions.

IV. SUMMARY

The purpose of this study was to compare performance using a membrane keyboard with that using a conventional keyboard for

different levels of typing proficiency. In general, the results indicate little difference in performance between keyboards for nontouch typists. For touch typists the results demonstrate better performance on conventional keyboards than on membrane ones. However, the results also demonstrate rapid learning to use the membrane keyboard. With just three hours of experience, performance differences between keyboards for touch typists reflected, at worst, a 27-percent advantage of the conventional keyboard over the membrane one and, at best, a 16-percent advantage.

These results provide human factors evidence that the difference in performance using membrane and conventional keyboards is much smaller than might be expected, given the absence of key travel feedback associated with each keystroke. Moreover, touch typists who are familiar with conventional, "full-travel" keyboards quickly learn to use membrane keyboards. For most touch typists, there remains an advantage for the conventional keyboard over the membrane keyboard after a limited exposure (only about three hours) to the novel keyboard.

These results are encouraging in terms of the application of membrane switch technology for some keyboard tasks. First, the particular membrane keyboard tested in this study does not necessarily represent an optimal membrane keyboard. Features of this particular keyboard that were identified as bothersome to subjects will be used in guiding the design of membrane keyboards for future evaluation. Second, extended practice with membrane keyboards may diminish residual differences in performance between conventional and membrane keyboards. Given the cost advantages and design flexibility afforded by membrane switch technology, such avenues may be well worth pursuing in the design of new products.

V. ACKNOWLEDGMENTS

The contribution of the following people to this research project is greatly appreciated: R. Rosebrock for providing the experimental keyboards; P. Vittorio for measuring actuation forces; D. Kelly for developing the hardware and software interfaces; D. Lewis for testing subjects and reducing data; A. Glance, J. Papay, and P. Keany for reducing data; and M. Katz for making helpful comments and suggestions.

REFERENCES

1. D. D. Lessenberry, S. J. Wanous, and C. H. Duncan, *College Typewriting*, Chicago: South-Western Publishing, 1965.
2. J. L. Rowe, A. C. Lloyd, F. E. Winger, *Typing 300, Volume One: General Course*, New York: McGraw-Hill, 1972.

AUTHOR

Karen M. Cohen Loeb, B.A., 1970 (Psychology and Sociology), Washington University, St. Louis; M.A., 1972, and Ph.D., 1978 (Experimental Psychology and Education), Harvard University; Bell Laboratories, 1979—. In the period 1972 to 1979, Ms. Cohen was a research associate and faculty member in psychology at the University of Denver and a research psychologist at the Denver Research Institute. Ms. Cohen's field of specialization is human visual information processing and cognition. At Bell Laboratories, she has done human factors research in support of product development, market research, systems engineering, and software development. Projects have included video-teleconferencing, terminal design, videotex trials, voice messaging systems, and business communication systems. Member, Human Factors Society.

Interface Design

Recent behavioral science work at Bell Laboratories and American Bell has focused on a portion of the user-machine interface, best characterized as dialogue design. This work has included the design of dialogues used for human-computer interactions, such as command languages, menu systems, and text editors. All of these are used in systems provided to customers and in the complex operations support systems used by telephone company employees. Dialogue design has also included new methods for interacting with advanced telephone services, such as call forwarding and teleconferencing.

The papers in this section focus on interface design. At best, from a human factors viewpoint, right from the start behavioral scientists were able to influence system design so that the most desirable interface was provided. In other cases, the functioning of machines was already determined, so that the behavioral scientist's job was to provide the best training or instructions to help users with their side of the dialogue.

The papers by Furnas, Landauer, Gomez, and Dumais, and by Streeter, Ackroff, and Taylor, are examples of research that guide the design of user-computer dialogue. The paper by Karhan, Riley, and Schoeffler describes the design of printed instructions to assist users of public telephones. The dialogue between people and telecommunications equipment often includes such interactions with printed material. The paper by Coke and Koether examines the readability of frequently used Bell System documents and the reading abilities of those who use them. Finally, the paper by Holmgren discusses the design of dialogues for users of automatic speech recognition devices.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July-August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Statistical Semantics: Analysis of the Potential Performance of Key-Word Information Systems

By G. W. FURNAS,* T. K. LANDAUER,* L. M. GOMEZ,* and S. T. DUMAIS*

(Manuscript received February 10, 1982)

This paper examines how imprecision in the way humans name things might limit how well a computer can guess to what they are referring. People were asked to name things in a variety of domains: instructions for text-editing operations, index words for cooking recipes, categories for "want ads," and descriptions of common objects. We found that random pairs of people used the same word for an object only 10 to 20 percent of the time. But we also found that hit rates could be increased threefold by using norms on naming to pick optimal names, by recognizing as many of the users' various words as possible, and by allowing the user and the system several guesses in trying to hit upon the desired target.

I. INTRODUCTION

Computer-based information management systems can store, manipulate, and transmit enormous quantities of information. They can allow almost unlimited organization, multiple indexing and cross referencing, and are capable of performing rapid and complex search operations. Thus, they can provide far more powerful tools for knowl-

* Bell Laboratories.

edge management than have been available previously. Such tools will be important to almost everybody: cooks wanting to find recipes, doctors needing patients' histories, managers tracking inventories, clerks filling orders and keeping records, and buyers looking for products. It would be fine if such systems could be used directly by inexperienced and occasional users, people whose main job or talents lie elsewhere. The hope is that new information-seeking tools could be operated with no more training than is needed to use a card-file or a book index, but with much greater success.

We believe that some of the greatest difficulties blocking this dream are psychological. Certainly, more available speed and capacity, and new algorithmic and data structure developments involving deep and difficult problems will be needed before the arrival of fully satisfying information systems. But, progress along these lines has already been enormous and continues at a brisk pace. Meanwhile, although even existing computational capabilities are very great, our ability to make them easily usable by nonspecialists has been quite limited. An important psychological problem is in understanding the relationship between what people say and what they want. This understanding is the key to designing systems that can infer what services or information users need from the input they provide. This basic capability is needed by information systems of many sorts, bibliographic reference search systems, business management databases, airline reservation systems, plant inventory and customer record systems, and even text processors. We believe that current systems generally fall very far short of the ideal of always knowing just what to give the user.

At this time although it is clear that there is a problem, very little relevant work has been done (see Carroll[1] and Landauer, Galotti, and Hartwell[2]). We do not even know the locus of the problem in the chain of actors and events. For example, the main problems may lie at the source; perhaps the human intellect is basically incapable of forming information specifications that are very precise. If so, perhaps no system could do much better than the current ones. But, it is also possible to believe that if systems could understand people as well as, say, close friends understand each other, there would be much less of a problem.

We have tried to get some idea of the extent of this problem by attacking a small but important part of it. We have asked people to give descriptions of various information objects, and analyzed their responses to determine how well the objects to which they refer can be inferred from what they say. We have begun by studying the referential properties of isolated words and short phrases.

Our goals in this research have been twofold; first, to advance understanding of the psychological processes by which human seman-

tic reference is generated, and second, to model and estimate the strengths and weaknesses of information systems that take human-generated descriptions of sought items as their input. Empirical observations of naming behavior provide the necessary data for both enterprises.

In this article we first describe four sets of object-description data, the way they were collected and reduced, and some of their more interesting features. Then we present a series of analyses in which we treat the collected descriptions both as representative of what users would provide as input to information retrieval systems, and as the source of the information that the system would use in determining user wants. The hypothetical systems we consider are limited to ones in which the user's initial entry or query consists of a single word or short phrase. We do not attempt to analyze the possible performance of systems that make use of sentential syntax, or linguistic or real world context. The reason for this limitation is largely pragmatic; it postpones analysis of many difficult complexities. However, the characteristics and limitations of single-word to object reference that we have investigated have strong implications for many access methods (by which we mean the access method provided for the user, not that used by the program). We are, of course, aware that there are data access methods that do not start with the user entering a key word or phrase; for example, there are strictly menu-driven systems, ones that rely on well-formed queries and restricted query languages, and also ones that attempt some form of natural language understanding. These other methods share some, but not all, of the same conceptual and practical difficulties of key-word methods, and each raises somewhat unique and interesting psychological issues of its own. Where our data bear on these issues in reasonably direct ways, we offer some comments, but we focus primarily on the key-word comprehension process and its ramifications. In the final section of this paper, we discuss some of the reasons why key-word access is as limited as we find it to be and consider several potential methods for overcoming the deficiencies.

II. DESCRIPTION OF DATA SETS

There are things a computer system has or does to which a user might wish to refer. These "information objects" are just the objects, e.g., operations or data sets, to which commands and queries apply. We have collected descriptions of information objects in four quite disparate domains, chosen in an intentional effort to achieve variety, and also for their relevance to a number of special problems that are outside the concerns of this paper. The four sets are: the verbs used in spontaneous descriptions of the operations needed to perform

manual text-editing operations, descriptions of named common objects designed to induce another person or a computer to return the name in a game like the *PASSWORD*™ game, superordinate category names for items available in a swap-and-sale listing similar to classified ads in newspapers, and index words provided for a set of main-course cooking recipes.

In this section we will describe how the object specifications were obtained from people, how they were reduced to single-word or short-phrase keys, and then summarize a few qualitative and statistical characteristics of the responses.

2.1 Text-editing operations

The first data came from language applied to text editing. The study was one of two conducted to explore "naturalness" in command names.[2] Forty-eight secretarial and high school students were asked to provide instructions to another hypothetical typist describing what operations needed to be performed on text marked by an author for correction. These corrections involved two examples each of 25 sorts of edits: five basic operations (insert, delete, move, change, and transpose) on each of five textual units (blanks, characters, words, lines, and paragraphs).

Preprocessing in this case reduced each response to the main verb or phrase in the instruction describing how to perform the editing operation. While this was in fact accomplished manually here, we believe that a simple parser and English word list could in principle have given nearly identical results. These expressions may be considered candidates for command names for editing systems.

Perhaps the most striking result was that there was extensive disagreement in the verbs people produced. This point is the main focus of the current article, and will be dealt with in considerable detail later. For now let us just make a few preliminary notes, e.g., that the three most popular names for each operation accounted for only 33 percent of the total number of responses. The intersubject agreement, the probability that any two people used the same verb in describing a particular text correction, is only .08. Since each of the 25 sorts of edits occurred twice, we also have a measure of within-subject (with 1200 observations) agreement. The probability that an individual subject used the same main verb in the two cases was .34.

What agreement there was did not favor the terminology used by our locally popular editor (the *UNIX** Operating System text editor ed). For 24 out of 25 of the types of edits, the name in ed was not the most frequent spontaneously given name. Use of the terms "delete" and "substitute" was quite rare, for example. (Landauer et al.[2] went

* Trademark of Bell Laboratories.

on to show, however, that this caused no problems initially in learning the basic editor.) People preferred "add" for the insert operations, "omit" for delete operations, and "change" for the replace operations. There was little consensus in describing the transpose and move operations.

2.2 Common object descriptions

These data were originally collected in a study by Dumais and Landauer[3] that examined how people naturally obtain information from one another. The information objects here were names of 50 common items chosen from 10 "categories." The categories were: cities, proper names, clothing, animals, food, household items, abstract words, a category of words with highly associated opposites (e.g., black, love), and two categories whose members were words selected in such a way that negation might figure strongly in the descriptions, e.g., to eliminate the unwanted set members. A total of 337 New York University students were asked to write down a description that would enable another person (or in half the cases a hypothetical computer) to guess the object. There were no restrictions as to the form or content of the descriptions, except that they could not contain the target word itself. Subjects also indicated whether they had any computer experience. Those with at least one computer course were classified as computer-"experienced" in the data summaries discussed below.

A subsequent study was conducted to evaluate the effectiveness of the descriptions generated in the first study. Twenty-five subjects (6 Murray Hill area homemakers and 19 employees of Bell Laboratories) were each given 150 descriptions randomly selected from those generated by the NYU students. They were asked to: (1) guess (without knowing the alternatives) the item being described, and (2) indicate on a five-point scale their confidence in their guess.

Principal results from the main study include the finding that when communications were intended for computers, people with computer experience were relatively more terse, and nonexperienced people were relatively more verbose than when communications were intended for people. However, there was no simple relationship between verbosity and effectiveness, i.e., guessing accuracy as indicated by the second study. People were somewhat more successful in guessing the target items when the descriptions had been provided by people without computer experience (81.2 percent vs. 78.5 percent), but this difference is not statistically significant.

A point of considerable interest here was the style of specification that subjects used. Subjects' descriptions were not very precise; typically they refer to a whole set of items, not just the intended target (although we have no good measure of this other than informal ratings

of denotative class size). Still, the average successful guess rate of the second group of subjects was over 80 percent. We will return to a discussion of this paradoxically high success rate towards the end of this paper. The most frequent way of specifying target items, used about 60 percent of the time, was to describe them in terms of a superordinate (sometimes followed by characteristics or attributes that distinguish the intended target from other members of the superordinate category). Another fairly common form of description (~20 percent) was the use of exemplars. For several of the target words (e.g., motorcycle, magazine, sports, games, science), subjects listed examples of more specific items falling into the target category (e.g., Harley, Suzuki ..., in the case of motorcycle) instead of attempting a more formal definition. Negations (and opposites) were used less than 50 percent of the time for the words we thought were particularly amenable to this form of description.

For the purposes of the analyses undertaken in the current paper, these descriptions were preprocessed to merge minor variations in as automatic a fashion as possible: uppercase was folded to lowercase, word endings were stripped (plurals, tense markers, etc.), and "non-content" words (including articles, imperatives, conjunctions, prepositions, pronouns, and tenses of the verb "to be") were removed.

An average of 8 words per description were in this way condensed to an average of 5.4 words. The first of the remaining words (i.e., first standardized content word) was tabulated for the statistical analyses.

2.3 Superordinate categories for swap-and-sale items

The major purpose of collecting these data was to develop empirical networks of "ISA" relations, that is, classification hierarchies based on user knowledge and representations, for a set of items to be incorporated in an experimental menu-driven information access system.[4]

The information objects were 64 items taken randomly from roughly 300 entries on a monthly bulletin board listing of items for swap and sale at Bell Laboratories. The subjects were 30 local New Jersey homemakers. Each subject worked with a random 32 of the 64 target items. They were told that they were participating in the study to find out how they classified various items being sold on local bulletin boards. The use of these categories in helping people in future computerized retrieval systems was mentioned. Subjects were instructed to complete successive "All ____ are ____" sentences. Beginning with the specific target they were to give its immediate superordinate (e.g., "all red Delicious apples for sale @10¢ ea. are apples"). Then they copied the first given superordinate ("apples") to the beginning of the next incomplete sentence and finished the new sentence with a still

more general category (e.g., "all apples are fruit"). They were to continue in this way (e.g., "all fruits are food") until they could go no further. They were then to go back and find some category that had another superordinate in addition to the one they had already cited, and list that category with its new superordinate. These data were also standardized by stripping off endings and discarding noncontent words.

On average, an individual subject produced 2.1 different chains of successively more general superordinate categories for each stimulus. The chains averaged 2.5 superordinates each, with superordinate categories named in phrases containing an average of 1.7 standardized (content) words. For the current paper, only the lowest level (most specific) category, from the first generated chain of superordinates, was used. The category name was used in its (standardized) entirety.

The categories given in this study make it apparent that the construction of a network that will faithfully match all users' conceptions of a domain is not an easy matter. People have difficulty in generating superordinates and show considerable disagreement as to how things should be grouped under those superordinates. Categorization and indexing schemes currently in use always depend on a user's either generating the same superordinate as the system knows about, or at least being able to choose the right one from a list. Perhaps the difficulty and lack of agreement among people in categorizing information objects account for much of the perceived deficiency of current menu-driven data access methods.

2.4 Recipe index words

The original motive for this data set was to study the effect of domain expertise (i.e., cooking skill) on indexing and key-word usage.[5] The information objects were 188 main-course cooking recipes (French, Italian, Mexican, and American cuisine) taken from 12 cookbooks of explicitly varied sophistication (ranging from a garden club's cookbook,[6] through *The New York Times Cookbook*,[7] to *The Art of French Cooking*[8]).

There were three groups of eight subjects each: experts, who taught cooking classes; and intermediates and novices, selected from local homemakers who came out at the high and low extremes of several self-rating scales on culinary sophistication.

Subjects were told their task was to describe each of the recipes in key-word form, selecting at least three but no more than seven descriptive words or brief phrases for each recipe. They were told that their job was similar to that of a librarian who is creating an index or card catalog, and that the descriptions should be useful to another person trying to locate that recipe in a large set of recipes. Half of the subjects

in each experience group were instructed to direct their descriptions to expert cooks using the index, the other half to novice cooks. The description task was self-paced by each subject in her home. Subjects required between 5 and 10 hours to complete the task.

Terms here were again preprocessed to remove word endings and noncontent words. All multiple-word productions were scored by two judges (the experimenters) to determine if the phrase could be decomposed into its constituent words and maintain its meaning. Subjects produced an average of 5.4 key words per recipe, the first, "most important" of which was studied here.

Again we found considerable diversity. For the 188 recipes, a total of 303 different word types were used by the 8 experts, 220 by the 8 intermediates, and 252 by the 8 novices. It is interesting to speculate on the reasons why these groups differ. Perhaps the experts have a large and specialized vocabulary and the novices have an unruly, haphazard one. In any case, there seems to be something more conventional about the word use of intermediates, a point to which we will return later.

2.5 General comments on the data sets

These data sets all pertain to information objects that one might want accessible on a computer. They were also all of modest size. Other than that, though, they tapped very different knowledge domains, they asked for specification in a number of different ways, and they were provided by different kinds of people. Moreover, the method of reducing the free-form descriptions given by our participants to single words and short phrases varied somewhat from one case to the other. This variety of data is important for our purposes. In order for results to have any pretense of robustness, it is important that they be obtained on a sufficient variety of cases to assure that it is not the particulars of the objects at hand that are responsible for the observed characteristics. We know of no way to actually sample data domains, descriptive methods, and reduction methods in a representative way. However, we believe that results that hold for all of the disparate sets that we have studied stand a good chance of holding for most others.

For each domain, our data can be represented as a table in which the rows are words provided by the subject, the columns are the objects to which these words were applied as descriptions, and the cell entries indicate the number of times each word was used in the description of a given information object. The questions we ask concern how the information contained in such a table might be used to guess from an input word what object is intended. Two partial tables are shown in Tables Ia and b. Table Ia is derived from the text-editing study (for five objects) and Table Ib from the common object data. The numbers

Table I—Word-object data

(a) Sample data from the text-editing study

Words	Objects				
	insert	delete	replace	move	transpose
change	30	22	60	30	41
remove	0	21	12	17	5
spell	4	14	13	12	10
reverse	0	0	0	0	27
leave	10	0	0	1	0
make into	0	4	0	0	1
...					

(b) Sample data from the common object study

Words	Objects					
	calculator	lime	Lucille Ball	pear	raisin	robin...
small	17	0	0	0	7	4
machine	4	0	0	0	0	0
green	0	18	0	7	0	0
bird	0	0	0	0	0	21
fruit	0	1	0	19	1	0
red	0	0	8	0	0	7
female	0	0	2	0	0	0
...						

in the tables represent the frequency with which a word was used to refer to an object.

In fact, there is an implicit third dimension to these tables, representing the person from whom the description was obtained, and sometimes a fourth dimension, representing which of several words of a multiple-word description provided by a given subject is involved. However, for most of the analyses we consider only the first word given, and the matrices are all very sparse, so we have chosen to collapse across subjects. It is worth noting that the tables are not sparse simply because we have failed to collect enough data. Word usage tends to resemble Zipf's distribution[9] (supposedly straight line relation between occurrence frequency and rank frequency when both are plotted logarithmically) in that a few words are used very frequently and many words (over 340 here) used only once (see Fig. 1). As more and more data are collected some cells increase in frequency, but the number of unique words also grows so that the sparseness of the table tends to be preserved. Moreover, most words refer only to a limited number of objects, so that such tables usually have a large number of empty cells.

III. INTRODUCTION TO ANALYSES TO DETERMINE REFERENT EFFECTIVENESS

In the analyses that we report in Section IV, we have been interested in how much information about referent object identity is contained

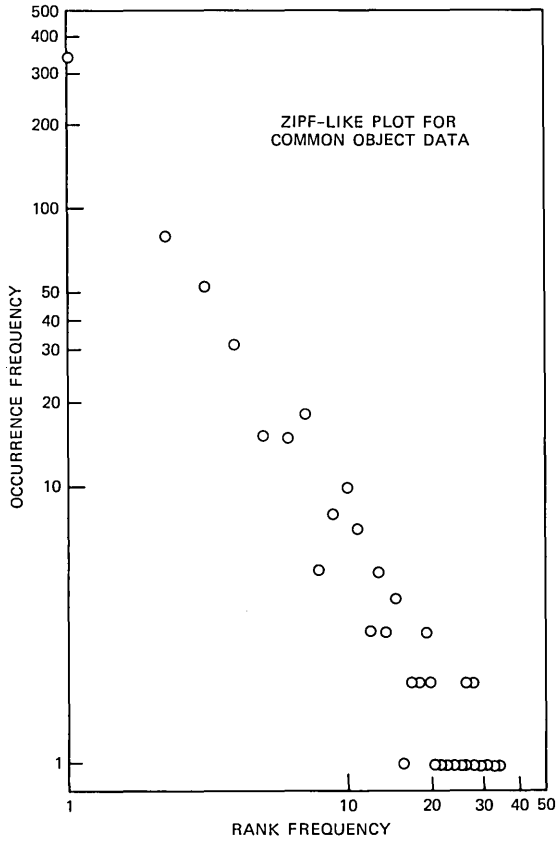


Fig. 1—Plot of common object word usage data.

in the words used to describe them. To consider how good a word or expression is as a reference to some object, it is necessary to suppose some sort of a mechanism by which the expression is comprehended. We can only estimate the value of inputs as determinants of output if we can specify the function that takes one into the other. The inputs of our problem are the words provided by people in specifying information objects. The output is a guess or set of guesses about which object was intended. Our approach has been to specify various ways in which the input can be made to yield the output. We have considered models or functions that are intended to mimic what happens in typical information systems and also ones that try to improve on current methods in various ways. We also develop models that allow us to estimate what ideal input/output mappings could accomplish. It is in this sense of ideal or asymptotic performance that we can appraise the limits of word reference.

The models that we examine have in common that they use the frequency table of users' word reference, and only this data, to guide the system in guessing users' intents.

3.1 The general model

In the sections below we consider a number of special cases of the following rather general baseline model. To retrieve a desired item the user is assumed to enter one (though in some instances, K different) key word(s). The system is assumed to make either one or M guesses as to the user's intent. Success is counted if the intended object is among the system guesses.

As an actual system, our baseline model would operate as follows. The user would enter one (or more) key words or phrases. If it could recognize the input at all, the system would return one or more potential objects according to some rule, perhaps relying on data in a table of observed word-use frequencies. (If it could not recognize the input, it would return nothing.) In cases where the system returns several alternative objects, it can be assumed that the user would be able to select from among them the one intended without error. This corresponds to a simple case of what might be called a "menu-on-the-fly" technique, in which the first entry is a freely chosen key word, and the next step is a choice among a menu of items selected on the basis of the system's knowledge about referents the human user might have intended by that key word.

We will now describe the models that we have analyzed, beginning with the simplest ones, those closest to the way typical systems are currently set up, and proceed to more sophisticated models. Each model is characterized by a set of assumptions or constraints on the input/output function of the system.

Sections 3.2 to 4.7 consider a number of models and analyses in considerable technical detail. Some readers may wish to skip these sections and continue to a summary of the highlights in Section 4.8.

To give a preview, the results convince us of three major points. First, the vocabulary problem is a serious one for untutored users of computers, particularly given the poor naming techniques typical of current practice. Second, the difficulties come from a severe and fundamental lack of consensus in the language community on what to call things. Finally, our research suggests that, by using a very non-standard approach, namely trying to make the best possible guesses on all user words, substantial improvements could be made in meeting the underlying objective of providing access to the things people want.

3.2 Formalizing the general model

Let us make things more formal. We have two sorts of entities, the user-generated inputs, referred to here as "words," "terms," or "de-

scriptors," and the system outputs, called "information objects," "targets," or just "objects." We are concerned with two relationships between these entities. On the system's side are the associations dictating how the system will respond to a user's words—which words will be associated with what system objects. On the user's side are the actual intended associations—what the user in fact meant by a given word. Both the system's and the user's relationships can be shown in the form of a table. The *system table* gives the input/output mapping of which object the system will retrieve (execute, etc.) for what words, with a 1 wherever a word is associated with a system object, and a 0 otherwise. The *user table* shows the frequencies with which people say what words for what objects. Examples of a user table were shown in Table I. This table can be used in the evaluation, and, if one wishes, in the design of the system table.

3.2.1 The system table

The system table has two aspects. First is what might be called the set of structural constraints. These concern how much and what variety of input the system will respond to, and are imposed by capacity limitations on memory, processing, data collection, or sometimes by the computer algorithms used. Second is the choice of a particular instance of the input/output system table, that is, the assignment of exactly which words the system will respond to, and with what objects. One design problem centers on whether assignments will be made in an effective or possibly optimal way.

Structural constraints in the system's input/output table correspond to general restrictions on which cells may be used in the mapping of input to output. We consider patterns based on limiting the number of cells in rows and columns (e.g., the number of words understood as referring to each object and the number of objects taken as possible referents for each word). In particular, we will be considering models where, in the system table, there are various restrictions on:

1. Total words recognized by the system (number of nonzero rows in the system table)
2. Number of words recognized for each object (column totals)
3. Total objects referenced (number of nonzero columns)
4. Number of objects referenced by each word (row totals)
5. Total word-object associations "understood" by the system (grand total of cells).

Once a general structure has been determined, one still must choose which associations to give the system, i.e., exactly which cells to include in the mapping. We consider three cases that give rise to different versions of each model. The first is a purely random assignment. Within the specified structural constraints, the cells defining

the system's associations between words and objects are chosen randomly. Clearly, this is not a realistic system model, but it is useful in the conceptual analysis of other models and as a baseline of comparison. The second case is a weighted random assignment, assigning a word to an object with the same probability that the word and object occur together in usage. That is, input/output associations are chosen with probabilities proportional to their frequency of use, as indicated by the user table. This is, in fact, a very important case to consider, since it approximates the way many systems are currently designed. We will often refer to this as the "armchair" method: a single human designer sits in his or her proverbial armchair and makes the name-object associations by some sort of intuitive guessing. The user table is in effect a compendium of many humans' "armchair" attempts to assign good words to the system objects. A weighted random sample from the table thus corresponds to a random person's "armchair" nomination. Such weighted sampling therefore allows us to estimate the effectiveness of "armchair" design. The last assignment technique is by the best available optimization procedure. Optimization methods make generous use of the empirical data in the user table to pick the best cells. For some system constraints we do not know a combinatorially feasible way to find the best configuration, but it is possible to improve substantially upon the weighted random method. Note that the success of any optimization attempt is also limited by the quality of the data available in the user table.

3.2.2 The user table

The user table is a data matrix compiled from the studies mentioned earlier in this paper. It records the frequency with which untrained users employed various terms in referring to the system objects.

We assume that the descriptions given by our subjects for the information objects in the various sets bear a close resemblance to the entries similar users would make in trying to specify the same objects in a database system. This assumption may be wrong in detail; in using actual systems people might give descriptions somewhat different from those induced by our instructions. However, when we varied the intended recipient of the descriptions, as in the recipe (experts, novices) and common object data (people, machines), there were only small changes in the descriptions. So we believe that descriptive language would change little in attempts to communicate with real, as compared to hypothetical, systems. (We cannot, however, estimate the effect that prolonged interaction with a given system might have on a user's vocabulary.)

We also assume that collection of user descriptions is a good basis for predicting actual user intent from key-word input. In approaching

a data access system, the user must have some information object in mind, and must give a description to the system. In actual use, it is difficult to know what the user really has in mind. Sometimes the user has no clear idea at all. It is not apparent whether this "diffuse target" case is easier or harder for a system to handle. A user with an unclear goal may or may not be more easily satisfied, but is probably less likely to give precise specifications.

It is obvious, however, that the diffuse target case is harder to study. Reliably inducing such a state in the user is difficult, and the system's success is not easily appraised. Therefore, we have collected data and investigated models based on the clear target case.

Thus, in these models it seems to us that the best information the system could start with is to know what typical users would give as descriptions, given the intended object was well known and clearly defined. Our assumption is that the descriptions people give of objects when we specify them will closely resemble the description they would give if they had thought of the specific objects themselves. This seemed a practicable and, we hope, realistic starting point.

3.2.3 Going through the models

Insofar as possible, we will use a consistent expository framework to describe each model. The framework is as follows:

1. Name: A mnemonic title, by which the model will be referenced.
2. Interpretation: What the model is all about.
3. Motivations: Where and why it might come up.
4. Structural constraints: The general form of the system's input/output table ("system table") for the model. This will be given as clearly as possible in English, with a more formal summary of the constraints for each model appearing in Appendix A. This appendix is a very useful reference source in comparing the models.
5. Analyses: For each of three methods of selecting cells for the system table (random, weighted random, and optimized), the following information will be given whenever possible.
   (a) Version: what the method means in this model.
   (b) Evaluation by: what statistic is used to assess its success.
   (c) Result: Statistics are computed for the four data sets:
      (1) Edit command data
         (a) All 25 Operation x Text-Unit combinations (abbreviated: "Ed25")
         (b) For purposes of comparison we include a second analysis of the same data: the five basic operations collapsed across textual units ("Ed5")
      (2) Common object data, all 50 objects ("CmOb")
      (3) Swap-and-sale data, the 64 sale items ("Swap")
      (4) Recipe data, the 188 main course recipes ("Recp")

The following statistics will be given:

Recall probability: how often the method succeeds in returning the desired object.

Mean number returned: when the system is able to make a guess about the user's desired object; this is a record of how many guesses it makes to achieve the reported probability of recall.

A few comments are needed about the "number returned" statistic. First note its importance in considering recall success probabilities. If the system returns a large number of guesses, it can obviously be expected to have a greater chance of including the target. Thus, comparisons of performance demand that this number be kept in mind. Second, note that these indicate the number of things returned when the system in fact recognizes the user's input and so is in fact able to hazard a guess. For some systems it will be common to have insufficient data and to make no guess at all for many user words.

We denote the number of objects, or columns, in the table by C, and the number of user words, or rows, by R. The former is in part well defined by the designers of the system but the latter, the vocabulary size, is usually an artifact of data collection, as it would tend to expand if more data were collected. Here it refers to the size of the vocabulary in the corpus of data we collected.

IV. ANALYSES TO DETERMINE REFERENT EFFECTIVENESS

4.1 Model 1: "One name per object."

4.1.1 Interpretation

Each object in the system is assigned a single term or name. The user enters one term. Success depends on the user's word coinciding with the system's.

4.1.2 Motivations

This approach is common in computer systems—each entity has one and only one name. The name is usually chosen by the designer or by an expert indexer. The designer hopes to establish a convention about what system objects will be called. Users must either learn or guess the names to make the system work. Such learning may be feasible in small systems or for highly practiced users. The growing community of novice and infrequent users, however, are often reduced to trying to second-guess the system, even to find documentation. This is often frustrating and we shall see that in principle the approach is far from adequate. Moreover, the real difficulties of such a scheme often go unremarked simply because traditional computer users have come to accept as normal the necessity of extended learning, repeated second-guessing, lengthy searches, or expert consultation in finding the correct names for programming commands, file names, or information categories.

4.1.3 Structure

The only constraint is that there be exactly one word for each object. Note especially that this model allows the possibility that any given word can be used to name more than one object. This is a situation that designers often try to avoid. It nevertheless arises in several situations, as when programs, commands, file names, or index categories are assigned by many independent users, or to highly similar objects (like bibliographic subject or author specifications, or the case of several functions being collapsed under a single command name). We evaluate models that disallow such "collisions" later. Note that for the likelihood of recalling a given object at all, the aspect of the problem on which we focus in these first analyses, assigning a word to more than one object as we do here, can only improve the expected system performance.

4.1.4 Analyses

4.1.4.1 Version: random. For each object one name is chosen randomly from the total vocabulary for that object, i.e., one cell from each column. That is, the name of each object is a random choice among all the total set of descriptors given to all the objects.

Evaluation: by theoretical value. In this case we can calculate the expected performance of the system exactly. The success rate is given simply by the ratio of t, the total number of cells included in the system mapping, to RC, the total number of cells in the matrix. A simple proof of this appears in Appendix B. See Results Table 1.

Results Table 1

recall probability =	$C/RC = 1/R$, i.e., the reciprocal of the total number of words that users use for all the objects
	Ed5 Ed25 CmOb Swap Recp
recall probability =	.012 .008 .002 .001 .001
mean number returned =	$1 + (C - 1)/R$

The "mean number returned" is a measure of the amount of ambiguity or imprecision in the terms as the system understands them. It is the average number of things the system knows by a given name that the system recognizes; the system must return all of these objects when trying to guess a target. For some of our models the number of objects that the system returns will be fixed ahead of time by design, but here it is the mean of a random variable, easily calculated to be $1 + (C - 1)/R$.*

* Under the pure random method, each object has one of the R words associated with it randomly and independently. Thus, for each object, any given word arising or not becomes a Bernoulli event with probability $1/R$. So having chosen a word for one object,

4.1.4.2 Version: weighted random. A system name is assigned to an object with the same probability that users attributed the term to the object. Here the user enters one key word; and the system has had a key word assigned to each object, based on one other person's nomination. This model mimics, more or less, a currently fairly typical situation for key-word systems (or program-name or command-name access systems) in which the system designer has provided an entry name for each object, obtained only from one usage datum (the designer or indexer's "armchair" introspection), and the user is required to enter just that name. We assume that system designers or indexers are like our subjects in their choice of names, so that the relative frequency with which a name was given to an object by our subjects provides an estimate of the likelihood that a designer would choose that name for that object. (One source of support for this assumption is that experts gave no more consistent names than novices; see below.)

Evaluation: by the "column repeat rate" statistic. We estimated how well such a system is likely to work by estimating the probability that a given word chosen randomly from our population (e.g., by one user on one occasion) would match another word chosen from the same population (e.g., by one designer). This probability is known as the repeat rate.[10] If we index rows (words) by i, and columns (objects) by j, in a word-by-object table, then repeat rate for a given object, rep_j is defined as:

$$r_j = \sum_i p_{ij}^2.$$

This formula can be understood by considering that a match between two randomly chosen words occurs when, for any given word first chosen, the second word is the same. Say the first word was $word_i$; the second word will match it with the probability of $word_i$ occurring in the population, p_{ij}. The probability of a $word_i$ being the first word is also p_{ij}, so the probability of $word_i$ being involved in a match is p_{ij}^2. Summing across all possible words, we get the equation given above.

An unbiased estimate of the population probability of such a match comes from calculating the true probability of such a coincidence in drawing from our sample without replacement, given by:

---

the number of the other $C - 1$ objects having randomly been assigned that same word is given by a binomial distribution with parameters $p = 1/R$ and $C - 1$. Thus the mean number of other objects with the same name as our given object is the mean of this distribution, $(C - 1)/R$. The system will return any of these objects plus the original object, or $1 + (C - 1)/R$ objects.

$$\hat{r}_j = \sum_i \frac{n_{ij}}{N} \frac{n_{ij} - 1}{N - 1},$$

where N is a column total and n_{ij} is the frequency of the ith cell in the column j.

Here we are interested in the average probability of success, given any particular target, throughout the table. So we decompose the overall probability by conditionalizing on columns, calculating the repeat rate for each column, and then average these success probabilities using weights proportional to the column total frequencies. (In our data these column totals are approximately equal by design.) These weighted average column repeat rates are given below for our four sets of data. These numbers may be interpreted as answering the following question: Given that all objects are equally often the desired target, what is the probability that the name given by a user trying to specify a target would match the name assigned to it by a designer? See Results Table 2.

Results Table 2

	Ed5	Ed25	CmOb	Swap	Recp
recall probability =	.07	.11	.12	.14	.18

While there is some variation among the values, they are all quite small. People do not agree with one another very well as to the first word or phrase with which to label an object. The probability that two typists will use the same main verb in describing an edit operation is less than one in fifteen. The probability that two people will use the same first key word for a recipe is less than one in five. (These numbers also tell us something about the size of the set of alternatives that people use in their disagreement. It is a property of repeat rate that the set of alternatives must be at least $1/r$, and can be quite a bit larger if they are not equally likely, as is the case here.)

Most of the interesting comparisons will be between the different models presented (e.g., this model with the subsequent ones), and not the different data sets. To facilitate model-to-model comparison, all results appear together in a summary table in Appendix C.* The reader is strongly encouraged to refer to this table throughout Section IV.

* Of course, some comparisons between data sets are also of interest. Note for example that the value for Ed25 is higher than for Ed5. This says that the set of words applied to the individual objects is more sharply restricted than for the collapsed classes. This is to be expected, since any diversity between objects in the pattern of terms applied becomes within-class diversity, when the objects are collapsed together, driving the column repeat rate down. Only if all objects in a class had identical naming patterns would the repeat rate not decrease.

The mimicked method is one in which the designer provides the system with only one entry word that it can understand, and the user enters just one key word. This is clearly unsatisfactory for untrained users. The usual solution has been for system designers to rely on users learning the chosen vocabulary, i.e., to try to force the user's table to adapt to a fairly arbitrary system table. When the system is small and the user's interaction frequent, this can work quite well. Indeed, Landauer et al.[2] have shown that using unrelated random names has little or no detrimental effect on initially learning to operate a small editor. But, if the system is large and its use intermittent and nonexpert (as for example in large-scale information retrieval systems like library catalogs, recipe files, or classified product catalogs), it is simply unreasonable to require users to learn a specialized vocabulary.* Despite the designers intentions, the uninitiated will try to make the system work without memorizing extensive naming conventions. Thus the problem remains a real one.

One approach we might consider at this point is to seek expert advice in choosing names. This is a fairly common approach, taken in the hope that experts in a given subject area know what things are, or should be, called and so might generate words of more general currency. Indeed, the indexes to books, libraries, user manuals, and other information sources are customarily created either by subject matter experts or by professional indexers.

We have collected some data relevant to this issue in the recipe study. Our key-word providers represented several levels of expertise. The situation is not unrepresentative; usually the indexes of cookbooks and recipe collections are created by cooking experts, presumably on the assumption that their characterization and labeling will be superior, even for less sophisticated users. We calculated repeat rates separately for the three groups of key-word providers (experts, intermediates, and novices) subdivided by whether they had produced the key words under instructions to make them appropriate to novices or to experts. The results are shown in Table II. The repeat rates shown were calculated in a special way. For each cell, the index words were provided by particular subsets of describers, and the proportion of matches was calculated on a pool of descriptor words provided by other subjects. Thus, the (ee, ee) cell estimates how likely a word provided by one expert for other experts was to match that provided by another expert with the same, expert, audience in mind. The (en, ne) cell estimates the probability that a word provided by an expert

* In intermediate cases, like program and command names for an operating system, the method may be satisfactory for expert users, while leading to dissatisfaction for others (see Refs. 11 and 12).

Table II—Repeat rate measures of agreement between indexers and users for the recipe data set

Group	ee	en	ie	in	ne	nn
ee	.11	.17	.14	.15	.10	.22
en	.17	.11	.18	.17	.16	.21
ie	.14	.18	.20	.20	.21	.23
in	.15	.17	.20	.19	.17	.26
ne	.10	.16	.21	.17	.16	.16
nn	.22	.21	.23	.26	.16	.32

for novice users would match the word provided by a novice for expert users. Clearly, the differences among marginal values (i.e., the averages of repeat rates for different index providers) are not large, and more clearly still, expert cooks do not provide better descriptions for the use of either other experts or novices. (If anything, novices do the best in using each other's words.)

The usual armchair approach, even if undertaken by subject matter experts, has only a small rate of success. The obvious step at this point is to seek explicitly optimal choices of names, treating this as the empirical question that it clearly is.

4.1.4.3 Version: optimized (best). If we want to use the name that has the greatest currency among subjects, we must choose the term that is in fact maximally used by subjects for each object. We pick the maximum cell in each column and use the corresponding term.

Evaluation: by various estimates of the range of expected performance. The lower bound is based on split halves analysis; the upper bound is based on a transformation of the column repeat rate and from an analysis that assumes that the sample data exactly reflect the population probabilities.

To know the performance of this model we need to know the true population magnitude of the maximum cell in a column. That cell is the one we would choose according to an optimum name assignment scheme, and its size would be the proportion of future users' terms for the given object that would coincide with our optimal choice. Problems arise in that there are no known distribution-free, unbiased estimators of the population magnitude of the maximum probability cell. We have, however, been able to devise a few techniques that let us put bounds on the performance of this system. The first uses a split-halves technique to give a lower bound estimate. The data in each column are split into two halves, and the maximum cell chosen on the basis of the first half (as though we had to design our "optimal" system on the basis of half the data). This cell is then matched against the second half to see how well it succeeds. Thus, the second half acts as a virtual experimental test of the performance of the "optimal" method. The split-half results shown here are the average performance of ten independent splits of each data set. See Results Table 3.

Results Table 3

	Ed5	Ed25	CmOb	Swap	Recp
recall probability =	.15	.19	.26	.26	.31

These numbers are an unbiased estimate of how well a system would do using this optimum strategy but constrained to a small amount of data (namely the amount in each half). It clearly underestimates how well one could do with more data. We have used two approaches to obtaining an upper bound. One is based on an interesting inequality relation that holds between the size of the maximum cell and the repeat rate statistic described above: It can be shown that the former is no greater than the square root of the latter.*

Thus, the performance of the optimum model is expected to be no greater than the square root of the performance of the weighted random (armchair) model. This statistic will overestimate performance to the degree that individual words other than the maximal one are also applied frequently to a given object. Since our own data suggest that this is commonly true, this upper bound is likely to be quite generous. It has an important pragmatic advantage, however, in that it is independent of sample size and easy to obtain. Other estimates of optimal performance require collecting detailed data on the precise pattern of naming. This estimate requires only that one observe the probability that two people use the same name (i.e., the repeat rate), without even having to note what particular terms they use. Taking the square root then yields an upper bound estimate for the best possible single name per object. For our data, these quantities are listed in Results Table 4.

Results Table 4

	Ed5	Ed25	CmOb	Swap	Recp
recall probability =	.27	.32	.33	.35	.42

The final way to estimate the performance is to let the data predict itself. The observed largest proportion falling in a cell is taken as the estimate of the population maximum. A familiar result in rank order

* This relation follows from two inequalities. First note that the repeat rate is the sum of the squared cell probabilities. This sum is clearly greater than or equal to any of its terms, since all are positive. In particular, it is larger than the square of the maximum probability cell. Thus, the expectation of the repeat rate is larger than the expectation of the squared maximum cell. Second, note that the expectation of any squared variable is always greater than or equal to the square of that variable's expectation. Putting these together, the expectation of the repeat rate is greater than or equal to the square of the expected magnitude of the maximum cell. Thus, an upper bound on the maximum cell is estimated by the square root of the repeat rate.

statistics is that in the presence of error the observed maximum of a number of observations is expected to be larger than it should be. Thus, using the maximum to estimate itself is likely to be an overestimate for any limited sample size. This is particularly true for small samples. In the extreme case note that if a column of cells has only one observation, the observed maximum will be in the one cell where the single observation fell, which will have an estimated probability of 1.0, regardless of the true underlying probabilities. This will become more relevant later when we calculate similar statistics on rows of the matrix, where many of them will involve small numbers of total observations in each row. For now, however, the samples are not too small, and the results are presented in Results Table 5.

Results Table 5

	Ed5	Ed25	CmOb	Swap	Recp
recall probability =	.16	.22	.28	.34	.36

4.1.5 Discussion

It appears that performance would be about twice as good when using an optimum naming strategy than when using the weighted random model (which we believe to be an approximation to typical current practice) in a one-word-per-object system. Still, the overall levels are not very impressive. It should be noted that these optimal strategies represent the best *any* single-name scheme could do. No expert, human or otherwise, could choose single names that would work better.

Is this the best that we can hope to do for people? The answer is no. There are other, more effective approaches. One is to use multiple names, sometimes called "aliases", for each object, which leads us to our second general structure for a system model.

4.2 Model 2: "Several names per object."

4.2.1 Interpretation

Each object in the system is assigned M terms or names. The user enters one term. Success depends on the user's word coinciding with any one of the system's M words.

4.2.2 Motivations

Giving things single names is not adequate for the uninitiated, so a reasonable next step is to try giving the system several names by which to recognize each object.

4.2.3 Structure

The constraint is that exactly M words are stored for each object.

4.2.4 Analyses

4.2.4.1 Version: random. For each object M names are chosen randomly from the total vocabulary, i.e., M cells are chosen from each column. The recall probability is M/R, and in our sample tables would range from .01 to .04 for three names per object ($M = 3$).

4.2.4.2 Version: weighted random. The first name is chosen with a probability proportional to its frequency in the given object's column of the table. Each successive name is then similarly chosen, without replacement, from the remaining cells. This is as though someone "sitting in an armchair" thought up several distinct names, any one of which the user could use with success. We make an independence assumption that the probability of a word being chosen is independent, except for renormalization, of the words already chosen. This is probably a faulty approximation for a single human generating a series of words, where the first word thought of may influence the next. If the words were separately proposed by different designers, this independence assumption might be more nearly correct.

Note that by reversing the roles of the system and the user we obtain a very interesting dual interpretation of this model. That is, let the system store a single word and the user make M different guesses. Success would be counted if any of the user's M words matches the system word. Either of these interpretations merely involves the probability that a single sample from the column will match one of M samples drawn without replacement from the cells of the same column (that is, without replacing the whole cell, not just the observation from the cell).

Evaluation: by "M-order repeat rate" statistic, within columns. These success probabilities can be estimated using an extension of the repeat rate statistic to this case of multiple samplings without replacement of types. The probability for column j can be shown to be estimated by:

$$\hat{r}_j^{[m]} = \sum_{i_1} \frac{n_{i_1 j}}{N} \frac{n_{i_1 j} - 1}{N - 1} \left[1 + \sum_{i_2 \neq i_1} \frac{n_{i_2 j}}{N - n_{i_2 j} - 1} \left[1 + \sum_{i_3 \neq i_1, i_2} \frac{n_{i_3 j}}{N - n_{i_2 j} - n_{i_3 j} - 1} \left[1 + \cdots \right. \right. \right.$$

$$\left. \left. \left. \cdots \sum_{i_m \neq i_1, i_2, \ldots, i_{m-1}} \frac{n_{i_m j}}{N - \left(\sum_{k=2}^{m-1} n_{i_k j} \right) - 1} \left[1 \right] \cdots \right] \right] \right].$$

This formula requires a calculation time that grows exponentially in M. We therefore limit results to $M \leq 3$. See Results Table 6.

Results Table 6

recall probability	Ed5	Ed25	CmOb	Swap	Recp
($M = 1$)	.07	.11	.12	.14	.18
($M = 2$)	.14	.21	.21	.25	.33
($M = 3$)	.21	.30	.28	.34	.45

Note that allowing the system (or, equivalently, the user) several words for each object in trying to match its partner achieves considerable gain. Performance almost doubles and triples as we go to two and and three guesses.

It should be remembered that the estimated probabilities of success for 1, 2, and 3 guesses were calculated on data from subjects' first responses only. Under the interpretation of this model in which subjects make repeated guesses at a system's single word, this is equivalent to assuming that successive guesses, given that previous guesses had failed, would resemble other first-provided words. To get a true estimate of the expected performance of a system that actually used this technique, we would need data on people actually making successive guesses. One would have to know how many such guesses can actually be made and with what quality before one could know what the maximum performance would be. However, if one supposes that the system, or perhaps the users' desires and abilities, limited input to three guesses, performance of such a system would not be likely to exceed about one chance in three of correct return.

4.2.4.3 Version: optimized. Choose as the M names for each object those terms that are maximally used by subjects for the object; that is, pick the highest M cells in each column and use the corresponding terms. This is a simple generalization of the single-name case.

Evaluation: by split halves and self-estimation. We again run into the problems involved in estimating maximum, and now also the nearly maximum, cells. The split-half and self-estimation procedures were used here to estimate upper and lower bounds. The split-half results represent the average performance of ten independent splits of each data set. The results are presented in Results Table 7.

Results Table 7

recall probability	Ed5	Ed25	CmOb	Swap	Recp	
($M = 1$)	.15	.19	.26	.26	.31	(split half)
	.16	.22	.28	.34	.36	(self-estimation)
($M = 2$)	.26	.32	.36	.36	.49	(split half)
	.28	.37	.41	.50	.56	(self-estimation)
($M = 3$)	.37	.42	.42	.45	.58	(split half)
	.38	.49	.48	.59	.67	(self-estimation)

4.2.5 Discussion

Considerable benefit is obtained both from using data to optimize choices and from giving the system several names for each object (or, equivalently, allowing the user several guesses). Though it was not undertaken here, it would be interesting to explore the possibility of letting both the user and the system use several names to try to match each other.

The improvement gained by storing multiple words has associated with it a potential cost in ambiguity. Nowhere in either this model or the previous, one-name model has there been any concern that the names be distinct. The names for each object were picked from the corresponding column, with no regard for what names were being chosen for other objects. This means that two objects could be assigned the same name, with consequent ambiguity should the user give that name. The system would be unable to tell which object the user intended, and would have to present the user with a menu of choices to differentiate among them. (Recall that at the outset we stated that we would be assuming such a system, and, moreover, that choices among the items on the subsequent "menu-on-the-fly" would be assumed error free.)

Naming choices will collide when two different objects have the same words applied to them. One might expect this to happen, for example, if two objects are very similar, so that the same words apply to them. We have collected some data that confirm the existence of this similarity effect. For both the recipe and the common object data we compared the probability of a naming collision for random and similar objects in the set. Subjects were given 5 "focal" objects and 25 "match" objects, all drawn at random from the set. For each of the focal objects, they were told to select the one match object that was most similar to it. Using the weighted random naming method, the naming collision rates were compared for these five pairs of similar objects and the same five focal objects paired with random members of the match set. That is, we calculated the probability that a word applied by one user to the first object would be the same as the word applied by another user to a second object. Using subject differences as a random error estimate, there was a highly significant increase in naming collisions for the pairs of objects judged to be similar [$t(14) = 3.86$ and $t(24) = 5.37$ for the recipe and common object data, respectively].

To examine the effects of trying to avoid collisions, we studied the next model.

4.3 Model 3: "A distinct name for each object."

4.3.1 Interpretation

Each object in the system is assigned a single term, with the proviso

that no term may be used more than once. Success depends on the user's spontaneously produced word coinciding with the system's distinct name for the intended object.

4.3.2 Motivations

Note that in typical naming circumstances the intent is often to establish conventions about terminology so as to avoid the ambiguity that would otherwise arise. Naming conventions are used not only to set by fiat the name by which an object will be known to a system, but also to proclaim that *only* that object will be so known. Often interaction will reach a stage where complete precision is needed, e.g., for actual command execution.

It may be the designer's motivation in selecting names that users learn terminology that allows them to be precise. As had already been pointed out, however, the designer's intent may not correspond with the user's reality; the untutored may try to deal with the system anyway. Thus, we explore the user-guess success for systems designed with the unique name constraint.

The models discussed here are just like those of the "one-name" case, except that words cannot be used more than once.

4.3.3 Structure

One distinct word is stored for each object, i.e., this is the same as Model 1, except that the words must all be distinct, so that at most one object is referenced by each word. This imposes strong constraints on the system table, on both row and column totals, and the number of rows and columns used overall. This high degree of constraint makes mathematical treatment difficult, as will be discussed shortly.

4.3.4 Analyses

4.3.4.1 Version: random. For each object, one name is chosen randomly from the total vocabulary. Once a vocabulary item is used, however, it is eliminated from any future consideration.

Recall probability would be $1/R$ and range from .002 to .014 for these data. The "number returned" is easy to give for this case, as it is set by the structural constraint, that each name have a unique referent. It is therefore 1 for all models having this structure. That is, when the system finds a target it returns exactly one candidate. However, there are many occasions when the user word matches no system word, and so no target is returned.

4.3.4.2 Version: weighted random. Here the name is chosen with a probability proportional to its frequency in the given object's column of the table. Then the chosen word is eliminated, and a name is chosen in an analogous way for another object from the remaining words, etc.

The results are clearly influenced by the sequence in which objects are dealt with. The approach used here was to choose in the following way: a word/object pair is chosen by a weighted random sampling from the whole table. This gives the right distribution within each column and allows appropriate representation of each column. Once such a cell is chosen, the corresponding object has been named and the word used, so both are eliminated, and the procedure is iterated on the table, now reduced by one row and column, until all objects have been named.

Evaluation: by Monte Carlo simulations with split halves. We could devise no way to evaluate this model analytically, so we used a Monte Carlo simulation on split halves. We divided the data in the user table in half and randomly picked names, according to the model outlined above, using the data from one half. The second half of the data was then used to evaluate the effectiveness of the names thus chosen. The results presented here are the average for ten split halves with ten independent Monte Carlo simulations of name selection in each. See Results Table 8.

Results Table 8

	Ed5	Ed25	CmOb	Swap	Recp
recall probability =	.07	.08	.11	.12	.09
number returned = 1.0 (by design)					

Note that, as might be expected, success is less than for the comparable model in which names did not have to be distinct. In some instances this decrease is small, as with the edit studies; in others it is large, as with the recipes. This should depend on whether there were a few high-frequency words that were used for many different objects.

4.3.4.3 Analysis of precision versus popularity of terms. In all of our data sets there was a slight trend for high-frequency words to be less discriminating than low-frequency words. In studying this relation quantitatively, the measure of discriminating power of a word was given by the repeat rate statistic, this time applied within each row. This row-repeat rate is a measure of the probability that two uses of a word refer to the same object. When this probability is high, the word is very discriminating. Below we present the correlation (Spearman r) of row-repeat rates with the marginal frequency of the words. See Correlations Table 9.

Correlations Table 9

	Ed5	Ed25	CmOb	Swap	Recp
corr:	−.28	−.30	−.21	−.24	−.16
(N words):	(74)	(84)	(301)	(145)	(167)

This correlation points to the cause of the difficulties run into by the constraint of distinct names. There is something of a conflict. If one chooses high-frequency names, they will be likely to collide. If one chooses less frequent names, the chances of collision will be somewhat less, but fewer users' queries will be handled. Unfortunately, there is no correspondence between the size of these correlations and the size of the decrease in performance for each data set. The reasons for this are not clear.

4.3.4.4 Version: optimized (though not best possible). The idea is to choose the distinct names such that the total probability in the cells chosen is maximal. The method used here is a greedy algorithm that is not truly optimal, but it is a reasonable improvement over the weighted random method. It begins by picking the highest cell in the matrix; then, after eliminating the corresponding row and column from the matrix, it iterates, picking the next highest cell, etc., until all objects have been assigned a name. Algorithms like this are called "greedy" because at each step they take the biggest possible chunk of what is left, without regard for what later problems that may cause. In this case, it is possible that an early choice will eliminate a word that would be very good for another object, when there was an alternate word that would have done almost as well at no such cost. Algorithms that would take such future complications into consideration, or equivalently consider so many possibilities at once, are typically combinatorially explosive. Thus we present the results of the straightforward greedy algorithm approach.

Evaluation: by split halves. There is no simple method to estimate the performance of such an approach. Again we turned to the split-halves technique of dividing the data in half, applying the algorithm to one half and then testing the chosen names on the other half. The results, averaged over ten independent splits, are given in Results Table 10.

Results Table 10

	Ed5	Ed25	CmOb	Swap	Recp
recall probability =	.14	.11	.23	.19	.11
number returned = 1.0 (by design)					

4.3.5 Discussion

As we expected again, the optimization attempt, even though imperfect, has a dramatic effect, in many cases doubling the performance of the system. Even in the best case, however, the system succeeds only about one quarter of the time, and typically little more than a tenth of the time. We note that the numbers here are substantially lower than in the case where there was no requirement that names be distinct. In fact this improved method really does no better than an

armchair (weighted random) version of the unconstrained model, and often worse. The lesson here is that adding the requirement of uniqueness, common in establishing conventions, hurts naive users quite a bit. The need for such conventions is not denied, it is simply asserted that auxiliary aids will be needed for systems that make heavy use of such conventions. It is worth noting that, for our data sets, the number of objects that are not disambiguated and would have to be presented for a second stage choice is always small, and the gain in recall always large. Thus, the "menu-on-the-fly" method implicit in several of the models presented here appears to be very promising as an aid to unsophisticated users.

4.4 Model 4: "Distinct names, augmented with M extra referents."

4.4.1 Interpretation

Each object is given a distinct name. $M - 1$ other referents for those words are also stored.

4.4.2 Motivations

Part of the inferiority of the distinct-name models, when compared to the unconstrained models, came from the fact that people often want to use the same names to refer to several objects. The motivation for the next model is to recapture access to some of those other interpretations of the term. A simple extension of the distinct-name structure is to begin with the situation where each object is associated with a distinct name, one that can be memorized and used unambiguously by experts, but also to store $M - 1$ other objects that the term applies to, explicitly for use in a "menu-on-the-fly" for the untutored. A reasonable system implementation might be to give the "distinct" name a special status (the "real" meaning of the term), and to admit the other interpretations as secondary, perhaps to be verified in the context of use.

This is the first model in which we explicitly design in more than a single system guess as to the intended object. To be sure, several guesses for the meaning of a term could (and would) have arisen in the uncontrolled name cases of Models 1 and 2, but here we predetermine exactly how many guesses the system makes and evaluate performance as a function of this parameter.

As more objects are returned, the likelihood of including the intended target is increased, but a certain cost is incurred—the cost associated with discerning the true target among all the returned objects. Data searches can fail not only by not giving access to a desired object, but also by returning too many unwanted objects. In information science the problem is familiar as the trade-off between *recall*—the number of desirable items returned—and *precision*—the

proportion of items returned that are desirable. This is also similar to the hit versus false alarm trade-offs of signal detection, familiar to psychologists and communication theorists. In the latter context the trade-off is sometimes examined by tracing out operating characteristics and interpreting them in terms of parameters of an underlying statistical theory. There does not currently exist a corresponding statistical theory for our precision and recall rates, but it is useful to be able to examine, or in this case, control, precision explicitly and see what gains in recall result. This yields an operating characteristic. Any version of the general baseline model in which the system is set up to return an explicit number of guesses provides a direct way to do this.

4.4.3 Structure

At least one name is given to each object, but each name that is used by the system also refers to M different system objects.

4.4.4 Analyses

4.4.4.1 Version: random. Here the primary cells are again chosen completely blindly, eliminating rows and columns already used in an iterative manner exactly as in the distinct-name case. After the C primary cells are chosen, the $M - 1$ additional or secondary cells are chosen randomly from each new row chosen.

The resulting probablity is M/R and would range from .001 to .035. The number returned is, of course, M by design for all versions of Model 4.

4.4.4.2 Version: weighted random. Cells are chosen at each step with a probability reflecting their magnitudes. Thus, first C distinct names are picked, as in the distinct-name model (Model 3). Then, in the row of each cell just picked, $M - 1$ more objects are chosen with a probability equal to their relative frequencies in the rows.

Again, the weighted random case is an approximation to what an individual designer might do without collecting data from other people. The process of coming up with additional interpretations requires a bit of further explanation. The asymmetries found in free association data suggest that people starting from terms and generating objects would yield probabilities quite different from those obtained from people starting with objects and generating terms.[13] This asymmetry means, in this model, that the additional interpretations cannot be thought of as the result of the designer sitting in an armchair and thinking up other interpretations for the terms. The efficacy of such an approach is not estimable from our data.

A scenario that might better satisfy the assumptions of this model would have the system memorize the first $M - 1$ nonstandard uses of its known terms that it comes across.

Evaluation: by Monte Carlo simulation on split halves. The data are given in Results Table 11.

Results Table 11

	Ed5	Ed25	CmOb	Swap	Recp
recall probability =					
($M = 1$)	.07	.08	.11	.13	.08
($M = 2$)	.13	.15	.15	.17	.15
($M = 3$)	.18	.21	.18	.20	.20

The number of guesses that the system is requiring to achieve this performance is of course, M, the number returned by design.

4.4.4.3 Version: optimized (though not best possible). Again it was not feasible to find the completely optimal choice of cells. Instead, the primary cells were found in the same "greedy" way used for the simple distinct-name "optimal" case. The subsequent choice of secondary referents for the words so selected can be optimized by choosing the $M - 1$ largest cells remaining in the row. Note that it is possible for some of these remaining cells to be larger than the primary cell, as when they refer to an object that was eliminated before the row was chosen. (It is the vagaries of the need for convention in names that brings about this ironic use of the term "primary.")

Evaluated: by split halves. In the now familiar procedure, the data are split and the first half used to choose the cells following the algorithm just outlined. Then the second half of the data is used to evaluate the choice of cells. See Results Table 12.

Results Table 12

	Ed5	Ed25	CmOb	Swap	Recp
recall probability =					
($M = 1$)	.14	.11	.22	.21	.11
($M = 2$)	.21	.20	.28	.27	.17
($M = 3$)	.27	.29	.31	.30	.22
number returned = M (by design)					

These results, for one to three total guesses, are plotted in Fig. 2. The model curves can be viewed as operating characteristics, giving total recall as a function of decreasing precision. The dashed diagonal lines represent expected chance performance if the system returns guesses at random. Note that in the case of the Ed5 data, the random performance is *better* than that of this model. This is possible because this model, like most current computer systems, makes no response at all to any but a few words. By failing on so many of the words it encounters, it can be surpassed by a system that only guesses, but at least guesses for all words. Admittedly, in the context of command execution any level of guessing without confirmation may be danger-

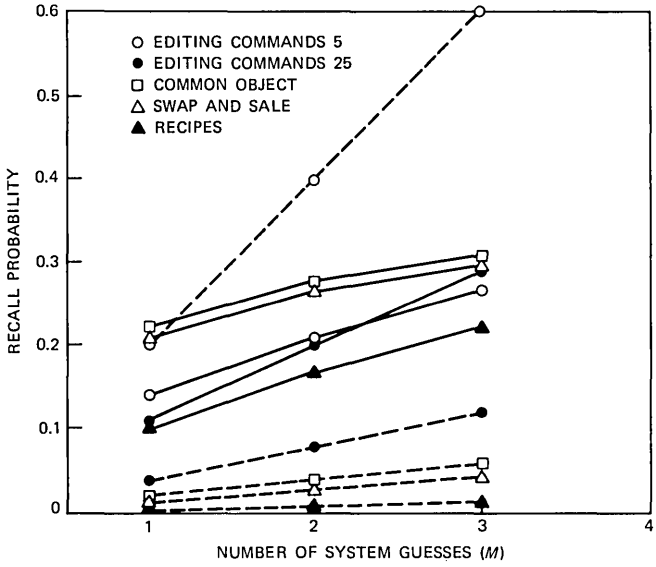


Fig. 2—Recall as a function of the number of system guesses for Model 4 ('distinct names, augmented with M extra referents') is shown by the solid lines. The dashed lines represent expected chance performance if the system returns guesses at random. (Different chance performance for the different data sets simply reflects the different number of objects in each set.)

ously out of place, but there are safer arenas, like help facilities, that could benefit. In any case the cost of ignoring so many user words is made clear. Notice that the total probability does not approach one, even in the case of command verbs for editing operations where there were only five objects. The reason is again that many words were used by our subjects to describe these objects, and the algorithm picked only a maximum of C of them to recognize. The modeled assumption was that if a subject supplied any word not recognized by the system, no choices would be returned and a failure would result.

4.4.5 Discussion

These models use the same number of cells as the multiple-names models, yet do not do as well. The implication, presumably, is that there is considerable cost in limiting oneself to a small number of words. The users distribute their descriptors too broadly for any such limitation to be very satisfactory.

Thus we are motivated to try a completely different approach. All the models up to this point have inherently focused on what the system brings to the interaction. The modus operandi has been to set all the system's objects before us and then try various ways to guess what users will call them, albeit with some improvements as we give

deference to the empirical characteristics of the user's language as seen in the user table.

Suppose we turn the tables, so to speak, and focus on what the users bring to the interaction: the terms they use when trying to specify things. Start with the terms, all of them, and concentrate on trying to guess to what object they refer. This brings us to our next model.

4.5 Model 5: "Recognize one referent for every word."

4.5.1 Interpretation

The system stores every word that it can, and for each has one guess as to the intended object.

4.5.2 Motivations

One of the principal problems with the approaches studied so far has been that a large proportion of the recall failures can be attributed simply to not recognizing many of the user's words. The systems modeled have paid too little attention to the wide variety of the terms people use spontaneously. Here we focus on this essential character of what the users bring to the interaction. For every word produced during the collection of the data for the user table, a guess is made. The different versions vary as to how these guesses are made.

It is useful to note that these models, and their generalizations in the next section, are really duals of the single- and multiple-names models. In those models the system had words for each object, and it succeeded when its name was the same as the word the user would use. Thus, to evaluate those models we calculated the conditional probability that, given a target object, the system and user names would coincide. Then we summed probabilities across objects, weighting each by its column's marginal probability (all essentially equal in our data, by design). Here we conditionalize the other way. Given the word used, we find the probabilities that the system and user associate the same object with the word. Then we sum across words (rows), weighting by the very uneven observed marginal frequencies for each word, to get the overall performance.

4.5.3 Structure

The system makes use of every word in the table, associating each with one, and only one, system object.

4.5.4 Analyses

4.5.4.1 Version: random. Here the system will associate random objects with the input words. That is, the system will simply make a

random guess for anything the user says. The recall probability is just $1/C$, and ranges from .005 for the recipes to .200 for the Ed5 data. The number of guesses being made to achieve these performances is predetermined by design to be 1 for all these models.

4.5.4.2 Version: weighted random. Here an object is associated with each word by a random method that takes each object with probability proportional to its relative frequency in the row. Arguments similar to those made for the augmented distinct-name model here imply that this is not the analog of having the designer sit down with the list of words and hazard guesses as to what they mean; the data used here go in the other direction. The more appropriate scenario is as if a system records the first encounter with a new word, and its intended referent, and makes a single pointer based on this single observation, forever freezing the meaning of the word. This is perhaps an unreasonable scenario, but the model is worth investigating because the statistics that result have other valuable interpretations.

Evaluation: by row-repeat rate statistic. The probability of success in this case is the probability that two users will mean the same thing by a given word. The results below are the average row-repeat rates, weighted by total row frequency. (Rows with a frequency of one were excluded, since the repeat rate is not calculable in that case.) This number is an unbiased estimate of the overall probability that any two occurrences of any word will be in reference to the same object. See Results Table 13.

Results Table 13

	Ed5	Ed25	CmOb	Swap	Recp
recall probability =	.41	.15	.52	.62	.13
number returned =1.0 (by design)					

We note how variable the probabilities are. They are low where the same words mean different things, and high where any one word refers to only one object. Apparently these domains differ in this aspect. There are at least two possible explanations. First is the similarity effect discussed in connection with Model 3. There it was shown that pairs of objects judged to be similar had higher overlap in the patterns of names applied to them than did random pairs. The resulting naming collisions meant that it was more difficult to find distinct names for similar pairs. Here we are explicitly interested in a different, though related, effect. We can use the previous experimental data (see Section 4.3.4.3) again, this time to demonstrate the similarity effect on a row-repeat rate measure. In this version we consider pairs of cells in each row: one cell from the target column, and one from either a column

rated similar or from a random control column. The repeat rate is calculated on the two cells and summed (weighted by row sum) down all the rows. Since a low repeat rate for a given word indicates that it does not discriminate well between the objects, we predict a lower repeat rate for the similar pairs of objects, and that is what we obtain. In the recipe data the mean repeat rate of the similar pairs was .73 and for the random pairs was .91 [$t(14) = -5.83$]. For the common object data, the similar pair repeat rate was .89 and the random repeat rate was .95 [$t(24) = -5.71$]. Thus similarity has a strong effect on repeat rate. Unfortunately, what we need to make sense of the varied results is some measure of the relative internal similarities of the various domains. Such data are neither available nor easily obtained. Still, it might be agreed that the set of common objects is more diverse than is a set of cooking recipes, or text editor operations, corresponding to observed differences in average repeat rates.

There is another factor that might be helping the swap-and-sale descriptors. Analysis of the other data sets was either strictly limited to single words, or else only to short phrases with few content words. Swap-and-sale descriptors contained an average of 1.7 content words, where there were fewer than 1.2 content words in the others. If these words tend to be used in a conjunctive sense and if they are not redundant, they should contribute to the high selectivity of the descriptors seen for the swap-and-sale data.

A final note should be made of the higher repeat rate for Ed5 compared to Ed25. Part of this is due simply to the fact that even with a purely random, undiscriminating distribution of name usage across objects, there would be an increase in repeat rate as the number of objects is decreased. If there are only two objects, people will have to mean the same object by any given term at least half the time. A more intriguing possibility is that the classes are more distinct entities than are the individual objects, and that this makes word usage more discriminating, above and beyond the chance effect just mentioned.

4.5.4.3 Version: optimized. In this case the user matrix is used to find the best possible choice for a word's referent. This is done by picking the maximum cell from each row, the object to which the word has most commonly referred.

Evaluation: by split halves, square root of the row-repeat rate, and self-prediction. Evaluation is again problematic. While there is no question that picking the maximum cell is the best possible strategy, there is no unbiased estimate of its true magnitude. Thus we resort to the same three methods used when similarly confronted in the optimal version of the one-name model (Model 1): the split-half technique, an unbiased estimate of how well one could expect to do with half the data; and two estimates of upper bounds. See Results Table 14.

Results Table 14

recall probability =	Ed5	Ed25	CmOb	Swap	Recp	
	.49	.18	.43	.35	.13	(split half)
	.62	.35	.65	.72	.28	(square root of row-repeat rate)
	.54	.26	.69	.81	.25	(self-prediction)
number returned =1.0 (by design)						

4.5.5 Discussion

There is considerable variability, both in the performance and in the range of these estimates. The largest range is for the swap-and-sale estimates. This is due to the very large number of descriptors that occurred only a few times. The data for each such descriptor are statistically unreliable, so the maximum cell cannot be accurately identified; this problem is severely aggravated by splitting the data and only using half to make the identifications. Note that this is not an experimental artifact. A very long tail on a descriptor distribution is a legitimate real-world problem, because it means that new terms will keep arising that the system will not have seen before, and about which it will not be able to make any educated guesses. A wide range in the estimate reflects the fact that very large amounts of data would be needed to approach asymptotic performance.

As for the overall diversity of scores from domain to domain, the arguments about similarity and number of words in the descriptors, given above for the repeat-rate case, are equally applicable here.

The most important point to be made about this optimized model is that it represents the best one could possibly do. We can do no better than to recognize every word users use and make an optimal guess as to its meaning. Clearly, if we have insufficient data to do this well, either preventing us from recognizing a word, or from being able to estimate the modal referent, performance will suffer. But the upper bound estimates represent the real, strict limits of performance. No other pattern of structural constraints could do better, except at the cost of precision. This trade-off with precision is explored in our final model, in which an explicitly prespecified number of multiple guesses is returned to increase the chance of including the user's intended object among them.

4.6 Model 6: "M referents for every word."

4.6.1 Interpretation

The system stores every word that it can, together with a set of M possible referents. Whenever the user enters a word, the system returns the M guesses, possibly ranked in some way.

4.6.2 Motivations

The "one referent for every word" model (Model 5), in its optimized version, gave the best possible performance for a system that hazards only one guess for a user's word. The only way to increase the chance of returning the user's intended object is to return more than a single guess. These guesses could be returned in the form of a menu of M items, among which users would choose.

As we mentioned, this is the dual of the multiple-name set of models, and several of the evaluation procedures differ only in that rows have traded roles with columns.

4.6.3 Structure

Like the previous model, the system makes use of every word in the table, but here it associates M system objects with each.

4.6.4 Analyses

4.6.4.1 Version: random. The system makes just M pure guesses, without replacement, from the set of system objects. The recall probability for these cases is thus exactly M/C, and the number returned is M, by design, for all versions of Model 6.

4.6.4.2 Version: weighted random. In a manner following the single referent weighted random model, the M cells are chosen with a probability that is proportional to their relative sizes. The M choices are required to be distinct, and so a cell is excluded from further consideration once it has been selected.

The scenario that this might correspond to would be one in which the system learns the first M distinct referents of the word that it comes across in use. Its encounters would be governed by exactly these probabilities.

Evaluation: by M-order repeat rate statistic, within rows. We want the probability that the user's single intended referent coincides with any of the system's candidates, when both sets are drawn from the same probability matrix. These success probabilities are estimated using the same extension of the repeat rate statistic used in evaluating success of the many-names-for-one-object model (Model 2, Section 4.2.4.2). Here though, it is applied to the cells within a row, rather than within a column.

It should be noted that the formula requires there to be M nonempty cells in the row, to prevent some of the denominators from going to zero. For example, it is not possible to give an unbiased estimate of how well three guesses would do if the data only show two objects. Note that while rows with highly discriminating words should in general have high M-order repeat rates, it is exactly such rows that will be less likely to satisfy this requirement, especially at lower

marginal frequencies. Thus, ignoring rows where this statistic cannot be calculated biases the average value of this statistic downward. Though it is not clear how to get an upper estimate, there is another way to get a lower bound, by recognizing that performance for M guesses is at least as good as performance for M-1 guesses. Thus, a lower bound on the average performance comes from using for each row the calculable repeat rate of highest order not greater than M. See Results Table 15.

Results Table 15

	Ed5	Ed25	CmOb	Swap	Recp
recall probability =					
($M = 1$)	.41	.15	.52	.62	.13
($M = 2$)	.66	.26	.65	.74	.21
($M = 3$)	.81	.36	.71	.80	.27

4.6.4.3 Version: optimized. This version takes the optimal strategy for multiple guesses. The user gives one word, and the system makes guesses that the user means one of the M most likely things the word has meant in the past, as deduced from the data in the user table. That is, it returns the objects associated with the M highest cells in the row corresponding to the word given. (A sequential version of the model would give the user the guesses in decreasing observed frequency, beginning with the maximum cell.)

Evaluation: by self-estimation and split halves. As in other cases where evaluation of performance depends on the estimation of true population ordered frequencies, we must resort to indirect methods to give a range. We use split halves to give a lower bound and self-estimate to give an upper bound. See Results Table 16.

Results Table 16

	Ed5	Ed25	CmOb	Swap	Recp	
recall probability =						
($M = 1$)	.49	.18	.43	.35	.13	(split half)
	.54	.26	.69	.81	.25	(self-estimation)
($M = 2$)	.70	.30	.53	.43	.20	(split half)
	.76	.42	.82	.92	.37	(self-estimation)
($M = 3$)	.83	.40	.58	.47	.26	(split half)
	.88	.53	.88	.96	.46	(self-estimation)
number returned = M (by design)						

The recall rates, as estimated by the conservative split-half procedure, are given in Fig. 3. The curves can be interpreted as operating characteristics; the distance between the curves and the corresponding dashed diagonal line reflects how much the implied system could be expected to outperform a simple menu system based on the pure random model given above, i.e., on a random choice of M objects. All

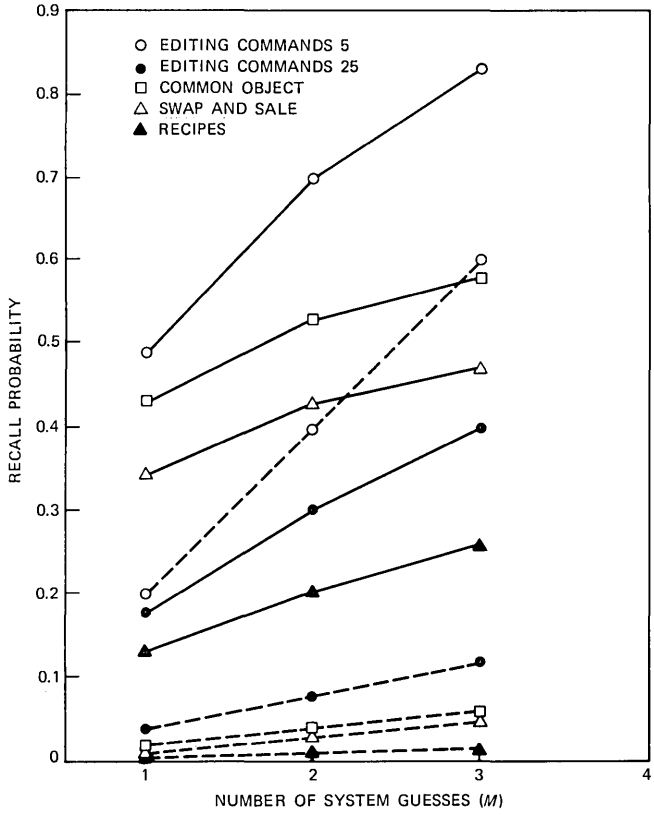


Fig. 3—Recall as a function of the number of system guesses for Model 6 ('*M* referents for every word') are shown by the solid lines (split-half estimation). The dashed lines represent expected chance performance if the system returns guesses at random. (Different chance performance for the different data sets simply reflects the different number of objects in each set.)

of our models do substantially better than chance. More importantly, they do much better than the augmented distinct-name model (Model 4) presented in Fig. 2.

4.6.5 Discussion

There are really two important points to be made about these results. The first is that these performances are variable and but moderately high. Note that the very high value of the three-system-guess case in Ed5 is rather vacuous, since there were only five objects to guess from. In the common object case, however, the high values are quite meaningful, since there were 50 objects.

The legitimate high values lead us to the second observation, that substantial improvement has been gained over the traditional, one-

distinct-name approach. The presumption is that this gain is due in part to the increase in capturing all the user's inputs. There are, after all, two ways for the system to fail in retrieving the user's target. One is by making wrong guesses, the other is by being unable to make any guesses at all. To give an idea of the size of this problem: for three of our data sets, no system using only C words (e.g., one for each object) could recognize even half the user's words. Thus, recall rates can only increase as the system is taught more words, and often this can be a sizable improvement. The only conceivable decrements (except those associated with time and space in index management) would be if the additional words were less precise, thus decreasing precision for a fixed recall rate, or decreasing recall for a fixed precision. But our finding of a consistent, albeit small, negative correlation between a word's frequency and its selectivity suggests that the opposite is true. As a greater number of lower-frequency words are included, precision will go up.

Recall also increases dramatically if the choice of algorithms is optimized, though it never exceeds the square root of the performance obtained by "armchairing", i.e., making pointers based on a single observation. The cost is in the trouble it may take to collect data. In some cases the variety of terms is so great that data collection must continue for quite a time before asymptotic performance is approached. Performance also increases as the system is allowed to make more guesses; this involves a direct trade of recall probability for reduced precision. In many circumstances, particularly in help facilities, this presents little problem. The user need only be given the various options and whatever additional information it takes to decide what is really sought.

4.7 Other possible models and some limits

The models presented so far have covered a large share of the space of simply constrained models where the user makes one guess, but they do not exhaust the ways in which a system could use empirical data on user descriptions to guess the user's desired objects.

One might consider, for instance, models that take advantage of redundancy in the tables. A good theory or description of what such tables are like could be used to augment the data to make better estimates of true population usage probabilities. For one example, if we knew that some single shape of distribution characterized all columns, we could use the data to estimate parameters of that distribution instead of individual cell probabilities. This would reduce the number of parameters that need to be estimated from the data, and consequently improve stability and reliability of predictions.

A related, and perhaps more interesting, idea would be to look for

latent structure in the similarity between objects, and between words, and use this to improve predictions. For example, if we knew or could infer from the data that two words were used almost identically, we might pool the cell frequencies for the two words; or if we knew that the objects and words related to each other by some more elaborate structure (e.g., a hierarchical tree), we could base predictions on calculated indirect reference paths.

While these ideas might be useful, the research presented in this paper puts strict limits on the success of any such approaches. As noted above in Models 5 and 6, the self-prediction evaluation procedure (determining the best guesses from the data and then using the same data to estimate success) and the "square root of repeat rate" evaluation procedure yield upper bounds on expected performance. Indeed, no analysis method of the kind we have just been discussing, no matter how clever, could improve on this result. The reason is as follows. Even using inherent structure to improve predictions could do no more than increase the accuracy of estimation of the population values for the input/output tables. Even with true population probabilities, however, the guessing decision rule would be the same; choose first the most probable referent of each word, and so forth. The input/output tables are inherently probabilistic, as a result of disagreement in word usage, not just of the estimation uncertainties. For a given user population, with a given degree of training, a certain amount of disagreement in word usage will occur. Even with perfect knowledge of the probabilities describing this usage, we could not predict individual referents perfectly. Now, the objective of having true population probabilities as cell entries is mimicked by our upper bound estimating procedure. It pretends that the values observed for each cell are exactly the probabilities that would exist in further sampling from the population. Thus, no method of "cooking" the data to reduce sampling error could achieve a better result than is illustrated by our upper bound values.

To improve performance beyond these limits, it would be necessary to construct systems that use different input, e.g., multiple words or interactive dialogues, or make the user learn to use more easily interpretable language. The last of these is the current default approach—make the user learn everything—and it has its limits for large systems or occasional users. Thus, we believe a more fruitful direction to explore further is multiple-word inputs.

The simplest models of multiple-word query might assume only content words (no syntax) and a statistical independence between words. Independence would allow the performance of such a system to be estimated from single-word data of the sort we have collected. Multiple words could be used disjunctively. In this case the system

would return any objects that matched any of the input words. This would result in an increase in probability of recall. Alternately, the words could be used conjunctively, so that the system would return only objects matching all the key words. The result would be higher precision. Mixtures of these two approaches (e.g., conjuncts of disjuncts) then clearly could be used to improve both precision and recall.

Of course, from our data we do not know how people can or do use multiple words, even without considering syntax. The popular balances of conjunction and disjunction are unknown, and spontaneous multiple guesses are no doubt not independent. But what sort of nonindependence is common? For example, do multiple terms tend to be more unrelated, or more redundant, than independence predicts? These questions must be explored before multiple-word systems can be theoretically evaluated, or optimized for the naive user.

The consideration of syntax leads to a whole new set of problems, those often encountered in Artificial Intelligence work on natural language understanding. We will not address them here.

4.8 Highlights

In the previous sections we used the tables of observed word usage to explore various models of how systems could name, and thereby give access to, their contents. We devised six general schemes for assigning names to objects. To test the schemes, we approximated user behavior by sampling appropriately from the tables.

Thus, for example, we simulated what word a designer might assign to an object (pick from the table), what word a user might choose (pick again from the table), and looked at how frequently the two words were the same. The results were taken to indicate how well a set of spontaneously generated single names for objects can be expected to work for untrained users. This particular example was called the "armchair model." The table is, after all, only a compendium of many people's spontaneous armchair attempts to give the best possible names for these objects. Thus, sampling from the table mimics designers picking names from their armchairs.

We used this general technique to explore variants on our basic model of interaction, in which we assume the designer builds in certain connections between names and the system's objects or services. These names were in some cases purely random, and in other cases were a simulation of a designer's best armchair guess; in still other cases the names were chosen more systematically, with a goal of optimality. In all cases, we assume that the user must "guess" the right name because we are concerned with untutored users who do not know the system words. The user's choice of words was always mimicked by sampling from the tables.

Of the models and variations, summarized in Appendices A and C, perhaps the most important are:

(1b) "One name per object: weighted random"
(1c) "One name per object: optimized"
(6c) "M referents for every word: optimized."

The first of these (1b) was the so-called "armchair" naming method just described. Expected results from the armchair method are shown in Fig. 4 model (1b). Note that despite the fact that the four domains differ dramatically in content, data collection style, and subject population, the results are quite consistent. They are all low. Two people—the designer and the user—will come up with the same name only about 15 percent of the time.

These results indicate that the current practice is deficient; it guarantees that it will be difficult for people to tell machines what

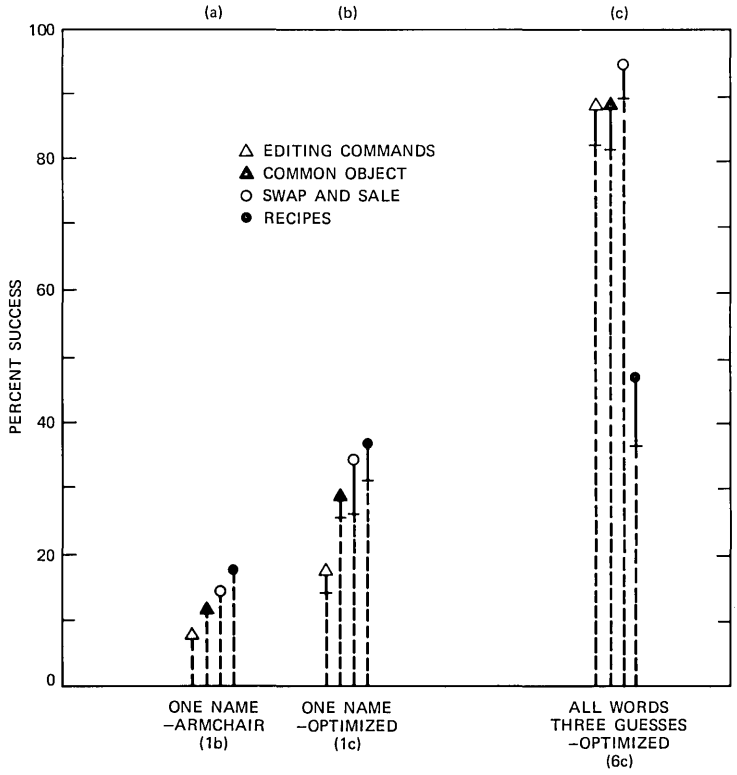


Fig. 4—Summary of expected success for three of the most important models presented. (a) "Armchair" model (1b); (b) Best possible single-name model (1c); (c) Model where the system recognizes all words and returns three guesses (6c).

they want, unless they already know what to say. Note that designers often have other, legitimate motives in assigning names—e.g., lack of ambiguity, memorability, or cuteness. But such naming practices only make the current problem worse; they lead to even less likely names.

The source of the problem is this. There are many names possible for any object, many ways to say the same thing about it, and many different things to say. Any one person thinks of only one or a few of the possibilities. Thus, designers are likely to think of names that few other people think of, not because they are perverse or stupid, but because everyone thinks of names that few others think of. Moreover, since any one person tends to think of only one or a few alternatives, it is not surprising that people greatly overrate the obviousness and adequacy of their own choices.

To improve performance of the "armchair" method we investigated whether experts in one of the content areas could pick better words. We had expert chefs choose key words for cooking recipes and found essentially no improvement: experts do not seem to do noticeably better picking terms from their armchairs than does anyone else (Section 4.1.4.2).

The next model of major interest (1c) substituted objective data for armchair suggestions. It used our data tables to identify the name that was most commonly used for each object. Thus, for example, for the operation we referred to as DELETE, the most commonly used term was "omit", with a frequency of 110, and that was the name used in this model. This empirical approach, occasionally advocated by human factors people, achieves the levels of expected success illustrated in Fig. 4 model (1c).

As discussed in Section 4.1.4.3, for statistical reasons we can only estimate certain upper and lower bounds on performance here. The lower one indicates how well one could do with a limited amount of empirical data from which to try to pick the best names. The upper one liberally estimates how well one could do with an infinite amount of data. The improvement over the armchair method is substantial, typically almost doubling the hit rate. This makes the point that the best name is a good bit better than the typical name. But it is also clear that even the best one is not very good.

It is critical tc note that these numbers represent the best one could possibly do by assigning a single name to each object. The problem is not solved by finding the right name. Different people, contexts, and motives give rise to so varied a list of names that no single name, no matter how well chosen, can do very well. Figure 5 is a plot of how many names are needed to account for a given percentage of the users' attempts, here for the common object data. (Here again we have the statistical estimation problem, so the boundary is double.) Note that

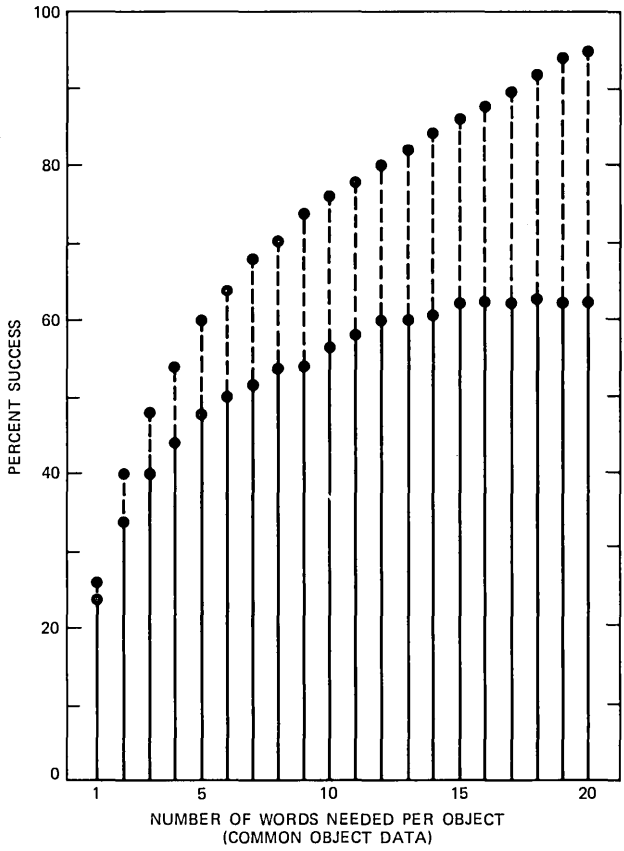


Fig. 5—A plot of how many different names (or "aliases") are needed for each object to account for a given percentage of success. This plot is based on data from the common object domain. The dashed lines indicate ranges.

even 15 words (word types or "aliases") per object account for only 60 to 80 percent of the words people apply. That is, even with 15 names stored per object, the computer will miss 20 to 40 percent of the time. The lesson clearly is that systems must recognize many, many names.

After considering several other models, we turned the problem around a bit. Instead of starting with system objects and looking for names, the new strategy was to begin with user's words and look for interpretations—that is, try to recognize every possible word that the users generate, and use empirical data to determine what they mean by that word.

We inject a cautionary note here: when word meanings are to be surmised from word usage data, there is always a risk that the same word may have been used in reference to several objects in the system.

In such a case, the system would not have a unique guess. (This was actually an implicit problem discussed for several of the earlier models.) In our simulations, we limited the system to returning just its single best (Model 5, Section 4.5) or best three (Model 6, Section 4.6) guesses as to what was wanted, given the word the user said. For discussion here we will mention only this last model (6c). The results in Fig. 4 show what can be achieved if all words are saved by the system, and it uses behavioral data to make the three best guesses as to what was meant. (Here too we can only present estimates of upper and lower bounds on performance.)

These results are encouraging. The big improvement is due primarily to the fact that the earlier methods, like most in current use, failed to recognize all the rare words. The unfortunate truth is that the majority of words used are "rare" words, and so it is a mistake not to reckon with them. Indeed, it turns out (see Section 4.3.4.3) that less common words tend to be more precise, more discriminating, and so are particularly valuable for the computer to know.

While we studied quite a wide variety of intermediate models, this last one really is the best. At the risk of belaboring the point, here it is summarized again. It requires: (1) the computer to keep all words that people use; (2) that data be collected on what users use each word to refer to; (3) that when someone uses a word, the computer goes to the data table and looks up the top few candidates to determine what its interpretation should be (what the word is most probably referring to); and (4) the computer presents the candidates to the user as a choice, since the interpretations may be in error. This model amounts, really, to a rich, empirically (i.e., behaviorally) defined cross index. Without it, performance is near abysmal, with it, potentially quite good.

Further improvements are shown to require the input of multiple words. This would entail means for representing and evaluating relations as well as reference, and matters of logic, syntax expression, and language comprehension. Detailed analyses of the complicated usability issues that would be involved are beyond the scope of this paper, but a number of related issues were addressed in Section 4.7.

V. DISCUSSION

We have seen that the object referent of a word cannot be predicted with great accuracy from knowledge of its past referential use. The use of statistical data on reference behavior can improve the choice of a single name by roughly a factor of two over the common procedure, in which only a single designer-chosen word is stored as an entry point for each object. Even the best system input/output function of the general kind we have described, while adding another factor of two or

so to performance, can be expected to perform at well below perfect reliability. In this section, we touch on some of the reasons for this limitation and suggest some avenues for further exploration of better means for understanding object reference.

Obviously, one of the main difficulties in predicting an intended object from a provided word is synonymy. There are many different words that can refer to the same object. Even though the receiver may know several of them, the communicant (or user) may choose another. It is the long-tailed distribution of the word usage for objects, as illustrated in Fig. 1, that is the villain. Unfortunately, however, this is only a part of the problem. Indeed, our upper bound estimates assume that this problem does not exist, that we know every word that will be used, and with what probability each will be applied to every object. Still our best upper bound predictions are quite error prone. Another part of the problem is polysemy; each word means many different things and can refer to many different objects. In our observed data, words that were frequently used tended to be applied to several objects. Clearly, if a word is used for two or more objects, there is no way that we can guess from the use of that word which unique object is its referent. The more similar the objects in the domain, the more likely it is that a single word will include more than one in its scope (Section 4.5.4.2). In general, then, one might expect that the more similar the objects in a set seem to people, the more difficulty they will have in describing them uniquely and the more difficulty a receiver of their descriptions would have knowing to what they refer.

How might one interpret descriptions of objects more accurately? Observe that in the common object experiment, human subjects were able to make a single guess as to the intended object with over 80 percent success. This is well above any of our performance estimates for systems based on the statistical information in single-word input/output tables. How do people do this? What other sources of information, either in the input description or in the receiver's mind, are brought to bear? What is needed to build a system that would approach human capacities?

One possibility is that the limits we observed may be due to utilization of only a single word or phrase from the input. Perhaps if multiple words, or the combined meaning implied by conjunctions of several words and their syntactic order, were utilized, much better reference could be achieved. The conjunctive use of words primarily serves to more narrowly specify the object of discourse. This presumably increases the precision of reference and would reduce the chance that we would guess an incorrect object from a provided description. But, under models that assume the comprehender can return several guesses as to the likely object, the recall provided by a full description

would not necessarily be much greater than that provided by the unrestricted meaning of its most important word. However, if users can provide many independent words, each an essentially new attempt, the recall rate could be raised substantially, as Model 2 suggests. Thus, allowing longer, more complex specifications and learning how to understand them is one promising direction to be investigated.

Recall also that using experts to generate armchair key words did not work well, at least in the recipe data where we tested it. We also note that in informal demonstrations, programmer subjects do not fare well in providing a name for a program to be matched by other programmers. Similarly, there have been a number of studies of indexer reliability in bibliographic indexing.[14–18] These come from quite favorable circumstances, in which the index terms are chosen from restricted vocabularies and the indexers are highly trained. The chances of one indexer choosing the same categories as another, even under these circumstances, are usually disappointingly low. Thus, the chance that professional indexers will agree with the first word entered by an untrained user does not seem to offer a promising route.

A possibility for improvement that seems worthy of further investigation is the study of the structure of the conceptions or mental representations of the objects in a domain to be referenced. The polysemy aspect of the problem arises because two or more objects are not linguistically separable. If we could learn how to group such objects into "super-objects," then we could potentially improve at least our ability to predict which of these "super-objects" is being sought. Similarly, informal impressions from the swap-and-sale superordinate data suggest that certain levels of superordination give rise to more consistent naming than others. If means can be found to reveal and represent strong hierarchical structure in the concepts being named, then possibly one can choose the levels or nodes in such structure that are best represented as the objects to be found in a data set. These "super-objects" also might be more amenable to automatic reference by the means we have modeled.

We have generated statistics relevant to this hypothesis, in the double treatment given to the editing terms. The five objects in the Ed5 data are in fact "super-objects", made by condensing the 25 editing operation–text unit pairs into categories involving just the editing operation. A look back at the numbers involved shows a uniform superiority in the retrieval and discrimination of "super-objects" over that of the individual constituent objects. In part this is just because there are fewer objects to be dealt with, but our hypothesis is that performance is further enhanced by the lower overall similarity at the super-object level. Suppose a system were built to capitalize on this kind of situation, that is by first empirically discovering the lowest

level of subdivision of a given domain at which natural descriptive language is adequate. Freely chosen key-word entries might well be effective for specifying objects at this level. But the whole problem of retrieval would not yet be solved. Users who wanted to access particular subordinates of these super-objects would have to be provided with some further mechanism. A hybrid approach in which early stages of specification are done by freely chosen keys and later stages by menu selection seems a promising approach.

An especially important matter, which we have so far neglected, is the prior probabilities of a user's intending various objects. Our data have been aimed at estimation of the input/output pointer functions of words to objects, and were collected with equal numbers of occasions for nomination of key words for each object. In real life, and in a real system, people seek different objects with different frequencies, perhaps with steep distribution functions resembling Zipf's law.[9] Our ability to predict what object a person has in mind by a word could be greatly improved by taking into account its prior, unconditional likelihood of being sought. Again, our informal impressions from the kinds of descriptions provided in the common object data is that human describers and recipients take great advantage of such frequency information. For example, in specifying the Empire State Building as a tall building in midtown Manhattan, the describer probably assumes that the receiver would choose, from the large set of possible objects, the one that is most likely to be the object of specification.

It will also require further work to see how such information could be incorporated in an automated access system. A start would be to consider an adaptive system that keeps track of the frequency with which objects are sought (as well as the frequency with which the particular input is satisfied by a particular output). Then object prior-probability data could be combined with input/output conditional probability data, like that we have investigated, by the use of Bayes' rule.[19] This would almost certainly yield much better predictions than our models estimate, or any that are currently achievable in available systems. Such a scheme might also take advantage of individual differences in cases where the same people will use the system repeatedly.

There are still other plausible means for circumventing the limitations our models suggest. There are other kinds of data access devices, such as menu-driven systems, query languages, or natural language understanding systems. These approaches all have something to offer, and probably can overcome some of the deficiencies of the pure key-word entry method. But each also has problems of its own. For example, menus cannot list a very large number of alternatives at once, so the desirable feature of turning the recall or production

problem into comprehension and choice is limited. (Note also that in menus the user must understand the system's terse descriptions of objects, much as the system has to understand those of the user of key words.) When there are many objects, a menu system must use a successive search method that relies on some kind of hierarchical tree or other presentation of the relations among the objects. How to do this in a way that leads to correct user choices at each level, to good overall performance, and to acceptable convenience are unsolved issues. Query languages generally require users to input well-formed relational algebra or Boolean expressions that are unlike anything seen in the data specifications provided in our password study. Such expressions require the kind of logical thought that is known to be extremely difficult for ordinary people.[20] Natural language understanding systems, as so far implemented at least, have yet to deal adequately with the lexical reference problem. They have usually glossed over the issue by restricting themselves to very limited domains and limited lexical input, for which they can store a reasonably adequate "hand-tailored" synonym list. We suspect that when such systems are developed for use in real data applications they will have to solve the synonymy and polysemy problems in some of the same ways (e.g., by the use of statistical input/output tables and prior object probabilities that we have been suggesting here. It is probably a mistake to take a too naively optimistic view of the value of "natural language" input to a computer device. For example, although our human subjects were much better than our model automatic systems at predicting the referent of a description, it is not obvious that their success was based on being able to "understand" the natural language of the input, at least if this is taken to mean successful syntactic parsing, etc. We believe that it is also possible that human success is based largely on statistical knowledge of the likelihood of objects and the likely referents of words, and that this is a matter that could be incorporated into a system without it doing highly intelligent natural language understanding.

VI. SUMMARY

The data we have collected on people describing objects have allowed us to estimate the likely performance of several mechanisms for understanding the references of words to information objects. We have found that input/output functions based on normative naming behavior of users will work much better than systems based on a single name provided by designers. We have also shown that a system that made several best guesses, and/or allowed the user to make several tries, and then returned a menu-like set of guesses to be chosen among, would substantially improve performance beyond current popular

methods. The best approach was to focus on what the user brings to the interaction, namely a great variety of words, and for each word have the system make one or more best guesses as to what the user meant.

But such a system for understanding references will still not perform nearly as well as a human receiver would. Thus, this model of the process of reference certainly does not fully capture what goes on in people's minds. Thus, we are clearly not yet ready to hazard a theory of how humans succeed as well as they do in this task, or to propose an automated method that would do as well or better. However, we believe that the evidence and analyses reviewed here lead to some promising suggestions for further exploration in both regards.

REFERENCES

1. J. M. Carroll, "Learning, using, and designing command paradigms," Human Learning: Journal of Practical Research and Application, *1* (January 1982), pp. 31–62.
2. T. K. Landauer, K. M. Galotti, and S. Hartwell, "Natural command names and initial learning: A study of text editing terms," Commun. ACM, *26,* No. 7 (July 1983).
3. S. T. Dumais and T. K. Landauer, unpublished work.
4. G. W. Furnas, unpublished work.
5. L. M. Gomez and R. Kraut, unpublished work.
6. *Our favorite recipes: Inverness Garden Club,* private printing, 1977–78.
7. C. Claiborne, *The New York Times Cookbook,* New York: Harper and Row, 1961.
8. J. Child, L. Bertholle, and S. Beck, *Mastering the Art of French cooking,* Vol. I, New York: Knopf, 1979.
9. G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology,* Reading, MA: Addison-Wesley, 1949.
10. G. Herdan, *Type Token Mathematics: A Textbook of Mathematical Linguistics,* Mouton: S-Gravenhage, 1960.
11. D. A. Norman, "The trouble with UNIX," Datamation, *27,* No. 12 (November 1981), pp. 139–50.
12. M. E. Lesk, "Another view," Datamation, *27,* No. 12 (November 1981), p. 146.
13. G. Keppel and B. Z. Strand, "Free association responses to the primary responses and other responses selected from the Palermo-Jenkins norms," in L. Postman and G. Keppel (Eds.), *Norms of Word Association,* New York: Academic Press, 1970, pp. 177–87.
14. R. S. Hooper, "Indexer consistency tests—origin, measurements, results, and utilization," IBM Washington Systems Center, Bethesda, MD, 1965 Congress Int. Federation for Documentation, October 10–15, 1965.
15. J. Jacoby, "Methodology for indexer reliability tests," RADC-IN-62-1, Documentation, Inc., Bethesda, MD (March 1962).
16. D. J. Rodgers, "A study of intra-indexer consistency," General Electric Company, Washington, D. C., (September 1961).
17. J. F. Tinker, "Imprecision in meaning measured by inconsistency of indexing," American Documentation, *17* (April 1966), pp. 96–102.
18. J. F. Tinker, "Imprecision in indexing (Part II)," American Documentation, *19* (July 1968), pp. 322–30.
19. V. E. McGee, *Principles of statistics: Traditional and Bayesian,* Englewood Cliffs, NJ: Prentice Hall, 1971.
20. P. C. Wason and P. D. Johnson-Laird, *Psychology of structure and reasoning: Structure and content,* Cambridge, MA: Harvard University Press, 1972.

APPENDIX A

Formal Model Structures

In Table III below we present the formal structures for each of the six models studied. The constraints critical to the definition of each

Table III—Summary of constraints defining each model

Model	Description	Words Used: Per Object	Words Used: Total	Objects Referenced: Per Word	Objects Referenced: Total	Total Word-Object Pairs
1	One name per object	**1**	$\leq R$	$\leq C$	**C**	C
2	M names per object	**M**	$\leq R$	$\leq C$	**C**	$M \times C$
3	Distinct name for each object	**1**	**C**	**1 or 0**	**C**	C
4	Distinct names, augmented with $M-1$ extra referents	$1 \leq n \leq M$	**C**	**M or 0**	**C**	$M \times C$
5	Recognize one referent for every word	$\leq C$	**R**	**1**	C	R
6	M referents for every word	$\leq C$	**R**	**M**	C	$M \times R$

Defining constraints for the models are in boldface. C = number of columns (objects), R = number of rows (words).

model are marked with asterisks. (Note that these constraints are to be met whenever possible, and for simplicity, the numbers below assume that this is always possible.)

APPENDIX B

Evaluating the Pure Random Versions of the Models

The success of any pure random model considered in this paper is given simply by the ratio of t, the number of cells included in the system mapping, to RC, the total number of cells in the matrix. To see this simply note that any structural constraints we might impose are only defined up to a permutation of the rows and columns. Thus, consider any table and its group, i.e., all its variants obtained by row and column permutations. The table that is the cell-wise total of all these tables must, for reasons of symmetry, be everywhere the same, and its grand total must be t times the number of tables in the group. The success of any individual table is given by the dot product of the user and system matrices (treating the matrices as vectors, i.e., sum the products of corresponding cells). The total success for the group is the sum of these dot products for the members of the group. But the sum of the dot product of one vector with several others is the dot product of that vector with the sum of the others, so the total success for the group is just the dot product of the user matrix and the total matrix. But since the total matrix is uniform, and we divide by the size of the group, we conclude that the average matrix then is just t/rc times the sum of user matrix. If the user matrix is in relative frequencies, the success is a probability.

APPENDIX C

Summary of Recall Probabilities

Table IV—Summary of recall probabilities

Model	Description	Version	M	Recall Probabilities Ed5 (n = 5)	Ed25 (n = 25)	CmOb (n = 50)	Swap (n = 64)	Recp (n = 188)
1	One name per object	WGT	1	.074	.106	.117	.142	.182†
		OPT	1	.151	.194	.257	.258	.312*
				.271	.322	.329	.353	.419†
				.163	.221	.279	.337	.362‡
2	Several names per object	WGT	1	.074	.106	.117	.142	.182†
			2	.144	.205	.210	.253	.332†
			3	.211	.295	.284	.337	.445†
		OPT	1	.151	.194	.257	.258	.312*
				.163	.221	.279	.337	.362‡
			2	.263	.319	.358	.358	.490*
				.281	.369	.406	.496	.564‡
			3	.365	.424	.420	.452	.578*
				.381	.492	.478	.595	.670‡
3	Distinct name for each object	WGT	1	.073	.076	.105	.124	.086*
		"OPT"	1	.140	.109	.226	.194	.105*
4	Distinct names, augmented with M extra referents	WGT	1	.067	.076	.109	.127	.084*
			2	.125	.145	.151	.172	.149*
			3	.175	.212	.181	.196	.200*
		"OPT"	1	.140	.109	.221	.208	.105*
			2	.214	.204	.277	.266	.170*
			3	.266	.287	.306	.296	.220*
5	Recognize one referent for every word	WGT	1	.413	.153	.516	.617	.128†
		OPT	1	.489	.176	.434	.351	.129*
				.620	.353	.647	.715	.276†
				.541	.258	.686	.813	.247‡
6	M referents for every word	WGT	1	.413	.153	.516	.617	.128†
			2	.656	.264	.646	.743	.207†
			3	.811	.358	.713	.795	.266†
		OPT	1	.489	.176	.434	.351	.129*
				.541	.258	.686	.813	.247‡
			2	.699	.302	.532	.427	.203*
				.761	.416	.819	.919	.373‡
			3	.830	.405	.577	.465	.259*
				.884	.531	.881	.956	.455‡

* Split halves
† One of the several repeat rate related statistics
‡ Self-prediction

AUTHORS

George W. Furnas, B.A. (Psychology), 1974, Harvard University; Ph.D. (Psychology), 1980, Stanford University; Bell Laboratories, 1980—. Mr. Furnas is a member of the Human Information Processing Research Department. He is doing research on cognitive aspects of person-computer interaction, with emphasis on access to information structures. Particular topics of recent interest include adaptive indexing, multidimensional scaling, and a "fisheye lens" viewer for large structures. Member, Psychometric Society, Classification Society, Society for Mathematical Psychology.

Thomas K. Landauer, B.S. (Anthropology), 1954, University of Colorado, Ph.D. (Social Psychology), 1960, Harvard University; Dartmouth College,

1960–1964; Stanford University, 1964–1969; Bell Laboratories, 1969—. During his first 10 years at Bell Laboratories, Mr. Landauer studied a variety of problems involving human memory. He worked on optimal scheduling of practice and study to maximize learning, did research aimed at understanding how humans retrieve information from their own memories, and developed a mathematical model of human memory that assumes it to be a content-addressable parallel access device that imposes no inherent structure on its content. For the last three years he has done research on psychological problems of human-computer communication. Fellow, AAAS; member, ACM, American Psychological Association, Psychonomic Society, Society for Cross-Cultural Research, Classification Society.

Louis M. Gomez, B.A. (Psychology), 1974, State University of New York at Stony Brook; Ph.D. (Psychology), 1979, University of California, Berkeley; Bell Laboratories, 1979—. Mr. Gomez is a member of the Human Information Processing Research Department. His current interests are in human factors of information retrieval, and individual differences in the acquisition of computer skills.

Susan T. Dumais, B.A. (Mathematics and Psychology), 1975, Bates College; Ph.D., (Psychology), 1979, Indiana University; Bell Laboratories, 1979—. Ms. Dumais is a member of the Human Information Processing Research Department. Her primary research interests are in the area of cognitive aspects of human-computer interface. She is currently involved in work on information retrieval, including menu selection, describing categories, and browsing for information. Member, AAAS, Midwestern Psychological Association, Metropolitan Chapter of the Human Factors Society.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

On Abbreviating Command Names

By L. A. STREETER,* J. M. ACKROFF,† and G. A. TAYLOR†

(Manuscript received May 21, 1982)

Test subjects' abbreviations of command names and randomly selected English words were examined for production regularities. Abbreviation rules based primarily on a word's number of syllables were devised to capture regularities observed in people's productions. This rule set was compared to two simpler abbreviation rules—vowel deletion and truncation. In subsequent learning experiments, separate groups of subjects learned the rule-derived abbreviations for words, while other groups learned the most frequently given abbreviation for each word. Subjects who studied rule-derived abbreviations remembered substantially more of them when prompted with full words than did subjects who studied the most frequently given abbreviations. Moreover, the rule-based abbreviations were superior even for those for which the rule-produced and the most frequently produced abbreviations were identical. When the task was reversed (recall the source term given an abbreviation), performance was best for vowel deletion abbreviations and worst for the rule set abbreviations. We suggest that both memorability of abbreviations and the probability that people will spontaneously produce a "correct" abbreviation are increased by: (1) selecting abbreviations using a vowel deletion rule for one-syllable words and an acronym rule for multiple-word terms, as well as (2) allowing variable length truncations of words.

* Bell Laboratories. † American Bell.

I. INTRODUCTION*

Zipf[1] observed that in a number of languages frequently used words tend to be short. Further, this relation appears to be a causal one: as a word or phase increases in usage, it becomes shorter. Examples of word or phrase abbreviation are plentiful in English (e.g., "television" becomes "TV"). Abbreviating is particularly prevalent in government, corporations, and disciplines sharing a specialized vocabulary. Often the abbreviation process is so complete that the antecedent term is forgotten and only the abbreviation is retained (e.g., "cathode ray tube" becomes "CRT" or "light amplification by stimulated emission of radiation" becomes "laser"). Given the frequency and importance of this phenomenon, it seems reasonable to ask whether the process is systematic. That is, given a word or phrase, can one predict how it most likely will be abbreviated? If one could formulate abbreviation rules rooted in natural abbreviating behavior, would these abbreviations be "better" than abbreviations not suggested by behavioral data?

Discovering ways to generate abbreviations that are easy to learn and use has practical significance, since in many circumstances brevity is needed or desired. For instance, while infrequent users of a computer command language may be content to enter complete command names and arguments, frequent users often prefer terse command strings. To avoid typing long command strings, mechanisms should exist for abbreviating commands and their arguments.

There are some common ways to provide abbreviation facilities. One, truncation, allows the user to enter only as much of the command name (from the beginning to some point less than the whole word) as necessary to differentiate it from all other commands in the language. Another provides the user with standard system abbreviations. It is also possible to use both of these methods.

Selecting "good" abbreviations is a complicated problem. There is probably no single measure of an abbreviation's goodness, since there are some competing requirements to consider. For instance, one might want abbreviations in a system to be unique, easy to type, memorable, and natural (i.e., users who knew only the command would be more likely to produce this particular abbreviation than any other). However, some of these requirements are at odds with one another. Consider "easy to type" and "memorable." As Experiment II shows (see Section III), abbreviations produced by a simple vowel deletion rule are easy to remember or generate. Thus, given the command "journal,"

---

* Some of the material contained in this paper was presented at the 52nd Annual Meeting of the Midwest Psychological Association, St. Louis, May 1980 and the 52nd Meeting of the Eastern Psychological Association, New York City, April 1981.

subjects can generate the requisite vowel deletion abbreviation, "jrnl," with nearly perfect accuracy. However, this abbreviation is probably not easy to input, since the user must spell the whole word and omit every vowel. While we have no production data, it is likely that using a vowel deletion abbreviation rule for long words is no faster than entering the complete word. In fact, it may be slower in many cases.

Consider "easy to type" and "unique." Truncation produces abbreviations that seem easy to type (at least no more difficult to type than the whole word for a skilled typist and probably easier than the whole word for a nonexpert typist), particularly if one truncates to only a few letters. However, the likelihood of collisions among commands with truncation increases as the number of letters used decreases. If one requires uniqueness of abbreviations and has a truncation rule, the user must learn the minimum number of letters required to distinguish a particular command from all other commands. An additional burden may be imposed on the user if the command lexicon grows to include words whose truncations suddenly conflict with old abbreviations. If the lexicon is large enough, conflicts between abbreviations are bound to arise. One way to resolve conflicts in a rule-based system is to allow exceptions to the abbreviation rule. However, to what degree rule-generated systems can tolerate exceptions is an open issue worthy of future investigation. That is, at what point do exceptions to the rule render the rule worthless? It could well be that rules are useful only when there are relatively few or even no exceptions.

While each of these abbreviation requirements is important and critical from a design standpoint, the present study focuses on abbreviation naturalness and memorability. We examined abbreviations to determine how much regularity existed in those produced naturally. Next we incorporated the observed regularity into a set of abbreviation rules and compared the learning of rule-based abbreviations to the learning of the most frequently produced abbreviations for each word. The learning process was also reversed to include recalling of a word from its abbreviation, when the abbreviation was produced either by rules or by consensus.

II. EXPERIMENT I: REGULARITIES IN NATURALLY PRODUCED ABBREVIATIONS

2.1 Introduction

To determine how people abbreviate, we solicited abbreviations for a number of command terms. We then looked for regularities and the amount of agreement in these productions.

There were already data indicating that long and short words are abbreviated differently. Hodge and Pennington[2] examined how people

abbreviate words that varied in letter length. They observed that subjects used two principal mechanisms. One method, "contraction" (omitting word internal letters), tended to be used to abbreviate short words, whereas the other method, "truncation," was favored in abbreviating long words. However, they did not examine other linguistic properties of words associated with each of these strategies.

To examine abbreviations, we partitioned them into classes based on whether a particular abbreviation was an instance of a vowel deletion rule, a truncation rule, or an acronym formation rule. (Here acronyms are formed from multiple words, in which the first letter of each word is taken for the abbreviation; e.g., "cathode ray tube" becomes "CRT.") We examined whether the proportion of abbreviations accounted for by each rule varied with the number of syllables or words in the abbreviated term.

2.2 Method

2.2.1 Stimuli

The vocabulary consisted of 81 command names and arguments that were to be used in a large computer system. Many of these terms, such as move, copy, and insert, are common to other computer systems. Others were unique to this particular application, such as usage billing number. The items varied from one to three words. Table I gives statistics for the commands and arguments used in the present study in terms of number of words and syllables and average number of vowels and consonants.

2.2.2 Subjects

There were two groups of subjects; one group consisted of 19 psychologists doing human factors work at Bell Laboratories, Holmdel, New Jersey, while the other set consisted of 30 adult females living in the Holmdel area. The Bell Laboratories employees volunteered their time; the other subjects were paid for participating.

2.2.3 Procedure

We gave each group a randomized list of the command terms and asked them to produce "a good abbreviation" for each, that is, one that would be easy to use and remember. Because the command list grew, the human factors psychologists saw a subset of 42 terms, whereas the other subjects saw 81 terms. The paid subjects produced the abbreviations and then performed a 30-minute text-editing task. After the intervening task, these subjects recalled as many of the command terms as they could. There was no free recall task for the psychologists.

2.3 Results and discussion

To determine the consistency with which subjects produced abbreviations, we examined the abbreviation that was most frequently produced for each item. Overall, the concurrence across the 81 terms was 37 percent; that is, on the average the most commonly produced abbreviation for each item comprised 37 percent of the abbreviations. There were no discernible differences between the psychologists and local residents in their abbreviations of the terms. The mean percentage concurrence was identical to the nearest percent for the two groups. The groups did not differ on other production measures, such as number of different abbreviations produced or the most commonly produced abbreviation. Consequently, data for the two groups were pooled in subsequent analyses.

Table II shows the average concurrence across items and the standard deviation both for all terms combined and separately by number of syllables. Concurrence decreased markedly as the number of syllables in a word increased and, correspondingly, the number of different abbreviations given for each term increased substantially. Thus, ab-

Table I—Composition of command names and arguments

	Overall N = 81	One-Syllable Words N = 15	Two-Syllable Words N = 42	Three-Syllable Words N = 12	Four-Syllable Words N = 2	Multiple Words N = 10
Mean number of:						
Consonants	4.11	2.60	3.80	4.58	4.50	7.10
Vowels	2.77	1.67	2.55	3.17	4.50	4.50
Syllables	2.19	1.00	2.00	3.00	4.00	3.50
Words	—	—	—	—	—	2.10

Table II—Experiment 1: Abbreviation production

	Overall N = 81	One-Syllable Words N = 15	Two-Syllable Words N = 42	Three-Syllable Words N = 12	Four-Syllable Words N = 2	Multiple Words N = 10
Mean percentage concurrence	37%	56%	36%	32%	28%	24%
Standard deviation	(16%)					
Mean number of different abbreviations	11.95	5.73	11.13	16.00	18.00	18.90
Standard deviation	(5.68)					
Average number of terms recalled	3.25	4.87	3.20	2.50	2.00	2.20
Standard deviation	(3.15)					

breviation production was more homogeneous in simpler linguistic environments.

The one or two most frequently elicited abbreviations for each term were studied to determine whether abbreviating could be characterized by some "rule-governed" process. If one ignored the linguistic composition of the words, the abbreviation process appeared unorderly. However, if the terms were partitioned into three classes (monosyllabic, polysyllabic, and multiple words), there was some consistency within each class. By inspection there appeared to be three different rules operating in people's productions—vowel deletion (delete word internal vowels), truncation, and "acronym formation" (for multiple-word terms, select the first letter of each word).

We derived rules to account for the regularity observed in the abbreviating behavior. Table III shows the set of rules that appear to describe the data. The application of a particular rule depended on the number of syllables and/or number of words in the command term.

Table IV compares the proportion of abbreviations accounted for by the three different rules (truncation, vowel deletion, and the rule set from Table III) with the most frequently given abbreviation from each term ("popular"). We combined all truncations from the first letter to one less than the number of letters in a word. We defined vowel deletion as the deletion of all vowels following the first consonant in a word ("a," "e," "i," "o," "u," and "y," when it functioned as a vowel). It is important to remember in interpreting Table IV that popular abbreviations and those produced by the three rules are *not necessarily mutually exclusive.* For example, for one-syllable words there is a great deal of overlap between the rule set and vowel deletion. The rule set for one-syllable words is essentially a vowel deletion rule with only a few other features. However, there was no overlap between the truncated one-syllable abbreviations and the rule set or the vowel

Table III—Abbreviation rules

- Monosyllabic words:
  1. Take initial letter of the word and all subsequent consonants.
  2. Make adjacent double letters single.
  3. If more than four letters remain, retain the fifth letter if it is part of a functional cluster (such as *th, ch, sh, ph, ng*); otherwise, truncate from the right.
  4. Delete the fourth letter if it is silent in the word.
- Polysyllabic words:
  1. Take the entire first syllable.
  2. If second syllable starts with a consonant cluster, add it.
  3. If first syllable is a prefix (such as *de, re, in*) add the second syllable.
  4. Make final double consonants single.
  5. Truncate to four letters (but always retain entire first syllable).
- Multiple words:
  1. Retain the first letter of each word as the abbreviation.

Table IV—Proportion of abbreviations generated for each rule (variable length truncation, vowel deletion, rule set) compared with popular abbreviation for each term

	Truncation	Vowel Deletion	Rule	Popular
		Overall, N = 81		
Mean	0.286	0.216	0.283	0.373
Median	0.286	0.143	0.267	0.367
Standard error	0.022	0.023	0.022	0.017
		One Syllable, N = 15		
Mean	0.272	0.519	0.556	0.556
Median	0.233	0.533	0.533	0.533
Standard error	0.053	0.052	0.036	0.036
		Two Syllable, N = 42		
Mean	0.313	0.192	0.208	0.358
Median	0.330	0.173	0.200	0.333
Standard error	0.025	0.024	0.023	0.018
		Three and Four Syllable, N = 14		
Mean	0.408	0.077	0.247	0.317
Median	0.388	0.051	0.224	0.265
Standard error	0.040	0.020	0.056	0.088
		Multiple Words, N = 10		
Mean	0.020	0.024	0.237	0.237
Median	0	0.027	0.173	0.173
Standard error	0.012	0.022	0.038	0.038

deletion set. For two-syllable words all rule set abbreviations were a subset of truncation abbreviations. In the case of multiple words, where an acronym formation rule applied, there was no overlap among the truncation, vowel deletion, and rule set.

Overall, truncation and the rule set were not significantly different from each other, but each was different from the most popular abbreviation (truncation vs. popular, $t_{80} = 3.507, p < 0.001$; rule vs. popular, $t_{80} = 5.316, p < 0.001$). For one-syllable and multiple words, the rule set was more often the same as people's natural abbreviations than was truncation. However, for two-, three-, and four-syllable words, truncation surpassed the rule set abbreviations.*

It is worth considering the results of allowing variable length truncation and the rule set abbreviations to be used in the same system. If both were allowed for this word set, what proportion of generated abbreviations would be subsumed? In this case the proportion equals 0.415 overall, 0.828 for one-syllable words, 0.313 for two-syllable words,

* In some cases the results of detailed statistical analysis are not presented. We have in all cases given variability estimates for each cell mean. For the most part, the differences among experimental conditions in all experiments were substantial and large relative to the observed variance.

0.408 for three- and four-syllable words, and 0.257 for multiple words. Thus, it is clear that allowing multiple abbreviations accounts for significantly more of the naturally produced abbreviations than any single abbreviation mechanism (truncation vs. combined rules, $t_{80} = 5.512$, $p < 0.001$; rule set vs. combined rules, $t_{80} = 9.846$, $p < 0.001$).

III. EXPERIMENT II: LEARNING RULE-GENERATED VS. POPULAR ABBREVIATIONS

3.1 Introduction

In the second experiment we examined whether learning a set of abbreviations generated by the rules was better than learning those abbreviations most frequently given by the subjects in Experiment I. That is, does placing abbreviations in an internally consistent set result in more memorable abbreviations than people's natural abbreviations?

3.2 Method

3.2.1 Materials

The materials consisted of the 81 command terms and arguments from Experiment I and their abbreviations. There were two abbreviation conditions:

1. Rule Condition: The 81 abbreviations produced by the rule set shown in Table III.
2. Popular Condition: The most frequently given abbreviation for each of the 81 terms.

(Note that for 65 percent of the terms, the abbreviations were identical in the two conditions. That is, for the rule set 65 percent of the terms produced the same abbreviation as the one most frequently given by subjects in Experiment I.)

3.2.2 Subjects

We paid 44 high school students from the Murray Hill, New Jersey, area to participate. There were 23 subjects in the "rule" group and 21 subjects in the "popular" group.

3.2.3 Procedure

Subjects were randomly assigned to either the rule or the popular group. The task was paired-associate learning of the terms and their abbreviations. The rule subjects saw each term paired with the rule-generated abbreviation, whereas the popular group saw each term paired with the most frequently given abbreviation. Subjects in each group were given a randomized deck of 81 cards, each of which contained one term-abbreviation pair. Each subject received a different random order. Subjects studied each card for five seconds, and at the

sound of a buzzer, flipped to the next study card. During an intervening period of 20 minutes, subjects solved logic problems. Then they were given a list with only the terms and were asked to supply the abbreviations they had learned. If unsure, they were to make their best guess.

3.3 Results and discussion

Table V shows the proportion of correct responses for the two experimental groups for all words combined, separately for the abbreviations that were identical for the two groups, and for those that were different for the two groups. It also shows the standard error of the mean calculated across terms. The difference between the rule and popular conditions was highly reliable ($p < 0.001$) in all three cases. Thus, learning an internally consistent set of abbreviations facilitated later recall or generation. (Note that if subjects implicitly or explicitly knew the rules, they should have been able to generate the correct abbreviation for a given term.)

We performed additional analyses to determine in which linguistic environments the rules most facilitated performance. Table VI shows the mean proportion correct as a function of experimental group for one-, two-, three-, and four-syllable words, and multiple words.

The ability of the abbreviation rules to predict the popular abbreviations was far from uniform across the syllable/word classes. For one-syllable words and multiple-word terms, the rule and popular abbreviations were identical. While there was still a recall advantage for the rule abbreviations, it was small and not statistically reliable. However, this was not the case for the other three word classes. Evidently, the abbreviation rules for one-syllable and multiple words are reasonably straightforward and either tacitly known or easily learned by subjects.

The rules and natural productions were in only moderate agreement

Table V—Proportion of abbreviations correctly recalled and/or generated as a function of experimental group

	Rule	Popular
Overall (81 terms)	0.70	0.54
Standard error of the mean	0.02	0.03
Abbreviations same in two groups (53 terms)	0.76	0.63
Standard error	0.03	0.03
Abbreviation different in two groups (28 terms)	0.54	0.37
Standard error	0.04	0.04

Table VI—Proportion correct in rule and popular conditions as a function of the linguistic class of term

	One-Syllable Words N = 15	Two-Syllable Words N = 42	Three-Syllable Words N = 12	Four-Syllable Words N = 2	Multiple Words N = 10
Mean proportion correct					
Rule group	0.78	0.62	0.59	0.50	0.97
Popular group	0.73	0.42	0.45	0.44	0.91
Proportion of items for which rule abbreviation and popular abbreviation were the same	1.00	0.45	0.58	0.50	1.00

for two-, three-, and four-syllable words, and for these word classes, using the rules to produce abbreviations had the greatest facilitative effect on performance. One could presume that there is more variation in terms of linguistic composition for these classes. Owing to this complexity, people's internal abbreviation rules may be inconsistent, incomplete, and therefore, more variable. However, if rules are formulated for subjects, the subjects can use them to learn abbreviations.

3.4 Conclusions

The results and conclusions to be drawn from Experiments I and II are relatively straightforward. First, the process by which we produce abbreviations is not entirely idiosyncratic, but is to a large degree regular. We derived abbreviation principles that characterized a majority of the most frequently elicited abbreviations for each term. The abbreviation rule produced abbreviations that were easier to learn than the most frequently produced abbreviations. There was a substantial advantage of learning abbreviations in a mutually consistent set, even when the abbreviation being learned *was the same* as would be frequently given by subjects. However, the generality of these results is limited, since the rules were derived from the same set of words that served as stimuli in the recall.

IV. EXPERIMENT III: ABBREVIATION PRODUCTION AND LEARNING OF RANDOM WORDS

4.1 Introduction

To test the generalizability of the rule set used in the previous experiment, we applied the rules to a random set of words selected to have the linguistic properties considered relevant in the rule set. A replication was needed, since it is possible that the terms used in the

command set studied in Experiments I and II have properties that differ in some unknown way from ordinary English words. A priori, it is conceivable that the command name words might be different semantically, phonetically, and grammatically from "ordinary" English words. Thus, to know whether rule-based abbreviations are better in general, the test words should be a random sample of words. Also, since the rules were generated from examining subjects' abbreviations of these words, the rule set may include conditions that are the result of idiosyncracies in the particular commands selected.

4.2 Method

4.2.1 Word selection

A total of 200 English nouns and verbs were randomly selected from Kucera and Francis[3]. Words were either one ($N = 40$), two ($N = 80$), or three ($N = 80$) syllables. For the two- and three-syllable words, 40 had prefixes (e.g., "de," "dis," "pro," "post," and "mis"), while 40 did not.

4.2.2 Production subjects

Thirty-two paid Rutgers University undergraduates supplied abbreviations for the random words.

4.2.3 Procedure

The word set was randomly divided in half with half of the subjects supplying abbreviations for 100 words.

4.3 Production results

We analyzed the data in the same way as in Experiment I. Tables VII and VIII show the proportion of abbreviation productions accounted for by variable length truncation, vowel deletion, the rule set, and the most frequently given abbreviation (popular). Table VII shows proportions for all words combined and separately for one-, two-, and three-syllable words. Note that the proportions in Table VII are reasonably close to those found in Experiment I.

Table VIII compares two- and three-syllable words with and without prefixes. The existence of a prefix decreased abbreviation agreement in the popular condition ($t_{79} = 2.790, p < 0.01$). Thus, prefixes affected production. However, it does not appear that prefix conditions in the rule set (as described in Table III) managed to capture any of the differences between words with and without prefixes. Abbreviation behavior was best represented by a truncation rule irrespective of whether the word had a prefix.

Table VII—Abbreviations generated for randomly selected English words

	Truncation	Vowel Deletion	Rule Set	Popular
		All words, N = 200		
Mean	0.396	0.234	0.304	0.406
Median	0.375	0.188	0.312	0.312
Standard error	0.013	0.014	0.015	0.015
		One-Syllable Words, N = 40		
Mean	0.228	0.483	0.473	0.512
Median	0.188	0.500	0.469	0.500
Standard error	0.027	0.033	0.034	0.027
		Two-Syllable Words, N = 80		
Mean	0.382	0.245	0.274	0.374
Median	0.375	0.250	0.281	0.375
Standard error	0.018	0.017	0.019	0.015
		Three-Syllable Words, N = 80		
Mean	0.493	0.099	0.250	0.383
Median	0.500	0.062	0.188	0.375
Standard error	0.019	0.008	0.024	0.017

V. LEARNING RULE-BASED AND POPULAR ABBREVIATIONS FOR RANDOM WORDS

5.1 Method

5.1.1 Subjects

A total of 90 paid adult recruits from the Holmdel area participated in the experiment.

5.1.2 Procedure

The procedure was identical to that reported in Experiment II with the exception that the interpolated task was underlining "important concepts" in a text-editing manual. There were three conditions. Thus, in separate conditions, each word was paired with (1) its rule set abbreviation, (2) its popular abbreviation, and (3) its vowel deletion abbreviation. There were 30 subjects in each of these three conditions.

5.2 Learning results

Table IX shows the proportions and standard errors across words for each of the three groups. In each row each condition was significantly different from every other condition. (The largest matched-paired t probability of occurrence was less than 0.02. However, this probability value is uncorrected, i.e., it does not take into account the number of post comparisons.) Thus, the ordering of generation or recall performance was best for the simple vowel deletion rule, second best for the rule set condition, and worst for the popular abbreviations.

Thus, the original results that rule-governed abbreviations were

Table VIII—Abbreviations generated for randomly selected two- and three-syllable words with and without prefixes

	Truncation	Vowel Deletion	Rule Set	Popular
		Two-Syllable Words, No Prefix (N = 40)		
Mean	0.425	0.259	0.326	0.411
Median	0.438	0.250	0.312	0.375
Standard error	0.028	0.025	0.030	0.021
		Two-Syllable Words, Prefix (N = 40)		
Mean	0.339	0.231	0.222	0.338
Median	0.312	0.188	0.250	0.312
Standard error	0.020	0.024	0.023	0.021
		Three-Syllable Words, No Prefix (N = 40)		
Mean	0.483	0.109	0.308	0.406
Median	0.500	0.094	0.344	0.375
Standard error	0.029	0.012	0.037	0.024
		Three-Syllable Words, Prefix (N = 40)		
Mean	0.503	0.089	0.192	0.359
Median	0.500	0.062	0.125	0.312
Standard error	0.024	0.012	0.029	0.025

better reproduced than abbreviations produced by consensus were strongly replicated. Note that a simple rule produced the best performance in this task. The vowel deletion rule appears to be a particularly easy rule for subjects to abstract. Thus, performance improved as the rules governing abbreviations became more straightforward.

VI. EXPERIMENT IV: RECALLING THE SOURCE WORD GIVEN ITS ABBREVIATION

6.1 Introduction

One of the anecdotal observations frequently made about new computer users is that they learn abbreviations for command terms without ever learning what some of the abbreviations mean. In this sense, the abbreviations represent an alternate name space. There is usually some care given to the choice of names for commands, devices, programs, etc, the assumption being that the use of appropriate names tends to make learning easier (but see Landauer, Galotti, and Hartwell[4], where random words assigned to text-editing functions were no easier or harder to learn or use than appropriate names). However, this care is not usually taken in choosing abbreviations. Most people's goal is to choose abbreviations that are easy to remember, given that the user remembers the right name in the first place. In this experiment we were interested in the decoding process: determining how well the abbreviations represented the words they abbreviated.

There is evidence that the decoding and encoding (determining an abbreviation, given a source word) are asymmetric processes. In par-

Table IX—Mean proportion of abbreviations recalled/produced correctly, given words

	Popular	Rule Set	Vowel Deletion
Overall (200)	0.466	0.650	0.811
Standard error	0.015	0.014	0.006
1-syllable (40)	0.579	0.634	0.875
Standard error	0.029	0.023	0.011
2-syllable (80)	0.428	0.622	0.815
Standard error	0.031	0.035	0.013
3-syllable (80)	0.446	0.685	0.774
Standard error	0.039	0.032	0.011

ticular, rule-based abbreviation schemes are not as effective for decoding purposes as for encoding purposes. With a well-formed rule it is possible to produce the abbreviation accurately when given the word, but the converse is not true. That is, knowing the abbreviation rule and the abbreviation does not necessarily allow one to generate the correct source word.

Rogers and Moeller[5] compared decoding of conventional Navy sonar abbreviations and rule-based abbreviations (produced by truncation). (The subjects in their experiments were sonar operators.) Reaction time to decode the abbreviation was measured. When the scoring criteria were strict, the rule-based abbreviations were worse than the conventional military abbreviations. The difference was due primarily to conventional abbreviations giving a better indication of the endings of words.

Using two 20-word lexicons, Hirsh-Pasek, Niedelman, and Schneider[6] examined decoding in a wide variety of conditions: truncation to four letters, vowel deletion with truncation to four letters, minimum number of letters to distinguish, a phonics system, and user-supplied abbreviations. Vowel deletion and the phonics system produced the fewest number of errors, whereas minimum to distinguish produced the most. Truncation was only slightly worse in terms of errors (perhaps not significantly) than vowel deletion and phonics.

Ehrenreich and Porcu[7] compared decoding of abbreviations formed by vowel deletion and by truncation. There were no reliable differences between these two methods with liberal scoring procedures—65 percent and 57 percent correct, respectively. With stricter scoring criteria, truncation abbreviations were easier to decode (54 percent) than vowel deletion abbreviations (48 percent).

6.2 Method

6.2.1 Subjects

We paid 90 residents from the Holmdel area to participate.

6.2.2 Procedure

We prepared cards using the 200 random English words used in Experiment III. One word and its abbreviation appeared on each card. There were three groups: one group learned popular abbreviations, another the rule set, and a third, the vowel deletion abbreviations. The procedure was the same as in previous experiments, except that after the 20-minute retention interval, subjects were given lists of *abbreviations* and asked to *write down the word* each represented.

6.3 Results and discussion

Table X compares the proportion of words recalled correctly in each of the three groups. Note that in Experiments II and III subjects recalled or produced abbreviations formed by the rule set more often than subjects who learned abbreviations arrived at by consensus. We attributed this improvement in performance to the fact that the existence of the rule lent some organization to the set of stimuli being learned. When we looked at ability to recall what the abbreviations represent, however, it appeared that the rule set seemed to interfere with subjects' ability to recall what the abbreviations stood for. Thus, there is an asymmetry between recalling the abbreviation from seeing the command term and recalling the command term from seeing its abbreviation. The superiority of the vowel deletion rule in this task is evident for all but the one-syllable words, and for this case the rule set is essentially a vowel deletion rule.

These results are somewhat at odds with those of Ehrenreich and Porcu.[7] First, in our experiment, the vowel deletion condition produced very high performance. Also, the overall level of performance was much higher in our study (79 percent here vs. 48 percent in the Ehrenreich and Porcu study). Ehrenreich and Porcu claim that when subjects produce words from vowel deletion abbreviations they are apt

Table X—Mean proportion of words recalled/produced correctly, given abbreviations

	Popular	Rule Set	Vowel Deletion
Overall (200)	0.626	0.467	0.788
Standard error	0.022	0.024	0.018
1-syllable (40)	0.662	0.635	0.641
Standard error	0.046	0.050	0.052
2-syllable (80)	0.628	0.500	0.799
Standard error	0.037	0.038	0.027
3-syllable (80)	0.605	0.350	0.850
Standard error	0.035	0.034	0.024

to make spelling errors. People who misspell a word are more likely to err on a vowel than a consonant. Their terms were Army terms—ours random words; their subjects were Army personnel—ours primarily spouses of Bell Laboratories employees. These differences may have produced the resulting performance differences.

It would seem that vowel deletion should have produced the best results, since consonants carry more information than vowels and the abbreviations in this condition were longer than in the other two conditions (vowel deletion = 4.43 letters, standard deviation = 1.13; rule condition = 3.38 letters, standard deviation = 0.62; popular = 3.94 letters, standard deviation = 1.08). Thus, the results vary directly with the number of letters in the abbreviation alone.

In summary, there was a large asymmetry between the encoding results for the popular and rule condition in Experiment III and the decoding results in Experiment IV. Rule abbreviations were better than popular for remembering the abbreviation given the word, whereas popular abbreviations were better than rules for remembering the word given the abbreviation.

VII. GENERAL DISCUSSION

While there is much variability in people's natural abbreviations, the process is far from random. People are most likely to abbreviate one-syllable words by deleting word-internal vowels, and multiword terms by deleting everything except the first letter of each word. Abbreviations of polysyllabic words are best characterized by truncation, that is, deleting letters from the end of the word. While these rules accounted for the majority of the most frequently generated abbreviations for each term, the proportion of all abbreviations covered by these rules was only about 0.30 in the two abbreviation production studies. Thus, there is much that people do when abbreviating that is idiosyncratic, variable, and not particularly amenable to rule-based descriptions.

We showed that if abbreviations are formed by applying rules derived from people's behavior, subjects can use these principles. Thus, subjects make use of the internal consistency of the set to learn the abbreviations. Performance with a set of rule-based abbreviations was better than with the corresponding set of most frequently produced abbreviations even when the two abbreviations were identical in the two conditions.

That people extract principles in complex stimuli is not a new finding. For instance Reber,[8,9,10] among others, has demonstrated that people's behavior when learning artificial languages in an experimental setting indicates that they have "learned" the grammatical rules of the language. However, very often they are unable to verbalize these rules

and seem unaware that they have extracted principles. It appears that rule formulation is often an automatic and unintentional process.

Throughout, we have explicitly assumed that rules based on behavior are better than arbitrary rules for abbreviation. One could argue that our own data contradict this assumption. That is, the vowel deletion rule used in Experiments III and IV produced the best performance in two tasks—recalling the abbreviation given the word and conversely recalling the word given the abbreviation. Don't the data then argue that one should abbreviate using vowel deletion? Yes, if the abbreviation is used to recall the term, that is, a decoding task. However, the ability to construct a correct abbreviation, given the word, is only one aspect of a good abbreviation. Another aspect that "abbreviate" itself denotes is that it should be short. Abbreviations produced by vowel deletion increase in length as the length of the source word increases, whereas abbreviations that people spontaneously produce and the ones produced by the rule set better maintain length constancy [e.g., a one-syllable word on the average is abbreviated to 3.4 letters in all conditions, whereas a three-syllable word is abbreviated to 5.2 letters (vowel deletion), 4.3 letters (popular), and 3.3 letters (rule)]. Thus, in terms of total number of keystrokes, the simple vowel deletion rule is least efficient.

A more serious failing for simple vowel deletion is that forming the abbreviation requires first producing the whole word, then systematically deleting the vowels, and finally outputting the remaining consonants. We suspect that this is difficult for long words. When a word is short, the entire word can be dealt with as a "chunk." For a single chunk it is probably not difficult to remove a vowel or two. Beyond simple, short words, the amount of mental bookkeeping increases. If one's goal is to enter command strings *quickly*, vowel deletion is probably not a serious candidate.

Truncation, on the other hand, has properties that make it the best abbreviation mechanism for polysyllabic words. Resulting abbreviations are short and easy to produce—the user does not need to pervert the assumed normal output strategies, just cut them short. It is a simple rule to teach[5–7] and users are most likely to abbreviate multisyllabic words in this way. Thus, it seems to have most of the desirable properties of an ideal abbreviation mechanism—natural, short abbreviations, easy to remember abbreviations, and easy to produce abbreviations. The one thing that may be sacrificed is uniqueness.

There have been suggestions on how to maintain uniqueness of abbreviations, which we now discuss. Many of the recommendations require the user to be flexible and relearn abbreviations from time to time, and some recommendations undermine the consistency of a rule-based set. Ehrenreich and Porcu[7] discuss two ways to resolve conflicts

among abbreviations: (1) use of alternate rules and (2) minimum number of characters to distinguish.

Using an alternate rule when conflicts exist has some serious problems. First, the user must know that there are two or more rules, and that one words most of the time, but when it doesn't, to use the alternate rules. Furthermore, the user has to know which one of the conflicting words requires which rule. The user doesn't know in advance when the primary rule will not work and, consequently, will probably have to learn by trial and error. When the number of exceptions to the primary rule is large, we suspect that the advantages gained by selecting abbreviations by rule will disappear. That is, the mixed rule set performance may approach or may even be worse than performance on abbreviations selected by consensus. However, the degree to which rule-based systems tolerate corruption warrants investigation.

In the case of minimum to distinguish, the user truncates to the number of characters required to differentiate that word for all other admissible words. For example, for "transport" and "transfer," the minimum to distinguish is in each case six characters—"transp" and "transf." There are a few problems worth noting with respect to minimum to distinguish. First, different words require different numbers of letters. Hirsch-Pasek et al.[8] have found in a paired-associates task that minimum to distinguish was more difficult to learn than the other abbreviation schemes they studied; truncation to four letters was the easiest. This may be only a relatively minor problem, which depends on the terms of the command set and the size of the lexicon. It could be dealt with by telling the user the minimum number of letters needed to distinguish all command names. However, a fixed minimum based on all commands will be unnecessarily long for many commands. The user could of course be told that fewer keystrokes will often suffice and learn the minimum to distinguish on a command-by-command basis. The second potential shortcoming occurs when new commands that produce collisions are introduced into the language. Truncations that worked previously are now ambiguous and require relearning. However, introducing software that informs the user of a conflict could solve these problems. If the input is ambiguous, the system could present the user with the collisions and ask the user to select one of the alternatives.

One way to minimize the likelihood of collisions is to make contextual information available to the system. For example, there may exist three words—"debug," "delete," and "define." If collisions were considered on the basis of the entire lexicon, three letters would be needed to differentiate these words. However, if "define" were a command while "debug" and "delete" were command options that occurred in

different environments, it could be the case that a "d" would suffice for either "debug" or "delete." Thus, if the command/option tables are structured hierarchically, such that a command points only to options meaningful in the context of that command, the potential range of conflicts is reduced from the entire command language to a small subset of the language.

In summary, considering all available evidence on producing abbreviations given command terms (encoding), truncation appears to be the best single abbreviation scheme. Truncation also best captures people's natural abbreviations in all environments except two—monosyllablic words and multiple-word terms. In these cases, we recommend using vowel deletion for the former and acronym formation for the latter. If, on the other hand, one's task requires generating full names, given abbreviations (decoding), vowel deletion abbreviations are better than other rule-based schemes.

REFERENCES

1. G. Zipf, *The Psycho-biology of Language*, Boston: Houghton Mifflin, 1935.
2. M. Hodge and F. M. Pennington, "Some Studies of Word Abbreviation Behavior," J. Exp. Psychol., *98*, No. 2 (1973), pp 350–61.
3. H. Kucera and W. N. Francis, *Computational Analysis of Present-Day American English*, Providence, Rhode Island: Brown University Press, 1967.
4. T. K. Landauer, K. M. Galotti, and S. Hartwell, "Natural command names and initial learning: A study of text editing terms," Commun. ACM, *26*, No. 7 (July 1983).
5. W. H. Rogers and G. Moeller, "Evaluation of Abbreviations in Sonar Displays," 52nd Annual Meeting of the Eastern Psychological Association, New York City, April 24, 1981.
6. K. Hirsh-Pasek, S. Niedelman, and M. L. Schneider, private communication.
7. S. L. Ehrenreich and T. A. Porcu, "Abbreviations for Automated Systems: Teaching Operators the Rules," *Directions in Human Computer Interaction*, A. Badre and B. Shneiderman (Eds), Norwood, NJ: Ablex Publishing Company, 1982, pp. 111–135.
8. A. S. Reber, "Implicit Learning of Artificial Grammars," J Verb. Lear. Verb. Behav., *6* (1967), pp. 855–63.
9. A. S. Reber, "Transfer of Syntactic Structure in Synthetic Languages," J. Exp. Psychol., *81*, No. 1 (1969), pp. 115–9.
10. Reber, A. S., "Implicit Learning of Synthetic Language: The Role of Instructional Set," J. Exp. Psych., Human Learning and Memory, *2*, No. 1 (1976), pp. 88–94.

AUTHORS

Lynn A. Streeter, B.A. (Psychology), 1969, University of Michigan; Ph.D. (Psychology), 1974, Columbia University; Bell Laboratories, 1969—. Ms. Streeter first joined Bell Laboratories in 1969 as a member of the then Human Factors Department in Holmdel. She rejoined Bell Laboratories in 1974 as a member of the Linguistics and Speech Analysis Department. The abbreviation data were collected when she was Supervisor of the Human Factors Engineering Group for the Advanced Communications Service project. Currently, she is working on mobile telephony interface design problems in the Communications Methods Research Department.

John M. Ackroff, B.A. (Psychology), 1971, Williams College; M.S. (Experimental Psychology) 1975, Ph.D. (Experimental Psychology) 1981, University

of Wisconsin–Milwaukee; Bell Laboratories, 1977–1982; American Bell, 1982—. While at Bell Laboratories, Mr. Ackroff worked on the Advanced Communications Service project, where he was responsible for the design and specification of the command language aspects of the user interface. At American Bell, he is responsible for developing an enhanced version of the command language parser.

Glen A. Taylor, B.S. (Psychology), 1969, Louisiana State University; M.A. (Experimental Psychology), 1975, Ph.D. (Experimental Psychology), 1977, University of Kansas; Bell Laboratories, 1979–1982; American Bell, 1982—. Mr. Taylor joined Bell Laboratories as a member of the Human Factors Engineering Group for the Advanced Communications Service project (now AIS/Net 1000 service). He was responsible for designing the user interface for the Net 1000 editor and for investigating various aspects of computerized electronic mail services. In 1981, he became a member of the User/Station Interface Software Group responsible for the development of on-line documentation retrieval software.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Designing and Evaluating Standard Instructions for Public Telephones

By C. J. KARHAN,* C. A. RILEY,† and M. S. SCHOEFFLER†

(Manuscript received July 30, 1982)

Difficulty finding and understanding information on public telephone instructions has led to problems using public telephones. To overcome these problems a new instruction card was designed. In this paper we discuss the information needs of the public telephone user, a conceptual solution to information presentation, a new standard instruction card design, and laboratory and field evaluations of the new design.

I. INTRODUCTION

Many people have reported that placing a call from a public (pay) telephone is a problem. Individuals and consumer groups have complained to their telephone companies and even their public utility commissions. A major reason for this problem is the difficulty of finding and using information on public telephone instruction cards. In this paper we describe a series of studies conducted to develop standard instructions for public telephones that are easier to use.

1.1 The public telephone operating environment

If all public telephones worked the same way, the task of designing

* Bell Laboratories. † American Bell.

instructions would be easy. However, there is considerable variation in the operating procedures of public telephones in the Bell System. In addition, new public telephone services are being introduced or will be introduced soon. For example, at some public telephones the user must deposit coins equal to the cost of a local call before dialing any call (coin-first telephones). From the user's perspective, the coins work like a switch to turn on the telephone, and they are returned if no payment is required for the call. At other telephones no coins are needed to turn on the telephone, and a large class of calls, such as collect, credit card, emergency, and operator assistance, can be dialed without depositing any coins (dial-tone-first telephones). On some other telephones coins cannot be used at all (Charge-a-Call telephones). At most telephones, coins are deposited before a local call is dialed, but at some telephones coins are deposited after the person at the other end answers (post-pay telephones). From some telephones you can dial long distance calls; from others the operator must dial for you. From some telephones you can dial in your credit card number; from others you must tell your credit card number to the operator. In some areas you must dial "1" before long distance station-to-station calls; others require no prefix. Furthermore, there are many variations in the numbers for emergency help, directory assistance, and repair service.

Because of these variations, no single instruction card can be used for all public telephones. More importantly, because of these variations, even the most experienced public telephone users must sometimes consult the instructions to place a call correctly.

1.2 The need for standard instructions

Prior to the standard developed from these studies, there was no standard for instruction card design in the Bell System. Various operating telephone companies have used different layouts, word choices, and color schemes to convey the same information. There are also variations within single telephone companies on the format and type of information included. Generally, telephones that operate differently, if only in minor respects, have instructions that vary in many ways that have nothing to do with operational differences. For example, instructions on how to place a long distance call may be different on a telephone where the initial deposit is 10 cents than on one where a 20-cent deposit is required, even though the dialing procedures are identical. There are also tremendous differences in the emphasis (location, color, and type size) placed on particular pieces of information. Finally, information is often provided in "telephonese."

The problem for users is not that all instruction cards are poorly

designed. While instruction cards in some locations may be particularly difficult to read and understand, the biggest problem for users is the inconsistency of design.

The use of a standard instruction card design for all public telephones should significantly reduce the problems of finding and understanding needed information. While identical instructions cannot be placed on all telephones because of differences in operating procedures, consistent language, graphics, and placement of information can be used. The information provided on the standard instruction card should match the information needs of the user. The instructions must be understandable, and the design should be attractive. A good design will help users; and, given a good design, a greater benefit for users will come from the universal application of this design.

1.3 Approach to the problem

The new instruction card design and the studies that supported the development of that design are described in the sections that follow. First, we studied public telephone users to identify their information needs, how they used instructions, and the telephone terminology they used and understood. Based on this information we formed a model of the instruction user. Second, we developed a conceptual solution to the problem. Third, we presented this solution (along with detailed information about public telephone operating procedures) to graphic designers at Henry Dreyfuss Associates, who created designs for the instruction cards. Fourth, we evaluated and refined the designs in several phases. First we used our own and the designer's judgment, and then conducted a laboratory test, and, finally, a field test. During all phases of the project we consulted with the AT&T Public Services Marketing organization.

II. THE INFORMATION USER

Before any attempts were made to design new instruction cards, we analyzed the information needs of the public telephone user.

2.1 Interview study

A total of 500 users were interviewed in Arizona, Florida, and New Jersey. The individuals interviewed were approached following the completion of a call from a public telephone. They were asked to participate in a study of public telephone service. The individuals who agreed to participate were asked a series of questions about instruction use, such as: "Did you read the instruction card?" "Did you look at the instruction card?" "What information did you need?" "Did you find the information you needed?" They also were asked if they had

any difficulty finding or understanding information, and were asked to define several telephone terms, such as "collect call" and "dial-tone-first telephone."

Few of the people interviewed (about 10 percent) reported that they read the instruction cards. These individuals looked for procedural information necessary to place a specific call. This information included the cost of the initial deposit, when to deposit money, how to dial local calls and various types of long distance calls, and specific numbers, such as directory assistance or repair service.

Some respondents reported that they were unable to find information that they needed on the cards. In almost all of these cases, the individual was looking for more specific information than could ever be placed on the instruction card, such as the area code for Los Angeles, or the cost of a particular long distance call, or the telephone number of the local veterinarian.

Our interviews showed that people generally approach the instructions with particular questions in mind, and that the questions have two general forms. The first form is one in which a specific fact or procedure is needed. Examples are "How much does a local call cost?" or "How do I place a collect call?" The second form of question is asked by the more sophisticated telephone user who is already familiar with alternative operating procedures. This user might ask, "I know there are two kinds of telephones, one where I can dial long distance calls myself, and one where the operator dials all long distance calls; which type of telephone is this one?"

A few people look at the instructions without needing any specific information. These users may also be characterized, however, as having a question in mind before reading (looking at) the instructions, which can be phrased as, "Is there anything (new) here that I need to know about?"

Interview questions about terminology showed that telephone users generally understand terms referring to types of calls, such as collect, credit card, person-to-person, and station-to-station. However, language on instruction cards referring to the type of telephone, such as "dial-tone-first" telephone, was not understood.

III. THE CONCEPTUAL SOLUTION

Most people do not read instructions on telephones; in fact, there are few circumstances in which anyone reads any instructions at all. (See Ref. 1 for a discussion of this issue.) Our goal was not to induce more people to read the instructions, but rather to make the instructions easier and more effective when they were read.

Our conceptual solution to instruction card design had several

components to simplify searching and understanding. The design should:

1. Provide the information that people need in words they understand.
2. Always place information about a particular aspect of a telephone's operating procedure in the same spatial location on the card. The card is thus like a form, with fields for each type of information. If there is no relevant information to be placed in the field, it should be left blank and not used for some other information.
3. Place frequently needed information on the upper card and detailed procedural information on the lower card.
4. Divide detailed information into related categories. Use the category labels to construct an index.
5. Provide visually obvious cues that are redundant with the text, e.g., colors, pictograms, and other symbols. This information should be visible at a glance and allow knowledgeable telephone users to answer the question "What type of telephone is this?" without reading the text. By systematically varying the visual characteristics of the instruction cards, telephones that have different operating procedures will have instructions that *look* different. Instruction cards on all telephones that operate the same way will look the same. Thus, information obtained at a glance can be informative. If the card "looks" different from cards the user is familiar with, then the instructions should be read.
6. As a final goal provide as much information as possible in a language-independent form to make the cards more usable by people who either do not read English or do so with difficulty.

IV. THE NEW STANDARD DESIGN

Figures 1 to 3 show examples of the new instruction cards, designed according to the principles described in the section above, and revised based on the results of the study described in the following sections. (Space constraints prohibit illustration of the cards through all the design iterations.) Figure 1 is an instruction card for a telephone that requires a coin deposit before any calls can be dialed, and from which operator assistance is required for all long distance calls. Figure 2 is for a telephone from which credit card, collect, emergency, and other "free" calls can be made without depositing any coins; long distance calls can be dialed by the caller. Figure 3 shows an instruction card from a "Charge-a-Call" telephone, from which no coin-paid calls can be made.

4.1 Design features

The instructions shown in Figs. 1 through 3 are for telephones that

Local calls Deposit 10¢ before dialing
Long Distance Dial 0
Operator will handle all Long Distance calls

SOS dial 0 for Emergency help

KEENE, N H -EXP 5

Charge and Person-to-Person	**Credit Card, Collect & Person-to-Person calls** . . . . . Operator
Station-to-Station calls	Local numbers beginning with: 239, 242, 352, 363, 399, 446, 563, 585, 756, 827, 835, 847, 876 . . . . . Number All other numbers . . . . . Operator
Free calls	**Directory Assistance** . . . . . Operator **Toll Free 800 Numbers** . . . . . Operator

AREA TYPE 2-EXP –5 Operator assisted rates apply to all toll calls from this telephone.

Fig. 1—Instruction card set for a telephone that requires a coin deposit before any call can be dialed, and from which the operator dials all long distance calls. Colored areas are tan.

differ considerably in operation; however, the common organization of the instructions is easily observed. Differences in visual appearance are also readily apparent. The features of the design are listed below:

1. Color: All instructions are printed on a white background; detailed instructions are in black type and emergency information is in red. Other colored sections—including the band across the top of the upper card, the "0+" and "1+" symbols, index headings and rules on the lower card—are color coded based on the basic operational procedure of the telephone. Telephones that require a coin deposit before any call can be dialed (coin-first) have tan instruction card coloring; telephones that do not require a coin for initial activation (dial-tone-first, including post-pay and Charge-a-Call) have blue instruction card coloring.

2. Upper-left symbols: Coin-first telephones have a white handset in a tan disk (coin); dial-tone-first telephones have a blue handset on a white background. The handset on the Charge-a-Call telephone has

Local calls Deposit 20¢ before dialing
Long Distance Dial all calls directly
0+ needed for Charge & Person-to-Person calls
1+ needed for Station-to-Station and Free calls

SOS dial 0 for Emergency help
SOS marque 0 para Emergencia

TSPS/DTF UPPER-EXP – 11

Charge and Person-to-Person calls	**Credit Card, Collect & Person-to-Person**	
	Within this Area Code	0+ Number
	Outside this Area Code	0+ Area Code + Number
Station-to-Station calls	Within this Area Code	1+ Number
	Outside this Area Code	1+ Area Code + Number
Free calls	**Directory Assistance**	
	Local	411
	Within this Area Code	1+ 555-1212
	Outside this Area Code	1+ Area Code + 555-1212
	Repair Service	611
	Toll Free 800 Numbers	1+ 800 + Number

TSPS/DTF STERLING LOWER-EXP – 11 Operator assisted rates apply to all toll calls from this telephone.

Fig. 2—Instruction card set for a telephone from which credit card, collect, emergency, and other "free" calls can be made without depositing any coins, and from which long distance calls can be dailed by the caller.

a broken horizontal stripe, representing credit card and speed. The functional meaning of these symbols is explained in the text immediately to the right of the symbol.

3. Dial tone/deposit sequence pictograms: The pictograms in the upper right corner illustrate the sequence for listening for dial tone (depicted as a handset with waveform) and coin deposit. Numbers, as well as spatial position, show the sequence. The post-pay instruction card (not shown) has the additional modifier "Deposit only after answer" between the handset and the coin deposit. The disk representing the coin also contains the initial deposit rate (cost of a local call).

4. Summary instructions: Summary instructions for placing local and long distance calls are located directly below the colored band (for the two-card design). In the long distance case, these instructions are intended to be sufficient for the knowledgeable user to identify the

Fig. 3—Instruction card from a "Charge-a-Call" telephone.

general procedure. The "0+" and "1+" represent the prefixes required before dialing certain types of calls, when needed. The colored word "Operator" indicates that long distance calls cannot be dialed by the caller. Experienced users should be able to recognize the symbols at a glance, and identify the procedure without reading the instructions.

5. Emergency: "SOS" is the internationally recognized emergency symbol, and it is used to identify the number to dial for emergency assistance. Its red color makes it very salient. The use of a unique identifying color makes possible useful public relations messages in non-English-speaking communities. For example, the Chinese press in New York could carry a message stating (in Chinese) "If you need emergency help, dial the red number." Where dual language instructions are required by regulation, the SOS instruction can be repeated in the second language. Where separate numbers for fire, police, and ambulance are required, the SOS is supplemented with specific symbols (flames, a shield, and a cross) paired with each number.

6. Local instructions: The space directly under the dial tone/deposit symbols is reserved for locally varying instructions. Operating companies may use this space for special instructions or to publicize a new service.

7. Lower card: Specific dialing instructions and service numbers are listed on the lower instruction card. The major sections are divided by a colored rule, and each section has an index heading to simplify searching.

4.2 Preliminary design differences

The instructions shown in Figs. 1 to 3 differ in several respects from the instructions used in the laboratory evaluation. Changes were made

based on the findings of that study, as we discuss in Section V. Briefly, the earlier designs differed from the new designs as follows: (1) There was no summary instruction for local calls; (2) the word "Operator" was not in color on the upper card for operator-dialed long distance calls; (3) all operator-dialed calls were combined into a category called "All other calls" on the lower instruction card; and (4) the local directory assistance number was listed in the same line as the heading "Directory Assistance" when that number was different from the "Within this area code" number. In addition, the handset with dial-tone pictogram and Charge-a-Call symbol were modified between the laboratory test phase and the field test.

V. LABORATORY EVALUATION

5.1 Study goals

The instruction cards were designed to satisfy two major goals. First, users should be able to search successfully and efficiently for the information they need. Second, knowledgeable users should be able to tell at a glance what services and procedures are available on a given telephone. The major purpose of the laboratory evaluation was to determine whether these design goals were met. We used an information search task to assess the new designs and identify needed improvements. The following questions were addressed:

1. Could users find the answers to questions about telephone operation?
2. Did users understand the information provided?
3. Were users able to find information quickly after some experience using standard instructions?
4. Did users understand the symbols and pictograms used on the instructions?
5. Did users infer the meaning of the redundant visual codes (e.g., color) that would allow them to get information at a glance?

In addition, several different formats were evaluated.

5.2 Method

5.2.1 Materials

Three formats for the upper instruction card and two formats for the lower card were used. The same content and design elements described in the preceding section were used in each format. The only differences between the formats were in the arrangement of the information. The formats of the three upper cards were:

1. Horizontal: the design illustrated in Figs. 1 through 3.
2. Two-panel: the information contained in the stripe on the horizontal design was placed in a square patch on the left third of the

card; the other information was listed on the remaining two thirds of the card.

3. Three-panel: the left third was the same as in the two-panel design; the remaining space was divided into two equal sections by a blue or tan vertical bar. The center section contained long distance dialing and local information, and the right-hand section contained emergency information.

The formats of the two lower cards were:

1. Ruled: illustrated in Figs. 1 through 3.

2. Unruled: the colored lines separating categories were removed, the colored headings were placed directly above the listings for that category, and the listings were centered on the card.

Instructions for eight different telephones were created. The operating procedures and available services varied on the eight telephones. Three of the telephones had coin-first service, two had dial-tone-first, one had dial-tone-first/post-pay, and two had Charge-a-Call. Other operating differences, such as initial deposit rate, long distance dialing procedures, emergency numbers, directory assistance numbers (local and long distance), and repair numbers were systematically varied across the eight telephones. Only combinations of services and procedures that actually exist in the operating telephone environment were used. A set of instruction cards was constructed in each of the formats described above for each of the eight telephones. (For this study, upper and lower housing cards were designed for Charge-a-Call.)

5.2.2 Experimental groups

Seventy-two people (six groups of 12) participated in the study. Six conditions were formed by combining each set of upper instruction cards with each set of lower cards. Each group of participants saw cards from only one condition.

5.2.3 Participants

Participants were recruited from the Holmdel, New Jersey, vicinity. The 67 female and five male participants had a median age of 34 years, and a median education of one year of college. They were paid to participate.

5.2.4 Procedure

An information search task was used. Participants in the study were asked to answer many questions about the operating procedures of telephones, based on information obtained from the instructions. Errors in responding to questions indicated that information was not

found or understood. Response times to questions were measured to assess the efficiency of information search.

After the series of main questions were answered, participants were asked to answer a second set of questions. The instruction cards shown in this series had all text removed. Thus, the second set of questions measured incidental learning of the meaning of colors, symbols, and pictograms that occurred in the first part of the experiment. Finally, participants were asked to complete a questionnaire about overall ease of use and appearance of the instructions, as well as further questions about the colors, symbols, and pictograms.

The participants sat at a table that contained a response panel and a 50-cm-square rear projection screen. The viewing distance was two feet, approximately the viewing distance when reading instructions on a public telephone. The response panel contained four buttons, labeled "Ready," "Yes," "No," and "Don't Know." Participants were instructed to place their index fingers on the "Yes" and "No" buttons, and a thumb on "Ready." They were told to imagine that they were traveling, and that they would need to make numerous telephone calls from public telephones. Since telephones have different operating procedures, they would need to use the instructions on the telephones to answer questions before placing the calls.

Each trial began with the presentation of a slide containing a question, such as "Does a local call cost 20 cents?" The participants were told to study the question until they understood and remembered it, and then push the "Ready" button. Pushing the button started a sequence that turned off the question slide, and turned on a slide containing a picture of a telephone with instruction cards on it. The pictured telephone and instruction cards were shown at actual size. Participants were told to look for the answer to the question they had just seen on the instructions, and answer either "Yes" or "No." They were told that all answers were available in the instructions, but if they could not find it to answer "Don't Know."

The experiment was run under computer control; responses and response times were recorded. Response times were measured in milliseconds from the time the picture appeared on the screen until the response button was pressed. There were 96 trials separated by a 2-second intertrial interval, with a 1-minute break between blocks of 24 trials.

5.2.5 Structure of the question series

The questions came from eleven categories: cost of local calls, dial tone/deposit sequence, types of calls that require coin deposit, emergency numbers, long distance calls, calls requiring a "1" prefix, dialing of credit card numbers, possibility of coin use, directory assistance

numbers, repair service numbers, and the address of the telephone (provided on the experimental instruction cards). The 96 trial series contained eight questions from each category except emergency numbers and dial tone/deposit sequence, from which there were 12 questions. For half of the questions the correct answer was "Yes." Questions from each category were paired with each of the eight telephones, except where inappropriate (e.g., no questions about the dial tone/deposit sequence were asked about Charge-a-Call telephones). Question order and left-right position of the "Yes" and "No" buttons were counterbalanced.

Following the 96 questions, 16 questions were presented using the cards with text removed. Participants were told to try to figure out the answers based on the information remaining. Questions were from four categories, emergency numbers, types of calls requiring coin deposit, long distance dialing, and possibility of coin use.

5.3 Results

5.3.1 Main question series

Mean response times were computed for each question category within each trial block (24 questions). An analysis of variance was done on these mean response times. Factors in the analysis were Upper Card Format, Lower Card Format, Trial Block, and Question Category.

No statistically significant ($p < 0.05$) differences in response times were observed between formats. The main effect of Trial Block was statistically significant ($p < 0.001$). Mean response times for trial blocks 1, 2, 3, and 4 were 13.4, 7.9, 6.3, and 5.8 seconds, respectively. Post hoc comparisons of the means showed that block 1 > block 2 = block 3 = block 4. Thus, participants responded more quickly to later questions, which indicates that they could find information more quickly once they had some experience using the cards.

The main effect of Question Category and the Question Category by Trial Block interaction were also statistically significant ($p < 0.001$). Mean response times differed widely across the different question categories. Slow response times, especially later in the question series, are an indication that users may be having difficulty finding or understanding information. In trial blocks 2 through 4, mean response times were relatively short (less than 6 seconds) for emergency numbers, possibility of coin use, repair numbers, cost of local call, and address; response times were slower (8 to 11 seconds) for "1" prefix for long distance, calls requiring coins, long distance dialing, dial tone/deposit sequence, directory assistance, and availability of credit card number dialing. Possible sources of difficulty will be discussed below.

The total number of errors in each trial block and in each question

category was computed for every subject. Both incorrect and "Don't Know" responses were counted as errors. The overall error percentage was 11.4 percent.

An Upper Card Format by Lower Card Format by Trial Block analysis of variance was done on the error data. Only the main effect of Trial Block was statistically significant. There were more errors in the first trial block (15.4 percent) than in the last three trial blocks (9.8 percent). The decrease shows the same facilitative effect of experience that was seen in the response time data.

An Upper Card Format by Lower Card Format by Question Category analysis of variance again showed only the main effect of Question Category statistically significant ($p < 0.001$). The pattern of errors is similar to the pattern of response times. The correlation between mean error rates and mean response times in the 11 categories was high, $r = 0.77$. High error rates are also an indication that users are having difficulty finding or understanding information on the instruction cards.

5.3.2 Incidental learning question series

The instruction cards used in this question series contained only the information that appears in color on the instruction cards. The purpose of this question series was to determine whether subjects had learned the meaning of this information during the main question series. All of the questions could be answered with the information available.

The total number of errors and mean response times were computed for each of the four question categories for every subject. An Upper Card Format by Lower Card Format by Question Category analysis of variance was done on the error data and on the response time data. In both analyses only the main effect of Question Category reached statistical significance. No differences between card formats were observed.

The questions on long distance calling ("Can you dial your own long distance calls?" or "Does the operator dial all long distance calls?") caused the most difficulty. The percentage of errors (58 percent) does not differ from the performance expected from guessing. The long distance category also elicited the highest percentage of "Don't Know" responses (7 percent). Clearly, the participants did not learn that the presence of the "0+" and sometimes "1+" symbols on the upper card indicated that long distance calls could be dialed, and that the absence of "0+" meant that the operator dials all long distance calls.

Performance on the other question categories was substantially better than what would be obtained by guessing, showing that some incidental learning had occurred. However, questions on the types of

calls that required coin deposit did have a fairly high (15 percent) error rate. The answers to these questions could be inferred from either the color of the card (blue or tan) or from the dial tone/deposit sequence.

The "SOS" emergency symbol was well understood. Responses were fast and few errors were made.

5.3.3 Questionnaire results

The questionnaire contained four rating-scale items. The instructions were rated on the ease of finding needed information, the ease of understanding the words in the instructions, the overall ease of understanding the instructions, and the attractiveness of the instructions. The responses were scored from 0 to 5, where 5 was the most positive response. An Upper Card Format by Lower Card Format analysis of variance was done on the ratings for each item. No statistically significant differences between formats were found for any item. Overall, participants rated the ease of finding needed information between "somewhat easy" and "easy" (mean rating 3.6), the ease of understanding words between "easy" and "very easy" (4.3), the overall ease of understanding between "somewhat easy" and "easy" (3.7), and the attractiveness between "somewhat attractive" and "attractive" (3.6).

When asked if any information seemed to be missing from the instructions, the majority of participants responded negatively (67 percent). Of the people responding affirmatively, most of the comments were about the lack of information on placing local calls or about the dial tone/deposit sequence. A few general comments were made, such as "Instructions too wordy," or "Instructions not detailed enough." Similarly, the majority of participants did not report any specific problems understanding the instructions in response to an open-ended probe. The problems that were reported were also about the lack of information on local calls, and confusion about when to deposit coins. In addition, some people were confused about the "0+" and "1+" symbols.

The majority of participants gave no suggestions for improving the instructions. Some general suggestions were made, such as "Make the instructions simpler," or "Make the instructions easier to understand." A few participants commented on the small size of the print. Most suggestions, however, concerned the variety of procedures described on the instructions. The most frequent comment was "Make the instructions the same on all telephones." The nature of the comments made by these participants implied that some thought the operating procedures were determined by the instructions, rather than the reverse.

In the last section of the questionnaire the participants were asked to explain the meaning of the colors and symbols. Most of the participants did not learn the meaning of the blue and tan colors but did learn the meaning of the Charge-a-Call symbol. Ninety-six percent of the participants correctly explained the meaning of "SOS," and most of these recalled that the "SOS" symbol was red.

The "0+" and "1+" on the upper instruction card were generally explained as "Dial 0 (1) and then the number." However, a third of the participants said that "0+" meant "Dial the operator" or "Operator dials long distance." A number of the participants who said that "1+" was for direct-dialed long distance also stated that "0+" was for operator-dialed long distance. Certainly these participants had not learned the distinction between 0+ and 1+ dialing.

The symbol sequence for the dial tone/dial/deposit coin sequence was confusing in some cases. The symbol sequence for coin-first operation was understood by 79 percent of the participants. In contrast, there was some confusion about the meaning of the dial-tone-first symbols. Fifty-four percent of the respondents correctly answered either "Dial tone first" or "No coins needed for dial tone." However, a third of the participants confused the typical dial-tone-first case (deposit coins before dialing local calls) with the post-pay procedure. This pattern of responses is consistent with the general confusion over the dial tone/coin deposit/dialing procedures observed in the experiment.

5.4 Discussion of problems and recommendations for improvement

The major findings in this study relate to the content of the instructions, not the formats.

The highest overall error rate in the main question series occurred on questions about directory assistance numbers. There were two major sources of difficulty. First, the listing for local directory assistance appeared in the same line as the heading, and was hard to find. Since no separate listing for "Local" was made on some of the instructions, the participants tended to infer that the number for "Within this Area Code" was always the local number. This problem was subsequently remedied by adding a separate line for "Local" on the lower instruction card. The other confusion has no obvious solution. The participants expressed confusion over the need for separate listings for "Local" and "Within this Area Code." The confusion may stem in part from the fact that all participants were from New Jersey, where only a single number is used for directory assistance within an area code.

The dial tone/deposit sequence questions contained two subsets of

questions. One subset contained questions about the temporal sequence of hearing the dial tone and depositing a coin, the other about the temporal sequence of depositing a coin and dialing a call. The problems occurred in the latter subset. Forty-eight percent of the participants responded that coins were deposited after answer for the typical (pre-pay) dial-tone-first case. Some people appeared to equate the pre-pay and post-pay dial-tone-first telephones. The post-pay telephone is uncommon (except in a few areas outside of New Jersey), so participants may have been confused about the two types of telephones. The cards tested in this experiment contained no written instructions on how to place local calls. Based on the difficulty seen in this area, some statement about when to deposit coins on local calls appeared warranted.

The same confusion appeared to affect the results in the category of calls requiring coin deposit. Most of the errors occurred on the question "Do you need to deposit any coins to make a local call?" Thirty-eight percent said "No." The additional written instruction described above should also help this problem.

The questions about long distance calling can also be divided into two subsets, one about whether customers can dial their own long distance calls, and the other about the specific digits to dial for long distance calls. Most errors occurred in the latter category, on telephones from which long distance calls were dialed by the operator. These instructions had no specific instructions for long distance on the lower card, just a listing for "All other calls" (Dial the Operator). Some improvement should be obtained if these instructions contained all categories, even though the instruction in every case is "Dial the Operator."

On the incidental learning question series, participants were asked whether long distance calls could be dialed. The answers to these questions had to be inferred from the "0+" and "1+" symbols. If these symbols were present, then long distance calls could be dialed; the absence of a symbol implied that the operator dialed the calls. The problem may have occurred because this upper card information was never used in the main question series. However, the problem may have occurred because of the lack of a positive indicator for operator-dialed long distance. The subsequent addition of such a cue should make the information on the upper card clearer.

Errors were made on 19 percent of the questions about the need for a 1 prefix on direct-dialed station-to-station long distance calls. These errors are hard to interpret since the prefix is not used in New Jersey. The other category that caused some confusion was about customer-dialed credit card numbers. As in the case of the 1 prefix, this information was explicitly stated on the lower instruction card when

appropriate; however, the new procedure was not familiar to New Jersey residents.

In summary, the improvements recommended were:

1. Add a separate line for "Local" directory assistance.
2. Add a written instruction on when to deposit coins for local calls on the upper card; the heading should be "Local calls" and be placed above "Long distance calls."
3. Add a symbol for "Operator dialed long distance" to contrast with the "0+" and "1+" symbols in the long distance section of the upper card.
4. Remove the heading "All other calls" from the lower card and replace it with the same headings used on the other cards "Charge and Person-to-person calls" and "Station-to-station calls."

These changes were incorporated into the instructions used in the subsequent field evaluation. The choice of format (horizontal and ruled) used in the final design was made by AT&T Public Services Marketing and the designer.

VI. FIELD EVALUATION

The major purpose of the field study was to assure that the instruction card design, modified based on laboratory results, was acceptable to telephone users, and that telephone users could use the instructions to find needed information. Because the overall success of the new instructions depends on the use of standard instructions throughout the Bell System, we could not expect to show all the advantages of the new instructions.

6.1 Field trial sites

The field trial of the proposed standard instruction cards was conducted in Illinois Bell Telephone (IBT) Company and New York Telephone (NYT) Company at three sites in each company: a major city airport, a large suburban shopping center, and a small, self-contained city. The sites were selected so that a variety of public telephone users would be exposed to the trial instructions: travelers, local residents, local callers, long distance callers, credit card callers, etc. In addition, a second shopping center was selected in each company. The new cards were not installed at the second shopping center.

A broad range of services and procedures was available at these sites. Among the variations in operating procedures were telephones requiring an initial deposit for all calls, telephones requiring a deposit only for coin-paid calls, and Charge-a-Call telephones. At some locations callers could dial their own long distance calls, at others operator assistance was required. At some locations a "1" prefix was required on calls to other area codes; on others it was not required.

6.2 *Instruction card content*

The information on the trial cards was the same as the information on the instructions they replaced, although some changes were made to conform to the new standards. The changes mainly involved the removal of specific place names. For example, on the existing instructions at Roosevelt Field shopping center in NYT, long distance dialing instructions were given for "Nassau and Suffolk" and "All other places." These specific names were replaced by the general, but equivalent, instructions for "Within this Area Code" and "Outside this Area Code."

6.3 *Interview study*

Interviews were conducted at selected public telephones in December 1979 (three months after card installation in New York, two months after installation in Illinois). People who had just used a public telephone were interviewed face-to-face by professional interviewers. The interviewer asked a series of questions about the call just made and the use of the instructions, then showed instruction cards identical to the ones on the telephone just used and asked a series of questions about how to place various types of calls. The interviews addressed the following areas:

—Did people look at the cards?
—What information was needed?
—Was the information easy to find?
—Was the information easy to understand?
—Could people find and report dialing instructions correctly for various types of calls?
—What was the overall judgment of the appearance of the card?
—Could people understand the pictograms used on the cards?
—Did people understand the meaning of "SOS?"

A total of 650 interviews were conducted at trial locations, 100 at each of the trial sites, plus an additional 50 interviews at Charge-a-Call telephones at O'Hare International airport. Two hundred additional interviews were conducted at the two shopping centers where the original instruction cards had been left in place. The locations provided a comparison of telephone users' acceptance of and problems using the trial cards. Since the advantage of the new cards depends upon their use in all locations, and since the old cards were familiar to the users, this comparison is not intended to be the proof that the new cards are better.

6.4 *Interview results*

For the following discussion, the normal approximation to the

difference between two observed and independent proportions was used to test the significance of the differences between results from the trial and nontrial shopping centers.

Less than 25 percent of the people interviewed reported looking at or reading the instructions cards at any location except at the Charge-a-Call telephones, where 63 percent of those interviewed reported reading the instructions. Charge-a-Call telephones look different from regular coin telephones; 88 percent of the Charge-a-Call users who read the instructions reported doing so to find out how to use the telephone, or because the telephone looked new or different. No significant differences were found between the number of interviewees who looked at or read the instructions at the trial and nontrial shopping centers.

The people interviewed reported needing a variety of types of information. The types of information requested were: how to use the telephone; how much money to deposit; how to make collect, credit card, or person-to-person calls; how to dial long distance; and what numbers to dial for service calls. Ninety percent of those interviewed who reported reading the instructions said that the information they needed was easy to find; 94 percent reported that the information was easy to understand. No significant differences were found between the trial and nontrial shopping centers.

Interviewees were asked to use a set of instructions cards identical to the cards on the telephone they had just used to find and report dialing instructions for the following types of calls: police emergency, local directory assistance, collect, and coin-paid station-to-station long distance calls. In all cases, most of those interviewed were able to correctly report the dialing procedures. Most of the errors occurred when the interviewee reported a dialing procedure that would be correct in other settings, for example, reporting "911" as the police emergency number at a location where "0" was the correct response. This type of error was made in both trial and nontrial locations. No significant problems in finding or understanding dialing instructions were identified with the new instructions. The interviewees had no trouble in finding the emergency information associated with the "SOS" symbol.

Pictograms were used on the upper instruction card to represent the dial tone/deposit sequence. The numerals "1." and "2." were used to indicate the sequence.* A coin-like disk indicated both the action of depositing and the amount of the local call deposit. A handset with an

* A separate study conducted by M. L. Viets reported that sequential steps should be indicated by numbers followed by a dot. Omitting the dot led people to interpret 1 and 2 as magnitudes rather than as ordinal steps in a sequence.

expanding waveform indicated the concept "listen for dial tone." Overall, a high percentage of interviewees (88 percent) were able to provide a correct or partially correct meaning for the dial tone/deposit pictograms. Thirty-nine percent gave correct responses that included the correct meaning of both pictograms and the correct action sequence. The remainder gave responses that included correct meanings for some of the elements.

Interviewees were asked to rate the overall appearance of the instruction cards on a six-point scale ranging from very attractive (a rating of 1) to very unattractive (a rating of 6). Overall, the trial cards were rated attractive; the median response was 2 at all locations except the New York shopping center, where the median response was 3 (somewhat attractive). The median rating of the cards at the Illinois nontrial location was 3; at the New York nontrial location the median response was 2.

Interviewees were also asked an open-ended question about their overall reaction to the instruction cards. Most of the comments were positive; however, about 20 percent of the respondents had some criticism of the instructions. The most common complaints were that the print was too small, the appearance of the card was cluttered, and the color unattractive. Overall, 9 percent of the trial location respondents complained about the print size (10-point type). In Illinois, the nontrial cards were dual-language cards. All instructions were printed in both English and Spanish. This required the use of smaller print (8-point type) than on a single-language card. Thirteen percent of the respondents at the Illinois nontrial location complained about the size of the print, and 13 percent more complained about the use of Spanish, indicating that it made the instructions harder to read. Overall, 6 percent of the trial location respondents complained that the cards were cluttered (too much information); however, less than 1 percent of the airport respondents had this complaint, compared to 11 percent in the small cities and shopping centers. Generally, public telephone users in the cities and shopping centers made local calls, and had little need for most of the information on the instruction cards. Seven percent of those who saw tan cards complained about the color of the cards; less than 1 percent disliked the blue cards.

6.5 Field trial conclusions

The proposed standard public telephone instruction cards were acceptable to users. They were rated as easy to use. Telephone users were able to use the trial instructions at least as well as the nontrial instructions, and no major problems were detected. The new card establishes a standard that satisfies users' needs.

VII. CONCLUSIONS

Based on the results of the laboratory evaluation and on the reception of the new instructions cards in the field, AT&T Public Services recommended to all Bell Operating Companies (BOCs) that the new instruction cards be used on all public telephones. The new cards have been introduced in most BOCs. In addition, the new design was made available to the Independent Telephone Companies for use in preparing their instruction cards. We anticipate that the widespread use of the new standard instructions will make public telephones easier to use.

REFERENCES

1. P. Wright, "Feeding the information eaters: suggestions for integrating pure and applied research on language comprehension," Instructional Science, 7 (July 1979), pp. 249–312.

AUTHORS

Christine A. Riley, A.B., (Applied Mathematics and Psychology), 1971, Brown University; Ph.D. (Psychology), 1975, Princeton University; University of Iowa, 1975–1977; Bell Laboratories, 1977–1982; American Bell, 1982—. At the University of Iowa Ms. Riley was an Assistant Professor of Psychology. In 1977 she began working for Bell Laboratories, and in July 1982 moved to American Bell Research and Development. She is currently a Supervisor in the Data Network Architecture Department. She has worked on a variety of human interface design issues, including the design of instructions for telephone services and the design of human-computer interfaces. She is currently supervising a group responsible for the user interface to AIS/Net 1000 Service.

Max S. Schoeffler, B.S. (Psychology), 1950, Rutgers University; M.A. (Psychology), 1953, Indiana University; M.A. (Mathematics), 1958, University of Michigan; Ph.D. (Psychology), 1954, Indiana University; University of Michigan, 1956–1960; Bell Laboratories, 1960–1982; American Bell, 1983—. At Bell Laboratories Mr. Schoeffler worked in the areas of human interface design for operator service systems; effects of system delay on user preference; design of dialing plans; and standardization of tones, symbols, instructions, etc. At American Bell he is supervising a group working on user needs for new services in office automation. Mr. Schoeffler has chaired the human factors working party of CCITT. He also holds four patents. Fellow, American Psychology Association; member, Psychonomic Society.

Cynthia J. Karhan, B.S. (English Literature), 1969, Indiana University of Pennsylvania; M.A. (Experimental Psychology), 1975, New School for Social Research; Bell Laboratories, 1974—. Ms. Karhan is a Member of Technical Staff in the Stored Program Control Network Services Department. She has worked on automated operator services, including Automated Coin Toll Service and an automated credit card service. She is currently working on the application of automatic speech recognition to network services.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

A Study of the Match Between the Stylistic Difficulty of Technical Documents and the Reading Skills of Technical Personnel

By E. U. COKE* and M. E. KOETHER*

(Manuscript received November 24, 1981)

The reading grade level of a sample of Bell System technical documents was assessed using the Flesch Reading Ease Index, which estimates the reading skill needed to cope with a document's writing style. The reading skills of two groups of Bell System craft and management trainees were also estimated from their performance on a standardized reading test. We found that most of these employees had sufficient reading skill to deal with the writing styles of the sample documents. This observed match between readers and documents resulted because most of our students were proficient readers, not because the documents' writing styles were easy. We compared the reading skills of students as they entered and completed training to see whether less able students were eliminated during training. Results did not support training selection as an explanation of our readers' proficiency. While results indicated that many Bell System employees have the reading skills necessary to cope with the technical documents they read, others may find that the writing styles of technical documents are real barriers to success on the job. We can identify a document that may place an undue burden on its readers by estimating the likelihood of a mismatch between the document's readability and the reading skills of its users.

* Bell Laboratories.

I. INTRODUCTION

How well technical documents communicate with their readers depends in part on the style in which they are written, since the *way* ideas are expressed can influence how easily they are understood. Readers may find a document particularly difficult to read when its writing style places an undue burden on their reading skill. For example, use of words that are unfamiliar to readers can obscure a document's message. Also, a document can be unclear if ideas are embedded in sentences whose syntax is too complex for its readers. Other characteristics of style are described elsewhere in this issue of the *Journal.*[1]

The stylistic difficulty of written communication can be appraised with the help of readability formulas.[2,3] These formulas predict the reading difficulty that may result from a document's writing style but not from its content, organization, or format. Readability formulas estimate the reading skill needed to cope with a document's writing style, as manifested in the writer's choice of surface structures. A readability index is calculated from measures of text features that are thought to be indicators of stylistic difficulty. For example, many formulas use the average length of a text's words and sentences to calculate readability. These two text measures, word length and sentence length, function primarily as indicators of the lexical and syntactic difficulty of the text. In English, shorter words are likely to be more familiar to readers than longer words, since shorter words tend to occur more frequently. Shorter sentences are apt to be easier for readers to process than longer sentences because shorter sentences tend to be less complex syntactically.

While readability indexes are often derived from relatively simple formulas, they can provide useful estimates of sytlistic difficulty.[4] For example, readability indexes tend to agree with human judgments of text difficulty,[5] and it is usually less time-consuming and costly to calculate a readability index than to collect human judgments of reading ease. Readability indexes have also been shown to predict behavioral measures of reading ease, such as reading rate[2,6] or duration of eye fixations,[7] in experimental settings. The readability level of documents has also been related to their effective use in work situations[8,9] and will be discussed later in this section.

Readability formulas have their limitations. Formulas only provide estimates of the stylistic difficulty of texts. These formulas are derived from correlations between text measures related to style and measures of reading difficulty such as comprehension test scores. Almost all formulas have been developed using a multiple regression technique to select the linear combination of stylistic measures that best predicts the scaled difficulty levels of a set of texts. Most of the commonly

used readability formulas, such as the Flesch and Dale-Chall formulas, produce estimates that are accurate to within ±1 reading grade level.[10] The Kincaid formula[11] is slightly less accurate, having a standard error of 1.41 as compared to the Flesch, which has a standard error of 0.81.[12]

Readability formulas are often criticized because editorial revisions that produce acceptable readability indices do not necessarily result in a more readable text.[3,13,14] For example, shortening words or sentences will lower a text's readability index without necessarily making it easier to read. This should not be surprising since readability formulas are derived from *correlations* between measurable text attributes and readers' performance on comprehension tests.[3] As pointed out in the preceding paragraphs, easily quantifiable text features such as sentence length gain their predictive power by indexing underlying contributors to text difficulty that are less easily quantified. Some of the important *determinants* of reading difficulty, such as syntactic complexity and the use of the passive voice, are discussed elsewhere in this issue of the *Journal*.[1]

Just how much difficulty a particular group of readers will have with a document's style depends on their level of reading skill. If a given reader's skill matches or exceeds that required by a document, its writing style alone is not likely to cause trouble for that reader, even though the document's subject matter may be difficult. Conversely, a reader with more limited skill is likely to find that the document's writing style alone presents a challenge.[2–4] By comparing a document's readability index (an estimate of stylistic difficulty) with an estimate of reading skill, we can predict the difficulty a reader might be expected to have with a document's writing style. An estimate of overall reading competence can be obtained from a person's performance on a standardized reading test. These tests sample a reader's behavior on tasks that have been found to discriminate between skilled and unskilled readers.

Estimation of the match between a document's readability and the reading skills of its audience can be helpful in evaluating the effectiveness of technical writing that contains job-related information. Many companies, such as the Bell System, rely heavily on written communication to support job performance. When the writing style of a job-related document is too difficult for its users, job performance may suffer in a number of ways. Studies have investigated the consequences of this mismatch between document style and reading skill. Jobs may take longer than usual to perform because needed information is hard to access.[9] Errors may increase because the descriptions of work procedures are unclear.[8] Perplexed readers may turn to co-workers for help, thus reducing their co-workers' productivity.[9]

In the present study, the match between the readability of docu-

ments and the reading skills of readers was estimated for a large sample of Bell System technical documents. These documents are a major source of information about the operation and maintenance of equipment. The readability of these documents was measured by computer and compared with the reading skills of two groups of Bell System employees who use the documents. Levels of reading skill were estimated from the employees' scores on a reading test.

II. METHOD

2.1 Readability survey

2.1.1 Documents

Readability measures were obtained for the text of Bell System technical documents called Bell System Practices (BSPs). These documents are a major source of information about the operation and maintenance of equipment. They are used throughout the system by both craft and management personnel. There are well over 250,000 BSPs so only a sample of these documents were analyzed. Most of these BSPs are typeset under computer control and consequently were available in computer-readable form.

The BSPs selected for analysis were those likely to be used frequently. To meet this goal, documents were chosen from three important areas of company activity—outside plant, customer equipment, and *ESS** switching equipment. The BSPs dealing with outside plant and customer equipment were taken in a random fashion from handbooks provided to craft personnel for use on the job. The BSPs about electronic switching systems were chosen in a random fashion from those describing the operation of two different systems, 1 *ESS* and 2 *ESS* switching equipment. Table I shows the number of BSPs analyzed from each category of company activity. Altogether, readability indices were obtained for 140 BSPs consisting of a total of 378,359 words.

2.1.2 Readability measure

The Flesch Reading Ease Index[12] was chosen to measure readability. The formula for calculating this index, given in Table II, uses two features of a text to predict stylistic difficulty: (1) the average length of a text's words in syllables, and (2) the average length of its sentences in words. These two measures function primarily as indicators of the lexical and syntactic difficulty of a passage. Available evidence[2-4] suggests that formulas using these simple word and sentence measures are satisfactory predictors of the stylistic difficulty of texts. The

* Trademark of Western Electric.

Table I—Number of BSPs in the readability survey from each category of company activity

Category	Number of BSPs	Total Number of Words
Outside plant	49*	101,162*
Customer equipment	48*	83,927*
ESS maintenance	50	213,373

* Seven BSPs that were in both the outside-plant and customer-equipment handbooks were included in the totals for both categories.

Table II—Formula for the Flesch Reading Ease Index and the algorithm for converting a vowel count to a syllable estimate

1. Flesch Reading Ease Index (FL):

   $FL = 206.84 - 84.6\ WL - 1.015\ SL$

   Where
   WL* is the average length of a word in syllables
   SL is the average length of a sentence in words

2. Algorithm for estimating the syllables (SYL) in a sample of words

   $SYL = 0.998\ VOW - 0.343\ W$

   Where
   VOW is total vowels (aeiouy)
   W is total words

* WL = SYL/W.

Kincaid formula[11] also uses these two measures and is the basis for the Writer's Workbench advisory on readability.[1] The Flesch formula was chosen for the present survey because it has been more widely used than the Kincaid formula and has been shown to be one of the most accurate predictors of readability.[10]

The Flesch Index ranges from 0 to 100 with a higher index representing a more readable text. A Flesch Index can also be expressed as a reading grade level, which is the average school grade of readers who would be expected to score 75 percent on a comprehension test for the reading material. Originally, a point on the Flesch Index scale corresponded to one-tenth of a grade,[12] but Flesch found that his formula underestimated the reading grade level of more difficult passages. He proposed a nonlinear relationship between the Flesch Index and reading grade level,[15] and this relationship is used in the present study to assign a reading grade level to an index value.

The Flesch Index was calculated by computer,[16,17] since the BSPs were in computer-readable form. A program was written to compute average sentence length for (complete) sentences in the text of the BSPs. The program was able to exclude tables, headings, and titles

that were flagged in the text. To calculate the average length of a word in syllables (WL), the program estimated the number of syllables in a word sample from a count of the number of vowels in that sample. This vowel count was then converted to syllables using the formula[18] shown in Table II.

When syllables are estimated from a vowel count, the presence of words without vowels, such as abbreviations, acronyms, and numbers, may be a problem. Their inclusion in the computer's estimate of syllables can lead to an underestimation of average word length and a consequent underestimation of reading difficulty. This is a problem when analyzing technical documents. In some of the BSPs selected for the present survey, over 11 percent of the words did not contain vowels. One solution to this problem is to base the syllable estimate on words that conform to the syllable estimation algorithm's assumption that a word contains at least one vowel. This solution was adopted in the present study and has been implemented in the Writer's Workbench. Only nonnumeric words with at least one vowel were used in the estimation of syllables from a vowel count. This solution seemed to work. Flesch Indexes were computed using hand counts of syllables for 16 BSPs that had a large number of words without vowels. These manually produced indexes were compared with computer-produced indexes. When words without vowels were omitted from the computer estimation of syllables, 11 of the 16 texts were assigned the same reading grade levels by the computer-produced and manually produced indexes. When words without vowels were included in the syllable estimation, only 4 of the 16 BSPs were classified at the same readability level by the computer-produced and manually produced indexes. This solution finds precedence in Flesch's suggestion[12] that counts of syllables be based on words whose pronunciation is unambiguous.

2.2 Reading skills survey

2.2.1 Reading test

The reading skills of Bell System personnel were estimated from their performance on a standardized, commercially produced reading test, the Nelson-Denny Reading Test, Form D.[19] This test had been standardized by giving it to a sample of over fifteen thousand high school students, representing the student population of the United States in 1972. The test performance of students in this sample was used to estimate the reading skill represented by a test score. Level of reading skill is often expressed as a score's reading grade level. A test score is assigned a reading grade level that corresponds to the *average* educational level of students who made that test score in the standardization sample. Thus, a reader's reading grade level is the school grade for which his or her test score is typical. It should be emphasized

here that the educational level of a person and the reading grade level of his or her test score are not identical.

The Nelson-Denny Reading Test was chosen for three reasons: (1) it can be given in a reasonably short period of time, (2) it covers a range of reading abilities appropriate for Bell System employees, and (3) it was well constructed according to standards for norm-referenced tests.[20]

The Nelson-Denny Test is composed of a vocabulary test and a comprehension test. The vocabulary test contains 100 multiple-choice items consisting of a sentence stem and four words that can complete the sentence. The comprehension test consists of 36 multiple-choice questions that relate to five short prose passages. Readers work on the vocabulary test for 10 minutes and on the comprehension test for 20 minutes.

Performance on the test is expressed as a total score. This score is the sum of two subscores: (1) the number of correct vocabulary items, and (2) the number of correct comprehension items multiplied by two. The reading grade equivalents for total scores are given in the Examiner's Manual of the Nelson-Denny Test.[19] In terms of accuracy, the total score has a standard error of measurement* of about 5.7.[19] Translated into reading grade level, this standard error is approximately a 0.5 reading grade level.

In the present study, total scores were used in all statistical computations. However, the results were also presented as reading grade equivalents of test scores. Reading grade levels were used because: (1) they are often a more meaningful way of expressing reading skill, and (2) the readability of the technical documents was expressed as reading grade levels.

2.2.2 Subjects

The Nelson-Denny Reading Test was given to 471 technical management employees and 210 craft employees. Both groups were students at Bell System training centers. The management students were taking a variety of technical courses covering such diverse subjects as switching systems, data management, building engineering, and finance. The craft students were taking the final course in an *ESS* switching equipment training program that involves approximately 700 hours of training.

These employees were chosen because they were likely to use doc-

* The consistency of a test can be expressed in terms of the standard error of measurement. Approximately two-thirds of the hypothetical "true" scores on a test would lie within one standard error of measurement from a score obtained on the test.

uments such as the BSPs sampled in the readability survey. In addition, students at the training centers are more readily available for testing than are employees in work situations.

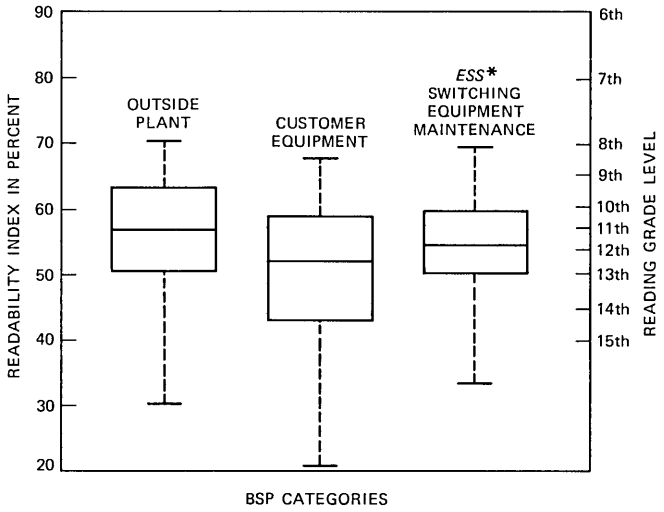
Each student took the Nelson-Denny Test and filled out a questionnaire requesting information about his or her age, sex, years of formal education, job level, years of service in the Bell System, and region of the country in which he or she worked. The study was conducted in the students' classrooms in groups ranging from 10 to 30 students. All students in the selected classes participated in the study although they were not required to do so. To ensure the anonymity of the participants, each student used a number to identify all test material and the questionnaire.

III. RESULTS

3.1 Readability survey

Figure 1 shows the distributions of Flesch Indexes for BSPs in each of the three categories: outside plant, customer equipment, and *ESS* maintenance. The distribution for each category is displayed as a box plot. The ends of a box are drawn at the upper and lower quartiles of the distribution, while the bar across the box represents the median. The extremes of the distribution are plotted as horizontal lines joined to the box by dashed lines.

Figure 1 suggests that considerable reading skill is required to cope with the writing styles of many BSPs. Thirty-three percent of all the



*TRADEMARK OF WESTERN ELECTRIC.

Fig. 1—The distributions of readability for the three categories of BSPs.

BSPs sampled had Flesch Indexes below 50. An appreciation of the stylistic difficulty of these BSPs can be gained by comparing their readability levels with those obtained by Flesch[12] for some representative magazine articles. Flesch found that academic journals such as the *American Scholar* or the *Yale Review* had Flesch Indexes ranging from 30 to 50, while quality magazines such as the *New Yorker* had Indexes ranging from 50 to 60. Over half of the customer-equipment BSPs were comparable to academic journals in their level of stylistic difficulty. The writing style of three-quarters of the *ESS* switching equipment maintenance BSPs was at least as difficult as that of quality magazines. In terms of required reading skill, the writing style of one-third of the BSPs would be likely to cause difficulty for readers with skills below the 13th reading grade level. The writing style of over half the documents (63 percent) would be likely to cause difficulty for those employees reading below the 11th reading grade level.

3.2 Reading skills survey

3.2.1 Reader characteristics

The questionnaires provided information about the students who took the reading test. Sixty-five percent of the technical management students were first-level management, 27 percent were second-level management, and the remaining 8 percent were third- (4 percent) and fifth- (4 percent) level management. As Table III shows, the technical management students represented a broader range of age, education, and years of service than did the craft students. The two groups differed primarily in educational level. More technical management

Table III—Age, formal education, and years of service

		Percentage of Students	
		Technical Management	Craft
Age Level	20–29	26	29
	30–39	39	52
	40–49	24	14
	50 or older	11	5
Educational Level	Some high school	1	3
	High school grad	30	55
	Some college	29	36
	College grad	26	6
	Some grad school	14	
Years of Service	1–5	19	9
	6–10	26	49
	11–20	30	30
	over 20	25	12

students (69 percent) than craft students (42 percent) had continued on to college after graduating from high school. There were also more women among the technical management students (30 percent) than among the craft students (9 percent).

3.2.2 Reading scores

The cumulative probability distributions of total scores on the Nelson-Denny Reading Test are shown in Fig. 2 for each group of students. The reading grade equivalents of test scores appear at the top of the figure. The median, upper, and lower quartiles of the distributions are also shown. The maximum total score is 172, while the highest reading grade level that can be assigned to a score is 15.

Test scores indicate that, on the average, both groups of students were competent readers. Mean scores were 81 and 94 for craft and technical management students, respectively. For comparison, the Nelson-Denny manual[19] gives mean scores for samples of freshmen, sophomores, juniors, and seniors in college. These scores were 75, 85, 93, and 98, respectively. Over half the craft and technical management students had test scores at or above the 14th (college sophomore) reading grade level. The reading grade levels of the mean scores were 14.0 and 14.9 for craft and technical management students, respectively, and this difference between student groups was greater than would be expected by chance ($t = 3.24$, $df = 671$, $p < 0.05$).

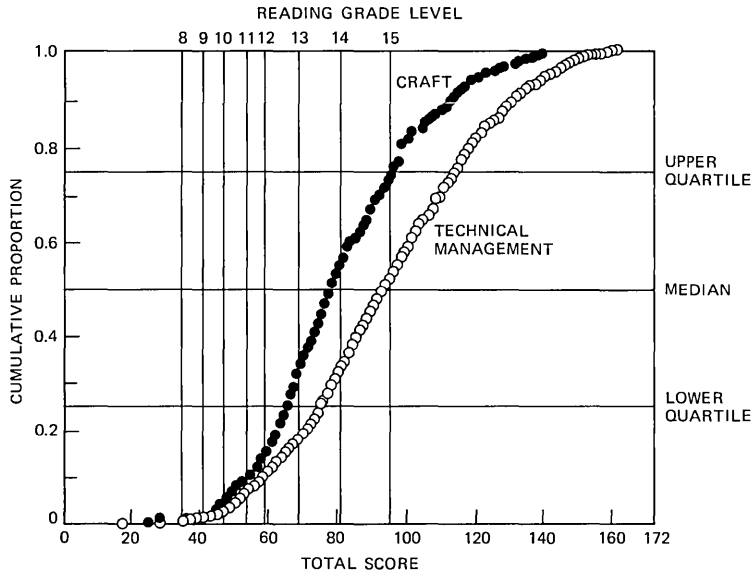


Fig. 2—Cumulative frequency distributions of reading test scores showing the median, upper, and lower quartiles of distributions.

3.3 Estimating the stylistic difficulty of BSPs

An estimate of the difficulty a group of readers may have with the writing style of a specific document can be made by finding the proportion of readers whose reading grade levels on a reading test are below the reading grade level of the document. If a document's reading grade level is 12, for example, and one third of its users have reading skills below the 12th reading grade level, then the likelihood of a mismatch is 0.33 (e.g., the probability that the users' reading skills will be below the 12th reading grade level).

This likelihood of a mismatch between audience and document can be extended to a set of documents. In this case, two measures are needed for each reading grade level represented by the set of documents:

1. The proportion of documents in the set whose reading grade levels are greater than a specified reading grade level (that is, the proportion of documents likely to be difficult for readers at that grade level);
2. The proportion of readers in the group at that reading grade level.

The likelihood of a mismatch is the sum of the products of these two estimates over all reading grade levels. This likelihood is the probability that a member of the audience will select a document from the set that exceeds his or her reading grade level.

The likelihood of a mismatch between BSPs and the technical management and craft students were calculated from the information in Table IV. This table shows the proportion of BSPs whose reading grade levels exceeded each level in the table. Also shown are the

Table IV—Proportions for calculating the likelihood of a mismatch between BSPs and two employee groups

Reading Grade Level	Proportion of Difficult BSPs	Proportion at Each Reading Grade Level	
		Technical Management	Craft
below 7.0	1.00	0.00	0.01
7.0–7.9	0.99	0.00	0.00
8.0–8.9	0.90	0.01	0.01
9.0–9.9	0.74	0.01	0.02
10.0–10.9	0.63	0.04	0.05
11.0–11.9	0.44	0.04	0.05
12.0–12.9	0.33	0.06	0.15
13.0–13.9	0.20	0.16	0.26
14.0–14.9	0.11	0.19	0.18
15 or greater	0.00	0.49	0.27

proportion of technical management and craft students at each reading grade level.

The likelihoods of a mismatch between BSPs and the student audiences was 0.14 for technical management students and 0.20 for craft students. This result suggests that technical management students were likely to have some difficulty with the writing style of about 14 percent of the BSPs in the sample, and craft students with about 20 percent. These estimates of difficulty suggest that the mismatch between students and BSPs is not great, but we do not really know how to interpret such results. For one thing, the Flesch Reading Ease Index assumes 75-percent reading comprehension. For technical documents, such as these BSPs, comprehension closer to 100 percent is probably desirable. In addition, although losses in job efficiency may result when employees have difficulty with a document, simplifying its writing style can be both costly and time-consuming. At present we know that it is useful to match documents to readers but we don't know how to calculate the benefits and costs involved in achieving this match.

IV. DISCUSSION

The results of the present study suggest that the writing style of many Bell System documents used on the job and in training requires considerable reading skill. Yet, the reading skill levels of the majority of the employees we tested matched or exceeded the estimate of required reading skill provided by the readability analysis. Our two samples of Bell System employees appeared to have fewer people with limited reading skills than might be expected from their educational levels. This was especially so for the craft students who had an average reading grade level (level 14—college sophomore) that was two grade levels *above* their average educational level (level 12—high school graduate).

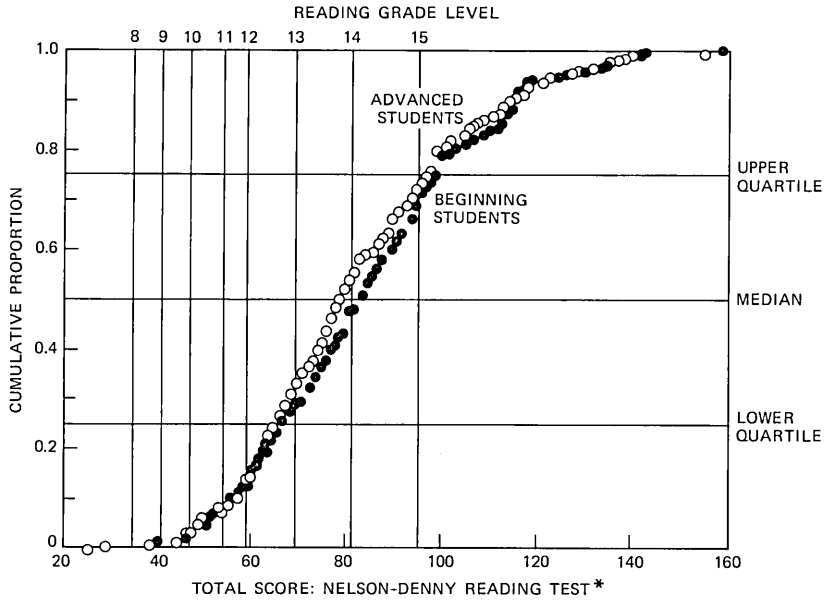
A number of selection factors may have acted to make our sample of Bell System employees such able readers. One of these factors may have been the documents that these students read during training. The craft students we tested had successfully completed a series of courses (approximately 700 hours of training) that required the use of highly technical documents, including BSPs. These technical documents may have acted as selective filters, tending to eliminate students with limited reading skills. This explanation assumes that students who have reading difficulties or who are averse to reading difficult documents will tend to leave training. If our filter explanation is correct, we would expect to find proportionately more skilled readers among students who were about to complete training than among students who were starting a training sequence. We explored this filter

explanation by comparing the reading-test performance of craft students who were just beginning an *ESS* switching equipment training sequence with students who were about to complete this sequence.

The Nelson-Denny Reading Test and a questionnaire were administered to 160 craft employees who were enrolled in introductory *ESS* switching equipment courses given at the training centers of several operating companies. Their performance was compared with that of the 210 craft students discussed earlier (Fig. 2), who were finishing their last course in the ESS training sequence.

The distributions of total scores on the reading test are shown in Fig. 3 for the beginning and advanced students. There was little difference between these two groups in their performance on the reading test. These results do not support the contention that training documents tend to filter out less able readers during *ESS* training, but rather they suggest that our sample of Bell System craft students was not biased by selective factors during *ESS* switching equipment training. Thus, their level of reading skill is likely to be representative of Bell System craft employees with similar training.

Training documents may not have worked as filters because of the procedures for selecting employees who enter *ESS* switching equip-



*MAXIMUM SCORE IS 172

Fig. 3—Cumulative distributions of reading test scores for advanced and beginning students.

ment training. These procedures, themselves, may lead to the choice of able readers. Operating companies undoubtedly choose their more able employees for *ESS* training because it is expensive and time-consuming. In addition, all those who start *ESS* training must complete a mini-course that helps to select employees who are likely to do well in *ESS* switching equipment self-instructional training. Even the perceived mental demands of operating and maintaining a complex switching system may discourage employees with limited intellectual skills. Although *ESS* trainees are not chosen explicitly for their reading skills, all the factors mentioned above could lead to the selection of competent readers as trainees.

V. CONCLUSION

In the present study, the reading skills of the majority of employees we tested were likely to be matched to the reading demands made by the writing styles of the majority of the technical documents they read. This match resulted because so many of the employees were proficient readers and not because the documents were easy to read. In many technical areas, such as *ESS* switching equipment, readers may have more reading skill than would be expected on the basis of their education. One implication of this finding is that demographic information about education may not be sufficient to predict an audience's level of reading skill. More direct estimates of reading skill may be needed, such as the Nelson-Denny Reading Test.

Although we found a reasonable match in grade levels between documents and readers in the present study, this does not ensure that the documents are easy to read. Documents can be identified as difficult but not necessarily as easy to read, because the readability index does not include such important determinants of difficulty as organization, content, or format. For this reason, our findings should not discourage Bell System writers from writing as clearly and simply as possible. For many employees, the writing styles of technical documents may be a real barrier to success on the job. Screening devices that combine information about the readability of documents and the reading skills of their audiences, such as the likelihood of a mismatch, can help identify documents that may be too difficult for their intended readers.

VI. ACKNOWLEDGMENTS

We are grateful to E. Z. Rothkopf for his helpful comments and suggestions and Western Electric at Winston-Salem, North Carolina, for providing us with the BSPs. We also wish to thank the management and instructors at the Bell System Center for Technical Education,

Lisle, Illinois, and the Bell System Technical Center, Dublin, Ohio, for their assistance. J. A. Leedy of Network Operations Training at AT&T and the management and instructors at the following Training Centers also lent us their assistance: Plant Training Center, Pennsylvania Bell; Birmingham Training Center, South Central Bell; Decatur Training Center, Southern Bell; Houston Training Center and St. Louis Training Center, Southwestern Bell.

REFERENCES

1. N. MacDonald, "*UNIX*™ Writer's Workbench Software: Rationale and Design," B.S.T.J., this issue.
2. G. R. Klare, *The Measurement of Readability,* Ames, IW: Iowa State Univ. Press, 1963.
3. G. R. Klare, "Assessing Readability," Reading Res. Quart., *10,* No. 1 (1974–75), pp. 62–102.
4. G. R. Klare, "A Second Look at the Validity of Readability Formulas," J. Reading Behavior, *8,* No. 2 (Summer 1976), pp. 129–52.
5. G. R. Klare, "Judging Readability," Instructional Science, *5,* No. 1 (January 1976), pp. 55–61.
6. E. U. Coke, "Reading Rate, Readability and Variations in Task-Induced Processing," J. Ed. Psych., *68,* No. 2 (April 1976), pp. 167–73.
7. E. Z. Rothkopf, "Copying Span as a Measure of the Information Burden in Written Language," J. Verbal Learning and Verbal Behavior," *19,* No. 5 (October 1980), pp. 562–72.
8. K. A. Johnson, R. P. Relova, and J. P. Stafford, *An Analysis of the Relationship Between Air Force Procedural Manuals and Discrepancies Involving Noncompliance with the Procedures,* Air University: Air Force Institute of Technology, (NTIS No. AD-750 917), 1972.
9. *Reading for Working: A Functional Literacy Anthology,* T. G. Sticht (Ed.), Alexandria, VA: Human Resources Research Organization, 1975.
10. R. D. Powers, W. A. Sumner, and B. E. Kearl, "A Recalculation of four Adult Readability Formulas," J. Ed. Psych., *49,* No. 2 (April 1958), pp. 99–105.
11. J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.* CNTT Research Branch Report 8-75, Naval Air Station, Memphis, Millington, TN: Chief of Naval Technical Training, 1975.
12. R. F. Flesch, "A New Readability Yardstick," J. Appl. Psych., *32,* No. 3 (June 1948), pp. 221–33.
13. A. Davison, R. N. Kantor, J. Hannah, G. Hermon, R. Lutz, and R. Salzillo, *Limitations of Readability Formulas in Guiding Adaptations of Texts* (Technical Report No. 162), Urbana-Champaign, IL: Center for the Study of Reading, 1980.
14. "Forum—Readability Formulas: Used or Abused?" IEEE Trans. on Professional Commun., *PC-24,* No. 1 (March 1981), pp. 43–54.
15. R. F. Flesch, *The Art of Readable Writing,* Revised Edition, New York: Harper & Row, 1974.
16. M. E. Koether and E. U. Coke, *A Scheme for Text Analysis Using Fortran,* Annual Meeting of the American Educational Research Association, New Orleans, February 28, 1973.
17. E. U. Coke, "Computer Aids for Writing Text," in *The Technology of Text,* D. H. Jonassen (Ed.), Englewood Cliffs, NJ: Educational Technology Publications, 1982.
18. E. U. Coke and E. Z. Rothkopf, "Note on a Simple Algorithm for a Computer-Produced Reading Ease Score," J. Appl. Psych., *54,* No. 3 (June 1970), pp. 208–10.
19. *Nelson-Denny Reading Test, Form D,* Boston, MA: Houghton Mifflin, 1973.
20. O. K. Buros, *The Sixth Mental Measurement Yearbook,* Highland Park, NJ: Gryphon Press, 1965.

AUTHORS

Esther U. Coke, B.A. (Zoology), 1951, Wellesley College; Ph.D. (Psychology), 1968, New York University; Bell Laboratories, 1957—. Ms. Coke is a member of the Learning and Instruction Research Department. Her research interests include human learning, stylistic factors that influence the effectiveness of technical writing, and computer techniques for evaluating writing. Member, APA, AERA, Psychonomic Soc., Sigma Xi, AAAS.

Mary E. Koether, B.A. (International Relations), 1951, Mount Holyoke College; Bell Laboratories, 1965—. Ms. Koether is a member of the Learning and Instruction Research Department. Her professional interests include the computer analysis of text and factors that influence text comprehension. She also participated in the design and programming of a computer-based course maintenance system. Member, AERA.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July-August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Toward Bell System Applications of Automatic Speech Recognition

By J. E. HOLMGREN*

(Manuscript received July 30, 1982)

Advances over the past few years in the field of Automatic Speech Recognition (ASR) have brought more attention to potential Bell System applications of this technology. Before reaching the point of ASR implementation, several human factors problems have to be overcome. This paper describes the central human factors issues, then summarizes the initial steps at Bell Laboratories in attempting to deal with those issues. Findings from observations of customers speaking credit card numbers to operators are described, followed by summaries of three studies investigating control of the speech of ASR system users.

I. INTRODUCTION

The potential of Automatic Speech Recognition (ASR) in telecommunication applications has long been recognized,[1] but until recently the state of ASR technology did not warrant effort beyond the existing laboratory activity. In the past few years, ASR has emerged from the laboratory into commercial use. Several systems are now available providing speaker-dependent word recognition.[2] Such a system must be trained by each user before it can recognize that user's speech.

* American Bell.

Most of these systems require isolated speech in the sense that each vocabulary item must be spoken with a short silent interval preceding and following each word, although a few systems allowing connected speech are now available.[2] Speaker-dependent systems have been applied primarily in industrial settings that require hands-free data entry, such as inspection and quality control.

Speaker-dependent ASR systems are inappropriate for many attractive telecommunications applications. Instead, speaker-independent systems, which need no prior training by users, are required. Various network operator services, such as credit card calling and directory assistance, are examples of applications in this category. Through the use of ASR, such services could be automated and used from any telephone that has access to the existing network. A few isolated-speech, speaker-independent ASR systems are already in use in the United States and Japan. They provide service to selected groups of users in applications such as private network call-routing and banking. Laboratory work is under way to develop a connected-speech, speaker-independent ASR capability.[3]

The use of ASR in universally accessible services raises many human factors questions. This paper summarizes Bell Laboratories initial human factors work leading toward the first network applications of speaker-independent ASR. Our work began in the context of considering the use of ASR in one particular network service: credit card calling. Today, most credit card calls require giving a credit card number (CCN) to an operator. A new service known as Calling Card Service (CCS) automates the handling of credit card calls from *Touch-Tone** telephones by allowing customers to enter their CCNs on the *Touch-Tone* telephone number pad. However, calls from rotary telephones must still be handled by operators. ASR would allow automation of all credit card calls.

Our first step in investigating the credit card application was to identify the critical human factors issues that require attention. Several of these issues are common to almost all potential network applications of speaker-independent ASR. A description of the common issues provides perspective on our subsequent human factors work.

II. THE CENTRAL HUMAN FACTORS ISSUES

Issues that surround the user-system dialog were selected as the focus of the work reported here. Although the other issues are only touched on in this paper, they are no less important to any successful network application of ASR.

---

* Trademark of AT&T.

2.1 User-system dialog

For the foreseeable future, ASR systems will be limited by several aspects of human speech, such as the vocabulary they can recognize, the maximum rate of speech they can handle, their ability to ignore extraneous words and sounds, and the accuracy of recognition per spoken vocabulary item. Thus, care must be taken in the design of any user-system dialog to overcome these limitations.

2.1.1 Instructions

Appropriate instructions are needed to control the speaking rate and vocabulary that untrained speakers use when they encounter an ASR system for the first time. These instructions also must allow experienced users to proceed without unnecessary delay.

2.1.2 Feedback

While current ASR systems are achieving impressive recognition accuracy for limited vocabularies, their accuracy is less than that of a human listener. Therefore, many applications, particularly those where errors are costly, require feedback to the user to ensure correct recognition by the ASR system. For many attractive telecommunications applications of ASR, user input will consist primarily of strings of digits (e.g., telephone numbers or credit card numbers). Several digit feedback provisions are possible. For instance, feedback could be given after each digit, after each group of digits, or after entry of the entire number. The optimal method depends on the nature of the application and the recognition accuracy of the ASR system.

2.1.3 Error correction

Methods are needed that allow users to correct both their own speaking errors and, when feedback is given, recognition errors on the part of the ASR system. Again, several options are available and the optimal method for each application is unclear.

2.1.4 Problem speakers

No matter how good an ASR system may be, there will always be some speakers whose speech cannot be reliably recognized by the system. Any service incorporating ASR will have to provide for some type of alternate treatment for such individuals; this will often mean transfer to an operator or attendant. Detection of problem speakers early in the dialog and swift alternate treatment will be necessary to ensure both service efficiency and customer satisfaction. The best way to detect these speakers has not yet been established.

2.2 Isolated vs. connected speech

Connected speech is preferable to isolated speech for use as input to an ASR system. However, when speaker-independent ASR systems that accept connected speech become available for use, isolated input will still be more accurately recognized. For this reason, in any application of ASR it will be necessary to decide whether the trade-off between the greater ease of use of connected speech and the greater accuracy of isolated speech favors the former or the latter.

2.3 Vocabulary choice and expansion

User and system considerations may often conflict when a vocabulary is selected for any given application of ASR. The most appropriate vocabulary for the speaker may be particularly difficult for the ASR system. For instance, while spoken, spelled input might be a natural way to specify a name to a directory assistance system, the spoken alphabet is a singularly difficult vocabulary for any ASR system to accept.[4] Words in a vocabulary such as the international word-spelling alphabet (Alpha, Bravo, Charlie, etc.) would be much more accurately recognized by machine, but much less convenient for most users. Closely related to vocabulary selection is the problem of vocabulary expansion. Careful selection of new vocabulary items is necessary because adding new words to a vocabulary may change the system's performance on the original set of words.

2.4 Integration of Touch-Tone *service with ASR*

For any network service using ASR, a large percentage of the customers will be calling from *Touch-Tone* telephones. Thus, it is necessary to consider the possibility of mixed *Touch-Tone* telephone and voice input to an automated service. This raises several questions regarding integration of the two, such as whether to provide both input options at every point in a service, whether to encourage the use of one option over the other, how to make voice input compatible in some sense with *Touch-Tone* telephone input when both are available, etc.

2.5 Template construction

In an ASR system templates represent the words to be recognized in a given application. Template construction is in many respects the most critical hurdle in applying speaker-independent ASR because it is largely the quality of those templates that determines the recognition accuracy of the system across the population of users.[5] Template considerations are included here because of the human factors problems involved in building the speech database needed to construct them.

2.6 System evaluation

The performance of a speaker-independent ASR system in any application depends not only on the characteristics of the system but also on the vocabulary used in the application, the set of templates constructed for that vocabulary, the transmission conditions, and the characteristics of the spoken input. Therefore, evaluating the adequacy of a system in an application involves more than simply obtaining some overall measure of recognition accuracy. Information will be needed about variation in recognition accuracy across segments of the user population, across vocabularies, and across transmission conditions. Other critical aspects of performance will be system response time and the rate of false recognition for words outside the application vocabulary.

III. TSPS OBSERVATION STUDY

To study user-system dialog issues in the application of ASR to credit card calling we gathered data from customers speaking their credit card numbers to Traffic Service Position System (TSPS) operators, who handle all nonautomated credit card traffic. Such data were needed to identify any customer behavior changes necessary to interact successfully with an ASR system.

The particular aspects of customer speaking behavior that we investigated were:

1. Customer segmenting of CCNs. Segmenting, as used here, refers to the tendency of most customers to break a CCN into spoken segments by pausing briefly after speaking a group of three or four digits. Segmenting is important because the per-item recognition accuracy of speaker-independent ASR systems that can accept connected speech is likely to be highly sensitive to the number of items in a connected sequence.
2. Customer vocabulary (e.g., "zero" versus "oh," "hundred" versus "zero zero," etc.).
3. Occurrence of words or sounds other than those used to give the CCN.
4. Frequency of customer mistakes in speaking the CCN and spontaneous correction of those mistakes.
5. Frequency of operator requests for repetition of a portion or all of the CCN. Our interest in this stems from the fact that reliable system performance is possible only when the customer corrects system-recognition errors using feedback from the ASR system; thus, it is useful to have information on the current frequency of operator-requested repetitions.
6. Speaking rates distribution for current customers.

3.1 Summary of results

3.1.1 Number of observations

A total of 3040 credit card calls were observed at three different TSPS offices, 1157 in Milwaukee, 742 in Louisville, and 1141 in Boston. The relatively low number of observations for Louisville reflects a low overall credit card call volume during the observation period. Since the card numbers for the observed calls at each site were not recorded, the number of distinct CCNs among the 3040 calls is not known.

3.1.2 Segmentation of the spoken CCN

The majority of CCNs were spoken in consecutive segments of 3, 3, 4, and 4 digits. Of all observations, 69 percent fell in this category. This is not surprising, since the format of most CCNs is NPA NXX XXXX XXXX and in the three operating companies visited the number is printed on the credit card with that segmenting.

The next most common segmentation was 3 3 4 3 1, used in 17 percent of the observations. The frequency of this segmentation probably reflects carryover from the previous 10-digit-plus-one-letter format used for CCNs up until two months before data collection. In most cases, the new 14-digit numbers were constructed from existing CCNs by adding an initial NPA and changing the terminal letter to a digit. Thus, despite the fact that the new number was printed on the credit card with the 3 3 4 4 segmentation, customers accustomed to speaking their old number with a 3 4 3 1 segmentation carried that habit over to the new number.

Only four percent of the observed numbers were spoken with the digits run together. A number was classified as run together if five or more digits were spoken without an intervening pause. This was the third most common segmentation category. The remaining 10 percent of the calls were distributed among several infrequently occurring segmentation categories.

3.1.3 Variation in vocabulary

Regarding the use of "zero" or "oh" when speaking the digit zero, 29 percent of the 3040 spoken CCNs contained a spoken "zero," 51 percent contained a spoken "oh," 9 percent contained both "zero" and "oh," and 11 percent contained neither. It is interesting that customers frequently say both "zero" and "oh" while speaking a single CCN. While figures are not available on the proportion of observed CCNs that contain multiple occurrences of the digit zero, it is evident that the percentage of multiple zero calls involving the use of both "zero" and "oh" is considerably greater than nine percent.

When speaking the CCN, customers used words outside the single-

digit vocabulary in only eight percent of all calls. Almost all of these calls involved some combination of three types of multiple-digit utterances. The most common type was a two-digit combination such as "fifty-six," "eighty-eight," "thirteen," etc. The next most common type was the phrase "double zero" or "double oh" for the digit combination 00, usually when the 00 was the leading pair of a segment. The word "hundred," usually used for a terminal 00 in a segment, was the third most common type.

3.1.4 Extraneous vocalizations

There were relatively few occurrences of extraneous vocalizations from the customer. Less than two percent of all calls included such events. The most commonly used word was "dash," spoken between segments of the CCN. This occurred because some operating companies print the CCN on the credit card with dashes between segments.

3.1.5 Customer correction of errors

The customer corrected an error in the spoken CCN before any request for repetition from the operator on less than three percent of all calls. The most common error was to leave off the area code when giving the CCN. Since the requirement to give the area code had been in force for less than three months at the time these data were collected, customers were still adjusting to the new CCN format.

3.1.6 Operator requests for repetition

The operator asked the customer to repeat some or all of the CCN on less than six percent of all calls. Usually the operator requested repetition of all 14 digits. This occurred because the operator can collect any number of digits entered on the TSPS console number pad only by cancelling all entered digits and beginning again. The three most common reasons for requesting repetition were operator keying errors, the operator misunderstanding or not hearing the customer, and the customer forgetting to give the area code.

3.1.7 Customer speaking rates

Customer speaking rates are of interest only for those calls on which the CCN was entered in a manner otherwise consistent with the constraints likely to be placed on users of any connected-speech, speaker-independent ASR system. Examining these calls allows one to determine if credit card customers were speaking too quickly for ASR systems when their spoken input was acceptable in all other respects. A call was classified as acceptable if it had the following characteristics:

1. Only single-digit words were spoken.

2. No extraneous vocalizations occurred.
3. The customer did not make an error in entry.
4. The operator did not request a repetition.

An operator may request repetition of a CCN, even though the input was acceptable to an ASR system. However, to be conservative, these calls were excluded from the acceptable set, which contained 80 percent of all calls.

As a function of location, the mean speaking time for a CCN is 5.16 seconds in Milwaukee, 5.33 seconds in Louisville, and 5.09 seconds in Boston. Although the variation among these means is small, an analysis of variance shows it to be statistically significant ($F = 11.79$, $p < 0.001$). The mean speaking times for males and females are nearly identical, being 5.17 seconds and 5.19 seconds, respectively.

The performance of connected-speech, speaker-independent ASR systems is likely to fall off rapidly as the rate of speaking connected strings increases beyond about 2.5 words per second.[6] For this reason, speaking rates within connected strings of digits are of more interest than total speaking times for CCNs. To compute a speaking rate for each call, the total speaking time was corrected for the pauses occurring between segments. This correction was made by assuming that for any speaker, the duration of such pauses was roughly equal to that speaker's mean speaking time per digit. Based on the listening experience of the two data collectors, this appears to be a reasonable assumption. The appropriate estimate of speaking rate for a call is therefore given by the following expression:

$$\text{Rate} = (14 + s)/t,$$

where s is the number of intersegment pauses and t is the total speaking time for the CCN. The value of s was based on the segmentation judgment made by each data collector on each call. Calls for which the digits were judged to be run together present problems for this rate estimate. Any call having five or more digits run together should fall in this category, so the number of intersegmental pauses is not constant within the category. For simplicity, it was assumed (based on the data collectors' observations) that the average number of pauses in a call classified as run together was one, so s was set to that value.

The distribution of speaking rates is given in Fig. 1, where, for 94.2 percent of the acceptable calls, the speaking rate was greater than 2.5 digits per second.

3.2 Implications for customer behavior

The above findings indicate two areas where the modal customer-speaking behavior is likely to require modification before customers

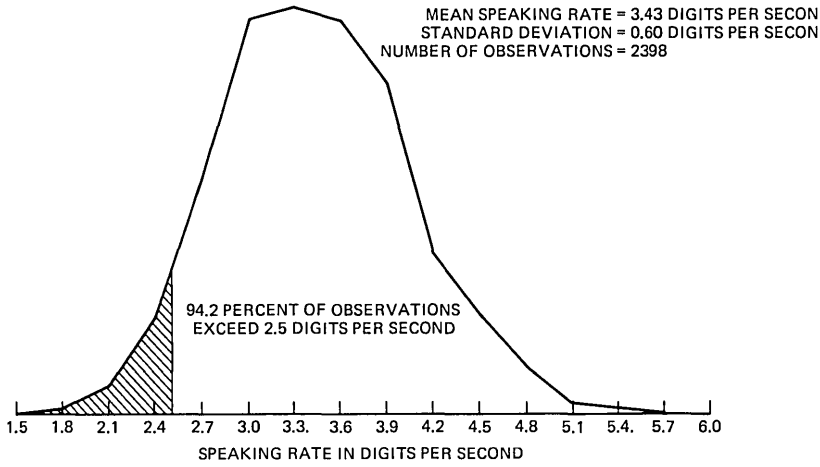


Fig. 1—Distribution of speaking rate for calls classified as acceptable.

could successfully interact with a speaker-independent, connected-speech ASR system. First, assuming that any such system will have considerably more difficulty recognizing "oh" than "zero" (due to the relatively small amount of energy in a spoken "oh" and to coarticulation effects), customers may have to learn to say "zero" in place of "oh". Second, assuming that 2.5 digits per second represent an approximate maximum acceptable speaking rate, most customers will have to lower their rates. Some means is needed of getting all customers under the critical rate without lowering the rate for many of them too much. In other words, both the mean and variance of the speaking rate distribution need to be decreased.

While these data provide some useful baseline information about spoken digit strings in a field environment, generalization of the findings to other situations must be done carefully. The CCN is a long, highly familiar number and credit card customers are experienced callers. Input characteristics may differ for shorter, less familiar digit strings or for a less experienced user population. Also, there are several important questions that could not be addressed in this study. For instance, while there were no important differences among locations on any of the variables investigated, the possibility remains that variation in speaker accent will present critical problems for any ASR system. Questions such as this cannot be answered without testing a system on actual speech samples from a field environment.

Despite its limited scope, this study was useful in guiding subsequent human factors work. The studies reported below followed directly from the questions produced by these data.

IV. CONTROL OF USER SPEECH

Having determined the speech characteristics of customers giving numbers to operators, we conducted a series of three studies designed to investigate methods of eliciting spoken numbers from users of automated network services that would be acceptable to ASR systems. These studies are summarized below.

4.1 Simulation study

The first study of the series involved a laboratory simulation of a connected-speech, speaker-independent ASR system in CCS. Each of the 30 subjects in the study was given a credit card with a 14-digit number printed with 3 3 4 4 segmentation. Subjects were then instructed to make a series of between 24 and 36 credit card calls that required voice entry of the credit card number to the simulated ASR system.

4.1.1 Prompts and feedback

After dialing a ten-digit number, most subjects heard a tone followed by a brief pause, then the following prompt:

> Please speak your credit card number in groups of four digits or less. Please wait for the numbers to be repeated back to you before speaking the next group. If the numbers are repeated incorrectly, say the word "error." Then when you hear the tone, repeat the last group of numbers spoken. Listen again for the response and then proceed with the next group of numbers. At the tone, please begin speaking your credit card number.

A similar prompt was used for another group of subjects that was not given feedback of the spoken digit groups. The instructions for this group included the phrase "say zero instead of oh." This phrase was not included when feedback was provided because we wanted to see the effect of feeding back "zero" when the subject said "oh." The subjects did not know that digit recognition was done by a human who keyed in the spoken segments of the number, triggering voice feedback of the spoken digits to those subjects selected to receive feedback.

We attempted to implicitly control speaking rate by means of the rate at which feedback was given; digits were fed back at either 2.50 or 1.25 digits per second. To investigate the effect of system recognition error rate on user input, two different digit recognition error rates of 1.8 percent and 5.4 percent were simulated. Each subject experienced at least two different combinations of feedback rate and error rate.

4.1.2 Summary of results

Subjects were initially told that once they thought they no longer needed to hear the prompt they could begin to speak after the first

tone. The typical subject would listen to the prompt on the first one or two calls, then begin to speak after the first tone on all subsequent calls. Despite the length of the prompt and the limited exposure to it, subjects had little difficulty following the instructions. All subjects in the feedback condition corrected errors smoothly and consistently. There were very few departures from the digit vocabulary (0.1 percent of all calls) and subjects nearly always used proper segmentation (99.8 percent of all calls).

While all subjects not given feedback followed the instruction to say "zero" instead of "oh," the feedback "zero" less successfully induced subjects who received feedback to say "zero". The proportion of those subjects saying "zero" on the first call was 0.3; this increased to 0.7 by the twelfth call.

The mean speaking rate across all conditions was 2.58 digits per second. While this rate is lower than the rate of 3.43 digits per second observed in the field, the speaking rate did not vary significantly as a function of either feedback (none, slow, fast) or system error rate. Subjects in the feedback conditions did lower their speaking rate slightly (by 0.2 digit per second) when correcting a system recognition error, but then returned to the higher rate on the next call. Thus, under the conditions of this study, user speaking rate was not sensitive to feedback, either in terms of feedback rate or recognition error rate. However, it should be noted that the simulated-recognition error rates in this study were independent of a subject's speaking rate. In a real ASR system, the system error rate would increase with increasing user speaking rate, possibly making users more sensitive to feedback.

4.2 Telephone prompt study

The simulation study showed that some aspects of subjects' speech could be effectively controlled (at least in a laboratory setting) with simple prompting. However, the instructions in that study did not attempt to control speaking rate. Therefore, the next study concentrated on the speaking rate problem. Instead of bringing subjects into a laboratory stimulation to evaluate prompts, we chose to contact subjects by phone in their own work environments and ask for their cooperation in a very brief experiment that required them to simply say their own home phone numbers, including the area codes. Participants heard one of a set of recorded prompts and responded by speaking their numbers. All subjects were employees at Bell Laboratories in Holmdel, New Jersey. Spoken home telephone numbers provided a stronger test of our ability to control speaking rate, since a highly familiar number is likely to be spoken more rapidly than an unfamiliar one.

4.2.1 Candidate prompts

Table I gives the four components from which six prompts were composed. Three of the prompts began with Component 1 in Table I; one of these was completed by adding only Component 2, while the other two were completed by adding both Component 2 and either the first or second sentence of Component 3. The remaining three prompts were the same as the above three, except that Component 1 was omitted.

This set of six prompts allowed separate evaluation of the effect on subjects of knowing that they were speaking to a machine and the effect of specific instruction on how to speak. The first sentence of Component 3 in Table I (rate instruction) was designed simply to lower the speaking rate; the second (isolation instruction) was designed to elicit isolated speech. Absence of either of those two sentences from the prompt provided a control condition. A total of 60 subjects provided data on these prompts, 10 for each of the 6 prompts.

4.2.2 Summary of results

Table II gives the mean speaking rate for each prompt, corrected for intersegment pauses. A 3 × 2 analysis of variance showed that informing subjects that they were speaking to machines significantly lowered speaking rates ($F = 6.04$, $p < 0.03$). Also, specific instruction about how to speak significantly affected speaking rate ($F = 7.83$, $p < 0.01$). The interaction between the machine information and specific instruction conditions was not significant ($F = 0.51$).

One other instruction evaluated in an early phase of this study deserves mention. This instruction was worded as follows:

> At the tone, please speak your telephone number. Say "zero" instead of "oh." Speak at the following rate: [Recorded voice speaking, "One, two, three (pause) four, five, six"].

This instruction produced the lowest mean speaking rate (1.31 digits per second) of any prompt evaluated. Prior to playing the prompt, the experimenter told the subjects that they would speak to a machine, instead of including that information as part of the prompt.

As can be seen in Table II, even when no machine information and no specific rate instruction is given, the speaking rate is still considerably lower than that observed in the field study. As in the simulation study, this probably occurred because the subjects knew they were

Table I—Components of telephone study prompts

1. This is a machine which recognizes human speech.
2. At the tone, please speak your telephone number. Say zero instead of oh.
3. Speak slowly and distinctly. OR Pause briefly after each digit.

Table II—Speaking rates in the telephone prompt study (digits/s)

	Machine Information	
Instruction	Absent	Present
Control	2.41	2.04
Rate	2.18	1.54
Isolation	1.56	1.36

participating in an experiment and were making a special effort to speak clearly. That special effort cannot be expected in an actual service environment. Nonetheless, the results clearly show that simple instructions can considerably lower the speaking rate.

Beyond the fact that the isolation instruction lowered user speaking rates, it is of interest to know if it also succeeded in eliciting isolated speech. Although the speech given in response to the instruction sounded adequately isolated to the listener, the data were not recorded in a form that allowed a more thorough treatment of the isolation question. This question is directly addressed in the next study.

4.3 Speech isolation study

Up to this point, the reported work has focused primarily on issues surrounding connected-speech recognition. For reasons related to a larger, ongoing ASR project, the focus of our human factors work now shifted to the problem of eliciting speech acceptable to an isolated speech recognition system. We wanted to see whether it was possible, through the use of prompts alone, to obtain isolated speech from subjects. While it is possible to force isolation through the use of a pacing cue or feedback after each spoken item, such techniques tend to produce slower input from the user than is required by the ASR system. Particularly in applications involving entry of long digit strings (such as credit card numbers), experienced users may find paced entry tedious. We adapted the procedure used in the previous study to the current needs. Besides attempting to develop in this study an initial prompt that would produce isolated speech, we also investigated the use of a reprompt to be used if the speech given in response to the initial prompt was not adequately isolated. Each of 90 subjects (60 Bell Laboratories employees and 30 from the surrounding community) heard one of a set of prompts over the telephone and responded with their home telephone numbers. Because of the concerns of the larger project, subjects were asked to give their numbers without the area codes.

4.3.1 The initial prompt

Based on the results of the previous study, we selected two variations

of each of two prompts for evaluation as the initial prompt. The candidates were:

1. At the tone, please say your number. Pause briefly between digits.
2. At the tone, please say your number. Pause briefly after each digit.
3. At the tone, please say your number. Pause between digits, like this. (Recorded voice speaking, "One, two, three")
4. At the tone, please say your number, as follows: (Recorded voice speaking "One, two, three")

The phrase "say zero instead of oh" was not used because reliable recognition of "oh" by an isolated speech system is not as difficult as it would be with a connected speech system, due to the absence of coarticulation effects. Also, in an attempt to keep the prompts as short as possible, we did not include a sentence telling customers that they would speak to a machine.

4.3.2 The reprompt

To evaluate the effect of a second attempt, a subject whose speaking time did not exceed 5.0 seconds on the first attempt received a second prompt, as follows:

We're sorry, would you please say your number again, but pause longer between digits.

4.3.3 Evaluation of isolation

Since the central question in this study was whether subjects were producing isolated speech, we needed a definitive test of isolation. This was provided by recording subjects' speech on analog tape and sending the tapes to the Bell Laboratories Acoustics Research Department at Murray Hill, New Jersey, where they were processed for end-point detection. Of primary interest were the number of isolated segments detected in each subject's speech. Since subjects spoke their own seven-digit home telephone numbers, seven segments should be detected on a number spoken with correctly isolated speech.

4.3.4 Summary of results

After gathering data from 40 subjects, 10 for each prompt, it was evident that Prompt 4 was unacceptable. Four of the ten subjects hearing this prompt responded with three digits or expressed confusion. Prompt 4 was therefore eliminated from further consideration. Data were then collected from an additional 30 subjects, 10 for each remaining prompt. After an initial evaluation of these data, 20 more subjects were added to better discriminate between Prompts 1 and 2.

For Prompt 1, 17 of the 30 subjects took longer than 5.0 seconds to say their telephone numbers on the first attempt and were therefore

not given the reprompt. The end-point detector found seven isolated segments in 14 of those 17 first attempts and six segments in each of the remaining two (one was missing from the tape). Of the 13 subjects given the reprompt, seven segments were detected on the first attempt for two subjects and on the second attempt for 12. The speech level was too low to segment on the second attempt of the remaining subject.

For Prompt 2, 22 of the 30 subjects were not given the reprompt. Of these, 18 achieved perfect isolation as determined by the end-point detector. Five segments were detected for two subjects, six for another, and again one subject was missing from the tape. None of the eight subjects given the reprompt spoke seven isolated segments on the first attempt. Six of the eight spoke perfectly isolated digits after the reprompt. One of the remaining two gave three segments and the other gave four.

For Prompt 3, 19 of the 20 subjects were not given the reprompt. Of these, 14 were determined to have spoken seven isolated digits. One spoke seven digits in six segments and another in five. The end-point detector found five digits in five segments for two subjects, with the speech level on the remaining two digits being too low to segment. For a third subject, there were six isolated digits, plus one with too low a level. The one subject given the reprompt spoke six segments on the first attempt and seven on the second.

The results of this study are an encouraging indication that the combination of an initial prompt with a reprompt can be effective in eliciting isolated speech. The differences in effectiveness among the initial prompts are not large, but are in favor of Prompt 3. However, since that prompt is considerably longer than either of the others, use of one of the comparable shorter prompts is preferable in any real application.

V. CONCLUSION

The work reported above represents the initial human factors steps toward eventual use of ASR in Bell System network applications. That effort has been concentrated on those human factors questions surrounding the user-system dialog. As indicated in Section II of this paper, there are many other human factors issues that must be faced before any network application of speaker-independent ASR will be possible. However, our work on those remaining issues is beyond the scope of this paper. Also beyond the present scope are those potential telecommunications applications of both speaker-dependent and speaker-independent ASR outside the network. This broader range of applications raises several human factors questions in addition to those considered here.

VI. ACKNOWLEDGMENTS

Several people deserve thanks for their contributions to the work summarized here. H. Holinka assisted in all of the studies. L. Chapman, a visiting student from the University of Texas at El Paso, helped plan and collect data for the laboratory simulation. T. M. Gruenenfelder was responsible for analysis of the data from that study. M. L. Viets was heavily involved in all phases of the telephone prompt and speech isolation studies. C. J. Karhan served as primary coordinator for the project of which the speech isolation study was a part. I would also like to thank L. R. Rabiner and J. G. Wilpon of the Acoustics Research Department at Bell Laboratories for evaluating the isolation of the recorded speech. Finally, special thanks are due E. A. Youngs, who initiated the human factors effort reported here and has been a constant source of ideas and encouragement.

REFERENCES

1. T. B. Martin, "Practical Applications of Voice Input to Machines," Proc. IEEE, *64* (April 1976), pp. 487–501.
2. J. M. Baker, "How to Achieve Recognition: A Tutorial/Status Report on Automatic Speech Recognition," Speech Technology, *1*, No. 1 (Fall 1981), pp. 30–43.
3. C. S. Meyers and L. R. Rabiner, "A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition," B.S.T.J., *60*, No. 7, Part 2 (September 1981), pp. 1389–1409.
4. A. E. Rosenberg, L. R. Rabiner, and J. G. Wilpon, "Recognition of Spoken Spelled Names for Directory Assistance Using Speaker-Independent Templates," B.S.T.J., *59*, No. 4 (April 1980), pp. 571–92.
5. S. E. Levinson, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-27*, No. 2 (April 1979), pp. 134–41.
6. L. R. Rabiner, private communication.

AUTHOR

John E. Holmgren, B. S. (Psychology), 1965, University of Wisconsin; Ph.D. (Mathematical Psychology), 1970, Stanford University; Bell Laboratories, 1979–1982; American Bell, 1982—. In the Human Factors Department at Bell Laboratories, Mr. Holmgren worked on potential applications of automatic speech recognition. His other major responsibility was in the area of audiographics teleconferencing. In July 1982, Mr. Holmgren transferred to American Bell, where he is a Group Supervisor in AIS Net 1000 Research and Development. Member, Psychonomic Society.

New Functions for Technology

This section describes the *UNIX*™ Writer's Workbench software, its philosophy (L. T. Frase), the programs (N. H. Macdonald), and people's response to the programs (P. S. Gingrich). The programs do many things that an editor does when proofreading and commenting on a paper, and they have stimulated much public interest. The programs have been discussed on the *Today* show, and in articles in *Time*, *Discover*, and other magazines and newspapers. Why are the programs so interesting? Part of the answer is that they relieve us from the drudgery of proofreading. But part of the answer is also that such programs come close to helping as a human would, without the embarrassment of personal criticism.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

The *UNIX*™ Writer's Workbench Software: Philosophy

By L. T. FRASE*

(Manuscript received December 17, 1981)

Technology has dramatically increased the number and complexity of written documents. The Bell System spends over $100,000,000 annually on technical documents, with much effort devoted to review and revision. The *UNIX*™ Writer's Workbench programs assist documentation by automating copy editing and proofreading tasks. These programs deliver detailed measures and comments about text readability, punctuation, word use, abstractness, and other features. Authors use the system to evaluate draft documents, and many feel the programs improve their writing skills. The programs have been used for quality control and text research, and in writing courses. This paper discusses principles used in system development, introduces the two following papers, and suggests possible uses for automated language analysis aids.

I. INTRODUCTION

Massive growth in the number of documents and in the sophistication needed to understand technical language has created challenges for many organizations, outside and within the Bell System. The Naval Air Systems Command, for instance, supplies technical manuals

* Bell Laboratories.

for 135 aircraft. There are over 25,000 manuals, totaling 3,000,000 pages. In 1950, the manuals for one aircraft contained fewer than 2,000 pages; today, the manuals for one aircraft contain nearly 300,000 pages.[1] In the Bell System, there are over 100 general categories of documents. Just one of these categories, the Bell System Practices, contains 400 broad document categories, with 35,000 active titles. Current issues are routinely distributed to 45,000 locations. But number of pages is not the only concern. In some states, if documents are relevant to the physical or financial well-being of individuals, laws dictate how difficult the language of a text may be. Recent "plain language" legislation, passed by six states, allows courts to rewrite texts that are unclear.[2]

The telecommunications industry, along with others, is faced with two substantial problems: (1) how to manage a large number of documents, and (2) how to make them understandable. Review and evaluation of technical documents is costly, time-consuming, subjective, and often tedious. Mechanized aids for reviewing documents, although not relieving humans of responsibility for document quality, can simplify and speed the documentation process. Efforts elsewhere in Bell Laboratories address the problem of managing a large number of documents.[3,4] The *UNIX** Writer's Workbench software was designed to help improve quality.

Until recently, document technology has consisted of two parallel, but separate, lines of work. The first consists of experimental research on what makes text easy or difficult to understand. This work, reflected in psychological and linguistic research, includes studies of the effects of text format, wording, and graphic displays.[5] The second line of work includes more intuitive or prescriptive bases for writing and text design. This involves the expertise of text and graphics designers,[6] such as technical editors and layout specialists, as well as rhetorical traditions that specify writing standards.[7] Both knowledge bases, one derived from experimental research and the other from nonexperimental sources, provide useful guidelines for effective text design.

Unfortunately, text researchers and designers have not always worked closely together. Psychological theories of what makes a document difficult to understand often seem oversimplified to designers, and designers' intuitions and prescriptions often seem vague to researchers. Designers have been concerned with creating good documents, while researchers have been trying to understand problems of text comprehension.

The use of computers for language analysis is beginning to stimulate

---

* Trademark of Bell Laboratories.

integration of technical knowledge from research and design areas. McMahon, Cherry, and Morris,[8] for instance, have shown how language analysis, such as vocabulary analysis or assigning parts of speech to words, can be done easily and accurately by computer. Humans find those tasks difficult and tedious. Models of decision processes used by experts have been incorporated into programs that make design recommendations, such as the most suitable column width to use for a particular text.[9] Knowledge-based expert systems,[10] developed by researchers in artificial intelligence, provide additional ways of expressing, in computer-readable language, expertise derived from different disciplines. Development of the Writer's Workbench system has required, and continues to encourage, this contact between different disciplines. The programs recognize the validity of principles derived from research and the validity of principles derived from intuition and practical experience. Hence, they help bring research and application closer together.

The Writer's Workbench system provides a simple set of commands that deliver many assessments needed in documentation work. These include editorial comments on punctuation, word use, spelling, and text abstractness, along with analyses of grammatical parts of speech and calculations of overall text readability (expressed as reading grade levels). The programs combine work done by the Computing Science Research Center and the Human Performance Engineering Department of Bell Laboratories.

Using routines already developed for *UNIX* software, such as spelling and parts of speech analysis programs, developers were able to build programs that expanded and went beyond existing facilities. For portability, the wide distribution of *UNIX* operating systems in Bell Laboratories was an advantage. An early version of the programs was tried on several *UNIX* systems for four months before the first general release within Bell Laboratories. During those months, program efficiency, outputs, and options were modified in response to user feedback. Although current versions of the programs incorporate further enhancements and pruning of esoteric programs, the system architecture has remained the same.

Text to be analyzed must be stored in the computer. Some authors prefer to compose text on line; others have it entered by clerical staff. Text should contain standard formatting commands, for instance, commands that define headings and other special elements. To analyze a text, the user logs in to the system and issues a command name followed by the name of the file that contains the text. The program then executes, prints analyses, and perhaps suggests changes to improve the text. Currently, the programs are appropriate for proofreading and stylistic analyses of expository and descriptive prose, but not

for procedural documents consisting entirely of lists or abbreviated step-by-step directions.

The Writer's Workbench system is limited by the text features that the programs can recognize, and also by our understanding of the relation of text features to reading difficulty. For instance, one program recognizes when the first word in a sentence is not capitalized; it is more difficult to develop a program that can recognize when important content is missing. Currently, the programs do not use linguistic parsing algorithms; they cannot "understand" the subject and object of a sentence, for example, as humans would. Nevertheless, the richness and variety of feedback delivered by the programs are different enough from commonly encountered word and text processing systems that we find it useful to refer to these expanded capabilities as "language processing" functions. Hence, we have used this terminology here and in the following two papers, perhaps as much to remind ourselves of what remains to be done as to characterize the current programs.

II. DESIGN PRINCIPLES

Six specific principles, reviewed below, guided Writer's Workbench program development. The principles provide a context for the detailed program descriptions in the paper that follows this.[11] To summarize, the principles assume that language analysis programs should be rational, diverse, evaluative, modifiable, specific, and informative.

2.1 Expert knowledge base

The Writer's Workbench programs are *rational* in the sense that they are based on research or expert consensus that justifies the relevance of program information for text comprehension or use. Some program outputs are based on psychological research, showing, for instance, that passive sentences are more difficult to understand than active sentences.[11] Other programs are based on advice from writing and style experts.[7,12] For instance, standard punctuation usage is evaluated by some programs. Knowledge about the features of good writing is scattered among disciplines; hence, no single discipline is entirely adequate for formulating language analysis programs. Currently, we have drawn from rhetorical and psychological literature to provide text assessments that people commonly use in document evaluation. To do more, contributions from artificial intelligence and related fields are required. In addition, since not all features of document design can be clearly supported by research or expert opinion,[13] programs contain messages directing users to information sources that explain certain evaluations.

2.2 Program diversity

Written communication is so complex that single measures, such as an overall index of text readability, can draw an author's attention away from other important text characteristics. Therefore, we wanted to provide *diverse* measures so users would think about a constellation of features. For instance, one program delivers over 40 measures of text characteristics, some measuring word characteristics, others measuring sentence characteristics.

We assume that authors have diverse needs, which cannot be foreseen entirely. Hence, the Writer's Workbench system contains many programs and options that can be used, or not, as an author desires. The diversity of programs works as a structured system to let users explore characteristics of words (such as syllables), sentences (such as length in words), or paragraphs (such as beginning and ending sentences).

2.3 Relative judgments

A text is good or bad depending on whether it is adequate for a particular task or audience. *Evaluative* judgments about a document usually relate to some standard of excellence. The Writer's Workbench system makes such comparisons explicit. Measures from one document are compared to measures obtained from other documents. For instance, evaluative comments about the readability of a text can be based on statistics derived from training or technical memoranda judged to be of high quality. The author is warned when the text being analyzed departs significantly from the standard documents. Ability to compare documents to various standards emulates an important component of human text judgments, and it is a distinctive feature of the Writer's Workbench system.

2.4 User modifications

Since it is not possible to anticipate the documentation needs of all users, programs should be *modifiable*. Three aspects of the programs can be adjusted by the user—standards, input, and output. For instance, one program calculates statistics for variables derived from texts supplied by an author. These statistics can then be used as standards to evaluate features of the author's subsequent texts, such as percentage of passive sentences. Other programs allow adding or deleting words or phrases that are detected by spelling and word usage dictionaries. In addition, input can be modified to include or exclude lists from analyses, and output format and length can be changed by appending simple descriptors to commands.

2.5 Text location

For authors to review exactly where language may have gone awry in a text, *specific* text locales must be identified. For instance, printing an average measure of sentence length is less useful for an author than printing sentences that exceed a specified length. Early work on identifying troublesome text locales for a reader was done by Koether and Coke.[14] Command options in the Writer's Workbench system provide summary statistics or review sentences that violate certain standards, such as length or word use. In the latter case, line numbers show where offending text can be found. This level of specificity is necessary if programs are to speed text revision.

2.6 Information resources

The Writer's Workbench system is intended to be *informative* rather than to make rigid decisions. Text statistics should be interpreted cautiously, and we suspect they will be used most effectively when an author can get advice related to those statistics. The Writer's Workbench system provides guidance, using information files, about word use, punctuation, and other features. Brief explanatory comments are also given in programs, with information about how to obtain more information.

III. RATIONALE AND DESIGN

"*UNIX*™ Writer's Workbench Software: Rationale and Design,"[11] describes program architecture, and discusses at length the literature behind programs and their options. Current literature on writing is rich in descriptions of text patterns that characterize faulty expression,[15] and contains descriptions of actions that a writer can take to correct those faults. These patterns and actions, which are the bases for program measures and recommended revisions, are described in the first part of the next paper. Next, the individual programs and their options are described. The programs are divided into those that proofread, those that comment on style, and those that comment on organization. Finally, features that help people use the programs—human factors issues—are surveyed.

IV. RESULTS OF A FIELD STUDY

The Writer's Workbench system has been used in Bell Laboratories for over three years; currently it runs on over 100 *UNIX* systems. Bell Laboratories users have responded favorably to the programs, and many believe the programs improve writing skills.[16]

Of course, there are many environments in which language analysis programs might be used. "*UNIX*™ Writer's Workbench Software:

Results of a Field Study,"[17] describes data collected from program trials conducted outside Bell Laboratories. The two outside locations included one human factors organization and one technical writing organization. The groups were engaged in different writing tasks and the staffs had different computer experience.

According to the field study, authors liked the detailed computer suggestions about how to revise their texts, and they were better able to detect errors with program help than without it.

V. CONCLUSIONS

The Writer's Workbench programs do not do everything humans do when evaluating text. They cannot determine whether important content is missing, nor do they assess format features or the meaning of texts, as humans would. Nevertheless, the programs speed editing, and they help detect features that authors might otherwise miss.

Language analysis programs have many uses. Components of the Writer's Workbench system have been used for research; for instance, to derive general quantitative descriptions of text style and to relate these styles to text difficulty.[18] They have also been used, by various groups at Bell Laboratories, to help develop technical documents for new software systems and for writing technical memoranda. Programs could be used to obtain management information, for example, to monitor changes in documents across time, and to determine the match between skills of readers and the texts they must read.[19] Finally, the programs can be used for instruction to provide detailed feedback to students about the characteristics of their writing, or by an instructor to evaluate the effects of instruction on student writing.

The Writer's Workbench programs have recently been used as an adjunct to a technical writing course sponsored by the Bell Laboratories Education Center. Such programs could be especially useful to authors whose native language is other than English. Use of language analysis programs in research, development, management, and instruction will no doubt grow as these programs increase in diversity and sophistication.

VI. ACKNOWLEDGMENTS

Programs created by L. L. Cherry contributed much to the development of the Writer's Workbench system. She also helped with program enhancements. W. Vesterman, of Rutgers University, stimulated early work on stylistic and diction programs. S. A. Keenan made important contributions to program development and documentation, as well as to Writer's Workbench system trials. J. L. Collymore, M. L. Fox, and M. F. Poller also contributed to development, and exchanges with W. F. Fox and J. J. Dever were helpful.

REFERENCES

1. W. G. Muller, "Useability Research in the Navy," in T. G. Sticht and D. Zapf (Eds.), *Reading and Readability Research in the Armed Services*, Alexandria, VA.: Human Resources Research Organization, 1976.
2. "Plain Language Laws," in "Simply Stated," Newsletter of the Document Design Center, Washington, D.C.: American Institutes for Research, No. 18(July, 1981), p. 1.
3. M. J. Rochkind, "The Source Code Control System," IEEE Trans. Software Eng., *se-1*, No. 4(December 1975), pp. 364–70.
4. M. H. Bianchi, R. J. Glushko, and J. R. Mashey, "A Software/Documentation Development Environment Built from the UNIX Toolkit," Int. Fed. Inform. Processing, Working Conf. Automated Tools Inform. Syst. Design and Development (January 1982), New Orleans, LA.
5. J. Hartley, "Eighty Ways of Improving Instructional Text," IEEE Trans. Professional Commun., *PC-24*, No. 1(March 1981), pp. 17–27.
6. E. S. Ferguson, "The Mind's Eye: Nonverbal Thought in Technology," Science, *197*, No. 4306(August 1977), pp. 827–36.
7. R. A. Lanham, *Revising Prose*, New York: Charles Scribner's Sons, 1979.
8. L. E. McMahon, L. L. Cherry and R. Morris, "UNIX Time-Sharing System: Statistical Text Processing," B.S.T.J., *57*, No. 6(July–August 1978), pp. 2137–54.
9. S. A. Keenan, "Computer Projections of the Cognitive Effects of Text Changes," Amer. Educ. Res. Assn., Boston, MA, April 7–11, 1980.
10. R. O. Duda and J. G. Gaschnig, "Knowledge-Based Expert Systems Come of Age," Byte, *6*, No. 9(September 1981), pp. 238–81.
11. N. H. Macdonald, "*UNIX*™ Writer's Workbench Software: Rationale and Design," B.S.T.J., this issue.
12. W. Strunk, Jr. and E. B. White, *The Elements of Style*, New York: Macmillan, 1972.
13. L. T. Frase, "Ethics of Imperfect Measures," IEEE Trans. Professional Commun., *PC-24*, No. 1(March 1981), pp. 48–50.
14. M. E. Koether and E. U. Coke, "A Scheme for Text Analysis Using FORTRAN," Amer. Educ. Res. Assn., New Orleans, LA, February 25–March 1, 1973.
15. L. Flower, *Problem-Solving Strategies for Writing*, New York: Harcourt Brace Jovanovich, Inc., 1981.
16. L. T. Frase, N. H. Macdonald, P. S. Gingrich, S. A. Keenan and J. L. Collymore, "Computer Aids for Text Assessment and Writing Instruction," N.S.P.I. Journal, (November 1981), pp. 21–4.
17. P. S. Gingrich, "*UNIX*™ Writer's Workbench Software: Results of a Field Study," B.S.T.J., this issue.
18. E. U. Coke, unpublished work.
19. E. U. Coke and M. E. Koether, "A Study of the Match Between Technical Document Difficulty and the Reading Skills of Technical Personnel," B.S.T.J., this issue.

AUTHOR

Lawrence T. Frase, A.B. (Philosophy), 1960, University of Miami; M.S., 1963, Ph.D. (Educational Psychology), 1965, University of Illinois; University of Massachusetts (Psychology Department), 1965–1967; Bell Laboratories, 1967–1976; National Institute of Education, 1967–1968; Bell Laboratories, 1978—. Mr. Frase has published research on human reasoning, reading, writing, and document design. While on leave from Bell Laboratories, he was head of the Learning Division at the National Institute of Education in Washington, D.C. Currently he is working on applications of computers to human information processing. He is a fellow of the American Psychological Association.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

The *UNIX*™ Writer's Workbench Software: Rationale And Design

By N. H. MACDONALD*

(Manuscript received December 17, 1981)

The *UNIX*™ Writer's Workbench software is a set of computer programs that help with two stages of document production: evaluation and editing. These programs analyze prose documents and suggest improvements. There are several types of programs: those that proofread, analyze style, and reformat the text in new ways, and those that provide information about the English language. This paper first describes the rhetorical and psychological writing principles that underlie the Writer's Workbench programs. It then describes the major Writer's Workbench programs and how they judge writing, based on these writing principles. Finally, it presents the human factors principles used in the design and development of the Writer's Workbench system.

I. INTRODUCTION

The previous paper in this issue of the *Journal*[1] pointed out the growing need for automated language processing, that is, for tools to help authors write clearly and understandably. Such tools are especially important for technical writing where the content is precise and a reader's failure to understand the text can be costly.[2] This paper

* Bell Laboratories.

describes how the *UNIX** Writer's Workbench software attempts to meet this need.

The first major section of this paper describes the principles of good writing that are incorporated into the Writer's Workbench programs. Many of these principles are supported by research showing that, indeed, different writing styles make a difference to the reader. The next section describes the major Writer's Workbench programs and relates them to the principles. The last major section describes the user considerations that guided the design of the Writer's Workbench programs.

II. SOME PRINCIPLES OF GOOD WRITING STYLE

This section presents some principles of good writing style. Research suggests that text that violates these principles is more difficult to comprehend. These principles belong to the word, sentence, paragraph, or document level, but of course, some principles overlap categories.

2.1 Word level

There are several word level principles besides the obvious ones such as using and spelling words correctly.

2.1.1 Wordiness

Style guides usually advise writers to avoid hackneyed, empty, or frequently misused phrases, such as "at this point in time" and "notwithstanding the fact that."

2.1.2 Definite, specific, concrete language

Strunk and White[3] remark that "the surest way to arouse and hold the attention of the reader is by being specific, definite, and concrete." They illustrate this principle by this pair of sentences, the first vague, the second concrete:

> A period of unfavorable weather set in.
> It rained every day for a week.

Psychological research on memory and the imageability (ability to create an image) of words suggests that texts with many abstract words will be more difficult to remember, and presumably to understand, than texts with concrete words.[4–8]

2.1.3 Word frequency and length

Coleman reported a high negative correlation between the average

* Trademark of Bell Laboratories.

frequency of content words in a passage and its difficulty to readers.[9] That is, texts made up of many infrequent words are difficult to comprehend. He also found that difficult passages contained more letters, syllables, and morphemes per word. Since frequency and length are highly correlated,[10] these two effects are possibly one effect. In keeping with these findings, Klare recommends Anglo-Saxon-based words over Latin-based words in English since the Latin form is usually longer, as in "go" versus "proceed."[11]

2.2 Sentence level

There are several obvious sentence-level principles, for instance, using correct punctuation and correct grammar. Some of the less obvious or more subjective principles are described below. These are not principles of right or wrong, but rather of better or worse.

2.2.1 Passive voice

One of the biggest writing problems, particularly in scientific writing, is the overuse of the passive voice. Historically the passive voice was used to indicate the objectivity of science and the scientist. The scientist did not state, "I found that...," but rather, "It was found that. . . ." This usage was dogma; as Einstein said, "When a man is talking about scientific subjects, the little word 'I' should play no part in his exposition."[12]

One problem with the passive voice in scientific materials is that its use has spread from obscuring "I" and "we" to many other cases as well, e.g.,

> A variable-gain control is included in this circuit.

Scientific objectivity is still served by stating:

> This circuit includes a variable-gain control.

Perhaps because of a change in scientists' perceptions of their role,[13] or perhaps because of the difficulty of the passive style, many scientists and scientific editors[14,15] now recognize and even promote the use of the active voice, including the use of first-person pronouns.

Why should we avoid the passive voice? The rhetoric books describe it as "dron[ing] like nothing under the sun," wordy and unclear,[16] and less direct and less vigorous[3] than the active voice. It may indeed be all those things, but in addition, psychological research has shown that the active versions of a sentence are recalled better[17] and verified faster.[18] Scientific texts written in the third person passive, as "It was concluded that ..." are remembered less well and appreciated less than the same content written in the active voice.[19]

Kirkman[20] took samples of scientific papers and rewrote the content in six different styles. He asked scientists and engineers to rate which version they found "most comfortable to read, easiest to grasp and simplest to digest." In three different surveys, he found they preferred "direct, active writing, with a minimum of specialist vocabulary, a judicious mixture of personal and impersonal constructions, short and uncomplicated sentences and liberal paragraphing."

These data do not imply that passive voice should never be used; at times it is preferable to

1. Emphasize the object of the sentence
2. Vary the rhythm of the text
3. Avoid naming an unimportant actor.

EXAMPLE: The mail was delivered.

However, the passive voice should be restricted to the useful and necessary cases, rather than used widely and indiscriminately.

2.2.2 Nominalizations

Nominalizations are nouns that have been created from verbs. They usually end in "ion," "ment," "ence," or "ance," e.g., "transformation," "establishment," and "admittance." The empirical case against using nominalizations is strong. Coleman[21] found that individual sentences without nominalizations were remembered better than their nominalized forms. A multiple-choice comprehension test on content failed to show a significant difference between the two versions. But on a memory task, subjects required significantly more exposures to memorize sentences containing two nominalizations than to memorize the same content written in active-verb form.

In a later experiment, Coleman[17] investigated ten different types of nominalizations and their active versions. He found that for those pairs of sentences in which the active version was phrased in two clauses and the nominalized version in one, the active was memorized more quickly. For example

ACTIVE: If he discusses the reason for the price-change, it will be appreciated.

NOMINALIZED: His discussion of the reason for the price-change will be appreciated.

Coleman[21] took passages from a psychology text and rewrote all the nominalizations, passive sentences, and adjectivalizations (verbs formed from adjectives) into active sentences. (Coleman[17] found no difference between adjectivalizations and their active forms.) He found

that students answered correctly 25 percent more questions from the rewritten texts than from the originals. In a similar experiment, subjects were asked to write the passages immediately after reading them. Subjects recalled the simplified passages significantly better than the originals.

2.2.3 Expletives

In grammar, the term expletive refers to a syllable, word, or phrase that adds no information. In particular "expletives" are words such as "it" or "there," which anticipate a later word or phrase. Thus, in "There are three solutions to this puzzle," "There" is an expletive anticipating "solutions."

Many times such expletives can be deleted, e.g., "This puzzle has three solutions," although sometimes they cannot, e.g., "It is raining." Although no research demonstrates that expletives make text more difficult, Brogan[22] argues that when the expletive deemphasizes the main verb inappropriately, the sentence should be changed. For instance,

> It is this necessity that adds to their complexity.

can be changed to

> This necessity adds to their complexity.

making "adds" more salient.

2.3 Paragraph level

2.3.1 Readability

The readability or reading grade score for a text predicts how many years of schooling a reader would need to understand it. (Units other than years of schooling are sometimes used.) The prediction is usually based on the length of the words in the text and the length of the sentences. Different readability formulas calculate the lengths differently and weight the factors differently.

As we mentioned in the previous section, the length of a word (highly related to its frequency) predicts its difficulty. Sentence length is related to sentence type, with complicated sentences usually containing more words than simple ones. Readability formulas predict reasonably well the difficulty of the text, not because sentence length and word length cause reading difficulty, but because they are highly correlated with features such as complexity and frequency, which do.

As with any predictor, these formulas can be fooled. The Dale-Chall formula[23] takes vocabulary items into consideration, but most formulas

do not and will provide incorrect readability scores for nonsense text. All the formulas will give the same values for text with the sentences input backwards or forwards. But, in general, the formulas give a reasonable prediction of text difficulty when presented with naturally written text.

Unfortunately, research has shown[11] that the comprehensibility of the text is not necessarily improved, although the reading grade score is, by simply shortening words and sentences. The best way to improve the comprehensibility of a text is to rewrite it following the principles of good writing.

2.3.2 Variation

In writing, as in most fields of endeavor, moderation is best. The previously described principles are not absolutes; some passives and nominalizations are reasonable, and in fact, variation in sentence length, structure, and type is necessary to make writing interesting and keep the reader's attention.[13] There are other more important reasons to vary sentence type, which usually varies length as well.

Writing instructors suggest that less important ideas should be grammatically subordinated to more important ones so that the grammatical structure emphasizes the logical structure. Two simple sentences can be joined by using a "that" clause or an adverb, such as "although," to subordinate one to the other. The less important sentence should be in the subordinate clause after the "that" clause or adverb. For example, the following sentences:

1. The short, simple sentence is the most comprehensible form for an individual sentence.
2. Overusing such sentences may make a document seem disjointed.

can be combined:

> Although the short, simple sentence is the most comprehensible form for an individual sentence, overusing such sentences may make a document seem disjointed.

The combined sentence subordinates sentence (1) to sentence (2), thus emphasizing that the information in sentence (2) is more important than that in sentence (1).

2.4 Document level

2.4.1 Organization

Most books on writing recommend that the first sentence of each paragraph present the topic of the paragraph or else provide a transition from the previous paragraph into a new topic.[3,13] If, for most paragraphs, the first sentence reflects the topic, then these sentences give the reader a reliable signpost to the meaning. Headings also provide signposts to topics and topic changes. A paper with good

headings and topic sentences can be skimmed easily and quickly, and even the person who reads every word will find it easier to follow.

2.4.2 Audience considerations

Style books strongly advise writers to know their audience and to write for them. This is particularly important for materials such as instruction manuals, which the reader needs to understand. Writing for the reader extends from questions of vocabulary and sentence structure to content and organization. For content and organization, Flower[24] gives particularly thorough advice.

III. PROGRAMS

Until recently, students of writing could use only books and teachers to help them. This has slowly been changing with the advent of computer programs to do some of the work. Most programs, however, have focused solely on checking and correcting spelling.[25] Several readability indices have also been automated, but in general, wire services, magazines, newspapers, and businesses still do not analyze their text with the computer, although it is often stored and edited in a computer.

Although not yet in widespread use, several text analysis systems do exist. Besides the Writer's Workbench programs[26,27] to be described here, other systems include EPISTLE,[28] an IBM project still in the research stage; JOURNALISM,[29] a University of Michigan system that provides feedback to journalism students; and CRES,[30] a Navy system to help improve the quality of technical manuals and training materials. The EPISTLE system is a business office system that will abstract the contents of incoming letters but will also correct grammatical errors in outgoing letters. JOURNALISM provides some proofreading, but because it is programmed with specific knowledge about the articles it evaluates, it is able to comment on the organization and content as well. CRES calculates the readability score for the text, flags uncommon and misspelled words and long sentences, and suggests simple replacements for difficult words and phrases.

The rest of this section describes some of the Writer's Workbench programs, focusing on those that most strongly reflect the principles discussed in Section II.

3.1 Proofreading: proofr

The most useful Writer's Workbench programs are in some ways the least interesting. Every writer can use proofreading help, since as we become more familiar with a piece of writing we become poorer at spotting errors in it.

The proofreading program, proofr, invokes five separate programs.

A three-line example of input text and its proofr output are shown in Fig. 1. Each of these five programs can be run individually, but are more conveniently run as a package. When run separately, some of the programs have capabilities not found when they are run as part of proofr. Each separate program will be discussed in turn.

3.1.1 Spelling: spellwwb

The spellwwb program, a spelling checker based on the *UNIX* system spell program,[31] allows users to have a personal dictionary of words.

```
INPUT:   Our report, "The Basic Fundamentals of Organizatonal Complexity",
         is enclosed.  Please send any recomended changes at your
         earliest convenience. thanks.

PROOFR   ********** SPELLING **********
OUTPUT:  Possible spelling errors in examplefile are:

         Organizatonal                recomended

         If any of these words are spelled correctly, later type
                   spelladd word1 word2 ... wordn
         to have them added to your spellidict file.

         ********** PUNCTUATION **********

         The punctuation in examplefile is first described.

         2 double quotes and 0 single quotes
         0 apostrophes
         0 left parentheses and 0 right ones

         The program next prints any sentence that it thinks is incorrectly
         punctuated and follows it by its correction.

         line 1
         OLD: Our report, "The Basic Fundamentals of Organizatonal Complexity",
         NEW: Our report, "The Basic Fundamentals of Organizatonal Complexity,"
         line 3
         OLD: earliest convenience. thanks.
         NEW: earliest convenience. Thanks.

         For more information about punctuation rules, type:

                   punctrules

         ********** DOUBLE WORDS **********

         For file examplefile:

         No double words found

         ********** WORD CHOICE **********

         Sentences with possibly wordy or misused phrases are listed next,
         followed by suggested revisions.

         beginning line 1 examplefile
         Our report, "The *[ Basic Fundamentals]* of Organizatonal Complexity",
         is enclosed.

         beginning line 2 examplefile
         Please send any recomended changes *[ at your earliest convenience]*.

         file examplefile: number of lines 3, number of phrases found 2

         ---------- Table of Substitutions ----------

         PHRASE                       SUBSTITUTION

         at your earliest convenience: use "soon" for " at your earliest convenience"
         basic fundamentals: use "fundamentals" for " basic fundamentals"

         ********** SPLIT INFINITIVES **********

         For file examplefile:

         No split infinitives found
```

Fig. 1—Input to and output from proofr program.

The program searches the input text and prints all words that are not in its dictionary or the user's dictionary.

In addition, the spellwwb program can be used interactively to correct spelling errors. The spellwwb program prompts the user with each possibly misspelled word. Responding to each, the user can:

1. Find all lines on which it appears in the file
2. Invoke another program to try to determine the correct spelling
3. Tell the program how to change the spelling
4. Leave it as it is
5. Save it in a personal dictionary of correct words.

The user can also specify that certain misspellings always be changed. For instance, a poor speller might store the correction "relevant" for "relevent." Then spellwwb makes all specified corrections and updates the user's personal spelling file.

3.1.2 Punctuation: punct

The punct program searches for simple punctuation errors. It recommends changes to:

1. Move commas and periods to the left of double quotes, and move semicolons and colons to the right of double quotes
2. Capitalize the first letter of sentences
3. Balance double and single quotes and parentheses.

The program enforces straightforward rules, not ones that require judgment, such as deciding whether a comma or semicolon is the appropriate mark. When punct finds an error, it prints the original line, followed by its correction of the line.

Some of the rules punct enforces are unfamiliar to many people, such as the relative position of double quotation marks with other marks of punctuation. Nevertheless, they are accepted standards of American English. The punct program directs a user who has made punctuation errors to use the program punctrules, which provides a list of pertinent punctuation rules. The user, thus, has easy access to reference information, which can be used to decide whether the suggested changes are appropriate.

3.1.3 Consecutive occurrences of the same word: double

Using context line editors for editing text increases the chance of having the same word twice in a row. The double program locates consecutive occurrences of the same word, which can be split across two lines.

3.1.4 Wordy phrasing: diction

The diction program, described by Cherry,[32] searches a text file for phrases that writing experts have classified as wordy or frequently

misused. The latest version of its dictionary also contains some phrases that may reflect a sexual bias. The program prints sentences containing such phrases and surrounds them with stars and brackets (*[ ]*). It then recommends substitutions for these phrases. For instance, for the phrase, "bring to a conclusion," it recommends using "conclude," "end," or "finish."

Users should have the flexibility of tailoring the word list to their environment. In general, "end," "stop," or "finish" may be good choices for "terminate," but for many people writing about software, "terminate" is the appropriate word. Such a user can direct the program to stop searching for "terminate" by creating a personal dictionary. Having a private dictionary of extra terms and suppressed terms increases the usefulness of the diction program.

3.1.5 Split infinitives: splitinf

The splitinf program uses a "parts of speech" analysis program, parts,[33] to find infinitives that are split by adverbs. Such split infinitives as in "to quickly decide" are the most common type. Since users may be unfamiliar with this error, those whose papers contain split infinitives are told to use splitrules, which will print grammatical information about split infinitives.

3.2 Stylistic analyses

The stylistic analysis programs provide information whose interpretation is less concrete than that given by the proofreading programs. Hence, the information is more difficult to use; following the advice faithfully can require a considerable amount of time.

3.2.1 Tabular stylistic information: style

The style program,[32] based on parts, provides 71 numbers describing the stylistic features of a text. The most important variables it reports are several readability indices (described more fully by Cherry[32]), information on the average length of the words and sentences, the distribution of sentence lengths, the grammatical types of sentences used, e.g., simple and complex, the percentage of verbs in the passive voice, the percentage of nouns that are nominalizations, and the number of sentences that begin with expletives.

The style output for this article, shown in Fig. 2, is more useful for research on the style of documents, however, than for helping inexperienced writers improve their writing style. The style table is difficult for many writers to interpret because

1. Users may not know the meaning of some terms, e.g., "expletive" and "nominalization."
2. Users frequently do not know whether the numerical values

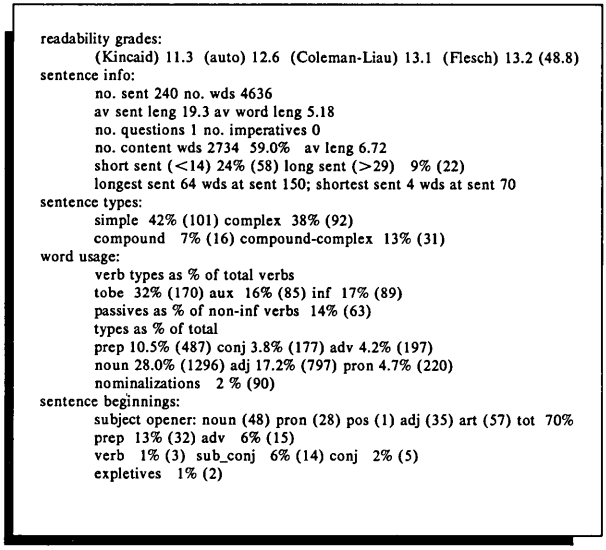


Fig. 2—Style program output for this paper.

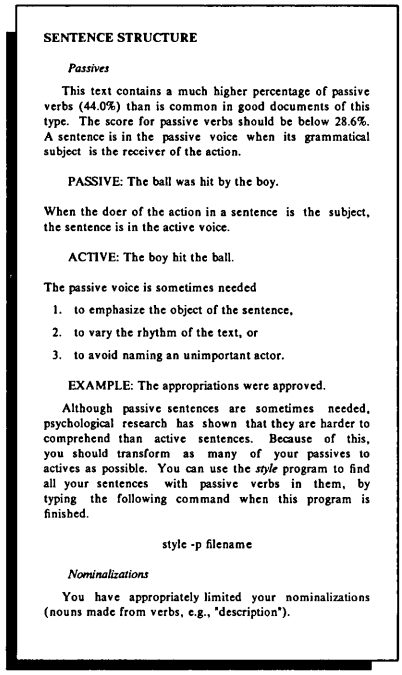
should be high or low, even for terms that are probably familiar, such as "complex sentence."

3. Users who know enough to minimize or maximize the use of some construction still do not know what numerical value is appropriate.

3.2.2 Interpreted stylistic analysis: prose

The prose program goes beyond the style program by providing the style statistics and an interpretation as well. The prose program compares the style values of the user's text against a set of standards and describes the differences in a two-to-three page output written in English sentences. A section of prose output is shown in Fig. 3.

Several sets of standards for comparison are available since texts are written for different types of readers and for different purposes. Users select which set of standards should be used in the interpretation of their text. Currently there are built-in standards for technical papers and prose training materials. These standards were set as follows. Department heads in the Bell Laboratories basic research area were asked to identify the best technical writers in their departments. These people were, in turn, asked to identify their best written technical documents (content aside). This process yielded twenty-eight documents. An editor in a Bell Laboratories training department provided 34 documents that he judged to be particularly well written. The technical and training documents were then run through the style program. The means and standard deviations of each of the style



SENTENCE STRUCTURE

Passives

This text contains a much higher percentage of passive verbs (44.0%) than is common in good documents of this type. The score for passive verbs should be below 28.6%. A sentence is in the passive voice when its grammatical subject is the receiver of the action.

PASSIVE: The ball was hit by the boy.

When the doer of the action in a sentence is the subject, the sentence is in the active voice.

ACTIVE: The boy hit the ball.

The passive voice is sometimes needed

1. to emphasize the object of the sentence,
2. to vary the rhythm of the text, or
3. to avoid naming an unimportant actor.

EXAMPLE: The appropriations were approved.

Although passive sentences are sometimes needed, psychological research has shown that they are harder to comprehend than active sentences. Because of this, you should transform as many of your passives to actives as possible. You can use the *style* program to find all your sentences with passive verbs in them, by typing the following command when this program is finished.

style -p filename

Nominalizations

You have appropriately limited your nominalizations (nouns made from verbs, e.g., "description").

Fig. 3—Section of prose program output for a poor paper.

variables were computed for each document set. These means and standard deviations now make up the standard for that document type. When the value of a user's variable is more than one standard deviation from the mean of that standard, prose recommends changes. The mkstand program can be used to create additional standards from a set of documents. This flexibility allows any writing group to tailor the standards for its particular audience.

If any of the input text's values is more than one standard deviation from the selected standard, prose explains why this may make the document hard to comprehend and how to rewrite the text to remove the problem. If scores on a variable are acceptable, and the variable is an important one, prose tells the user that the text achieved an appropriate score.

3.2.3 Stylistic problems in context: findbe

Although the style program can isolate individual sentences that contain passive verbs, expletives, or nominalizations, it is frequently difficult to know which sentences to change and how to change them without seeing the surrounding context. The findbe program automates part of a prescription for revision given by Lanham, who tells the

revisor to circle all forms of the verb "to be" and to try to replace them.[34,35] This advice agrees with some of the information presented earlier since "to be" occurs with passives, e.g.,

The difficulty of the passive voice *IS disregarded* by many writers.

with many nominalizations, e.g.,

Coleman's *discussion* of nominalizations *IS* comprehensive.

and with expletives, e.g.,

There ARE many ways to avoid expletives.

The findbe program underlines and capitalizes all forms of "to be." The text is then formatted as usual, providing the user with a paper that is easy to read, since all sentences are in context and problem areas are highlighted. This turns out to be a useful way of looking at the first draft of a paper.

3.2.4 Checking text abstractness: abst

The abst program[36] indexes the conceptual abstractness of the input text. Abstractness is defined here as the percentage of words in the text that also occur on a list of 314 words rated as abstract in psychological research.

When the percentage of abstract words is over 2.3 percent, the program suggests that concrete examples be introduced to make the document more understandable. (This cutoff was derived from the collection of good documents used to develop the prose standards.) The abstract words found in the text are saved in a file for the user to review.

3.3 Organization

3.3.1 Judging organization: org

The organization of a text is important since an appropriate structure makes it more comprehensible. For a computer program to analyze the organization of a text fully, it would need to abstract the content. Without a parser for English, or some other way of interpreting the meaning of a text, our programs cannot give feedback on the quality of the content and organization.

The org program, however, was designed to give a writer a different perspective on a text as an aid in evaluating its organization. The program formats the text and preserves headings and paragraph

boundaries, but prints only the first and last sentence of each paragraph.

For writers who follow the traditional topic sentence and concluding sentence format for paragraphs, the output can be a good abstract of the paper. Even for writers with a more casual style, seeing the overall structure of a long paper can help to improve it.

IV. USER CONSIDERATIONS

In discussing human-computer interactions, two issues should be considered: the ease of using the programs and the quality of the computer's responses.[37] This section will describe how the Writer's Workbench system attempts to optimize both aspects of the interaction.

4.1 User interactions

4.1.1 Program power

Perhaps the most important organizational decision made was to design the Writer's Workbench system hierarchically. One command, wwb, runs the most commonly used programs, the proofreading program, proofr, and the English-language stylistic program, prose. The wwb program is easily remembered as the acronym for the Writer's Workbench system, and so the casual user of the system finds the system simple to access.

For the experienced user of the Writer's Workbench programs, proofr and prose can be used individually, as can the separate components of proofr, thus allowing such users all the power they need. For one program, proofr, there is an alternate spelling, proofer, because based on the pronunciation it was often misspelled this way.

Many of the programs allow users to have their own dictionaries to tailor the output. Rather than requiring the users to type the names of the dictionaries as part of the command line, these programs use a particular file if it exists. Of course, users can still override this when they choose. Furthermore, commands were created to add words to these dictionaries, rather than requiring the users to edit them and keep them in sorted order. Thus, casual users can create personal spelling dictionaries and access them automatically with just two commands.

4.1.2 Documentation

For new users, there is now a substantial amount of paper documentation that describes the programs, how to use them, and their relation to good writing. But because users frequently do not own the paper documentation or do not have it near them, there are many on-line aids as well.

As with all *UNIX* system commands there are manual pages for each Writer's Workbench program, which are helpful to experienced *UNIX* system users. For the casual user, there is an on-line introductory system, which describes the uses of each Writer's Workbench program command.

Two other commands give on-line help. The wwbinfo command provides a list of all the commands and their functions. The wwbhelp command takes a key word as an argument and lists all the programs that have anything to do with that key word. These are words such as the following: "syllable," "prose," "passive," "sentence," and "organization."

Each Writer's Workbench program can also be run with a "flags" option, which prints that program's format and options. Further, since many of the programs have default options, these options are also echoed back with the user-specified options, as described in Section 4.2.2.

Finally, the output of many of the programs suggests other programs that would be useful. These suggestions are based on the analyses of the input text, and can thus remind users of programs that are useful for that text. For instance, users with spelling errors are told how to add correctly spelled words to their personal spelling dictionary, and the prose output suggests other appropriate programs. For example, users with too many passives are told how to find all sentences with passive verbs by using the style program.

4.2 Computer responses

What users have to remember and what they have to type are certainly important variables to consider when evaluating a system. But the quality of the messages the computer provides is also important.

4.2.1 Output length

The Writer's Workbench programs have attracted many new *UNIX* system users. For such users, the traditional "silence" of *UNIX* system commands is unfriendly,[38] e.g., if the spell command finds no misspelled words it simply stops, and the user receives a prompt (not a pat on the back). This silence is exactly what the regular *UNIX* system user wants, as McIlroy states, "Canned chitchat, like the plastic announcements on airplanes, may please newcomers, but it annoys old hands."[39] The obvious problem is that regular and casual users share machines, and what is right for one is frustrating for the other.

The first versions of most of the Writer's Workbench programs were verbose, irritatingly so, for experienced *UNIX* system users. We added "-s" options (for short) to most programs, which removed the

chitchat and most of the "pats on the back." Recently, we changed all the programs so that users can specify once what length of output they want in the future. Since expert users are best able to change this default, users who do not specify any length receive the long version.

4.2.2 Feedback

The Writer's Workbench programs inform the user when the program has started. Even experienced users may want reassurance that everything is proceeding on a heavily loaded system. Most commands echo the command line and include any options that were not chosen by the user but came about through default. This gives users a record of the exact command run; it may also remind them that they do not want a particular default, and it may alert them to options they did not know existed.

The Writer's Workbench programs cannot correct incorrect entries, but they try to provide complete, informative, and accurate error messages.

4.3 Needed improvements

To date, proofr is a major proofreading package, geared toward users who print the output on paper and make any recommended changes themselves. Since the output refers to a text problem by labeling it with the line number in the unformatted file, a user who has someone else type the text and only has the formatted output does not find these line numbers useful. For users with CRT display terminals, the output can be too lengthy to fit in the terminal's memory so that by the end of the program the first part of the output has disappeared. This makes it difficult to notice all the problems and change them.

We are currently designing two new versions of the proofr program. A version for word processing center customers will provide all the proofreading comments superimposed on the formatted output. For CRT users, we plan a completely interactive version, which will move linearly through the file and display possible corrections for the user to accept or ignore.

V. SUMMARY

This paper described some principles of good writing and some experimental results that show that readers prefer writing that embodies these principles and find it easier to understand. The paper then described a set of computer programs called the *UNIX* Writer's Workbench software. These programs help a writer isolate problems with general style as well as with individual sentences, phrases, and words. The final section of the paper described some of the human

factors principles that guided the design of the programs. The following paper in this journal[40] will describe the reception of the Writer's Workbench programs at Bell Laboratories and at two trial locations.

REFERENCES

1. L. T. Frase, "The *UNIX*™ Writer's Workbench Software: Philosophy," B.S.T.J., this issue.
2. J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel," Navy Training Command Research Branch Report 8-75 (February 1975).
3. W. Strunk, Jr. and E. B. White, *The Elements of Style*, New York: The Macmillan Co., 1959.
4. A. Paivio, "A Factor-Analytic Study of Word Attributes and Verbal Learning," J. Verbal Learning and Verbal Behavior, *7,* No. 1 (February 1968), pp. 41–9.
5. G. Frincke, "Word Characteristics, Associative-Relatedness, and the Free-Recall of Nouns," J. Verb. Learn. Verb. Behav., *7,* No. 2 (April 1968), pp. 366–72.
6. W. A. Winnick and K. Kressel, "Tachistoscopic Recognition Thresholds, Paired-Associate Learning, and Immediate Recall as a Function of Abstractness-Concreteness and Word Frequency," J. Exp. Psychol., *70,* No. 2 (August 1965), pp. 163–8.
7. A. M. Gorman, "Recognition Memory for Nouns as a Function of Abstractness and Frequency," J. Exp. Psychol., *61,* No. 1 (January 1961), pp. 23–9.
8. R. C. Anderson and R. W. Kulhavy, "Imagery and Prose Learning," J. Ed. Psychol., *63,* No. 3 (June 1972), pp. 242–3.
9. E. B. Coleman, "Developing a Technology of Written Instruction: Some Determiners of the Complexity of Prose," In E. Z. Rothkopf and P. E. Johnson (Eds.), *Verbal Learning Research and the Technology of Written Instruction*, New York: Teachers College Press, 1971, pp. 155–204.
10. G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Cambridge, MA: Addison-Wesley, 1949.
11. G. R. Klare, *A Manual for Readable Writing*, Glen Burnie, MD: Rem Company, 1953.
12. A. Einstein, *Essays in Science*, New York: The Philosophical Library, 1949, cited in Mills and Walter, p. 29.
13. G. H. Mills and J. A. Walter, *Technical Writing*, 4th ed., New York: Holt, Rinehart, and Winston, 1962.
14. R. A. Day, *How to Write and Publish a Scientific Paper*, Philadelphia: ISI Press, 1979, p. 119.
15. M. O'Connor, *The Scientist as Editor: Guidelines for Editors of Books and Journals*, New York: John Wiley and Sons, Inc., 1979, p. 163.
16. S. Baker, *The Practical Stylist*, New York: Thomas Y. Crowell Co., 1962.
17. E. B. Coleman, "Learning of Prose Written in Four Grammatical Transformations," J. Appl. Psych., *49,* No. 5 (October 1965), pp. 332–41.
18. P. B. Gough, "Grammatical Transformations and Speed of Understanding," J. Verb. Learn. Verb. Behav., *4,* No. 2 (April 1965), pp. 107–11.
19. R. D. Ramsey, "Grammatical Voice and Person in Technical Writing: Results of a Survey," J. Tech. Writing Comm., *10,* No. 2 (1980), pp. 109–13.
20. J. Kirkman, *Good Style for Scientific and Engineering Writing*, London: Pitman, 1980.
21. E. B. Coleman, "The Comprehensibility of Several Grammatical Transformations," J. Appl. Psychol., *48,* No. 3 (June 1964), pp. 186–90.
22. J. A. Brogan, *Clear Technical Writing*, New York: McGraw-Hill, Inc., 1973, pp. 147–8.
23. E. Dale and J. S. Chall, "A Formula for Predicting Readability: Instructions," Educ. Res. Bull., *27,* No. 1 (January 1948), pp. 37–54.
24. L. Flower, *Problem-Solving Strategies for Writing*, New York: Harcourt Brace Jovanovich, Inc., 1981.
25. J. L. Peterson, "Computer Programs for Detecting and Correcting Spelling Errors," Commun. ACM, *23,* No. 12 (December 1980), pp. 676–87.
26. N. H. Macdonald, "Pattern Matching and Language Analyses as Editing Supports," The Amer. Educ. Res. Assn., Boston, April 1980.
27. N. H. Macdonald, L. T. Frase, P. S. Gingrich, and S. A. Keenan, "Writer's

Workbench: Computer Aids for Text Analysis," IEEE Trans. Commun., Special Issue on Communications in the Automated Office, *30,* No. 1 (January 1982), pp. 105–10.
28. G. E. Heidorn, K. Jensen, L. A. Miller, R. J. Byrd, and M. S. Chodorow, "The EPISTLE Text-Critiquing System," IBM System J., *21,* No. 3 (1982), pp. 305–26.
29. R. L. Bishop, "The JOURNALISM Programs: Help for the Weary Writer," Creative Computing, *1,* No. 2 (January/February 1975), pp. 28–30.
30. J. P. Kincaid, J. A. Aagard, J. W. O'Hara, and L. K. Cottrell, "Computer Readability Editing System," IEEE Trans. on Professional Commun. *PC-24,* No. 1 (March 1981), pp. 38–41.
31. M. D. McIlroy, "Development of a Spelling List," IEEE Trans. Commun., Special Issue on Communications in the Automated Office, *30,* No. 1 (January 1982), pp. 91–9.
32. L. L. Cherry, "Writing Tools," IEEE Trans. Commun., Special Issue on Communications in the Automated Office, *30,* No. 1 (January 1982), pp. 100–5.
33. L. L. Cherry, "PARTS—A System for Assigning Word Classes to English Text," Computing Science Technical Report, *81,* Bell Laboratories, Murray Hill, N. J. 07974, 1978.
34. R. A. Lanham, *Revising Prose,* New York: Charles Scribner's Sons, 1979.
35. R. A. Lanham, *Revising Business Prose,* New York: Charles Scribner's Sons, 1981.
36. L. T. Frase, P. S. Gingrich, and S. A. Keenan, "Computer Content Analysis and Writing Instruction," Amer. Educ. Res. Assn., Los Angeles, CA, April 17, 1981.
37. R. W. Bailey, *Human Performance Engineering: A Guide for System Designers,* New York: Prentice Hall, 1982.
38. D. A. Norman, "The Trouble with UNIX," Datamation, *27,* No. 12 (November 1981), pp. 139–50.
39. M. D. McIlroy, unpublished paper.
40. P. S. Gingrich, "The *UNIX*™ Writer's Workbench Software: Results of a Field Study," B.S.T.J., this issue.

AUTHOR

Nina H. Macdonald, A.B. (Linguistics), 1971, A.M. (Linguistics), 1974, and Ph.D. (Linguistics), 1979, University of Michigan; Bell Laboratories, 1976—. Ms. Macdonald joined Bell Laboratories first as a Resident Visitor in the Linguistics and Speech Analysis Department, studying the role of pauses in the perception of sentences. In 1979 she joined the Human Performance Engineering Department, where she has worked on the development of the Writer's Workbench programs. Member, Linguistics Society of America, Association for Computational Linguistics.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

The *UNIX*™ Writer's Workbench Software: Results of a Field Study

By P. S. GINGRICH*

(Manuscript received December 17, 1981)

This paper describes a study in which the *UNIX*™ Writer's Workbench software was used by writers to analyze and revise the texts they wrote. For ten weeks, two groups of writers were observed: those for whom writing was their principal activity, and others for whom writing was only one of their responsibilities. These participants used the Writer's Workbench programs to produce documents related to their jobs. The results indicate that both types of writers can use the programs without changing their accustomed modes of operating. Writers found the programs helpful; they liked the immediate feedback the programs provided and the detailed suggestions on how to revise their texts. In addition, when editing prepared texts, writers found more errors using Writer's Workbench output than when they had no output available. Finally, participants thought they spent less time editing their documents when they used the Writer's Workbench programs, although total time spent on writing and editing did not change.

I. INTRODUCTION

The *UNIX*† Writer's Workbench software was developed in two major stages. The first version of the system contained 23 programs

* Bell Laboratories.

† Trademark of Bell Laboratories.

ranging from those that searched for proofreading errors to those that summarized stylistic features. These programs were used within Bell Laboratories for nearly a year, during which time data were collected on patterns of program use. A second, expanded version was then developed to reflect the experience gained from the first version.

To date, there is little formal evidence about user interaction with Writer's Workbench programs. We monitored the program on eight *UNIX* systems and observed that the programs were used most when they first became available, and then later they were used on a constant, but less frequent basis, amounting to an average of six to seven program runs per week per user. Informal discussions with users indicated that the peaks in use were due to users experimenting with a variety of input texts, good and poor, to evaluate the potential of the programs. Once they understood the potential, use dropped to a more normal level.

The present study examined the way experienced writers at two Bell System locations used the Writer's Workbench programs as a tool in performing their jobs. We were interested in examining how using the programs affected people's attitudes toward the system, the documents they write, and their modes of writing.

The Writer's Workbench programs provide feedback about many characteristics of writing. But do writers want the feedback? Which programs are used most often? Does using the programs make a difference in the way people write or in their attitudes about what constitutes good writing? How might the programs be modified to make them more useful to writers? This study was the first attempt to address these issues systematically.

II. FIELD STUDY

2.1 Background of the participants

Fourteen participants from each of two Bell System locations took part in the study. The 14 participants in Group 1 were course developers and instructors with a mean of 10.5 years service in the Bell System, and most had been in their current jobs between one and three years. Forty-three percent of the Group 1 participants had completed four years of college. Before the study began, these participants said they spent about 75 percent of their time writing and editing documents.

Of the 14 participants in Group 2, 11 were technical staff and 3 were clerks. For the technical participants, writing documents was only one of many job responsibilities; prior to the study, they reported spending about one-third of their time writing and editing documents. They also helped to design, develop, and assess systems. The three clerks were included because they processed documents for three managers

from whom we collected data. The technical participants had a mean of 13.6 years service in the Bell System, and most had been in their current jobs between one and three years. Sixty percent of these participants had completed four years of college, and half of those had master's degrees. All three managers had completed four years of college; none of the clerks had.

The two groups of writers differed in their approaches to writing documents and their knowledge of the *UNIX* operating system. Group 1 participants were more knowledgeable about the *UNIX* system and more inclined to compose, edit, and complete drafts on the terminal without the services of clerks. Such differences between groups could lead to different patterns of using the Writer's Workbench system.

2.2 Design of the study

2.2.1 Trial procedures

The study included two days of training, a one-week period for practice using the Writer's Workbench programs, and 10 weeks of using the programs in normal work activities.

During the two-day training sessions, participants completed questionnaires, learned about the Writer's Workbench system, and then used the programs. After we trained the participants, we tested them for mastery of 11 important concepts of the Writer's Workbench system, including characteristics of a document's style, specific program functions, and program options. Our training successfully conveyed the hierarchical arrangement of the major Workbench commands and the most frequently used options. Participants did have difficulty, however, in naming specific commands for given functions, perhaps because they had not yet used all the commands. All the commands, however, were explained during training.

After the training was completed, participants were encouraged to explore the Writer's Workbench programs during one week of warm-up before the official on-line record keeping began. We provided structured exercises that participants could use to run each program, as well as a handbook that contained a tutorial introduction to the Writer's Workbench programs, and documentation describing each program.

For the next 10 weeks participants used the Writer's Workbench programs. We gave them guidelines, which suggested primary programs to run on each document they wrote, and we encouraged them to use additional programs as well.

2.2.2 Data collection instruments

2.2.2.1 Questionnaires. Participants completed questionnaires before the study began, at the midpoint, and after the study ended.

These questionnaires gathered demographic information about the participants, assessed their attitudes toward computers, and determined characteristics of their type of work and mode of writing. Some questions recurred on two or three of the questionnaires to measure changes in attitudes and modes of operating.

2.2.2.2 On-line data collection. Every time a Writer's Workbench program was run, the program automatically recorded its name, the date and time, who ran it, and the name of the text file it analyzed. Some programs, such as the prose analysis programs,[1,2] also saved data about the text characteristics. We used this information to describe program use across time and also to note changes in text files.

In addition, an interactive program prompted users to rate the helpfulness of each program's output on a five-point scale. These data were collected to determine users' opinions of how well the programs analyzed each text file.

2.2.2.3 Revision tasks. During the study, participants completed four document revision tasks, which involved editing in ten minutes a 300-word passage containing planted errors. Each participant edited the four different passages once. Each passage appeared equally often in the following sessions:

Session 1. Edit one passage before training begins (no Writer's Workbench program output).

Session 2. Edit one passage immediately after training (no Writer's Workbench program output).

Session 3. Edit one passage at the end of the study. (Appropriate Writer's Workbench program output was attached to each passage.)

Session 4. Edit one passage three weeks after the end of the study (no Writer's Workbench program output).

These revision tasks provided a controlled environment in which to compare the participants' ability to improve a text at various stages of the study.

2.2.2.4 Interviews with participants. Each participant was interviewed for 25 minutes at the end of the study. We asked 11 general questions to elicit the participants' comments on how they used the Writer's Workbench programs and what they liked most and least about the system.

2.3 Results

2.3.1 User acceptance of the programs

We examined user acceptance of the Writer's Workbench programs in several ways. Program use was recorded automatically, and following each use, participants rated the helpfulness of the output. In interviews, participants also reported what they liked least and best

about the system. We found no significant difference in the way the two groups of participants used the programs or their acceptance of them. Consequently, their data were combined in most of the following sections.

2.3.1.1 Program use. Although the way participants used the programs varied greatly, on the average they ran about six programs a week.

Table I shows the number of times each program was run during each week of the study. The wwb program, along with its two components proofr and prose, and the spellwwb program were run most often. Use of the programs diminished over the course of the study, but use of the informational "help" programs, such as punctrules and wwbinfo, dropped to zero some weeks before the end of the study. It seems that participants used the "help" programs only as they learned about the Writer's Workbench programs.

2.3.1.2 Helpfulness ratings. Participants rated the helpfulness of the output each time they ran a Writer's Workbench program and generally found the programs helpful. The average rating was 3.8 on a scale of 1 (not helpful) to 5 (very helpful). The wwb and prose programs, those most frequently used by Group 1 and Group 2, respectively, had average ratings of 4.0 and 3.7 on the helpfulness scale. Table II shows the mean rating for each program by group.

2.3.1.3 What participants liked best. During the final interviews and on questionnaires, many participants said that the Writer's Workbench programs were most valuable because they gave immediate, objective criticisms on concrete problems in their text. They liked the advice from the proofr component because they found it accurate and complete.

For the prose component, participants reported that they felt the advice was objective and specific, but they often were not sure of how to make the changes it suggested. Nevertheless, many participants reported that the stylistic information from prose was an improvement over the often vague advice and subjective opinions of human reviewers.

Participants commented on three other general aspects of the Writer's Workbench system. First, it saved time (both by the speed with which it reviewed text and by eliminating delays caused by looking for a colleague to review the document). Second, it was completely private, giving the writer a chance to improve the document before anyone else saw it. Third, participants reported they were more aware of principles of good writing and of how to change their style to suit the purpose and audience of a particular document.

2.3.1.4 What participants liked least. Two criticisms of the programs surfaced during final interviews and from questionnaire responses.

Table I—Average individual program use during training, warm-up week, and trial weeks

Program	Week of Trial													
	TR	WU	1	2	3	4	5	6	7	8	9	10	11†	Total
wwb programs														
*wwb	128	111	52	43	29	22	22	38	22	21	6	9	9	512
proofr	11	29	14	15	3	6	7	8	3	1	0	2	0	99
spellwwb	22	58	11	27	13	10	20	42	20	10	20	5	4	262
punct	17	10	2	0	0	1	0	1	2	0	0	0	0	33
double	9	4	0	3	1	1	0	0	1	0	0	0	0	19
dictplus	0	3	2	6	2	0	2	0	1	0	0	1	0	17
diction	0	4	2	0	0	1	0	0	4	0	0	3	1	15
suggest	1	23	21	0	1	1	7	10	0	0	0	0	4	68
splitinf	0	2	0	0	0	0	0	0	0	0	0	0	0	2
prose	44	47	22	25	9	8	31	7	9	13	2	4	0	221
parts	1	4	1	1	1	0	2	0	1	1	0	0	0	12
style	32	27	13	5	9	5	23	5	5	3	3	2	6	138
Help programs														
spelltell	9	3	2	8	2	1	1	6	8	0	0	0	0	40
worduse	14	22	17	6	5	11	16	8	2	1	0	0	5	107
punctrules	6	5	0	0	1	0	0	0	0	0	0	0	0	12
splitrules	6	3	1	0	1	0	0	1	0	0	0	0	0	12
wwbhelp	16	8	2	1	1	0	1	0	0	0	0	0	0	29
wwbinfo	26	4	1	2	1	1	2	0	0	0	0	0	1	38
wwbstand	10	4	0	1	1	1	1	2	1	0	0	0	0	21
Other programs														
*abst	10	12	3	53	12	0	0	13	1	2	0	0	0	106
*acro	5	7	2	0	1	3	0	0	0	2	0	0	0	20
chunk	0	4	1	1	0	2	1	0	0	1	0	0	1	11
dictadd	9	17	7	2	0	1	0	2	4	0	0	3	0	45
*findbe	16	26	12	12	4	0	3	2	3	2	0	1	2	83
match	4	4	0	2	3	0	3	0	2	1	1	3	1	24
org	2	3	0	1	1	0	5	0	0	1	0	0	2	15
*sexist	10	45	6	22	2	1	0	1	0	1	0	0	0	88
spelladd	20	13	6	10	0	28	2	7	1	0	0	0	0	87
syl	4	1	1	3	0	0	0	0	1	0	0	0	0	10
topic	0	2	1	1	0	0	1	0	0	1	0	0	0	6
wwbmail	6	3	5	11	2	1	1	2	4	1	0	0	0	36
Total	438	508	207	261	105	105	151	155	95	62	32	33	36	2188

* Programs whose use was highly recommended in the trial guidelines.
† Week 11 includes only use by Group 1.

Table II—Users' ratings of Writer's Workbench programs

Program	Group 1		Group 2	
	Mean Rating of All Participants	Number of Program Runs	Mean Rating of All Participants	Number of Program Runs
wwbhelp	NR*	0	5.0	3
punct	3.0	4	5.0	2
splitrules	3.5	2	5.0	1
punctrules	NR	0	5.0	1
spelladd	3.6	41	4.9	12
wwbinfo	NR	0	4.7	6
sexist	3.0	7	4.7	33
spelltell	2.0	3	4.7	23
wwbmail	NR	0	4.6	28
spellwwb	3.9	123	4.4	56
dictplus	NR	0	4.4	14
double	NR	0	4.2	6
worduse	3.3	43	4.1	27
wwb	4.0	156	3.9	92
proofr	4.0	3	3.9	41
prose	2.5	12	3.8	94
abst	2.1	16	3.7	67
style	4.0	42	3.7	23
wwbstand	NR	0	3.6	7
acro	2.7	3	3.6	5
findbe	3.1	21	3.4	8
dictadd	3.4	9	3.4	8
match	2.5	8	3.0	4
topic	4.0	1	3.0	2
org	2.9	8	3.0	2
parts	2.3	4	2.7	3
chunk	1.0	3	2.7	3
syl	NR	0	2.3	4
diction	2.4	8	2.0	34
suggest	2.9	42	NR*	0
splitinf	NR	0	NR	0
Average (Total)	3.6	559	4.0	578

* NR means no rating.
NOTE: The number of ratings does not agree with the number of programs run shown in Table V because participants did not begin rating program output until after the warm-up week. Also, some end data was not incorporated from the last week.

First, some participants reported that they found it difficult to apply the advice they received from the prose and diction programs. They either were unfamiliar with the principle the program checked or were unsure of how the principle was being evaluated.

Participants also criticized the length of some program outputs. For example, some said that for the wwb command, the default output was too long and wwb – s output was too short. An output of intermediate length was suggested.

2.3.2 Effects on documents

What effects does using the Writer's Workbench programs have on the documents produced? Do writers locate more errors using Writer's Workbench output to edit a document? Does using the programs for

a period of time later enable writers to edit more effectively on their own?

2.3.2.1 Revision tasks. Results from the revision tasks show how the participants' ability to improve a document differed at various stages of the study. The revision tasks were completed: (1) before the study, (2) after training, (3) after 10 weeks of using the Writer's Workbench programs, and (4) three weeks after the study ended. Output from the Writer's Workbench programs was available only in the third session.

Passages used in the revision tasks were adapted from stories in the *Bell Laboratories Record* and traced the development of new Bell System operations. The average length of the passages was 300 words. Table III describes the twenty errors planted in each of the four passages.

Of the nine categories of inserted errors shown in Table III, the Writer's Workbench programs cannot provide feedback on using a word that looks like the correct word (category g), mistyping one word that produces another (category h), and making a mistake in the agreement of subject and verb (category i).

Table IV shows the results of the revision tasks. More errors were detected in Session 3 than in any other session. Having the Workbench output available improved performance compared both with where the user had no experience with it (Sessions 1 and 2) and where the user

Table III—Type and number of errors in each revision passage

Type	Number of Error(s)	Item
a	4	Spelling errors, which were not legitimate words and would be found by the spell program
b	2	Punctuation errors that the punct program could identify
c	6	Wordy phrases included in the diction phrase dictionary
d	1	Immediate repetition of a word
e	1	Split infinitive
f	3	Instances of sexist language
g	1	Case of a word substituted for another that looked similar
h	1	Misspelled word that was a proper English word and would not be found by the spell program
i	1	Case of subject verb disagreement

Table IV—Proportion of errors detected for each session*

	Session 1	Session 2	Session 3	Session 4
Mean	0.271	0.275	0.408	0.309
Standard error	0.006	0.007	0.007	0.007
Number	21	21	21	21

Session 3 vs Session 1—$t_{(20)} = 3.35, p < 0.005$
Session 3 vs Session 2—$t_{(20)} = 3.41, p < 0.005$
Session 3 vs Session 4—$t_{(20)} = 2.49, p < 0.05$

* Data are reported for those 21 participants who completed all four sessions.

had employed the Writer's Workbench system in the past but did not have its output available while proofreading (Session 4).

The data are broken down by category in Fig. 1. We see from this figure, which shows the percentage of detected errors for each session (bars 1 to 4) for each category, that Session 3 is clearly distinct from the others.

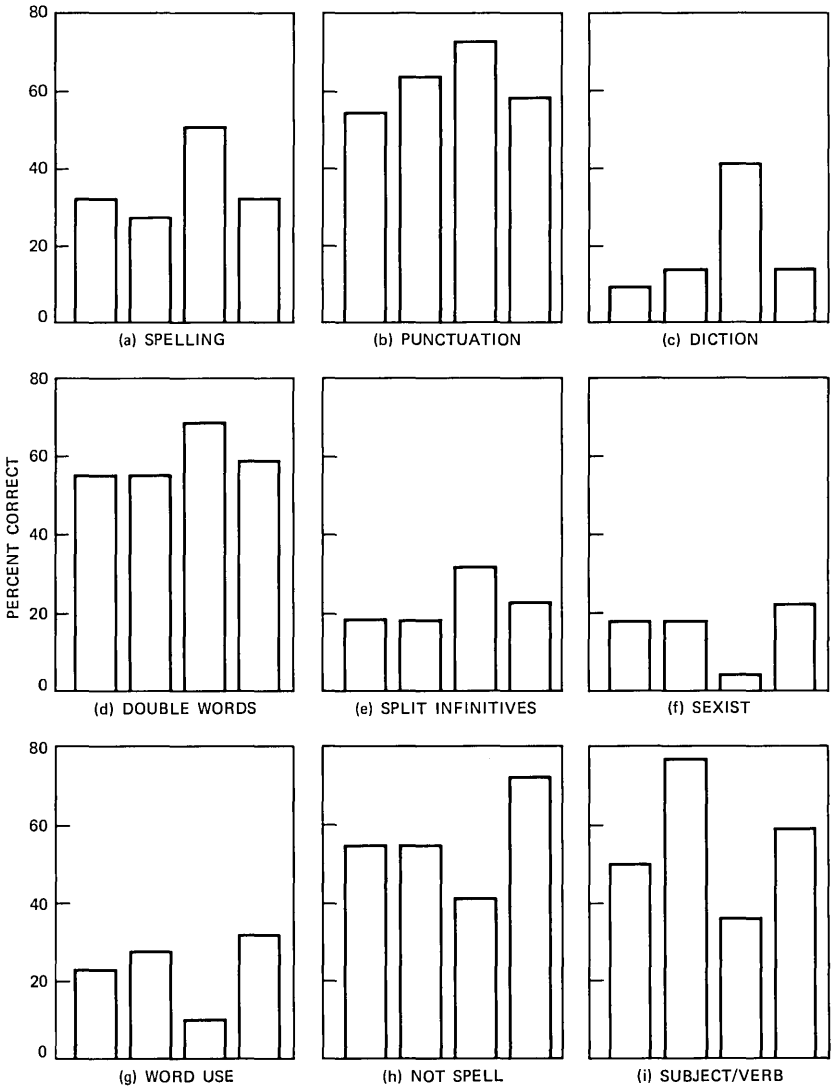


Fig. 1—Percentage of errors detected for each error category (a through i) in Sessions 1 through 4: (a) spelling errors; (b) punctuation errors; (c) wordy diction; (d) double words; (e) split infinitives; (f) sexist language; (g) incorrect word use; (h) spelling errors not found by spell program; and (i) subject/verb disagreement.

We included the errors in categories g, h, and i because we wanted to determine whether the participants had become so dependent on the Writer's Workbench analysis that they would fail in Session 3 to notice the errors that it does not find. The results of Session 3 support this supposition; participants found many of the Writer's Workbench errors and overlooked others.

It appears that participants used the time in Session 3 to find and mark the errors given by the Writer's Workbench programs, but did not have a chance to proofread the paper for other errors. Indeed, several participants remarked that they spent their time reviewing the output and locating the corresponding error in the passage, rather than proofreading the passage as a whole. As a result they may have run out of time. This seems likely in view of the low rate of marking sexist usage, the output for which came last in the collection of Writer's Workbench program outputs given to participants. As Fig. 1f shows, participants marked far fewer sexist usages in Session 3 than in any other session. Even so, overall performance was best in Session 3.

2.3.3 Effects on writers' activities

The writers completed questionnaires on how the Writer's Workbench programs influenced the way they prepared documents and assessed their stylistic features.

2.3.3.1 Mode of writing. The questionnaires asked participants how they write: Do they write on-line or on paper? Do they make their changes on-line or on paper? As we described above, before the study began participants in Groups 1 and 2 differed in their methods of writing and editing. To test whether their methods changed during the study, we performed a 2×3 (group $\times$ time) analysis of variance on each of these two items. (The range of responses to both questions represented the degree of involvement with a terminal during the task of either writing or editing. A value of 1 represented the most direct involvement with the terminal and a value of 4 represented the least.) Again, time (pre-, mid-, and post-study) was the within-subject repeated measure; and group (Group 1 or Group 2) was the grouping factor. The results show only a significant group effect for the writing method [$F(1,15) = 7.32$, $p < 0.05$], as well as for the editing method [$F(1,15) = 21.0$, $p < 0.001$]. Participants in Group 1 were less likely to have a clerk do the terminal work for them ($M = 1.43$ for writing and $M = 1.19$ for editing) than were participants in Group 2 ($M = 2.47$ for writing and $M = 2.37$ for editing). There were no significant time effects or time $\times$ group interactions.

2.3.3.2 Estimates of stylistic features. We wanted to know whether using the Writer's Workbench programs would affect the way participants evaluated the stylistic errors in their documents. The pre-,

mid-, and post-questionnaires included questions about nine features of writing style that are measured by the Writer's Workbench programs. Participants estimated whether their draft documents typically had

—too few
—the right amount
—too many instances

of each feature, or they indicated that they could not evaluate the amount by saying either

—there was no right or wrong answer, or
—they could not answer the question.

The most marked difference was in how participants evaluated their proportion of passive sentences before and after the study. Before the study, about 85 percent of the Group 1 participants said there was no right or wrong number of passives sentences, or said they could not answer. After the study, no one answered in either of those categories; half thought they had too many passive sentences in their drafts, and half thought they had the right number. Before the study, more than 50 percent of Group 2 said they either had the right number of passive sentences, or they could not answer. By the end of the study, 60 percent of Group 2 thought they either had too many, or too few, passive sentences.

2.4 Discussion

Participants were able to use the programs frequently, liked them, and found the output helpful. By reviewing the results of the different measures together, we come to a clearer understanding of what effects the Writer's Workbench programs have on writers.

First, the data on program use reveal that, on the average, participants ran six Writer's Workbench programs per week. (This figure is comparable to data collected from Bell Laboratories *UNIX* systems we have monitored in the past.) Over time, there was a reduction in the total number of programs run, possibly because the novelty of using the programs wore off, or because of the cyclical nature of the documentation process. Writing a document generally entails collecting information, organizing it, and then writing, editing, and revising. Participants may have confined their use of the Writer's Workbench programs to the first stage of the editing process and to the final version. In addition, documentation efforts vary with the phases of product development; thus, the activities that the Writer's Workbench programs can support are cyclical. Observing program use over a much longer time period, perhaps a year, would give a clearer understanding of the long-term frequency of use, one less subject to seasonal cycles of writing demands.

Even when program use declined over time, the programs were still rated as helpful. The average helpfulness rating was 3.8 on a scale from 1 (not helpful) to 5 (very helpful). The use of programs was more strongly determined by individual needs than by our recommendations. For example, use of the acro program, which locates and prints all acronyms in a document, was highly recommended in the study guidelines. Yet, users in Group 1 did not find this program helpful (2.7 helpfulness rating) because it gave them too much output. Hence, no one used it after the third week of the study. The participants did, however, continue to use the programs that took the tedium out of editing, such as the proofreading programs. That the participants continued to use many of the programs indicates that they found these programs helpful.

The results of the revision tasks clearly show that the programs are helpful in locating errors when the person is under pressures of time. Using the Writer's Workbench system in Session 3, participants found significantly more errors in the same amount of time than in the other sessions. To produce their typical error-free final products, participants would need to spend far more editing time without Writer's Workbench system than with it. The results of the Session 4 revision tasks also show that one must use Writer's Workbench programs continually to get this benefit; it is not enough to have used them in the past. The computer is just much better at finding certain errors than our participants were. However, we do not know whether error types not located by the programs would tend to be missed by the human proofreader more or less often as a result.

When we look at how using Writer's Workbench programs affects a writer's mode of writing, we see few changes. Those who used clerks to type their texts before the study continued to do so during the study. Those who typed their texts themselves continued to do so. Participants did not need to change how they wrote to use the Writer's Workbench programs effectively, nor did it increase or decrease the personal contact individuals had with terminals.

In the current study, we were limited in the degree of control we could impose on the participants. There were no rewards for using the Writer's Workbench programs and incorporating the suggestions into later drafts, nor were there any punishments for not using them. We could not control how much or how little participants wrote during the 10-week trial. We had no independent judges to evaluate the quality of the documents produced with the aid of the Writer's Workbench programs, nor did we have comparable people writing comparable texts without the aid of the programs so we could compare our participants' texts. Thus, we were limited in the issues we could address.

Currently, studies with university English composition students are addressing some of the unanswered questions. In composition classes, students are using the Writer's Workbench program output to edit and revise their essays. Instructors and independent judges will evaluate the essays and compare the quality of these essays to those of control students not using the Writer's Workbench programs. In addition, these new studies include unskilled writers rather than experienced adult writers, which will allow us to determine whether using the system helps students learn how to write better. These findings will enable us to evaluate more fully the effectiveness of the Writer's Workbench programs.

III. ACKNOWLEDGMENTS

I want to thank the people who contributed to the work of the study reported here. Nina Macdonald conducted the activities for participants in Group 1. Once the study began, Stacey Keenan monitored both sites and analyzed much of the data. Mary Fox summarized responses from hours of taped interviews. Merle Poller assisted in developing the questionnaires. I also want to thank those who participated in the Writer's Workbench system study.

REFERENCES

1. N. H. Macdonald, "The *UNIX*™ Writer's Workbench Software: Rationale and Design," B.S.T.J., this issue.
2. N. H. Macdonald, L. T. Frase, P. S. Gingrich, and S. A. Keenan, "The Writer's Workbench: Computer Aids for Text Analysis," IEEE Trans. Commun. (Special issue on communication in the automated office), *COM-30*, No. 1 (January 1982), pp. 105–10.

AUTHOR

Patricia S. Gingrich, B. S. (Education), 1970, Pennsylvania State University; M.S. (Education), 1973, Hunter College of the City University of New York; Bell Laboratories, 1979—. Before joining Bell Laboratories, Ms. Gingrich taught at Hunter College. At Bell Laboratories she worked on the development and testing of the Writer's Workbench programs. She is currently a member of the Product Strategy Department, where her work includes product evaluation and planning.

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Current Contents/ Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Science Abstracts (Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts), Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.


Bell System