Using Subjective Ratings to Select Independent Variables
in the Design of Telephone Inquiry Systems
by

Peter Jay Merkle, Jr.

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Master of Science
in
Industrial Engineering and Operations Research

APPROVED:

Robert C. Williges, Chairman

John G. Casali

Beverly H. Williges

August 31, 1988
Blacksburg, Virginia

Using Subjective Ratings to Select Independent Variables
in the Design of Telephone Inquiry Systems
by

Peter Jay Merkle, Jr.

Dr. Robert C. Williges, Chairman
Industrial Engineering and Operations Research
(ABSTRACT)

This thesis describes a two part research program in which the applicability of subjective ratings to the selection of independent variables was evaluated. The first portion of the research reviewed a case study involving the application of complex system investigation to the development of a telephone inquiry system. A telephone inquiry system is one in which users seek information in a data base by calling the system, listening to information presented by a synthetic voice, and directing movement through the database with commands on the telephone keypad keys. The complex system investigation method used included identifying the independent variables by brainstorming, then reducing the list by subjecting the variables to literature review, feasibility analysis, relevance analysis, and subjective ratings of the factors based on a prototype system. Variables which were not likely to have an immediate impact on human performance in the system were set to a constant value. The use of subjective ratings to select independent variables stems from the need to reduce large numbers of independent variables to a list which can be used as cadidates for a screening study. The result of the case study was a list of 19 candidate factors suggested for implementation in a screening study. The second portion of the research describes an experiment in which 5 independent variables (number of steps in a search, adapting speech rate,

transaction summary, native/non-native, and sex of the voice) were chosen to represent the 19 candidate factors in an experiment testing the validity of the subjective ratings technique. The results indicated that the subjective ratings of the prototype system were effective in predicting performance and subjective ratings. The impact of these results on the methodology and telephone inquiry systems is also discussed.

ACKNOWLEDGEMENTS

This author would like to thank his chairman, Dr. Robert C. Williges, who was patient and resourceful in guiding this research. This author would also like to thank the other members of his graduate committee, Dr. John G. Casali, and Ms. Beverly H. Williges, for their unfailing support throughout this research. The author wishes to acknowledge the support of the National Science Foundation under Contract No. IRI-860-4793. For the conduct of this research Dr. H. E. Bamford served as technical monitor. The author also wishes to thank Mr. Calvin L. Selig for assistance in the development of software for this research and Mr. C. R. Arrington for allowing the author to borrow his Apple Macintosh and invade his home. The author would like to thank by Dr. Jesse C.Arnold for his contribution to the statistical analysis. This research is distinguished by the fact that it was conducted simultaneously with two other experiments. This author would like to thank his research colleagues Mr. Douglas B. Beaudet, and Mr. David W. Herlong. The intellectual and personal contributions made by these two gentlemen are immeasurable. Finally, this author would like to thank his family and friends for faith in his ability and encouragement when his faith dwindled. Specifically, he would like to thank his parents Mr. and Mrs. Peter J. Merkle and Ms. Michele M. Motosko.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

INTRODUCTION

In 1947, Fitts suggested that data collection (research) serves a three-fold purpose in human factors: (1) to identify the most important variables, (2) to determine quantitative relationships, and (3) to apply to equipment design. Fitts emphasized the need to conduct research which not only describes the seminal variables, but is also directly applicable to equipment design. According to Fitts, data should not be an end, but a means by which equipment design is improved. In principle, Fitts' outlook has been adopted by the profession. Sanders and McCormick (1987, p. 20) cite the applicability of research results to equipment design as the feature which separates human factors research from other types of behavioral research.

Forty years later, the purpose of data collection remains a germane issue. In 1987, the Human Factors Society published a series of three commentaries in the *Human Factors Society Bulletin* (Smith, 1987, Simon, 1987, and Rouse 1987) and held a panel discussion at the 31st Annual Meeting, all of which addressed the ability of human factors research to develop useful data applicable to equipment design. Smith (1987) initiated the series by predicting the complete demise of the discipline unless researchers tailored programs more toward design engineers. In remedy, Smith suggested increasing the generalizability of results and implementing a data bank for these results. Simon responded with an article which highlighted his previous work and addressed Smith's solutions. Like Smith, Simon (1987) supported an increase in ability to generalize results, but disagreed with the idea of a data bank on the basis that such an effort would fail because of lack of cooperation. Rouse replied stating that the real problem was not a lack of data, but a lack of tools and organizational structures which permit the data to be used.

Like Fitts, Rouse (1987) states that data are not an end. To Rouse data are "a means whereby we test our hypothesis about human abilities, limitations, and preferences." It is Rouse's viewpoint that data provide limited information which assist in design, and that data extrapolation is necessary to answer designer's questions.In support Rouse cites the results of structured interviews with designers in eight aerospace companies (Rouse and Cody, 1986, 1987). In these interviews, designers reported instances when human factors data were required, but that these were outnumbered by the times when novel solutions were required.

Many of the complaints about data are directed at the inability to answer basic questions frequently presented to design engineers. About such basic questions as "How much fidelity in a simulation is enough?", Smith (1987) comments

> What is particularly perplexing is that at this very moment, the bulk of the empirical community in human factors is conducting research related to most of these issues. What is even more perplexing and frustrating is that this has been the case for 25 years...But this covert ignominy will not last forever, and when the dam breaks there are words in the English language such as ineffable and egregious just waiting to be used.

The 1987 articles are not the only ones critical of human factors methods and data. The most prolific critical writer has been Simon, who reports that the greatest problem in human factors methodology is the use of small experimental designs which account for little of the overall performance. Simon (1973) reviewed 239 experiments published in Human Factors between 1953 and 1972 to determine the state of the methodology. In over 60 percent of the experiments, the purposefully varied factors accounted for less than half of the total performance variance. The number of factors varied ranged from 0 to 7 with a median of 2 factors. Experimental designs with 4 of more factors accounted for less than 8 percent of the studies. Similarly the median number

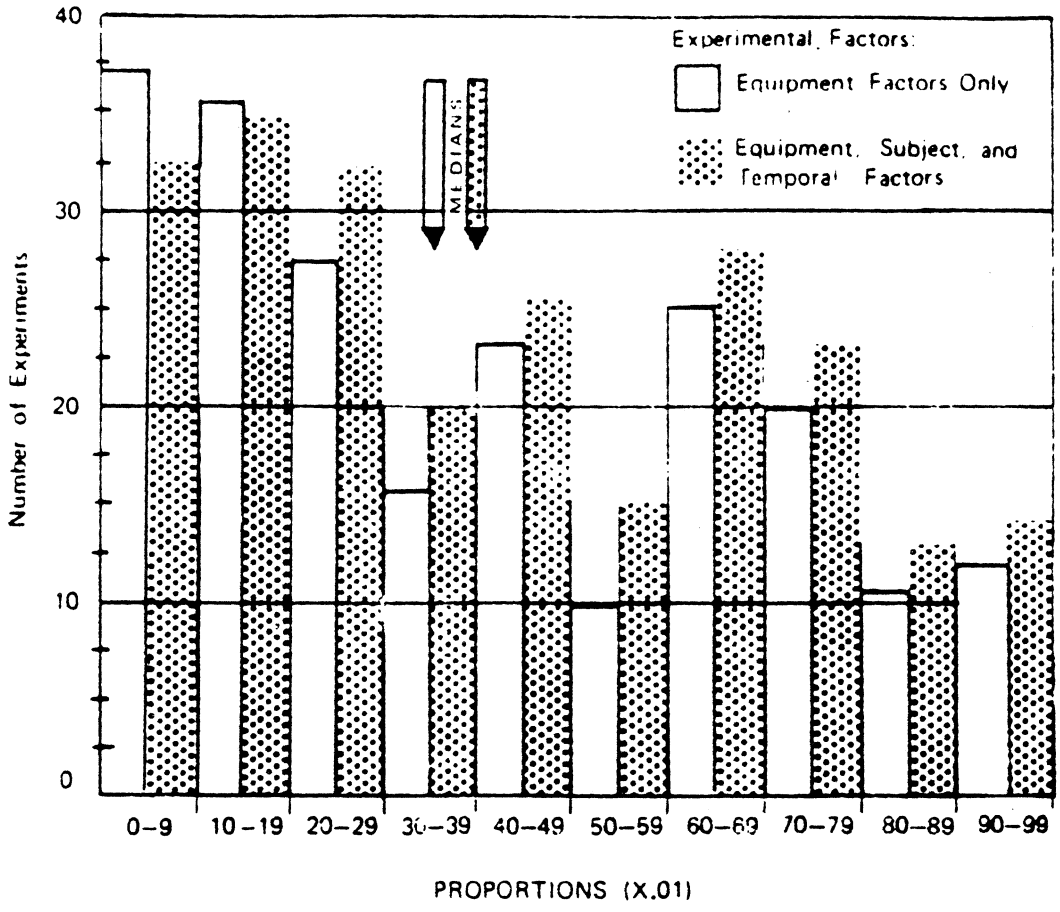
of levels for an equipment factor was 3. Figure 1 and Table 1 present summaries of Simon's analysis.

Simon (1977a) also reports a similar review of 374 human resource articles published between 1973 and 1979 conducted by the General Accounting Office of the United States. In this review, field personnel reported that only 164 of the 374 publications which had intended to support changes in policies, manuals, training programs, and equipment were accually used. In fact, 39 of the 210 unused articles contained questionable results.

Chapanis (1963) reached a similar conclusion regarding questionable results in a review of engineering psychology. In his words, "A distressing amount of literature in engineering psychology is not very good. Moreover, the flaws are not minor methodological faults, but are serious methodological ones which often invalidate the authors conclusions."

Overall, the criticisms focus on: (1) dubious methods which produce questionable results, and (2) sound methods which produce results unusable by design engineers. The criticism of dubious methods is difficult to resolve, since any solution would require a method of forcing researchers to select correct methods. Not only would this be difficult because of philosophical differences in methods, but would also meet with significant political resistance. A more viable technique for improving human factors research would be to develop research methods which directly and economically address the needs of design engineers. Economy has been added since most design engineering is preformed in an industrial setting where economy is at a premium.

Sequential experimental design is a set of methodological and organizational tools which has been developed to provide quickly and economically the data which design engineers need. Sequential experimental design describes a collection of economical statistical designs gathered under



Distribution of proportions of variance accounted for by experimental factors in 239 experiments.

Figure 1. *Summary of Simon (1973) analysis.*

TABLE 1.

Summary of Simon (1973) Analysis.

Analyses of 236 Experiments Published in Human Factors[a]

(1) Number of[b] Equipment Factors in a Single Experiment	(2) Number of Experiments in this Category	(3) Median Number of Levels per Experiment	(4) Median and Maximum Number of Replications Based on				(5) Total Number of Observations in Experiments	
			Subjects	(Max)	Trials	(Max)	(Median)	(Range)
1	71	4	6	(30)	1	(70)	72	18-1120 (7680)[c]
2	93	3	10	(64)	1	(10)	180	15-1944 (2016)
3	55	3	6	(36)	1	(12)	192	24-5184 (9600)
4	13	3	6	(18)	1	(2)	768	48-3888
5	4	2	6	(10)	1.5	(5)	1200	192-3000

[a] There were also three other experiments not included here: 2 zero-factor and 1 six-factor study.

[b] "Equipment" factors are those associated with the equipment, system, and environment.

[c] Numbers in parentheses refer to the upper limit of the total number of observations when the effects of trials represented a meaningful factor such as learning, fatigue, etc. The upper limit of the total number of observations not in parentheses refers to those experiments when trials were not a factor but treated simply as a form of replication.

an organizational plan for the investigation of complex systems. The statistical designs include modifications to analysis of variance designs such as, blocking, partial (fractional) factorials, pseudo factoring, and single subject/cell designs. Most important are the empirical models of human performance in the system (Williges, 1981). The goal of empirical model building is to provide a prediction of performance rather than a description alone. These models are aimed at providing the design engineer with a tool for predicting performance with multiple factors. The key organizational elements are screening studies and a modular, flexible experimental plan which focuses the experimental resources on the most informative portions of the problem domain.

The mathematical techniques of sequential experimentation originated in the work of applied statisticians such as W. Cochran and G. Box, who both contributed heavily to the application of statistics to experimental design in industrial settings. Cochran's avocation as a statistical consultant, exposed him to a wide variety of statistical applications, such as agriculture and human performance (Rao and Sedransk, 1984). Box, most well known for the his work in empirical modeling, developed statistical procedures for the evaluation of chemical processes (Box, Hunter, and Hunter, 1978). In addition to mathematical contributions, Cochran and Box advanced applied statistics by recognizing the need for organizational practicalities in research.

Cochran contributed most in the areas of sampling, the design of practical experiments, and the need for flexibility in research (Rao and Sedransk, 1984). To Cochran, all experiments should be viewed as sequences of experiments which iteratively refine the data to reach the research objective(s). Before undertaking large or long term experimental projects, Cochran would evaluate samples from the experimental space in an attempt to predict, and, thus avoid insignificant results. Samples were also evaluated

during long term experiments to redirect the effort toward a more informative outcome. The success of an experiment was judged by the balance between the constraints and objectives of the experimenter and mathematical elegance. Cochran's pragmatism received both praise and criticism. Those who have praised him are also those who have grappled with the same problems in applied statistics. Criticisms of his work have been cautions against allowing realistic constraints to shadow rigorous experimental controls (Rao and Sedransk, 1984).

Much of the current state of empirical modeling is a result of the work of Box, whose writings in this field have been extensive and pervasive. As mentioned previously, Box was particularly interested in the modeling and refinement of chemical processes. In these processes, it was essential to model as many factors as possible with few data points (Box, Hunter, and Hunter, 1978). As a result, Box popularized the use of response surface methodology and central-composite designs. Since the introduction of empirical modeling to human factors, it has enjoyed a rather limited yet successful use (e.g., Simon, 1977b; Williges and Baron, 1973; Williges, 1981).

One of the most comprehensive adaptations of sequential experimental design to human factors has been developed by Simon (1977b). In this report, Simon states the five principles on which this paradigm is based. The first principle is equivalence sampling theory which states that the better the experimental representation of the operational environment, the better the experimental data will predict operational performance. This leads the experimenter to identify as many of the salient variables and define their operational ranges; thereby, creating the ability to maximize initial sampling of performance in the problem domain. The second principle is the pareto maldistribution theory. Although many variables may exist which affect

performance, only a few account for most of the variability. This principle is the basis for the use of screening studies to cull the independent variables. Only those variables which account for appreciable performance variability are retained. The principle of parsimonious models of performance states that the most simple model of perfomance should be used when developing an model. In terms of linear regression, why create a third order model when a second order model is sufficient? Trivial error variance principle proports that the residual variance is often confounded with discernible real effects. These real effects should be identified and removed from the residual variance to refine the error variance as much as possible. Thus, the refined error becomes a purer representation of performance error. The fifth principle is minimum replication. Replication is one method to estimate variability, but certainly not the only method. When human subjects are used, replication is often costly and difficult. Alternate sampling techniques should be evaluated to determine the most economical method within the researchers' objectives for acquiring the information. Consider a study for the evaluation of equipment for a small and specialized population. One must first find the subjects, then either transport them to the experiment, or take the experiment to them. Minimizing replication would certainly decrease the cost and increase the probability of successful completion.

In summary, Simon (1977b) states that strategies for implementing these principles should strive to maintain relevance, generality, modularity, and economy in the research plan. Besides the principles, Simon has developed a guide for the implementation of this plan. The plan consists of 5 phases to a research program listed below.

Define the problem

Identify the critical variables

Develop the response surface

Refine the equation

Verify the experimental results

Figure 2 describes the relationship between the each phase and its goals, location. approach, method, and analysis. Since the purpose of this thesis is the identification of critical independent variables, the latter three phases which describe the development of the empirical model will not be reviewed in detail. The reader is referred to Simon (1977b) for a full description of these phases. The most components of this plan which are most pertinent to this thesis, are Simon's guide for the problem definition and critical variable identification.

Williges and Mills (1982) have developed a similar approach to the complex systems research. In their approach, the research is divided into seven phases.

Problem definition

Study Plan

Pretesting

Experimental Strategy

Data Collection

Data Analysis

Interpretation

Another important contribution made by Williges and Mills (1981) was the division of the factors into constants, parameters, and variables. Constants are those factors which are identified as relevant to a problem, but are not experimentally manipulated. Unlike other strategies which might ignore these factors, Williges and Mills suggest reporting each factor and its corresponding

RELATIONSHIP AMONG PHASES, GOALS, AND METHODOLOGY AS THE EXPERIMENT PROGRESSES

	1	2	3	4	5
PHASE:	Defining the problem	Identifying critical variables	Approximating response surfaces	Equation refinement	Verification
GOAL:	Exploring and limiting the problem	Building a quantitative data base			Evaluating
LOCATION:	Field*	Laboratory or Field			Field
APPROACH:	Undesigned (No control; measure)	Systematic (manipulation, control; measure)			Systematic or undesigned
METHOD:	•Literature search •Interview •Observe •Experience •Measurement	Fractional factorials: screening -- group, individual measure	Central-composite designs; refinement points	Replication; iteration	Test-residual analysis
MODEL:	-	Res. III**	Res. IV	Res. V	Res VI +
ANALYSIS:	Correlational: •factor analysis •ridge regression •cluster analysis •etc.	Analysis of Variance: • mean differences • eta squared • ordered graphics • etc.	Correlational: ridge canonical regression		Correlation; significance

*May not be possible if system doesn't exist; simulator may serve instead.

**Res. = Resolution. The Roman numeral indicates which sources of variance are isolated and aliased.

Figure 2. Simon (1977b) Methodology.

level throughout the research program. Factors that are believed to affect performance across many components of the research plan are classified as parameters. Parameters are factors of such great importance that they are manipulated in all experiments during the research program. Finally, factors which affect performance across a limited scope of the research plan are classified as variables and are manipulated in one or more experiments.

To define the problem domain, Simon (1977b) suggests a less systematic investigation and stresses the need for expert judgement. Williges and Mills (1982) suggest a more systematic approach which is directed at defining the problem for the needs/considerations of the design engineers who will eventually use the data. Problems are defined so investigation will yield data to support design specifications, trade-off analyses, and decomposition of systems into subsystems based on performance.

In both paradigms, by end of this phase the researcher should have identified the critical components of the problem, and determined which of these components can be examined in an experimental setting. This requires gathering information regarding the performance of the system in either the field, or a simulation (possibly a prototype). The sources of information for this phase are (Simon, 1977b):

Literature review

Interview

Direct observation

Personal experience

Data collection

The goal of the next phase is to cull the list of variables and identify those which will contribute the most to the model. Screening studies are the main tool used to select independent and dependent variables. Simon suggests designs

which include 100 independent variables, whereas Williges and Mills (1982) suggest limiting the designs to 30 or less. The restriction of screening studies to 30 or less variables is practically more appealing. First, the cost in time and money of developing an experimental environment where 100 independent variables can be manipulated at several levels seems well beyond the resources available to most researchers. Second, the object of a screening study is the testing of variables not the development of an empirical model, the cost of a 100 variable study would be difficult to justify, since it would contribute little to the construction of an empirical model. By design, screening studies include many factors at a few levels. Therefore, such studies do not contain the necessary resolution to build an empirical model. A more rapid and cost effective method for systematically reducing the number of independent variables would be beneficial. One such method is using subjective ratings by experimenters and end users of the importance of potential independent variables

This application of subjective ratings is based on the premise that given a detailed description of the systems (simulation/prototype), users and experimenters are able to identify factors which will affect performance. The use of user ratings to evaluate equipment design has been popular for some time. Common protocols collect ratings from current users of an existing system or target users of a new system, then use these ratings to support decisions on an iterative design (Meister, 1985).

Purpose

The purpose of this thesis is twofold: (1) to review a sequential experimental design case study in which subjective ratings of independent variables were used to select candidates for a screening study, and (2) to

present an experiment which tested the validity of the subjective ratings technique to predict independent variables.

SELECTION OF CANDIDATE VARIABLES

This section describes a case study in which subjective ratings were used to select candidate independent variables for a screening study. Overall, the research described is directed at developing a systematic methodology for experimentation with complex systems. In particular, the program is directed at experimentally determining the factors of importance to the design of a telephone inquiry system. In a case study manner, this chapter outlines the problem definition phase and the methods by which the problem was reduced to a feasible screening study. The role of the subjective ratings was to provide the final reduction of the variable list to candidate variables for use in a screening study.

Naturally, in the systematic development of an independent variable list, several stages preceded the use of subjective ratings. Figure 4 is a time line summarizing the efforts which have preceded and followed the use of subjective ratings. In this diagram, the stages are enclosed in boxes. The input/output product of these processes is the independent variable list marked as vertical lines. For the reader, three lists are noted which represent the waypoints during the project when the list was held relatively constant for working purposes. The following sections contain a description of each of the stages and the changes to the variable lists.

Problem Domain

To test a method for evaluating an emerging technology, one must identify and select a technology which is novel and unresearched. In light of these criteria, a telephone inquiry task which uses a synthesized speech display was selected. Telephone inquiry systems use a touch tone telephone and synthesized speech to provide users with functions such as data selection, data ordering, and voice mail. The display is text spoken by a rule based

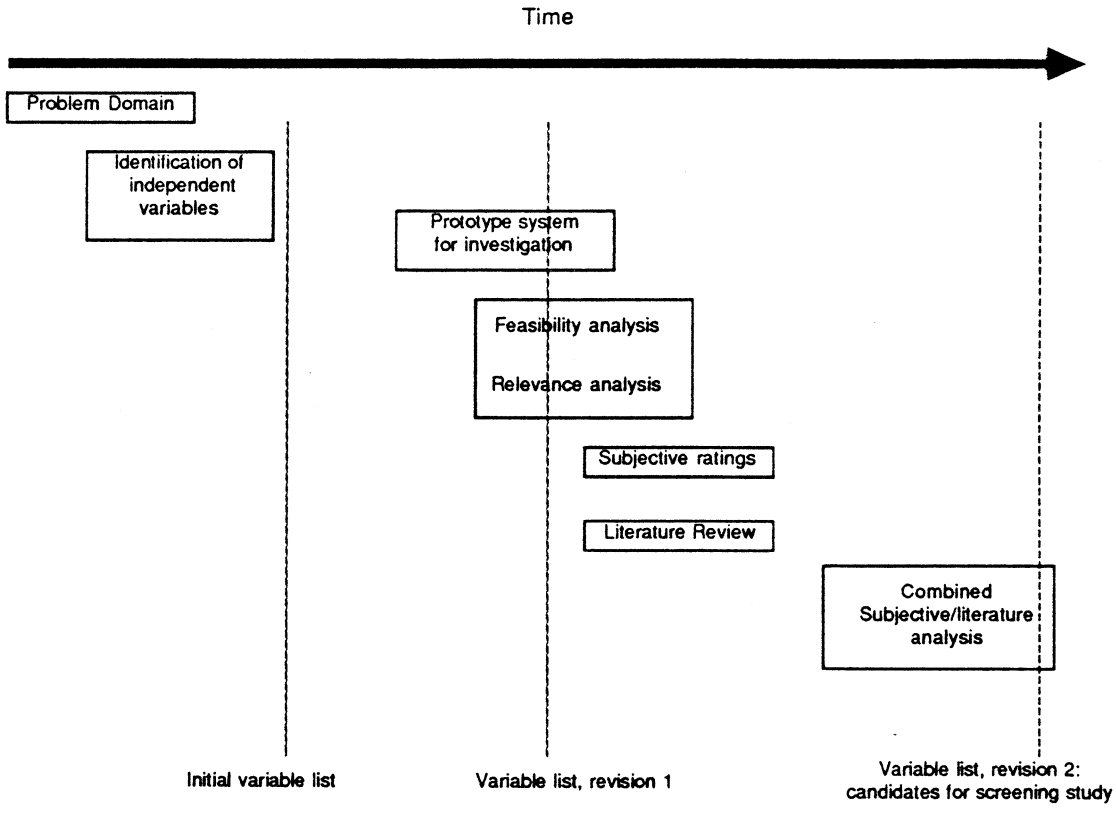


Figure 3. *Time line of events preceding variable selection.*

speech synthesizer, and the controls are command actions attached to keys and key sequences on the telephone keypad. These systems are promising as a means of permitting remote office functions and data base access. One of the largest analogous systems provides preflight weather information to general aviation pilots (Gerald, 1984). Because of the use of synthesized speech and the nonconvertible display/control of a telephone inquiry system, the criteria of novelty is well satisfied.

Identification of Variables: Brainstorming

Once the problem domain was defined, the experimental team identified variables likely to affect operator performance. Experimenters first identified variables by "brainstorming." During the brainstorming process the experimenters met as a group, naming, defining, and recording as many variables as possible with little criticism regarding the variable's value. Only those variables which were clearly irrelevant or could not be clearly defined were rejected. The object of the exercise was to identify the broadest range of variables by stimulating creative thought about the problem domain. After the brainstorming session, the variable list was reviewed to eliminate synonymous variable names. The technique produced a list of 95 variables encompassing: speech display quality, dialogue design, dialogue control by the user, user characteristics, task characteristics, and environmental factors as summarized in the first column of Table 2. Because certain variables were eliminated during later analyses, some of the columns contain missing data points. These missing data points are marked with a minus sign (-).

The rational behind using the experimental team for variable identification is similar to the rationale underlying variable selection in traditional experimentation. Experimenters in a general technical field are

TABLE 2.
Summary of Candidate Variable Selection.

VARIABLES FOR SPEECH SYNTHESIZER RESEARCH
IDENTIFIED DURING LITERATURE REVIEW

INDEPENDENT VARIABLES:	Lit Fnd	Lit Eff	R	F	Int'l Subjects Spd	Int'l Subjects Acc	Int'l Subjects Acp	Amer Subjects Spd	Amer Subjects Acc	Amer Subjects Acp	Expert Spd	Expert Acc	Expert Acp
SPEECH QUALITY													
Speech Rate													
Overall Rate (pause between words)	Y	Y	Y	V	4	4	5	4	5	4	5	5	5
Pauses between Syllables	N		-	N	-	-	-	-	-	-	-	-	-
Pauses between Phrases	N		Y	S	4	2	5	4	3	5	4	3	5
Pauses between Sentences	N		Y	V	2	0	3	4	5	4	3	3	4
Amplitude	Y	CS	-	N	-	-	-	-	-	-	-	-	-
Harmonic Structures (Type of Voice)													
Base Pitch	Y	Y		N	-	-	-	-	-	-	-	-	-
Mean Pitch	Y	Y	Y	V	0	1	1	2	1	2	0	2	4
Range of Pitch	Y	Y	Y	V	2	3	3	1	0	1	0	2	5
Smoothness	Y	Y	Y	V	2	2	3	4	3	5	3	3	4
Assertiveness	Y	Y	Y	V	1	1	2	2	1	5	2	3	3
Breathiness	Y	Y	Y	V	3	2	4	4	4	4	5	4	4
Richness	Y	Y	Y	V	1	1	2	2	2	4	1	2	3
Head Size/Resonance	Y	Y	Y	V	0	0	3	1	3	5	0	0	1
Gain, Fricatives	Y	Y	Y	V	1	1	2	2	2	5	0/3	0/3	0/3
Gain, Asperatives	Y	Y	Y	V	1	2	3	2	3	4	0/3	0/3	1/3
Gain, Voicing	Y	Y	Y	V	1	1	3	3	3	4	0/3	0/3	0/3
Gain, Nasal	Y	Y	Y	V	2	1	5	3	4	5	1/3	1/3	1/3
Age	N		Y	S	2	3	4	5	4	5	4	2	5
Sex	Y	Y	Y	V	1	1	1	1	0	2	0	1	0
Prosodics													
Stress	Y	Y		S	-	-	-	-	-	-	-	-	-
Inflection	Y	Y		S	-	-	-	-	-	-	-	-	-
Timing	N			N	-	-	-	-	-	-	-	-	-
Regional Accent	N		-	N	-	-	-	-	-	-	-	-	-
Use of Exception Dictionary	N		Y	V	5	5	5	4	5	5	5	5	5
DIALOGUE DESIGN													
SPEECH DISPLAYS													
Vocabulary													
Size of Vocabulary	Y	Y	M	V	3	3	4	5	5	5	3	4	3
Familiarity of Words	Y	Y	Y	S	4	3	4	4	5	4	5	5	4
Length of Words	N		M	V	3	3	2	3	3	3	3	4	2
Use of Jargon	N		N	S	3	2	3	5	5	5	3	3	4

TABLE 2 cont.
Summary of Candidate Variable Selection.

VARIABLES FOR SPEECH SYNTHESIZER RESEARCH
IDENTIFIED DURING LITERATURE REVIEW

INDEPENDENT VARIABLES:	Lit Fnd	Eff	R	F	Int'l Subjects Spd	Int'l Subjects Acc	Int'l Subjects Acp	Amer Subjects Spd	Amer Subjects Acc	Amer Subjects Acp	Expert Spd	Expert Acc	Expert Acp
Syntactical Structure													
Active vs. Passive Voice	N		Y	S	4	0	5	5	3	4	3	3	3
Simple vs. Complex Sentences	N		Y	S	3	2	4	5	5	5	4	4	4
Length of Sentences	Y	Y	Y	S	4	2	4	5	3	4	5	5	2
Order of Inf. in Sentences	Y	Y	Y	S	5	3	4	5	5	5	4/4	4/4	1/4
Semantic Structure													
Length of Messages	Y	Y	Y	V	4	1	4	4	4	3	4/4	4/4	2/4
Order of Messages	Y	Y	Y	V	5	3	5	5	5	5	4/4	4/4	1/4
Inf. Coding/Diff. Voices	N		Y	V	4	4	5	1	1	3	4	5	4
Adapt Speech Rate	N		Y	V	5	5	4	3	3	3	3	5	4
KEYPAD INPUT													
Disambiguation of Input													
Double Keying	Y	CS	N	V	4	4	5	4	3	4	4	4	3
System Intelligence	Y	CS	-	S	-	-	-	-	-	-	-	-	-
Input Echoing Level	Y	Y	Y	V	4	1	4	2	2	3	5	3	5
Menu Dialogue Style													
Menu Length	Y	CS	Y	V	3	2	3	4	3	4	4/4	4/4	3/4
Menu Depth	Y	CS	Y	V	5	1	3	5	5	5	5	4	4
Methods for Traversing Menus	N		Y	V	3	3	3	4	4	4	5	3	4
Command Dialogue Style													
No. of Commands	Y	CS	Y	V	3	3	5	5	5	5	4	5	5
Length of Commands	Y	CS	N	V	4	4	5	5	4	5	5	3	4
Command Abbreviations	N		N	V	4	4	4	5	4	5	5	3	3
Command Synonyms	N		N	V	3	3	3	5	5	5	4	3	3
ERROR HANDLING													
Error Detection													
Level of Error Detection	Y	CS	N	V	4	4	4	5	5	5	3	4	5
Error Recovery													
Undoing Actions	Y	CS	M	S	5	4	5	5	5	5	4	5	4
Input Timeouts	Y	CS	Y	V	5	2	5	5	5	5	4	1	5
Change Wording or Phrasing of Speech Display	Y	CS	M	S	3	4	5	5	5	5	2	5	5
Change Rate of Speech Display	Y	CS	N	S	-	-	-	-	-	-	-	-	-

TABLE 2 cont.
Summary of Candidate Variable Selection.

VARIABLES FOR SPEECH SYNTHESIZER RESEARCH
IDENTIFIED DURING LITERATURE REVIEW

INDEPENDENT VARIABLES:	Lit Fnd	Eff	R	F	Int'l Subjects Spd	Int'l Subjects Acc	Int'l Subjects Acp	Amer Subjects Spd	Amer Subjects Acc	Amer Subjects Acp	Expert Spd	Expert Acc	Expert Acp
USER ASSISTANCE													
HELP													
Initiation of HELP	Y	CS	-	S	-	-	-	-	-	-	-	-	-
Selection of HELP	N		-	S	-	-	-	-	-	-	-	-	-
Content of HELP	N		-	S	-	-	-	-	-	-	-	-	-
Access of HELP	Y	CS	-	S	-	-	-	-	-	-	-	-	-
Organization of HELP	N		-	S	-	-	-	-	-	-	-	-	-
Embedded Training/CAI	N		Y	V	4	4	4	4	4	4	4	4	5
Other User Aids													
Transactions Summaries	N		Y	V	2	2	2	3	2	1	2	1	2
Wallet Guide	Y	CS	Y	V	4	3	4	4	4	4	0	1	2
Hardcopy Summary	N		M	V	0	1	4	3	2	3	5	5	5
Human Assistance	N		M	V	3	3	4	3	3	2	4	4	4
DIALOGUE CONTROL BY USER													
Speech Quality													
Amplitude Control (speakerphone only)	N		Y	V	1	1	1	2	3	3	1	3	3
Frequency Control	N		-	N	-	-	-	-	-	-	-	-	-
Speech Display Pacing													
Repeat Speech Display	Y	Y	Y	V	3	3	5	5	5	5	4	5	2
Repeat Word or Sentence	N		-	N	-	-	-	-	-	-	-	-	-
Spell Out Speech Display	N		Y	V	4	4	4	4	5	5	5	4	3
Pause/Resume Speech Display	Y	CS	Y	V	2	2	3	3	4	4	3	3	5
Interrupt Speech Display with Control Input	Y	CS	Y	V	4	0	5	4	2	4	4	4	5
Sequence of Events	N												
Suppressing Menu Prompts	N		-	N	-	-	-	-	-	-	-	-	-
User-Defined Markers	N		-	S	-	-	-	-	-	-	-	-	-
USER CHARACTERISTICS													
Experience													
Experience with Info. Sys.	Y	N	Y	V	4	4	4	4	4	5	5	5	3
Experience with Computers	Y	N	N	V	2	2	3	3	2	4	4	3	4
Experience with Synthesis	Y	Y	Y	S	4	3	5	5	5	5	4	4	4
Experience with Other Speech	Y	N	M	S	4	3	5	5	4	5	1	3	1

TABLE 2 cont.
Summary of Candidate Variable Selection.

VARIABLES FOR SPEECH SYNTHESIZER RESEARCH
IDENTIFIED DURING LITERATURE REVIEW

INDEPENDENT VARIABLES:	Lit Fnd	Eff	R	F	Int'l Subjects Spd	Acc	Acp	Amer Subjects Spd	Acc	Acp	Expert Spd	Acc	Acp
Demographics													
Age of User	Y	N	Y	V	2	1	1	3	5	5	3	3	2
Sex of User	Y	N	Y	V	0	0	0	1	1	1	0	0	1
Hearing Impairments	Y	Y	Y	S	5	5	5	5	5	5	5	5	5
Native vs. Foreign	Y	Y	Y	V	5	5	5	5	5	4	3	5	4
Educational Level	N		Y	V	-	-	-	-	-	-	-	-	-
5. TASK CHARACTERISTICS													
Response Time													
System Response Time	N		Y	V	3	1	3	5	5	5	5	2	3
Database Structure													
No. of Keywords	Y	CS	Y	V	1	3	1	4	4	4	4	4	2
No. of Information Nodes	Y	CS	Y	V	2	2	4	3	4	4	5	3	2
Organization of Database	Y	CS	Y	V	0	0	1	3	3	3	0/0	0/0	0/0
Type of Data	Y	CS	Y	V	5	4	5	5	5	3	3	2	2
Task Complexity													
Number of Steps in Search	N		Y	V	4	1	3	5	4	3	5	4	5
No. of Searches per Session	N		N	V	4	3	3	5	5	5	4	2/4	2/4
Complexity of Searches													
Logical Operators	N		N	S	-	-	-	-	-	-	-	-	-
Multiple Targets	N		Y	V	4	2	0	5	5	5	5	4	3
If/Then target selection	N		N	V	5	2	4	5	5	5	4/5	3/5	3/5
General vs. Specific Search	N		N	N	-	-	-	-	-	-	-	-	-
Type of Competing Tasks	N		N	S	4	2	1	5	3	4	5	5	4
6. ENVIRONMENTAL FACTORS													
Competing Speech	Y	Y	Y	V	5	4	2	5	5	3	5	5	3
Noise	Y	Y	Y	V	5	5	4	4	4	4	5	5	2
Background Music	N		N	V	3	4	3	3	4	3	3	3	2

considered sufficient experts to select relevant factors from a problem domain. This assumption might be challenged on the basis that the current research is aimed at a new technology for which few or no experts exist. In response, one could argue that new technologies often contain components analogous to existing systems; thus expertise gained from existing systems can be applied to the new system. An expansion of this argument would be that general technical training

would enable the experimenters to identify factors pertinent in many problem domains.

Experimental Prototype of System

Next, a prototype of the telephone inquiry system was developed. The primary purpose of the prototype was to provide an example of the target system for both the experimental team and the users. The experimental prototype provided a valuable source of information for the feasibility, relevance, and subjective ratings.

The early development of the prototype aided the selection of variables in three ways by: (1) increasing the knowledge about the technology and its applications, (2) generating information about the feasibility of the experimental manipulation of the variables, and (3) generating information about the relevance of variables. From the time and level of effort required to implement features in the prototype, more realistic estimates of experimental resource expenditures could be made for further development and experimental efforts. Experimental resources used included: computer time, programmer time, and computer hardware. This information was most useful when benefit/cost comparisons were later made.

Aside from the general information gain, specific knowledge about the capabilities of the speech synthesizer and the computer system caused later

modifications to the variable list. As a result, the variables "pauses between syllables," and "variation of voice across time" were both eliminated because of the limitations of the speech synthesizer in terms of manipulating these features. The subjective voice qualities ("trust," "authority," "businesslike"), the "user help dimensions of clarity," and "user tailoring" were removed because of difficulties in operational definitions.

The decision to use a keyword interface also modified the variable list. Under the keyword system, the need to trim menus of extraneous speech became moot since the menus had been reduced to only the essential words. Since it was decided that the telephone inquiry system would be limited to one style of interface, the type of command input was also eliminated as a variable. Finally, variables in the database structure section were redefined from "number of variables" to "the number of keywords", and the "number of observations" was changed to the "number of information nodes". Although these changes to the variable list were decided on, they were not implemented until after the feasibility and relevance analysis.

Feasibility and Relevance Analysis

A feasibility analysis was performed to identify variables which would be unrealistic to implement due either to time or cost. Simultaneously, a relevance analysis was also performed to eliminate any variables which had become inconsequential as a result of the development of the experimental prototype. These variables were "designed out" of the system as a result of development decisions. Both the feasibility and relevance analysis were conducted using the information gained from the prototype development, and expert judgement.

Feasibility was broken done into overall feasibility and the level of effort required to develop the software to implement the variable. Feasibility was rated as:

(N)ot-Unable to implement on the available computer hardware.

(S)omewhat-Able to implement on the available hardware, but may require extensive software development with questionable success. Also given to variables which the ability to implement experimentally could be confounded with other variables. For example, the age of the voice is a composite of other listed voice parameters such as assertiveness, breathiness, and head size.

(V)ery-Conceptually very easy to implement. May require major software development, but high likelihood of success. In a group meeting, each variable was presented and the rated verbally by the group. Any disagreements were discussed and resolved at the time.

At the same time as the feasibility rating, a relevance rating for each variable was taken. Relevance was rated as:

(Y)es-Clearly pertinent to the problem domain and the prototype.

(M)arginal-Limited relevance to the problem domain. Possibly limited by the detailed nature of the prototype

(N)ot-A variable, that through further investigation of the problem had been determined to be not relevant to the problem domain or the specific implementation.

Variables identified as infeasible or irrelevant were eliminated from the consideration for experimental manipulation and were marked as constants. However, these variables would not be absent from future research. Henceforth, these variables and their corresponding levels would be reported,

but not changed. Potential variables were marked for exclusion from further investigation on the basis of one of three rules:

1. Not relevant
2. Not feasible
3. Somewhat feasible, but major software development.

For both the feasibility and relevance analyses, the letter within parentheses in the definitions was used as the symbol for the category in Table 2 in the "F" and "R" columns.

A summary of the ratings is listed in Table 2. By the end of these analyses the independent variable list had been reduced to 95 variables, 48 of which were identified as infeasible and/or irrelevant, leaving 47 remaining. Although the feasibility/relevance analysis reduced the variable list by half, 47 variables were still too many for a screening study. The next stages of the variable culling process were the literature analysis and the subjective ratings.

Literature Analysis

A traditional component of any problem domain analysis is a literature review. A structured literature review of the 95 variables listed in Table 2 was conducted to determine whether each variable had been referenced, and if so, was it a part of an opinion paper, case study, or experiment. Opinion studies were those papers in which a system design had been described, but no data were collected. A case study was one in which a design was specified, and opinion or performance data were collected, but no variables were experimentally manipulated. Experiments were those studies in which data were collected on a specific design under experimentally controlled manipulations of the independent variables.

The review identified 29 articles considered directly applicable. Please refer to *Appendix A* for a bibliography. Of the 95 independent variables

identified, the literature review showed only 30 manipulated in an experiment, and another 21 referenced in case studies. It should be noted that many of the articles categorized as case studies were descriptions of commercial products for which evaluation strategies are questionable. For each of the 95 variables in revision 2 of the variable list, the literature analysis was summarized by listing whether the variable had been considered (specified) in an article, and if so had the variable affected performance.

In Table 2 the columns marked "Lit" indicate the results of the literature analysis. A "Y" was placed under the subheading "fnd" for each variable which was found in the literature; an "N" indicated that the variable was not considered. All variables marked with an "n" in this category were classified as "not found in the literature." Likewise, a "Y" in the subheading of "eff" indicated that the variable had affected performance in an experiment. Variables which had been suggested to have an effect in a case study were marked with a "CS". If a variable had not been found in the literature or had been suggested as affecting performance, the variable was marked with an "N".

Subjective Ratings

The use of subjective ratings in the selection of independent variables is based on the assumption that users and experts are capable of identifying factors which affect their performance. Subjective measure have a long history in test and evaluation of systems, in which the users assist in determining the value of a system feature (Meister 1985). Similarly, subjective ratings have been used to determine user preference of the inclusion of features in a system under development. Geiselman and Samet (1982) conducted a study in which users were allowed to set display parameters and then performance was measured using these settings. Geiselman and Samet concluded that allowing subjects to customized their screen design had profound effects on information

processing in the experiment by enhancing the subjective mental organization. Following these results, if users are able to set the display parameters and maintain performance given a list of system parameters, then users should also be able to select factors which are likely to affect performance. Collecting data on these variables would create another source of information which experimenters could use to cull the list of variables.

Method: subjects. A total of 15 subjects, partitioned into 3 groups of 5, participated in the ratings. Each group of 5 represented a different segment of the population, introducing breadth to the survey. The first group consisted of the 5 researchers on the project, and represented the opinions of subject matter experts. The other 10 subjects represented users and were divided equally between native and non-native English speakers. Non-native English speakers were included to assessfor any possible interaction between native language and synthesized speech.

The 10 users were administered and passed a screening test for hearing considered acceptable for the study. During the screening test subjects acknowledged a 3 second pulsed tone presented to either the left or the right ear. A series of 5 tones was presented sequentially at 30 dB(A) at the following frequencies: 800, 1000, 2000, 4000 Hz. All 10 subjects passed the criterion of acknowledging a tone 2 out of 3 presentations.

Method: apparatus. The experimental prototype was used to introduce the subjects to the system and its operation. The system was implemented on a Digital Equipment Corporation (DEC) VAX/11 750 mini-computer running the VMS operating system. Additional hardware included a DECTalk 2.0 rule based speech synthesizer, and a Panasonic speaker telephone. Standard telephone lines were used to connect the speaker-phone to the computer. All

the text was spoken by the DECTalk across the telephone line and reproduced over the speaker phone.

A community recreational activity data base served as a prototype application for the telephone inquiry system. The system was designed as a hierarchy of keyword menus. Each keyword described a group of related items. For example, a group of movie titles (*Dune, Gone with the Wind, Brazil North by Northwest, and From Russia with Love*) was gathered under the keyword "movies." Users desiring information about movies would select the movies keyword, which would initiate the presentation of the movie titles menu. By selecting a sequence of appropriate keywords, users located the item of interest. By selecting the item of interest, users heard a brief message relating to that item.

Within each menu, keywords were presented serially with a 5 second pause to allow command input from the user. If the user did not press a key representing a command function, the system continued by speaking the next keyword in the menu. Menus were presented in a circular format. Following the last keyword in the menu, the system presented the help keyword, paused, then the system automatically spoke the first keyword in the menu. Each menu was presented twice. If the user did not take any control action before the end of the second menu repetition, the system "hung-up" assuming the user had done the same. The system initiated hang-up was necessary because of the inability of the speech synthesis hardware to detect a telephone hang-up.

During the control input pause, three commands were available to users:

Select - Press the # key. Selects the keyword preceding the pause, then begins a new menu or presents an information message.

Back-up - Press the * key. Returns the user to the previous menu in the hierarchy.

Restart - Press the 0 key. Returns the user to the first keyword in the main menu.

Figure 4 depicts the structure of the data base used in the experimental prototype. Menu length (number of keywords on the menu) varied from 2 to 7 items. Menu depth (number of selections required to reach the information message) varied from 2 to 4 with a the number of items heard before reaching an information node varying from 3 to 10.

The standard male voice for the DECTalk version 2.0, "Perfect Paul," was used because of its high intelligibility (Green, Manous, and Pisoni, 1984). Because of the difficulty that the speech synthesizer had with pronouncing proper names, manual phoneme and stress coding was used to enhance intelligibility in the data base. Using manual stress coding, one can force the speech synthesizer to stress a particular syllable. Likewise, manual phoneme coding forces the speech synthesizer to speak a particular phoneme. Using these two techniques the designer can instruct the speech synthesizer to modify the rules by which the word is pronounced. Without these improvements, names such as Brazil would have been unrecognized when pronounced.

The subjective rating scales were administered using a paper and pencil format. Subjects rated each of the variables on whether they believed the variable would affect the speed and accuracy of using the system and the acceptability of the system. The question for each variable consisted of the variable name, a definition, and the set of rating scales. *Appendix B* contains a complete listing of the subjective rating scale instrument.

Because the feasibility/relevance analysis had not been completed when the subjective rating scales were designed, only 22 of the 95 variables had been eliminated, leaving 73 variables to be rated. By the end of the subjective rating questionnaire design and the ensuing data collection, the

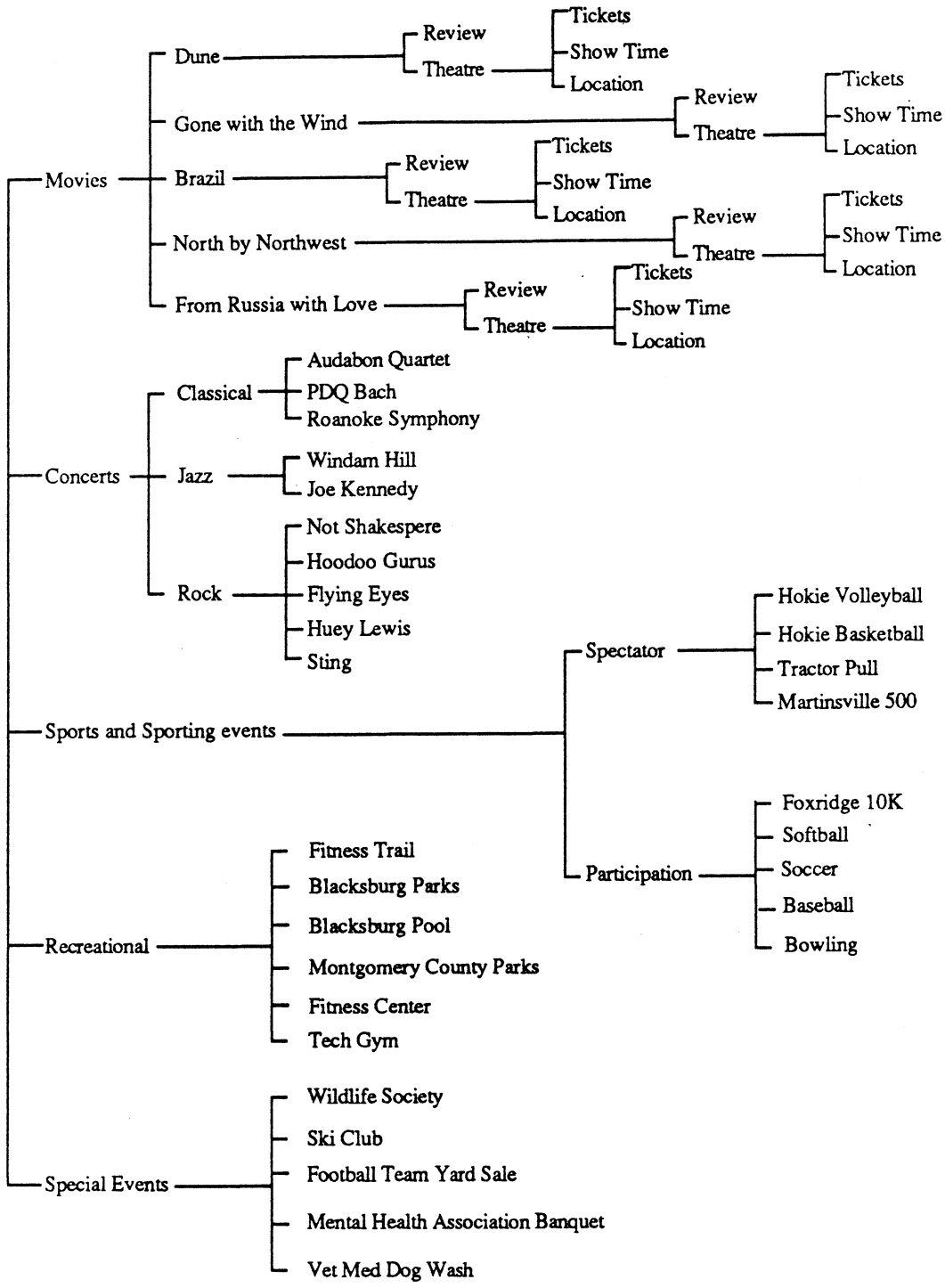


Figure 4. *Diagram of prototype database.*

feasibility/relevance analysis had been completed and not all the variables used in the subjective data collection were used in the analysis. Variables designated to be set as constants were not used in the analysis. Nineteen variables were eventually marked as constants leaving 54 remaining and used in the analysis.

Method: procedure. Prior to the experimental task, subjects were requested to complete an informed consent (*Appendix C*), a hearing screening test (refer to the "subjects" section for a detailed description), and a demographic survey (*Appendix D*). Next, the subjects were presented with the experimental task instructions in typed and tape recorded formats (*Appendix E*). After the instructions, the experimenter answered any questions pertaining to the experimental procedure.

Subjects were introduced to the prototype system by asking them to search for five information targets. The targets were presented on a flip chart located in front of the subject on the table. *Appendix F* lists all 5 targets. The target set included three searches for a single information message and two searches for more than one information message in a single search. The subject searched for two information messages in both multiple information message searches. Information messages were between one and three sentences long. Subjects were asked to record the portions of the information relevant to the target question on an answer sheet . The answer sheets were not scored.

Finally, each subject completed the variable rating questionnaire with 5 minute rest breaks after questions 17, 37, and 64. Participation required approximately 2 hours, for which the subject was paid 5 dollars per hour.

Results and Discussion. Table 2 presents the results of the subjective ratings for the Native English users, the non-native English users, and the

experts groups under the headings "Amer", "Intl", and "Expert", respectively. The minus signs (-) indicate that the variable was not rated because of being eliminated as a result of the feasibility/relevance analysis. These numbers represent the number of "effect expected" ratings for the variable. On twelve of the variables, a portion of the experts felt they were unable to rate the variable. For these twelve variables, the number of variables rated as "effect expected" ratings are on the left hand side of the slash (/) with the total number of ratings on the right hand side of the slash. After reviewing the data, It was determined that the data from the 10 users and the 5 experts would be aggregated for the analysis. The decision to aggregate the ratings was based on the desire to include as much information as possible which could be used to reduce the independent variable list.

Variables were then divided into three categories of "effect expected," "no effect expected," and "no clear opinion." Variables were categorized as "effect expected" if they were rated as increasing or decreasing (speed, accuracy, and acceptability) on two or more of the three dimensions by seven or more of the subjects. Similarly, variables rated by seven or more of the subjects on two of the three rating dimensions as having no effect were categorized as "no effect expected". All the variables which were not included in either of the two preceding categories were classified as "no clear opinion". The implementation of this classification scheme yielded 30, 14, and 10 variables in the "effect expected", "no effect expected", and "no clear opinion" groups, respectively. Table 3 lists all the variables by their categories.

Next each of the categories was cross referenced with the literature analysis. From the literature, each variable was classified as (1) not referenced, (2) referenced, but not experimentally manipulated, (3) referenced and experimentally manipulated, but no significant effect, or (4) referenced,

TABLE 3

Listing of Variables by Subjective Rating Category.

Rating category	*Variables*
No Effect	Mean pitch, Range of pitch, Assertiveness, Richness, Head size - Resonance, Gain - Fricatives, Gain - AspirativesGain - Voicing, Gain - Nasal, Sex of Speaker, Transaction summary, Hardcopy summary, Amplitude control on telephone, Sex of listener (users)
No Clear Opinion	Length of sentences, Order of information, Keypress echoing, Information coding with different voices, Menu length, Menu depth, Spell out speech display, System response time, Number of keywords, Number of message
Effect	Overall speech rate, Smoothness, Breathiness, Familiarity of words, Active vs. passive sentences, Simple vs. complex sentences, Length of messages, Order of information in messages, Double keying, Number of commands, Error detection, Undoing actions, Embedded training, Repeat display, Hearing impairment, Native vs. foreign listener, Competing speech, Noise, Music, Pauses between phrases, Pauses between sentences, Age of speaker, Exception dictionary, Adapting speech rate, Wallet guide , Pause/Resume, Interrupt speech, Number of steps in search, Multiple targets, Input timeouts, Experience with information systems

experimentally manipulated, and found to have a significant effect. Cross-referencing the four literature categories with the three subjective rating categories created a total of 12 cells for the 54 variables included in the ratings. Refer to Table 4 for a summary of these results.

Given the information available from the literature review and the ratings, one could conceive of several methods for reducing the number of variables to a reasonable size for a screening study. Methods would vary in the criterions used including a variable based on the literature analysis, the subjective ratings, and the combination literature review and the subjective ratings. Across all methods, the ability to gain information about the effect of a variable would be the most useful criterion for inclusion in the screening study. Variables for which an effect is known would be eliminated from the screening study because of the limited information which would be gained from including it. Uncertainty about the effect of a variable would cause the variable to be included in the study.

To meet these criteria using literature analysis alone, one could eliminate the variables experimentally manipulated on the assumption that the current literature is valid. Retaining the variables not found or not manipulated in the literature would reduce the list to 22 variables by eliminating 32 variables. This would trim the list to 40% of its original size. Although this represents a substantial reduction in the list size, it is not an adequate reduction.

Alternatively, one could use the data from the subjective ratings independent of the literature review. In this approach, it is assumed that the subjective ratings are more sensitive to the effects of the variables in the present system than either general or highly related literature. Using the new information criterion and great confidence in the ratings, only the variables for which no clear opinion was formed would be suggested as candidates for the

TABLE 4
Subjective Ratings Cross-Referenced with the Literature Analysis.

Subjective ratings	*Literature review*				*Row totals*
	Not found	*Case study*	*Experimental*		
			No effect	*Effect*	
Effect opinion	7	5	0	18	30
No effect opinion	3	0	1	10	14
No clear opinion	2	5	0	3	10
Column totals	12	10	1	31	54

screening study. Such a strategy would assume that the subjective ratings of "effect expected" are accurate, therefore variables of that class can be eliminated. Using only the "no clear opinion" variables would eliminate 44 variables leaving 10, thus, reducing the list to 19% of its original size. A screening study could easily be constructed with only 10 variables. If one has great confidence in the subjective ratings, the technique could have reduced the size of variable list to a managable size without the aid of a literature review.

However, accepting the ratings of "effect expected" as completely valid and eliminating them from the screening study is questionable. A more conservative method would be to include the "effect expected" rating and "no effect" literature review variables and eliminate the "no effect expected" rating variables. Variables would be included based on the possibility of finding an effect and eliminated on the probability of not finding an effect. This technique would produce a list of 40 variables eliminating 14 and retaining 75% of the variables. With a stricter criterion for use of the opinion data, the subjective ratings were less effective at reducing the number of variables than the literature review.

Without assuming great confidence in the subjective ratings, the literature review would be a more powerful tool to reduce the list of variables. Although the literature review dramatically reduced the variable list, it should be remembered that this case studyis only one example. In this particular problem domain, few references directly relevant to the specific application were found, but in the articles available the parameters of synthetic speech quality were well investigated. Because the variables pertaining to synthetic speech quality accounted for a large number of variables on the list, eliminating these variables greatly reduced the list size. The efficiency of the literature review to select variables would be decreased in a field where less relevant literature

existed. Even in the present example the literature review did not reduce the variable list to a manageable size. If the efficiency of the literature review were reduced further, the need for alternative methods for eliminating variables would increase.

Similarly, it would be inefficient, if not imprudent, to use the subjective ratings without the benefit of the literature review. Ignoring the literature would unnecessarily eliminate valuable information and cause too great a reliance on the subjective ratings which are still unproven in this application. Depending on the criterion chosen, the efficiency of the subjective ratings ranged from good to poor in reducing the number of variables for a screening experiment in this application.

The method used for reducing the variable list in this project combined both the subjective ratings and the literature review. The list was reduced in two stages. First, the variables which had been experimentally manipulated in the literature (n=32) were eliminated since the effects were known. Second, the remaining 22 variables were reduced again by eliminating the variables subjectively rated as "no effect" (n=3). This method produced a candidate list of 19 variables. All variables to be held constant in the screening study with their corresponding constant levels are listed in Table 5.

Using the both the data from the literature review and the subjective rating scales, the subjective ratings added little power to the variable reduction process because of the great number of variables eliminated by the literature review. In other problem domains where the existing literature is less appropriate the subjective ratings are likely to be more important.

As a preliminary evaluation of the validity of the use of subjective ratings, the variables eliminated by the literature analysis should be examined. A logical method for examining these variables is to analyze the agreements and

TABLE 5

Constant Values of Variables Eliminated from the Screening Study.

Variable	*Level*
Mean pitch	Paul (120 Hz)
Range of pitch	Paul (100%)
Assertiveness	Paul (100%)
Richness	Paul (20%)
Head size - Resonance	Paul (100%)
Gain - Fricatives	Paul (73dB)
Gain - Aspiratives	Paul (70dB)
Gain - Voicing	Paul (71dB)
Gain - Nasal	Paul (69dB)
Sex of Speaker	Paul (male)
Transaction summary	
Hardcopy summary	none
Amplitude control on telephone	none
Sex of listener(user)	male/female
Overall speech rate	Paul
Smoothness	Paul
Breathiness	Paul
Familiarity of words	familiar
Active vs. passive sentences	active
Simple vs. complex sentences	mixed (report)
length of messages	1 sentence
Order of info in messages	end
Double keying	none
Number of commands	3 to 12, 4 recommended
Error detection	none
Undoing actions	none
Embedded training	none
Repeat display	keywords(0), messages(0)
Hearing impairment	none
Native vs. foreign listener	both
Competing speech	none
Noise	ambient
Music	none
Length of sentences	noun verb-phrase object
Order of information	end
Keypress echoing	none

TABLE 5 cont

Constant Values of Variables Eliminated from the Screening Study.

Variable	*Level*
Size of vocabulary	report
Length of words	report
Use of Jargon	none
System intelligence	none
Methods of transversal	keyword
Length of commands	one
Command abbreviations	none
Command synonyms	none
Error handling	none
Change wording of phrasing	none
Change rate of speech display	none
Initiation of help	user
Selection of help	user
Content of help	annotation
Access of help	direct
Organization of help	linear
Human Assistance	none
Frequency control	none
Repeat word or sentence	none
Suppressing menu prompts	none
User defined markers	none
Exper w/ computers	report
Exper w/ speech synthesis	report
Exper w/ other speech	report
Age of user	report
Organization of data base	conical
Type of data	verbal and numerical
No. of searches/session	two
Logical operators	none
If/then target selection	none
General vs. specific search	specific
Type of competing tasks	none

disagreements between the subjective ratings and the literature review. An agreement is defined as a variable rated the same by both the literature review and the subjective ratings. A disagreement is an "effect" rating by one analysis and a "no effect" on the other. The variables subjectively rated as "no clear opinion" (n=3) are excluded from the discussion since no comparisons can be made. This leaves 29 variables for the agreement/disagreement discussion.

Of the 29 variables, the results from the subjective ratings and the literature analysis agreed on 19 (66%) variables. In addition, no variables were listed as "no effect" from the literature review but rated as an "effect" by the users.

The two information sources disagreed on the ratings of 10 of the variables (34%). After examining the nature of these variables, it is likely that the conflict occurred because of the highly technical meanings of the variables. All 10 of these variables were related to the components which control the parameters of the synthesized speech in the DECTalk version 2.0. These variables were of of such a highly technical nature that their meanings and effects were probably not understood by the raters. Even the experts expressed difficulty in estimating the effects of these variables. The subjective ratings were based on a brief system demonstration and written descriptions of the variables. The levels of the variables were not manipulated for the user. This method forces users to speculate on the effect of each variable. Although cost and time prohibited allowing users to manipulate the level of each variable, it may be necessary to demonstrate the levels of certain variables. This seems especially true with unfamiliar or highly technical variables. A prototype which manipulates these variables at various levels is likely to increase the quality of the subjective ratings. Without the demonstration of these variables, the accuracy of the subjective ratings suffers.

Summary

From the problem domain, 95 variables were initially identified. The variable list was first refined by setting those variables which were infeasible or irrelevant. This reduced the list to 54 variables to constants. The 54 variables were analyzed by structured literature review and subjective ratings. The combined literature review and subjective ratings reduced the list to 19 candidate variables (Table 6) for a screening study. The final 19 variables were suggested as candidates for a screening study because: (1) no known effect/no effect existed, and (2) the subjective ratings indicated "effect expected", or "no clear opinion." The variables rated as "no clear opinion" were included because of the possibility that an effect might be found in the screening study.

In this application, the literature review was more powerful in reducing the variable list than was the subjective ratings. Had less relevant literature existed, the value of the subjective ratings would have increased greatly. A preliminary evaluation of the validity of the subjective ratings focused on the agreement between the subjective ratings and the literature review. Agreement was found in three of the four possible conditions with a 66% agreement on the ratings of the variables.

Although this approach was useful in reducing the initial list of 95 variables, the validity of subjective ratings was not evaluated. Other than the traditional means of variable list reduction such as, feasibility, relevance, and literature review, the decision to include or exclude a variable was based on the subjective ratings. The subjective ratings are based on the untested assumption that users and experts are able to select factors which will affect performance. If users cannot make valid assessments of the effects of independent variables, then this approach is questionable. The purpose of the

next chapter is to review an experiment that tested the validity of the subjective ratings technique.

TABLE 6
Candidate Variables for the Screening Study

Variables
Pauses between phrases
Pauses between sentences
Age of speaker
Exception dictionary
Number of steps in a search
Multiple targets
Experience with information systems
Adapting speech rate
Wallet guide
Pause/resume
Interrupt speech display
Input timeout
Information coding with different voices
System response time
Menu length
Menu depth
Spell-out speech display
Number of keywords
Number of information nodes

VALIDATION OF THE RATINGS TECHNIQUE

Although the subjective ratings technique was helpful in reducing the size of the variable list, it is based on the untested assumption that users are able to select factors which will affect performance with the actual system. When the subjective ratings were combined with the literature review, the assumption that the literature analysis is a more accurate predictor than the subjective ratings was made. This assumption might not always be true due to conflicting research results or a limited generalizability of the literature to the specific system being developed.

To assess the validity of the subjective ratings technique for selecting variables for screening studies, five categories of variables were tested. Each of these categories corresponded to a condition in Table 7. The effect in bold-face represented the predicted outcome based on the decisions made during the selection of the candidate variables. Three of the conditions contained variables where the effect of the variable could not be determined from the literature; the predicted effect was then based solely on the subjective ratings. The other two conditions were effects predicted from the literature and the ratings either agreed or disagreed.

Conditions 1 and 2 were predicted to affect performance based on the subjective ratings. Variables in these conditions are listed in Table 7 and were suggested as candidates for the screening study. Finding the predicted effect would support the assumption that users are able to identify variables which affected performance. Failure to find an effect would discredit the use of the subjective ratings method.

The subjective ratings suggested that condition 3 would not affect performance. Finding significant effects in this category would suggest that the rating technique is likely to eliminate significant variables improperly. Not

TABLE 7
Experimental Conditions in the Validation Study

Condition	*Literature review*	*Subjective rating*
1	not found	**effect**
2	not manip	**effect**
3	not found	**no effect**
4	**effect**	no effect (disagreement)
5	**effect**	effect (agreement)

finding an effect does not unequivocally support the ratings technique since such a statement would require accepting the null hypothesis. Variables in this category were eliminated from the proposed screening study based on the subjective ratings and are found in Table 6.

Conditions 4 and 5 contained variables which were eliminated from the proposed screening study because the effects known from the literature, and the subjective ratings either agreed or disagreed with the literature. Relying on the literature as the primary source of information, one would expect that variables in category 4 would significantly affect performance. Failing to demonstrate these results would imply that the subjective ratings were more sensitive to the effects in the present system than a review of related literature. Finally, condition 5 represents an agreement between the ratings and the literature review. Failing to demonstrate this effect would suggest that another technique should be investigated since both the ratings and the literature review failed to predict the effect of the variable on the specific subsystem.

Table 8 depicts the variables selected to represent each of the five validation conditions. Because of time and cost constraints, a single variable was selected to represent a category of variables. It was assumed that each variable within a condition would have the same outcome and that each variable was equally likely to demonstrate that outcome. Therefore, selecting any one variable should have the same effect as selecting any other variable. This assumption is based on main effects alone and not interactions.

Method

Independent variables. This factor represented validation condition 1. Because this variable was not found in the literature, and the subjective ratings of the prototype indicated that an effect would be found, the validation

TABLE 8
Independent Variables by Condition.

Condition	*Variable*
1	Number of steps in a search
2	Adapting speech rate
3	Transaction summaries
4	Sex of voice
5	Native/non-native

experiment hypotheses was based on the subjective ratings of the prototype that the variable would have an effect on performance.

The number of steps in a search was equivalent to the number of keyword selections necessary to access the information node. This implied that the number of keywords was dependent on the hierarchical structure of the data base. To maintain the same keywords and the same logical categorizations in the hierarchy, it was necessary to vary the organization of the data base. The most convenient structures for varying organization and maintaining the same number of steps, the same keywords, and the same logical categorizations in the hierarchy were symmetrical. In a symmetrical structure, the path to each target was of equal length. Additionally, the greater the depth of the data base, the greater the number of selections to find a target information node.

Two organizational schemes were developed for the data base. The 2x6 hierarchy (Figure 5) contained six levels of menus with two keywords in each menu. A menu depth of six corresponded to six steps in a search. The second organization was a 8x2 hierarchy (Figure 6) with two levels of eight keywords each. This represented a two step search. Both data base organizations contained 64 keywords.

Adapting speech rate permitted the user to increase or decrease how fast the voice spoke based on the ease of executing the task , and represented condition 2 (found but not manipulated in the literature, and rated as having an effect during the prototype evaluation). The experimental levels were implemented as the presence or the absence of the capability to adapt speech rate. This feature was predicted to significantly affect performance based solely on the subjective ratings of the prototype.

The initial speech rate was set at 180 wpm based on the results of Merva (1986). The upper and lower limits of 250 wpm and 150 wpm were dictated by

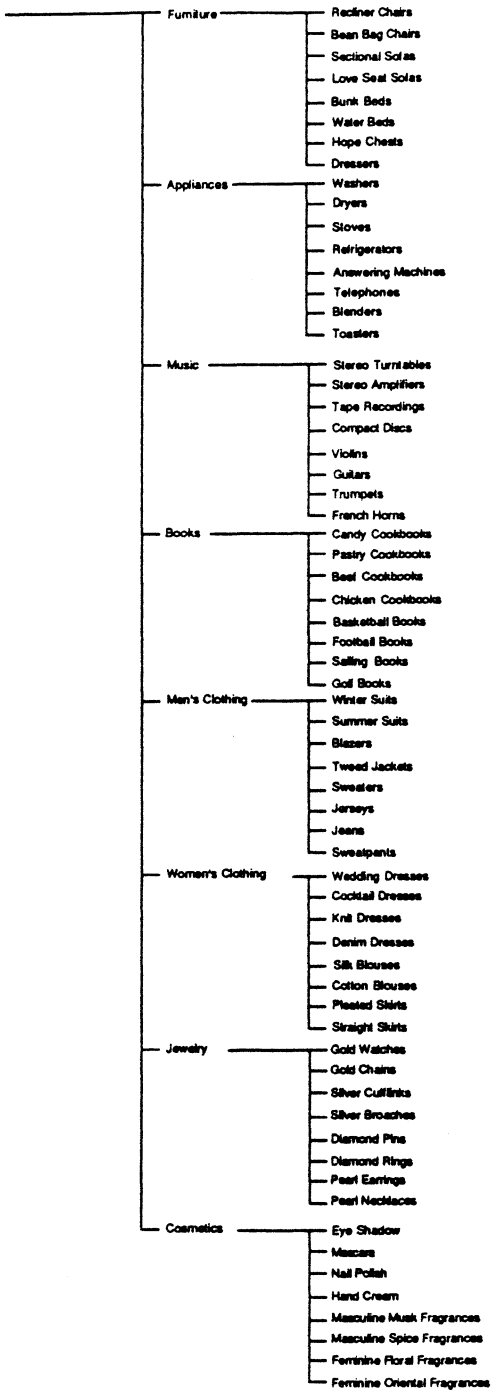


Figure 5. *The 2x6 hierarchical data base*

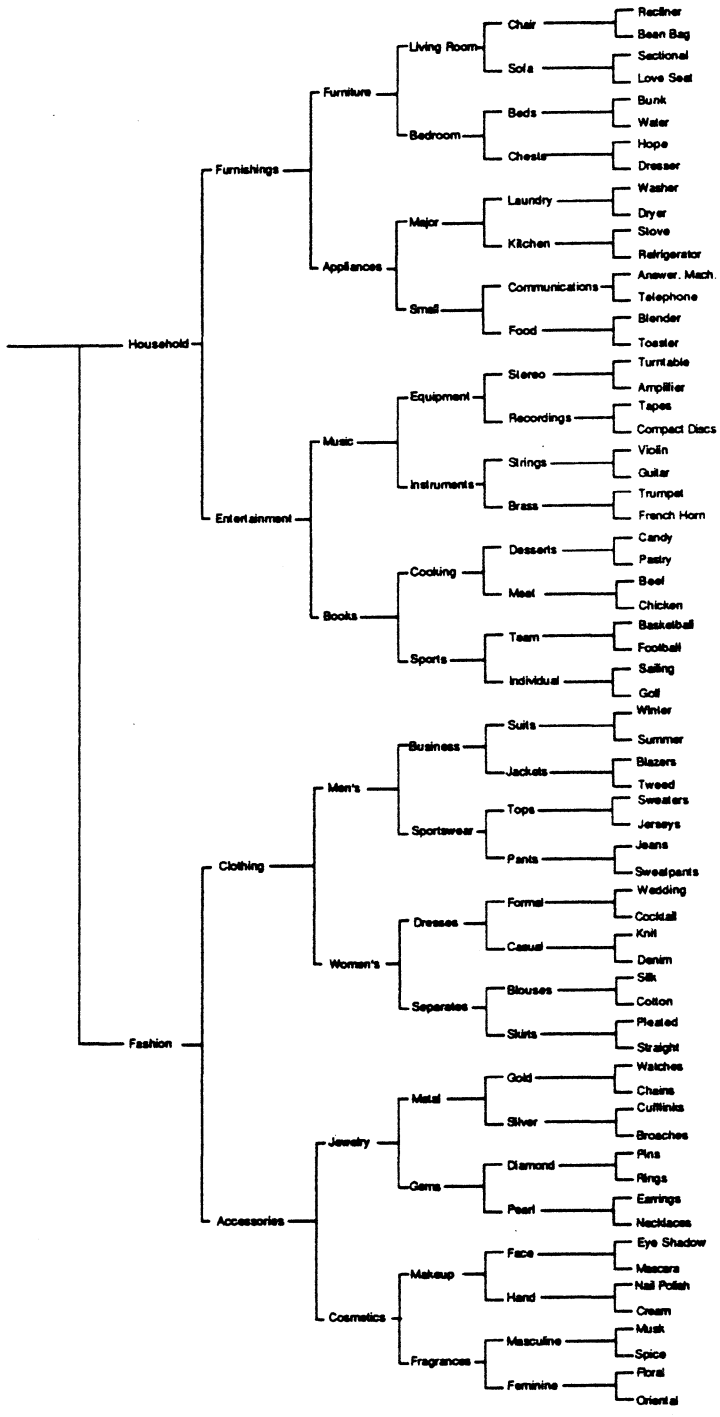


Figure 6 *The 8x2 hierarchical data base*

the capabilities of the speech synthesizer (DECTalk version 2.0). Subjects had the capability to vary the speech rate in 10 wpm increments throughout the entire range of the speech rate.

At the end of each target search the subject was given the opportunity to change the speech rate. The options were "slower", "same", and "faster." The default selection was "same." If the speech rate reached either limit in the range, a choice which would have exceeded the limit was not presented. For example, if the current speech rate was 150 wpm, only the "same", and "faster" options were displayed.

Using the transaction summary facility, subjects could review their path to the current location in the data base. After selecting the transaction summary, a subject heard the list of keywords which had been selected, along with the main, and backup path commands. This feature was designed for persons conducting a lengthy search during which they might become confused. This feature represented condition 3, and was not predicted to affect operator performance significantly based on the subjective rating data. This factor was represented as either present or not.

The sex of the voice represented condition 4, a disagreement between the subjective ratings and the literature review. The literature predicted an effect, whereas the subjective ratings predicted no effect. The voice was implemented as the DECTalk standard male (Perfect Paul) and female (Rough Rita) voices. Under this condition, the literature review prediction, which was that the variable would affect performance was selected as the experimental hypothesis. Not finding the effect would be consistent with the subjective ratings suggesting that the subjective ratings might be more sensitive to effects on operator performance than the literature review.

This effect represented condition 5, which in turn represented an agreement between the literature review and the subjective ratings. Following the results of the literature review and the subjective ratings of the prototype, The validation experiment hypothesis was that native language would have an effect on performance. Not finding the predicted effect would indicated that under some conditions neither the literature review nor the subjective ratings would be sensitive to the actual effect. The effect of Native language was be implemented using native anglophones and Hindis (Indians). The Hindis were chosen as the non-native group because of the relatively homogeneous exposure to English and the availability of a large subject pool at Virginia Tech. The homogeneity of exposure to English was presumed because of the use of English as an associate official language in India. Although not native anglophones, Indians receive instruction in English and frequently conduct business in English.

Native language was determined by the subjects' self report at the beginning of the experimental session. English proficiency necessary to participate in the experiment was determined by the subject's self report of English proficiency certification by Virginia Tech. Non-native subjects were required to be certified English proficient as stated in the Virginia Tech Graduate Catalog (1986). Native anglophone were exempt from this requirement. According to the Virginia Tech Graduate Catalog (1986), for an international student to be certified competent in English,

> Students who have not received their baccalaureate degree from an American, Anglophone Canadian, or British university are required to take an English placement test, which is administered by the Graduate School.... Students who pass the test are certified as proficient. Students who demonstrate a need for remedial instruction will be required to enroll in special English course(s) immediately.... Successful completion of the course constitutes certification.

Experimental Design. With five factors at two levels each, a full factorial design would contain 32 experimental treatments. Because no interactions were expected and only main effects were of interest, a more economical fractional factorial design was used. By confounding the effects of the interactions which were of no interest to the validation, unnecessary data collection was eliminated. A 1/4 fractional replicate which confounds interactions with the main effects reduced the number of experimental treatments to eight. Table 9 presents the eight experimental treatments. All factors were implemented as between subjects factors with four subjects in each of the experimental treatments. Table 10 shows a summary of the factors, the identity relationships, and their aliases.

The decision to use four subjects per experimental cell was supported by a power analysis. Power is the probability that a statistical test will reject the null hypothesis when the null hypothesis is actually false. Alternatively, it is also the

probability that the alternative hypothesis will be accepted when it is true. A power analysis can be used to determine the sample size by setting the power limit and solving the equation in reverse to determine the sample size needed to achieve the set level of power. A power analysis for the *F* test was performed knowing that experimental design could easily be analyzed with an ANOVA technique.

To perform a power analysis based on the *F* ratio, it is necessary to specify the level of power desired and estimate both the treatment and the error effects. Keppel (1982) states that there are no established conventions for selecting a power level. Cohen (1977) recommends a power level of .80, but Keppel disagrees with Cohen's logic for selecting this level. Keppel (1982) suggests two sources for the treatment and error effects (1) data from other

TABLE 9
Experimental Treatments.

Treatment	*Variable*				
	Number steps	*Adapting speech rate*	*Transaction summary*	*Native non-native*	*Sex of voice*
1	6	yes	no	non	male
2	6	yes	yes	native	male
3	6	no	yes	non	female
4	6	no	no	native	female
5	2	yes	yes	non	feamle
6	2	yes	no	native	female
7	2	no	no	non	male
8	2	no	yes	native	male

TABLE 10
Factors and their Aliases

Factor	*Alias*
Number steps(S)	RxV, SxTxNxV, RxTxN
Adapting speech rate(R)	SxV, RxTxNxV, SxTxN
Transaction Summaries(T)	NxV, SxRxTxV, SxRxN
Native/non-native(N)	TxV, SxRxNxV, SxRxT
Sex of Voice(V)	SxR, SxRxTxNxV, TxN
SxT	RxTxV, SxNxV, RxN
SxN	RxNxV, SxTxV, RxT

Identity Relationships

C_1: SxRxV
C_2: TxNxV
C_1+ C_2: SxRxTxN

related studies, or (2) pre-testing. Because no related experiments had been performed in this field, the pre-testing method was chosen. One subject was tested in each of the eight conditions according to the experimental procedures described in the following section and scored for percent transcription accuracy under a strict scoring criterion.

Per the discussion in Keppel (1982), the single subject's score in each condition was used to estimate the treatment mean. Each treatment mean was then subtracted from the overall mean, squared, and summed with the other treatment means. As shown in Equation 1, the sum of the squared means is divided by the number of treatment conditions (a), and multiplied by the sample size per treatment condition (s'). Table 11 shows the percent transcription accuracy scores and the deviations from the population mean by subject.

$$\Phi_A^2 = \frac{s'\left[\sum_{i=0}^{a}(\mu-\mu)^2\right]/a}{\sigma_{s/a}} \qquad (1)$$

Two methods were used to estimate the error variance. Because power analysis is an estimate not an conclusive quantity, two methods of error estimation were used (one conservative, one generous) to construct a range of the power estimation. This range was then used to calculate an average power estimate. The first, which was suggested by Keppel (1982), used the sample variance as an estimate. This approach is considered a conservative estimate since the sample variance does not extract any effects from the error term and is likely to be larger than the error variance in the experiment. The second error variance estimation method was based on a single subject ANOVA. Because

TABLE 11
Treatment Means for Power Analysis.

Treatment	*Errors*	*% Score*
1	19	70
2	12	81
3	9	85
4	5	92
5	14	78
6	3	95
7	18	71
8	8	87
Mean	11	82.81

no subjects/treatments variability exists, this method pools the interactions as an estimate of the error variance. This method of error effect estimation is less conservative but provides an upper bound for the power estimate range. The summary table for the single subject ANOVA can be found in Table 12.

Applying the error estimates previously described to Equation 1, two sets of power estimates were generated (Table 13). The power estimate for the *F* was determined from Table A-2 of Appendix A in Keppel (1982) by using the result of the equation (phi), and the degrees of freedom from the numerator and denominator of the the *F* ratio. Table 13 contains the summary of the power analysis using both methods of error estimate. Figure 7 graphically represents the relationship between the sample size and the power estimates. It should be noted that power levels were visually estimated from the chart and do not represent exact quantities. An average estimate of the power was constructed by summing the power estimates generated by the different error estimates and dividing by two. It was decided that a sample size of 4 should be used since this sample size was the first average power to exceed the .80 criterion.

Subjects:Thirty-two (16 Anglophone, 16 Hindi) college students were compensated for volunteer participation in the experiment. Native language was determined by the subject's self report.

Subjects were administered a hearing screening test to eliminate subjects who were "hard of hearing" (American National Standards Institute, 1973). During the screening test, subjects were asked to acknowledge a 3 s pulsed tone presented to either the left or the right ear. A series of five tones was presented sequentially at 26 dB at the following pure-tone frequencies: 800, 1000, 2000, 4000 Hz on a Beltone Model 109 audiometer. Had a subject failed to acknowledge a tone on two out of three presentations, the subject

TABLE 12
Summary Table for the Single Subject ANOVA.

Source	*df*	*SS*	*F*
Number of Steps in a Search (S)	1	1.22	0.07
Native/ Non-Native(N)	1	312.38	17.68
Adapting Speech Rate (R)	1	19.56	1.11
Sex of Voice (V)	1	206.35	11.68
Transaction Summary (T)	1	1.22	0.07
S x T	1	0.001	0.01
S x N	1	30.46	
Error †	2	35.34	

† Pool of interactions

TABLE 13
Summary of the Power Analysis

Sample Size	*Sample*		*Single N*		*Average Power*
	Phi	*Power*	*Phi*	*Power*	
2	1.32	0.35	2.86	0.97	0.66
3	1.62	0.50	3.50	0.99	0.75
4	1.87	0.74	4.04	0.99	0.87
5	2.09	0.84	4.51	0.99	0.92
6	2.29	0.88	4.95	0.99	0.94
7	2.47	0.94	5.34	0.99	0.97
8	2.65	0.96	5.71	0.99	0.98
9	2.81	0.98	6.06	0.99	0.98
10	2.96	0.99	6.38	0.99	0.99

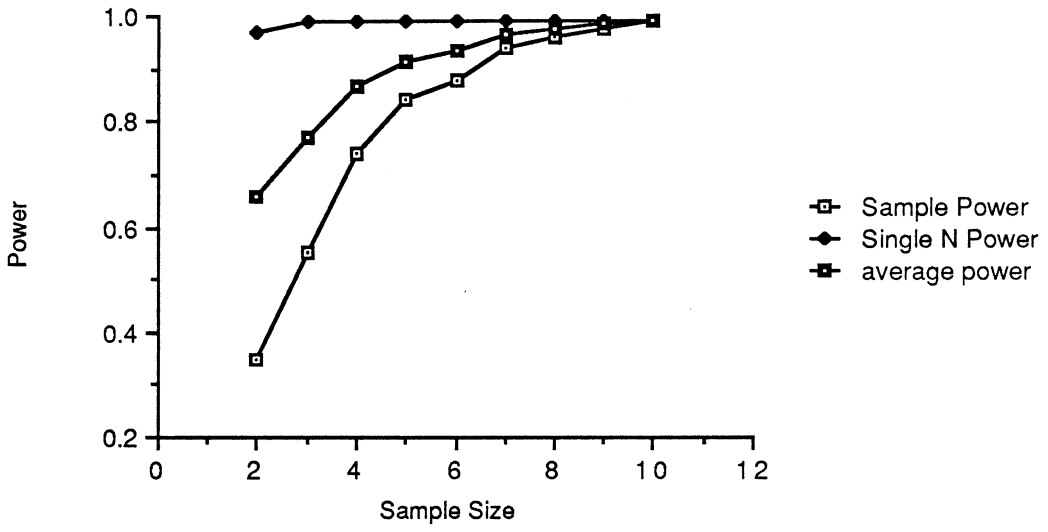


Figure 7 *Graphical summary of the power analysis.*

would not have meet the acceptable criterion, and their data would not have been used in the study. None of the subjects volunteering in the experiment failed the hearing criterion.

Apparatus: As in the subjective ratings study, the telephone inquiry system was implemented using the following hardware: DEC VAX/11 750 mini-computer, DECTalk 2.0 rule based speech synthesizer, and a Panasonic speaker telephone. The experimental prototype was extensively modified to allow the manipulation of more variables and more rigid control of the experiments.

The "select" and "backup" commands were not changed. To clarify the function, the "restart" command was renamed "main." Two new commands were added to accommodate the experiment. "Transaction summaries" was assigned to the "5" key on the telephone keypad. To indicate the beginning of the transcription task described in the procedure section, the "2" key was assigned the "begin transcription" function.

The system used the same commands and menu style as the prototype evaluation study. Keywords and information messages were entered without any manual phoneme or additional vocal stress cues. All text was spoken from standard spellings. However, the information database used was new.

The new information database was a balanced 2x6 or 8x2 tree hierarchy. The structure created a 64 possible information messages. A department store (Hokie Wholesale) was used for the information content in the database. Keywords represented items and classes of items commonly found in a department store.

Information messages related to specific store items and were one of four types: general, location, price, and availability. Each message was a single sentence and was of a controlled syntax in the form of *adjective-noun-verb-*

preposition-object. The verbs and prepositions used in the information messages are shown in Table 14.

Procedure: The entire experimental procedure sequence is summarized in Figure 8. Prior to the experimental task, subjects were requested to complete an informed consent document (*Appendix C*), a hearing screening test (refer to the "subjects" section for a detailed description), and a demographic survey (*Appendix D*). Next, the subject was presented with the experimental task instructions. Subjects first received the general instructions (*Appendix G*) presented in a paper format and spoken by the DECTalk version 2.0 spoken over the telephone. These instructions were intended to familiarize the subject with the operation of the speaker telephone, the telephone inquiry system, and the experimental task. When the subject finished reading and listening to the general instructions, the same instructions were presented in a videotaped version. The videotape consisted of a reading of the instructions with a concurrent demonstration of the telephone inquiry system. After the videotaped instructions, the subject received specific instructions describing the command keys operational during their experimental session (*Appendix H*). These instructions were also presented in written and DECTalk versions. The experimenter then asked the subject to recapitulate the instructions to verify the subject's comprehension of all of the instructions. To ensure consistency of instruction presentation across all subjects, no further assistance was provided regarding the operation of the system.

During the experimental task, each subject was asked to complete 2 warm-up trials and 16 search trials. A pretest determined that subjects frequently missed the first one or two targets. Two warm-up trials were to reduce the possibility of missing the search trial solely because of a lack of understanding of the task environment. The 18 information targets are

TABLE 14
Information Messages Format

Information Type	*Format*			
LOCATION:	*Adjective subject*	is/are	near in on	*object*
PRICE:	*.Adjective subject*	is/are reduced	for by	*object*
	Adjective subject	is/are sold	for by	*object*
AVAILABILITY:	*Adjective subject*	is/are available	with at by in	*object*
INFORMATION:	*Adjective subject*	is/are offered	with for to	*object*
	Adjective subject	is/are required	within for on to	*object*

WELCOME AND ORIENTATION (~ 15 mins)

- Informed Consent
- Subject Information Questionaire
- Hearing Test

INSTRUCTIONS AND PRACTICE (~ 20 mins)

- Introduction (audio - written)
- Instructions (audio - written)
- Video Instructions
- Telephone Key Instructions (audio - written)
- Subject Recapituation of Instructions
- Practice Targets (n=2)

EXPERIMENTAL TASK (~ 30 mins)

- 8 Experimental Targets
  - Target Search
  - Transcription
  - Target ratings (inc. Adapting speech rate)
- Break (minimum 1 minute)
- 8 Experimental Targets
  - Target Search
  - Transcription
  - Target ratings (inc. Adapting speech rate)
- Post Experimental Ratings

POST EXPERIMENTAL SESSION (~ 15 mins)

- Debriefing
- Payment and Dismissal

Figure 8. *Experimental procedure sequence.*

presented in *Appendix I.* At the beginning of a trial, a target (store item) was presented on the computer screen adjacent to the speakerphone and remained for 15 s to allow the subject to read the target. After the 15 s target familiarization period, a "ready..." message appeared on the screen indicating the beginning of the trial. The subject then phoned the system to begin the search.

When the subject reached the information level in the data base, the message "At store item *keyword*" was spoken. If the subject selected an incorrect information message, the message continued with "continue searching." The subject was then forced to continue searching until the correct information message was found. If a subject had searched for more than 10 minutes to locate a target, the experimenter would have terminated the task. No such incidences occurred.

Selecting the correct information message resulted in a precursor message of "At store item *keyword*", press the 2 key to hear the information message (*Appendix J*)". Pushing the 2 (begin transcription) key transferred control to the transcription task. During the transcription task, subjects heard the information message, then typed the information message on the computer terminal.

The transcription difficulty of the 16 information messages was determined by pretesting. Four subjects transcribed a set of 77 messages with the Perfect Paul voice speaking at 180 wpm. Prior to these sentences the subjects practiced transcription of synthesized speech on five Harvard Psychoacoustic Sentences (Egan, 1948). The message transcription accuracy ratings were computed by the method described in the dependent measures section. Information messages were eliminated if no mistakes were made or if all four subjects incorrectly transcribed the sentences. The remaining messages

contained between one and three transcription errors across the four pretest subjects.

After transcribing the message, the subject rated the certainty of his or her transcription, the difficulty of the transcription, and the difficulty of the search trial. These measures are described in detail in the dependent variables section.

After completing 16 search trials, all subjects rated the ease of use and speech intelligibility of the telephone inquiry system. Subjects receiving the adapting speech rate and transaction summary features rated how essential were each of these features. All subjective ratings were on a 7-point anchored scale, displayed on the screen. The scale was graphically displayed on the screen with the verbal anchors at "1" and "7." Subjects entered their rating by pressing a number and the "return" key. Once the post-experimental ratings were made, the subjects had completed the experimental task.

After completing the task the experimenter debriefed the subject in a post-experimental interview. The interview was structured according to Appendix K. During the debriefing, the experimenter followed the structured interview and discussed any peculiar occurrences which arose in the session. Subjects who received but did not use the adapting speech rate and/or transaction summary features were asked why the feature was not used. After participation, subjects were compensated for their time and thanked for their participation. At this time the experimenter answered any of the subject's questions about the intent of the experiment.

Dependent measures. During the engineering analysis phase of the project, 12 relevant dependent measures (5 objective, and 7 subjective) were listed. The objective measures were automatically recorded in real time by automated metering embedded in the telephone inquiry system.

Target search time ratio was the time required by the user to locate the target divided by the minimum time necessary for an expert to locate the target. An analysis of pilot subject data revealed that input timeouts (the pause between keywords), which was set a 4 s varied up to 0.5 s on some occasions. It was determined that these variations were the result of the operating system routines in the VAX 11/750 VMS operating system. It was also determined that the DECTalk text-to-speech synthesizer also varied slightly in the time to speak a particular keyword or information message. Because of these uncontrollable variations in system performance, it was necessary to use average expert task completion times generated by executing 4 iterations of a simulated expert using the experimental task. The simulated expert was implemented by having the experimental software execute the target searches by accessing input from a file rather than the telephone. The file contained the keypresses necessary to find the target without error. The expert selection time of 0.57 s was obtained from the American Institutes for Research Data Store (Munger, Smith, and Payne, 1962). Munger, Smith, and Payne (1962) described this value as the time to press a single pushbutton when a cue light was extinguished. In a telephone inquiry system, an expert would know the telephone key to push. The expert would use the beginning of the command input pause as the cue to input the command. The main advantage of using the simulated expert was that it accounted for and averaged the system variations.

Target search efficiency ratio was the ratio of minimal keypresses necessary used to find the targetto the user's number of keypresses to find the target. A minimal (also optimal) search strategy would result in a target search efficiency of one. Any keystrokes over the minimum strategy would decrease the ratio thus indicating the level of inefficiency.

The average number of invalid keypresses used was a measure of the average number of telephone keypresses per target which were completely unrelated to the task (i.e. selecting a telephone key not assigned to a function). The purpose of this metric was to extract inefficient keypresses which were not attributable to search strategy. These keypresses may be due to telephone keypad design, function mapping, fatigue, etc.

The number of transaction summaries requested measured the number of times the person used the transaction summary feature. The intent of this measure was to collect a performance measure which would assist in determining the need for the feature.

The purpose of the message transcription accuracy score was to assess the user's ability to understand the spoken messages. As established by Merva (1987), each transcribed message was scored on two dimensions: (1) exact transcription, and (2) intent of the message (synonym scoring). Four words were scored in each information message - two words at the beginning and two words at the end. On the exact transcription dimension, answers were scored by assigning one point for each correctly transcribed beginning and end word. On the synonym dimension, accurate beginning and end words were scored as correct along with synonyms even though the synonyms were not spoken in the message. For example, substituting handbag for purse would be scored as correct.

Recorded after each target search, message transcription certainty rating was a measure of the subject's certainty in understanding the information message. Transcription certainty was rated on a 7 point anchored scale (1 = very uncertain, 7 = very certain). Each subject was given an overall transcription certainty score by assigning the median value for all sixteen trials.

Recorded after each search target search, message transcription difficulty rating was a measure of the subject's view of the difficulty in understanding the information message. Transcription difficulty was also rated on a 7 point verbally anchored rating scale, (1 = very difficult, 7 = very easy). As with the transcription certainty rating, each subject was assigned a median score based on all 16 search trials.

Recorded after each target search, search trial difficulty rating was a measure of the subject's view of the difficulty in locating the information message. Search trial difficulty was rated on a 7 point verbally anchored scale rating scale, (1 =very difficult, 7 = very easy). Again, this measure was recorded as the median of all 16 search trials.

Several subjective ratings were collected after the target search and transcription portions of the task were completed. Adapting speech rate usefulness rating was a post-task rating of the usefulness of the adapting speech rate feature. Rated on a 7 point verbally anchored scale (1 = not essential, 7 = very essential). Transaction summary usefulness rating as a post-task rating of the usefulness of the transaction summary feature. Again, this feature was rated on a 7 point verbally anchored scale (1 = not essential, 7 = very essential). Ease of use rating was a post-task rating of the system's ease of use rated on a 7 point verbally anchored scale (1 = very difficult, 7 = very easy). Speaker intelligibility rating is a post-experiment rating of the subject's perception of the intelligibility of the voice used in the telephone inquiry system and was rated on a 7 point anchored scale (1 = not understandable, 7 = very understandable). These ratings allow the experimenter to determine the effects of speech display characteristics on the subject's impression of the overall system intelligibility.

RESULTS

The data were analyzed in two sections: (1) the objective performance measures, which met the criteria for parametric analysis, and (2) the subjective ratings which were also analyzed by parametric methods even though the data were ordinal scale. The rationale for using parametric analyses for ordinal scale data was based on the tendency of rank sums to approximate the normal distribution as sample size increases. This decision is discussed in detail in the subjective ratings section of the results. The decision was reached after consulting a statistician at the Virginia Tech Statistical consulting center. The parametric analyses discussed herein were performed using the Statistical Analysis System (SAS) software implemented on the Virginia Tech IBM mainframe computer (VTVM2). The overall type I error for each analysis was set at p =0.05. The null hypothesis for each test was that no significant difference exists between the levels of each the factors tested.

Objective Measures

A Multivariate Analysis of Variance (MANOVA) was performed to determine any statistically significant effects of the five main effects and two interactions on the five objective dependent measures. Because of the confounding of higher order interactions, the usual error term of subjects within the fifth order interaction does not exist. For this reason the residual sum of squares cross product matrix was used as the error term. The statistical significance of each effect was tested with the Wilk's likelihood ratio which was also converted to the familiar F statistic. A summary of this analysis is presented in Table 15. The number of steps in a search $F(5,20)=20.67$; p =0.0001 and native/non-native ($F(5,20)=6.62$; p =0.0009) were revealed as significant. None of the other main effects or interactions were statistically significant in this analysis. In keeping with a standard interpretation of

TABLE 15

MANOVA Values for Main Effects and Interactions on all Objective Dependent Measures

Source	d_V	df_H	df_E	*U*	*F*	*p*
Between Subjects						
Number of Steps (S)	5	1	20	.1622	20.67	.0001*
Adapting Speech Rate (R)	5	1	20	.8556	.68	.6472
Transaction Summary (T)	5	1	20	.8587	.66	.6589
Native/Non-native (n)	5	1	20	.3765	6.62	.0009*
Sex of Voice (V)	5	1	20	.8601	.65	.6642
S x T	5	1	20	.8874	.51	.7672
S x N	5	1	20	.7446	1.37	.2764

* significant at $p = 0.05$

where d_V = number of dependent measures

df_H = degrees of freedom in the treatment effect

df_E = degrees of freedom in the error effect

U = Wilk's likelihood ratio statistic $\frac{|E|}{|E + H|}$

where $|E|$ = determinant of the sum of squares and cross-product matrix for the error term

where $|E + H|$ = determinant of the sum of squares and cross-product matrix for the error, and the sum of squares and cross-product matrix for the treatment term

the MANOVA, only the significant effects found in this analysis (number of steps in a search, and native/non- native) were considered in the separate analyses on each of the dependent measures

A separate ANOVA was performed for each of the five objective dependent measures to determine which of these was affected by the main effects of number of steps in a search and native/non-native. The results of these analyses are presented in Table 16 for each of the dependent measures. A listing of each of the separate ANOVA tables for the objective measures can be found in Appendix L.

Table 16 demonstrates that the target search time ratio was significantly affected by the number of steps in a search $F(1,24) = 78.92$; $p = .0001$, and native/non-native $F(1,24) = 9.21$; $p = .0056$. The subject target search time ratio was higher using the database with two levels of eight item search (8x2) than the subjects using the database with six levels of two items each. The mean target search time efficiency ratio for two steps in the search was 0.91, whereas, the mean target search time efficiency ratio for six steps in the search was 0.65. The non-native English speakers had a significantly lower mean target search time efficiency (0.73) than the native anglophones (0.83). None of the other main effects or interactions showed statistically significant results.

The results of the search path efficiency ratio were similarly to those found for the search time efficiency ratio. Again, the significant effects were the number of steps in a search $F(1,24) = 47.69$; $p = 0.0001$, and native/non-native $F(1,24) = 9.51$; $p = 0.0051$. Just as in the search time efficiency ratio, the mean efficiency of the search path ratio was higher for the database with two steps (0.94) than was the database with six steps (0.40). The native anglophones were also significantly more efficient in locating the target ($\mu = 0.88$) than were the non-native English speakers ($\mu = 0.80$).

TABLE 16

Summary Table of Significant Results from Individual ANOVAs on Objective Measures

Factor	*Objective Measure*	*F*	*p*
Number of Steps	target search time	78.92	0.0001
	target search effeciency	47.69	0.0001
Native/Non-native	target search time	9.23	0.0056
	target search efficiency	9.51	0.0051
	invalid keypresses	7.41	0.0119
	transcription acc. (strict)	26.89	0.0001
	transcription acc.(synonym)	19.55	0.0002

As shown in the Table 16, the native/non-native factor was also significant (F (1,24) =7.41; p =0.0119) on the average number of invalid keypresses metric. The native anglophones had a lower invalid keypress average of 0.004 in comparison to the non-native English speakers who had an invalid keypress average of 0.08.

As described previously, the message transcription accuracy was scored by two methods: (1) a strict scoring regime, and (2) a synonym scoring regime. Under the strict message transcription accuracy scoring, only the native/non-native factor F (1,24)=26.89; p =0.0001 affected the score. The native anglophone had the higher mean strict message transcription accuracy score of 3.6 (out of 4) words correctly transcribed per sentence. On average, the non-native English speakers accurately transcribed 3.1 (out of 4) words per sentence.

In the synonym method of scoring message transcription accuracy, transcribed words that were not spoken in the original message but that retained the meaning of the message were scored as correct. Correctly transcribed words from the original message were also scored as correct. Not surprisingly, the results of the synonym message transcription accuracy score were similar to the results under the strict scoring regime. Only the Native/non-native factor F =19.55; p = 0.0002 affected the synonym score. Again, the native anglophones scored higher with a mean message transcription accuracy of 3.71 than did the non-native English speakers with a mean message transcription accuracy of 3.28.

Subjective Ratings

As discussed previously, the subjective ratings were analyzed using parametric statistics. With 7 dependent measures, the most desirable inferential statistical test would be a single nonparametric test which tested all the main

effects while controlling the type I error for the repeated testing necessary resulting from the multiple dependent measures. Such a test would be a nonparametric analogue to the objective measures MANOVA, but no such test exists to the best of the author's knowledge. Therefore, one must determine whether it is more prudent use a MANOVA to control type I error while violating the assumption of interval data, or to use a nonparametric test with reduced power as a result of adjusting the type I error for the number of comparisons Given the parameters of this investigation, a statistical consultant for the Virginia Tech recommended using the MANOVA. The consultant's recommendation was based on the fact that the distribution of ranked sums (the subjective ratings) would tend to approximate the normal distribution even with a sample size of 4 subjects per cell in the present experimental design. It was the consultant's opinion that it was more advantageous to use the more powerful MANOVA, thus, parametric tests were used to analyze the subjective ratings.

The subjective ratings MANOVA was performed and interpreted in the same manner as the objective measures MANOVA. The MANOVA was used to determine if any statistically significant effects existed in any of the five main effects and two interactions on the 7 subjective dependent measures. As with the objective measures MANOVA, the confounding of higher order interactions eliminated the usual error term of subjects within the fifth order interaction, therefore, the residual sum of squares cross product matrix was used as the error term. The statistical significance of each effect was tested with the Wilk's likelihood ratio which was also converted to the familiar F statistic. A summary of this analysis is presented in Table 17. Only the native/non-native factor was significant (F (7,18)=2.63; p = 0.0467). None of the other main effects or interactions were statistically significant in this analysis. In keeping with a

TABLE 17

MANOVA Values for Main Effects and Interactions on all Subjective Ratings

Source	d_V	df_H	df_E	*U*	*F*	*p*
Between Subjects						
Number of Steps (S)	7	1	18	.6117	1.63	.1899
Adapting Speech Rate (R)	7	1	18	.6178	1.59	.2013
Transaction Summary (T)	7	1	18	.8364	.50	.8202
Native/Non-native (n)	7	1	18	.4947	2.63	.0467*
Sex of Voice (V)	7	1	18	.6548	1.36	.2824
S x T	7	1	18	.6680	1.28	.3153
S x N	7	1	18	.7799	.73	.6526

* significant at $p = 0.05$

where d_V = number of dependent measures

df_H = degrees of freedom in the treatment effect

df_E = degrees of freedom in the error effect

U = Wilk's likelihood ratio statistic $\frac{|E|}{|E + H|}$

where $|E|$ = determinant of the sum of squares and cross-product matrix for the error term

where $|E + H|$ = determinant of the sum of squares and cross-product matrix for the error, and the sum of squares and cross-product matrix for the treatment term

standard interpretation of the MANOVA, only native/non-native factor was considered in the separate analyses on each of the dependent measures The individual analyses of each of the 7 subjective measures demonstrated the main effect of native/non-native on three measures: (1) transcription certainty (F (1,24)= 6.42; p = 0.0182), (2) ease of use (F (1,24)= 6.11; p = 0.0209), and (3) intelligibility(F (1,24)= 5.11; p = 0.0331). A summary of these results can be found in Table 18. The complete ANOVA summary tables can be found in Appendix L. The native anglophones had a higher mean transcription rank of 20.312 than did the non-native English speakers (μ rank = 12.688). Likewise, the native English speakers ease of use mean rank (20.531) was higher than the non-native English speakers (12.469). The intelligibility was also rated and higher by the native anglophones(μ rank = 19.375) than by the non-native English speakers(μ rank = 12.625)

The results of the subjective rating scales are presented graphically with each scale represented by the median rating by independent variable level. The graphs with average ratings were not used in the analyses, but are presented to give the reader a parsimonious view of the data. The transcription certainty (tcert), ease of use (ease), and intelligibility (intel) were rated for all the main factors. The results of these ratings for the native/non-native factor are graphed in Figure 9. For the factors transaction summary, and adapting speech rate, subject's receiving the feature rated how essential they perceived the feature. Figures 10 and 11 present these data

TABLE 18

Summary of Subjective Ratings which Significantly Affected the Native/non-native Factor as Result of Individual ANOVAs

Subjective Rating	*F*	*p*
transcription certainty	6.42	0.0182
ease of use	6.11	0.0209
intelligibillity	5.11	0.0331

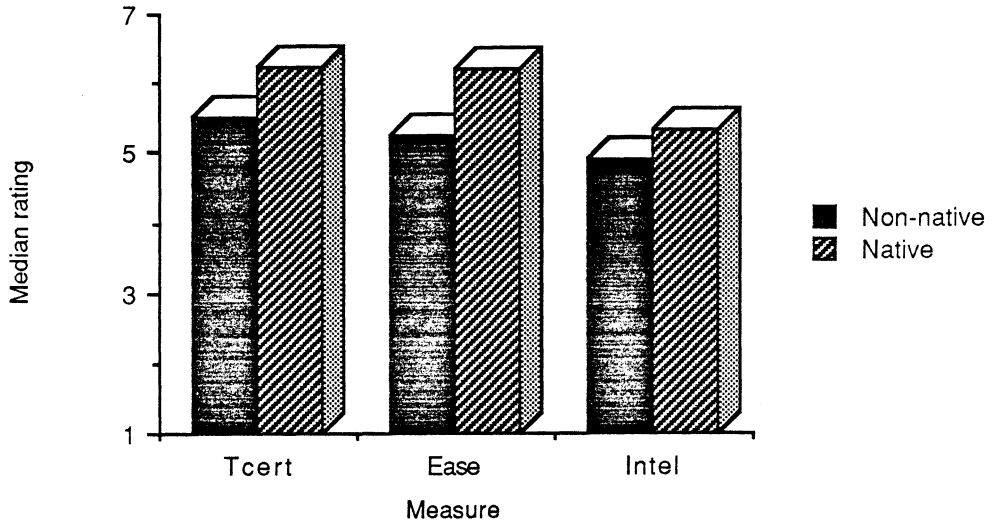


Figure 9. *Subjective ratings on native/non-native.*

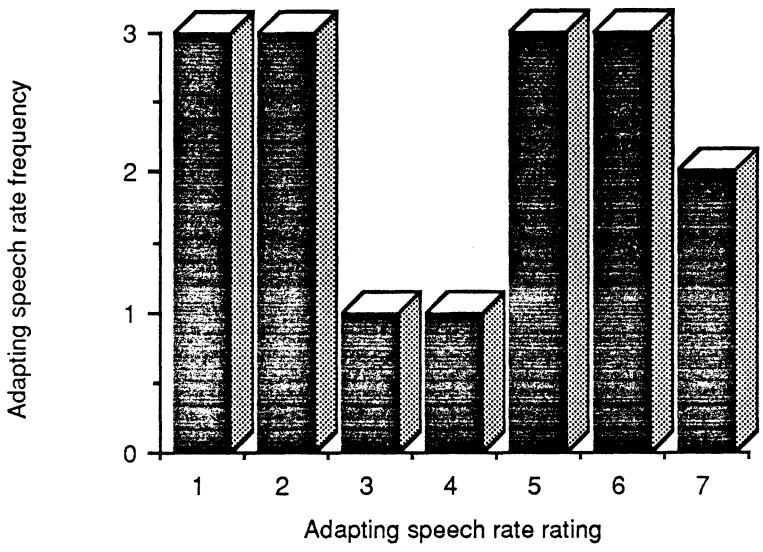


Figure 10. *Essentialness rating on adapting speech rate.*

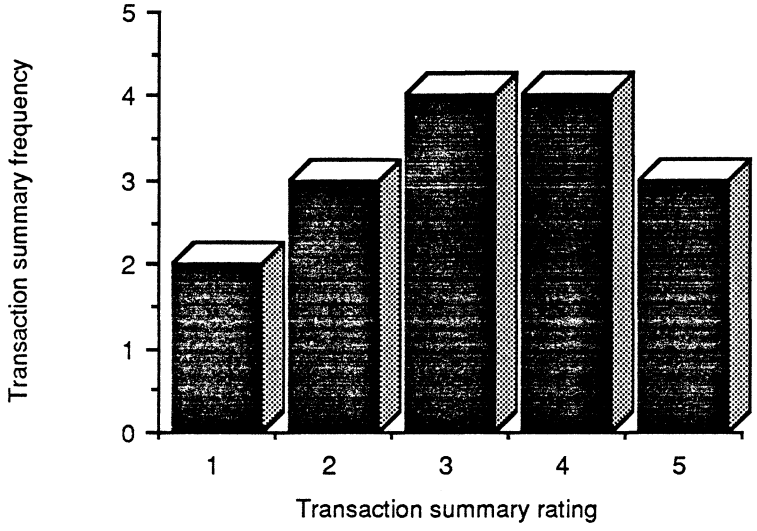


Figure 11. *Essentialness rating on transaction summary.*

DISCUSSION

In terms of the experimental hypotheses, two factors (Number of steps in a search, and Native/non-native) were found to have the predicted effect on performance and the subjective ratings. The absence of a performance or subjective ratings effect followed the experimental predictions for the transaction summary variable. The number of steps in the search affected both the search task dependent measures of target search time efficiency, and target search path efficiency. Referring to the prototype evaluation the number of steps in a search represented the group of variables which were not found in the literature and were rated as important by the subjects evaluating the prototype. As discussed previously, the efficiency ratios for each subject are relative to the simulated expert time. The use of these ratios eliminated system differences in the menu organization and implies that subject performance and preference for the 8x2 database is superior to the performance and preference on the 2x6 database. This investigation did not research the issue in sufficient detail to make design recommendations, but the results indicate that the subjective ratings are useful in identifying independent variables not found in the literature which would prove significant in subsequent investigations.

These results are completely opposite to results of depth and breadth studies of visual menus. Kiger (1984) and Miller (1981) found that and 8x2 menu organization produced significantly higher performance rates than did a 2x6 menu organization. The difference between the results of this study, and the results from the studies of visual menu organization can be attributed to the difference between the primary sensory mode used to perceive the menu. In a visual menu, multiple items can be displayed simultaneously. Users can scan many items and select an appropriate choice. Therefore, it follows that an 8x2 organization would improve performance. In a telephone inquiry system,

audition is the primary sensory mode, and audition is primary temporal relying on rehearsal. An 8x2 menu structure would require the subject to remember up to 8 items to make a single select. Alternatively, a 2x6 organization requires the subject to remember at most two items to make a selection. It is reasonable that the menu organization (2x6) which is less taxing to memory would promote higher efficiency and fewer errors in a telephone inquiry system.

In contrast the adapting speech rate feature, which was reported in a case study, and subjectively rated as having an effect in the prototype evaluation did not significantly affect performance or preference in the experimental validation. This result is likely an artifact precipitated by the rigid design of the transcription sentences, or the small number of trials used in the experiment.

Subjects receiving the adapting speech rate feature were instructed to adjust the speech rate in response to the difficulty in using the system. The information messages used for the transcription portion of the task were designed and pretested to induce one to three errors per sentence across four subjects. This restricted the difficulty to the middle of the range eliminating both the easy and difficult extremes. Had the sentences been very difficult, subjects might have decreased the speech rate to increase intelligibility. Alternatively, had the sentences been very easy. Subjects might have increased the speech rate to finish the task more rapidly. Given the restricted range of sentence transcription difficulty and a greater number of trials, the need to adapt the speech rate might have increased. This increase in training with synthetic speech might have also made the system less difficult, thus, motivating subjects to increase the speech rate.

Another explanation for not finding the predicted effect on the feature adapting speech rate is that the speech rate of 180 was optimum as suggested

by Merva(1987). When subjects were permitted to adapt the speech rate, the highest speech rate used was 200 wpm, whereas,the lowest speech rate used was 150 wpm. The average speech rate ranged from 182.5 wpm to 150 wpm. If 180 wpm is the optimum speech rate, then the subjects' choice of not changing the speech rate maintained the rate within range optimum range.

The results from the transaction summary feature indicate that the subjective ratings predicted that no effect would be found. This is a somewhat tenuous observation since it requires accepting the null hypothesis, but as with many decisions it is based on the weight of the evidence. In this case, the variable was not found in the literature. During the prototype evaluation, subjective ratings predicted that no effect would be found. The validation experiment confirmed the prediction. As further evidence, the transaction summary was activated only once by one subject in the entire experiment. Post-experiment interviews with the subjects revealed that the subjects were aware of the feature and familiar with it benefits, but choose not to use it because they did not find an occasion when the feature would have aided their search. These results support the premise of using subjective ratings of a prototype to eliminate certain independent variables from consideration based on a prediction of no effect.

The results of the adapting speech rate and the transaction summary variable should be discussed in terms of the implementation of these factors as voluntary features. Although experimental debriefings indicated that subjects were aware of the utility of these features, neither feature was exercised frequently. As stated previously, the transaction summary feature was selected only once during the entire experiment. This implies that performance in the voluntary selection level of these features would be expected to be very similar to performance in the not available level of these features. Essentially, these

features were not manipulated. Therefore the results should not be interpreted that these feature would never affect performance, but that giving the user control to exercise these options in a similar system is not likely to affect performance. Had the subjects been forced or motivated to exercise these features, performance differences might be found.

Another experimental hypothesis was reaffirmed by the results of the native/non-native factor which was predicted by both the literature review and the subjective ratings of the prototype. The native/non-native factor represented the variables for which the literature review and the subjective ratings of the prototype agreed on the predicted result. The replication of the literature results in the validation experiment supports the use of subjective ratings to select independent variables by demonstrating that subjective ratings are able to predict replication experiment results. The impact of the native/non-native factor on message transcription accuracy (strict and synonym) and target search time efficiency ratio was expected. It is likely that listening to speech synthesizer speak a foreign language is likely to increase the time to comprehend the message, as well as increase the chance of incorrectly transcribing the message.

The effect of the native/non-native factor on the target search path efficiency is less clear. Path efficiency measured the subject's ability to select correctly and keywords leading to a particular target - a skill which is mostly dependent on the ability to transverse a hierarchical menu. It is unlikely that being raised in an Indian culture alone would cause a marked decrease in the ability to search hierarchical menus. It is more likely that the subjects had difficulty understanding the keywords, therefore, causing them to deviate from the optimal path. This decrease in target search path efficiency may be attributed to the intelligibility of the synthetic speech, or it may be attributed to

the nature of the vocabulary used in the database rather than skill with the English language. All the Indian subjects were required to be proficient in English, but this criterion was designed to screen for general English usage. The database used in the experiment was composed of items commonly found in American department stores. The post-experimental interview revealed that the Indian subjects were confused by some of the categories because they were uncertain about the classification of the item. For example, a majority of the Indian subjects indicated that they were confused by the hope chest item. They were certain it was a piece of furniture, but were uncertain as to where the item would be located in the home. All the Indian subjects were certain of the British term for the item, a wooden trunk. Regardless of origin of this effect, implementing a telephone inquiry system across language/culture boundaries obviously requires more investigation.

The sex of the synthetic voice had no significant effect on either the performance measures or the subjective ratings. This result was not predicted by the experimental hypotheses which favored the literature review, but was predicted by the subjective ratings of the experimental prototype. This indicates that the subjective ratings may be more sensitive than the review of the literature when predicting performance in specific system. In the validation experiment the structure of the database, and the task of locating information on items commonly found in a department store provided an increased level of context for the transcription sentences. Yet, the results indicate that when context exists, performance differences attributable to the sex of the voice (i.e. voice type) become non-significant. From these findings, it is impossible to determine whether context is the cause, and if so, how much context is necessary to reduce performance differences. The design of the validation experiment is similar to a screening study in that it attempts to discern overall

effects with low resolution. The information needed to answer specific design questions is not available. The purpose of the experiment was to test the validity of the use of the subjective ratings to predict experimental results.

Finally, the existence of several significant inverted *F*s tests indicates that the assumption of homogeneity of variance made in the ANOVA may have been violated. This assumption was tested by evaluating the significance of the inverted value of the original *F* ratios less than one. If the inverted value was significant then the assumption of the homogeneity of variance may have been violated. Of all the inverted *F* ratios which were significant in an individual ANOVA, non of the significant inverted *F* ratios were on factors which had been found significant in the corresponding MANOVA. On the objective measures, neither the number of steps in a search nor, the native/non-native factors had a significant inverted *F*. Likewise, the native/non-native factor did not result in a significant inverted *F* on any of the subjective ratings individual ANOVAs.

Winer (1971) states that the ANOVA is robust with respect to the violation of the assumption of homogeneity of variance, but reassurance that the tests were not invalidated by these results does not explain their existence. A significant inverted *F* test implies that the mean squares of the the error term was significantly greater than the mean squares of the treatment term, thus, uncontrolled variability was significantly greater than treatment effects. Winer (1971) stated that other than the variability due to treatments two sources of variance exist: (1) differences that existed prior to the experiment, and (2) uncontrolled variability introduced during the experiment which was not related to the treatments. Although uncontrolled variability may have been introduced during the experiment, it is more likely that the error terms contained preexisting, systematic variability. It is very likely that the heterogeneity of

variance was introduced by one or more of the other variables affecting performance in a telephone inquiry system.

CONCLUSIONS

In review, the purpose of this thesis was to use subjective ratings of a prototype to select independent variables, then to validate the use of the ratings.This research developed from a need for a valid effective method for limiting large numbers of possible independent variables to a smaller set which can be implemented in screening studies. Therefore the results of the study do not generate human factors design recommendations, but define the limits of the method and point to new research issues.

From the results of the validation experiment there is clear evidence that subjective ratings of prototype systems can be successful in identifying significant and nonsignificant results. Most importantly, the results support the the assumption that users are able to make valid assessments of possible performance differences based on subjective ratings collected on a prototype. Table 19 reviews each of the main factors in the validation experiment by the predictions from the literature review, subjective ratings, and the experimental hypotheses. Figure 12 is a graphical presentation of the results of Table 19. Table 19 also contains the results of the validations experiment. As evident from the results, the subjective ratings predicted greater percentage of the actual results than did the literature review or the experimental hypotheses. The experimental predictions were based on the literature review when available, and the subjective ratings when the no literature existed. The experimental hypotheses were accurate on three of the main effects. The literature review was accurate on one of the main effects, and the subjective ratings were accurate on four of the main effects. Relying on the literature in lieu of the subjective ratings was less accurate than relying on the subjective ratings alone.

The results can be interpreted that the subjective ratings were more sensitive to effects on performance in the prototype system than the general literature. This is most likely due to the lack of relevant literature and poor generalization of the literature to a particular system. The number of steps in a search factor is an example where the

TABLE 19
Grand Table for Hypotheses and Results

Independent Variables	*Llt. Review*	*Subj. Ratings*	*Exp. Results*	*Dependent Measures*	
				Objective	*Subjective*
Number of Steps	Not Found	**Effect**	Effect(s)	stim, seff	
Adapting Speech Rate	Not Manip'd	**Effect**	No Effect		
Transaction Summary	Not Found	**No Effect**	No Effect(s)		
Sex of Voice	**Effect**	No Effect	No Effect(s)		
Native/ Non-Native	**Effect**	**Effect**	Effect (l,s)	stim, seff strict, synon	tcert, ease intel

Effects in **bold** indicate the experimental hypotheses.

Letters in parentheses indicate the analysis (literature review, subjective ratings) which predicted the results.

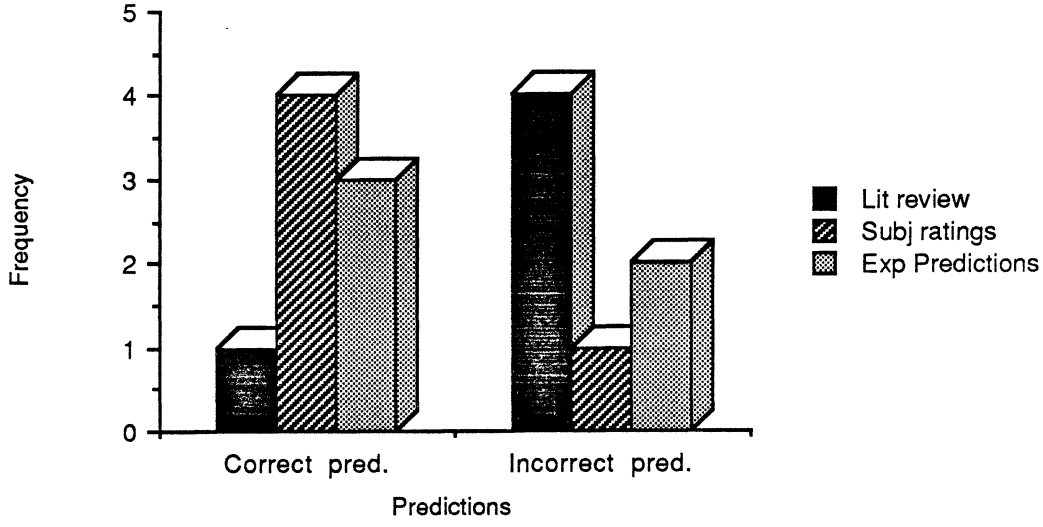


Figure 12. *Summary of hypotheses and results*

literature did not address the issue. The subjective ratings indicated the effect as possibly significant, and the experimental results confirmed the significance. The results from the adapting speech rate factor contradict this conclusion, but there is the possibility that the structure of the validation experiment may have precluded this effect. The results from the sex of the voice factor can be interpreted as an example of how the basic research does not translate well into an applied system. In this case the literature predicted performance differences when the subjective ratings did not.. The results of the validation experiment concur with subjective ratings.

In Simon's (1977b) methodology for investigating complex systems, he suggests that screening studies consisting of up to 100 independent variables are reasonable. As previously stated, such large screening studies seemed unreasonable based on the expense and feasibility of implementing the factors in an experimentally instrumented system. A practical obstacle is the ability to represent the independent variables simultaneously. Implementing five independent variables which were seemingly orthogonal proved difficult in this experiment. The outcome of the adapting speech rate is most likely a cause of the rigidly controlled difficulty in the transcription sentences and the brevity of the task. Changing either of these would have had an adverse impact on the experiment by introducing a difficulty artifact, or increasing the cost of the experiment. Adapting speech rate is not easily studied in combination with other factors; other such factors must also exist. Attempting to implement these factors in conjunction with other factors has the potential of reducing the power of the experiment, which in a screening study could eliminate a significant factor from further investigation.

Recommendations and Research Issues

Design of telephone inquiry systems. The results of the validation experiment indicate that the number of steps in a search (database organization) and native/non-native factors should be investigated further. There is evidence from the prototype evaluation and the validation experiment that indicate both of these factors significantly affect performance. Unfortunately, the results of this experiment do not have sufficient resolution to recommend design guidelines. Because of the problems encountered in implementing the adapting speech rate factor it is advisable to consider this factor for further investigation. Strong evidence also exists to dissuade further research on the transaction summary.

Whenever addressing speech quality issues in the design of a telephone inquiry system, the transmission of the speech signal from the synthesizer to the user should be considered. In the apparatus sections of the subjective ratings of the prototype and the validation experiment, the DECTalk speech synthesizer, telephone line, and speaker telephone were listed. The design of each of these components could impact the transmission of the speech signal, thus, impacting performance. For example, the synthesizer may have the ability to produce a high quality speech which when transmitted across a telephone line or converted to sound by the telephone could be significantly changed. When designing a telephone inquiry system, one should fully investigate the ability of the systems to transmit the speech signal from the source to the user. Likewise, telephone inquiry system research should fully consider the effect that transmission of the speech signal could have on performance.

Use of subjective ratings. The outcome of the sex of voice issue raises interesting research issues. The literature reports that a difference exists, whereas the validation experiment indicated that no such difference existed. Is

this difference caused by the increased context in the database?. If so, how much context is necessary to nullify performance differences due to the sex of the voice?

Although the results of the validation experiment are positive, they are not conclusive. The validation is based on only five factors and even the best predictive analysis, subjective ratings of the prototype, was only accurate 80 percent of the time. Further investigation of this technique is necessary to define the limits of its useful application, and methods of improving the subjective ratings. The most obvious method for improving the subjective ratings would be to provide the prototype evaluation subjects with examples of the levels of each independent variable. This could be accomplished either by demonstration or direct subject manipulation. It seems that direct subject manipulation of independent variables would yield the best results, but would drastically increase the cost.

Finally, recommendations for the use of subjective ratings in the selection of independent variables should be made given the results of the validation experiment. In the case study, 54 variables were cross-referenced on the literature search and the subjective ratings. The decision rules used in the case study weighted the literature review greater. Variables which were experimentally manipulated were first set to constants. These variables were set to constants because information on the variable's affect on performance was known. Next, the variables which had been subjectively rated as having no effect were set to constants. These variables were eliminated because the other variables were considered more likely to affect performance, therefore, more information could be gained by studying the variable. This left 19 candidate variables to be included in a screening study. Given the results of the

validation experiment, there is evidence that supports weighting the subjective ratings greater than the literature review.

Altering the decision rule used in the case study, It is recommended that the variables be divided into three categories: (1) variables that should definitely be set to constants, (2) variables that should considered for setting to constants, or otherwise included in a screening study, and (3) variables that should definitely be included in a screening study. Figure 13 visually divides the cross-reference of the subjective ratings and the literature review into these three categories.

All variables for which the subjective ratings predict an effect or no effect should be set to constants, unless there is a disagreement between the subjective ratings and the literature review. Using this rule, one selects all of the variables for category one and sets them to a constant for the screening study. Category two variables represent the variables for which the subject ratings and the literature review disagree about the variable's affect on performance. These disagreements occur when either the subjective ratings predict one outcome and the literature review predicts the opposite outcome, or when the subjective ratings have no clear opinion and the literature review predicts and outcome. Each variable in this category should be considered individually either to include the variable in the screening study, or set the variable to a constant. Expert judgement should be used to decide the costs and benefits of including the variable. The third category represents variables for which no clear rating was found, and no literature was found. These variables must be included in a screening study since no information exists about the variable's effect on performance.

Reflecting on the case study, category one would have included 34 variables. Twenty variables would have remained in categories two and three.

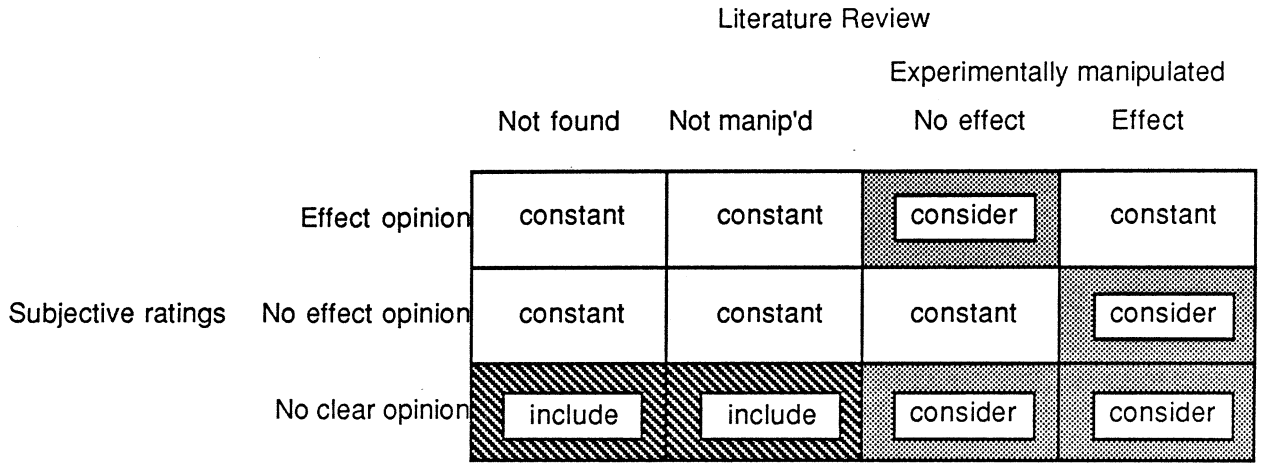


Figure 13. *Recommendations for the use of subjective ratings and literature review to select candidates for a screening study.*

Category two would have contained 13 variables, and category three would have contained seven variables. By including all of the variables in categories two and three, the new rules would have suggested 20 variables as candidates for a screening study. Certainly, expert judgement could be used to decide to set some of the variables in category two to constants which would reduce the list even further. If it was determined that all of the category two variables could be set to constants, the recommended method for suggesting candidate variables would have reduced the list to seven which is a much smaller list than the 19 suggested in the case study.

Considering that the new rules would have reduced that number of candidate variables to between 20 and seven, the recommended technique for selecting candidate variables is at least as powerful as the technique used in the case study. At best, the recommended technique is considerably more powerful than the technique used in the case study. Given the results of the validation experiment, one can be confident in applying the subjective ratings as a powerful tool for the selection of variables for a screening study.

REFERENCES

American National Standards Institute. (1973).*Psychoacoustical Terminology* . New York, NY: Author.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978).*Statistics for experimenters*. New York, NY: Wiley.

Champanis, A. (1963). Engineering Psychology. *Annual Review of Psychology. 14*, 285-318.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope, 58,* 955-961.

Fitts. P. M. (Ed.). (1947). Psychological research for equipment design. *Army Air Force Aviation Psychology Program ,* Research Report No. 19. Washington, DC

Geilselman, R. E., and Samet, M. G. (1982). Notetaking and comprehension for computer displayed messages. *Personalized versus fixed formats*. In Proceedings of Human Factors in Computer Systems (pp. 45-50). Washington DC: National Bureau of Standards.

Gerald. J. A. (1984). A voice response system for general aviation pilots.*Speech Technology, 2* (3), 33-37.

Kiger, J. I. (1984). The depth/breadth trade-off in the design of menu dreiven user interfaces. *International Journal of Man-Machine Studies, 20,* 201-213.

Meister. D., (1985). *Behavoiral Analysis and Measurement Methods.* New York, NY: Wiley.

Merva, M.A. (1987). *The effects of speech rate, message repetition, and information placement on synthesized speech intelligibility.* Unpublished masters thesis: Virginia Polytechnic Institute and State University, Blacksburg, VA.

Miller, D., P. (1981). The depth/breadth tradeoff in hierarchical computer menus. In Proceedings of the Human Factors Society 25th Annual Meeting (pp. 296-300). Santa Monica, CA: Human Factors Society.

Munger, S. J., Smith, R. W., and Payne, D. (1962). *An index of electronic operability: Data Store.* Pittsburg, PA: American Institute for Research.

Rao, S.R.S., and Sedransk, J. (1984). *W.G. Cochran's Impact on Statistics.* New York, NY: Wiley.

Rouse, W. B. (1987). Much ado about data.*Human Factors Society Bulletin, 30* (8), 1-8.

Rouse, W. B., and Cody, W. J. (1986, November).*Cold water and empty guns: A report from the front.* Paper presented at the Meeting of the Department of Defense Human Factors Engineering Technical Advisory Group.

Sanders, M.S., and McCormick, E.J. (1987). *Human factors in engineering and design.* New York: McGraw-Hill.

Smith, L. L., (1987). Whyfore Human Factors? *Human Factors Society Bulletin, 30,* (2), 6-7.

Simon, C. W. (1971). *Considerations for the proper design and interpretations of Human Factors Engineering Experiments* (Tech. Report ARL-71-27/AFOSR-71-11). Culver City, CA: Hughes Aircraft Co.

Simon, C. W. (1973). *Economical multifactor designs* (Tech. Report P73-326). Culver City, CA: Hughes Aircraft Co.

Simon, C. W. (1977a). *Design, analysis, and interpretation of screening studies for human factors engineering* (Tech. Report CWS-03-77). WestlakeVillage, CA: Canyon Research Group.

Simon, C. W. (1977b). *New Research Paradigm for applied experimental psychology: a system approach* (Tech Report. ADA056984). Westlake Village, CA: Canyon Research Group.

Simon, C. W. (1987). Will egg-sucking ever become a science? *Human Factors Society Bulletin, 30,* (6),1-4.

Virginia Polytechnic Institute and State University. (1986). *Graduate School Catalog.* Blacksburg, VA: Author.

Williges, R. C. (1981). Development and use of research methods for complex system/simulation experimentation. In J. Moraal, and K. F., Kraiss (Eds.). *Manned systems design: Methods, equipment, and applications* (pp. 59-87). New York, NY: Plenum.

Williges, R. C., and Baron, M. L., (1973). Transfer assesment using betweeen-subjects central composite designs.*Human Factors, 15*, 311-319.

Williges, R. C., and Mills R. G. (1982).*Catalog of methodological considerations for systems experimentation* (Tech. Report). Wright-Patterson Air Force Base, OH: Aeronautical Systems Division.

Winer, B. J. (1971). *Statistical principles in experimental design.* New York, NY: McGraw-Hill.

Appendix A. Literature Review Bibliography

Ainsworth, W. A. (1984). Performance of a speech synthesis system. *International Journal of Man-Machine Studies, 6,* 493-511.

Anderson, D. P. (1984). A talking computer give weather forecasts by telephone. In *Proceedings of the1st International Conference on Speech Technology*, (pp 98-103). Brighton, UK: North-Holland.

Berstein, J. (1982). Evaluating synthetic speech. In *Proceedings Workshop on Standardization for Speech Technology*, (pp 91-98). Gathersburg, MD: National Bureau of Standards.

Cox, A. C., Cooper, M. B. (1981). Selecting a voice for a specified task: the example of telephone announcements. *Language and Speech, 24* (3), 223-243.

Edman, T. R., and Metz, S. V. (1983). A methodology for the evaluation of real-time speech digitization. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 104-106). Santa Monica, CA: Human Factors Society.

Greene, B. G., Manous, L. M., and Pisoni, D. B. (1984). Perceptual evaluation of the DECtalk: a final report of version 1.8. *Research on Speech Perception, Progress Report Number 10,* Indiana University.

Greenspan, S. L., Nusbaum, H. C., and Pisoni, D. B. (1985). Generalization of training with synthetic words and sentences. In *Research on Speech Perception Report No. 11* (pp. 361-376). Bloomington, IN: Indiana University.

Gould, J. D., and Boies, S. J. (1984). Speech filing - An office system for principles. *IBM Systems Journal, 23* (1), 65-81.

Hise, H. H., and Lundin, F. J. (1985). Text-to-speech quality in a telephone information system. *Journal of American Voice I/O Society, 2,* 65-74.

Kidd, A. L. , (1982). Problems in man-machine dialogue design. In *IEEE Proceedings of the Sixth Conference on Computer Communications* (pp. 531-536). Amsterdam: North Holland.

Logan, J. S., Pisoni, D. B., and Greene, B. G. (1985). Measuring the segmental intelligibility of synthetic speech: Results from eight text-to-speech systems. In *Research on Speech Perception Progress Report no 11* (pp. 3-32). Bloomington, IN:Indiana University.

Manous, L. M., Pisoni, D. B., Dedina, M. J., and Nusbaum, H. C. (1985). Comprehension of natural and synthetic speech using a sentence verification task. In *Research on Speech Perception Report No. 11* (pp. 33-57). Bloomington, IN: Indiana University.

McPeters, D. L., and Tharp, A. L. (1984). The influence of rule-generated stress on computer-synthesized speech. *International Journal of Man-Machine Studies, 20,* 215-226.

Merva, M. A. (1987). *The effects of speech rate, message repetition, and information placement on synthesized speech intelligibility.* Unpublished masters thesis: Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Nakatani, L. H., and O'Conner, K. D. (1980). Speech feedbacking for touch-keying. *Ergonomics, 23,* 643-654.

Nusbaum, H. C., and Pisoni, D. B. (1984). Constraints on the perception of synthetic speech. In *Research on Speech Perception Report No.10* (pp. 153-168). Bloomington IN: Indiana University.

Pisoni, D. B. (1979). Some measures of intelligibility and comprehension. In *Research on Speech Perception Report No.5* (pp. 3-47). Bloomington, IN: Indiana University.

Podgorny, P. (1985). Telephone as computer terminal. In *The Official Proceedings of Speech Technology 1985* (pp. 103-109). New York, NY: Media Dimensions.

Rosson, M. B. (1985). *Listener training for speech- output.* (Research Report RC11029 #49529). Yorktown Heights, NY: IBM Watson Research Center.

Rosson, M. B., and Cecala, A. J. (1985). *An analysis of listener's reactions to synthetic voices.* (Research Report RC11398 #51318). Yorktown Heights, NY: IBM Watson Research Center.

Rosson, M. B., and Mellen, N. M. (1985). *Behavioral issues in speech-based remote information retrieval.* (Research Report RC11028 #495287). Yorktown Heights, NY: IBM Watson Research Center.

Sanders, M. S., and McCormick, E. J. (1987). *Human factors in engineering and design.* New York: McGraw-Hill.

Schmandt, C. (1985). Voice access to an electronic mail system. In *The official Proceedings of Speech Technology 1985* (pp. 89-91). New York, NY: Media Publications.

Schwab, E. C., Nusbaum, H. C., and Pisoni, D. B. (1985). Some effects of training on perception of synthetic speech. *Human Factors, 27,* 395-408.

Simpson, C. A., and Marchionda-Frost, K. (1984). Synthesized speech rate and pitch effects on intelligibility of warning messages for pilots. *Human Factors,* 26, 509-517.

Slowiaczek, L. M., and Nusbaum, H. C. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors, 27,* 701-712.

Waterworth, J. A. (1983). Effect of intonation form and pause durations of automatic telephone number announcements on subjective preference and memory performance. *Applied Ergonomics, 14,* 39-42.

Waterworth, J. , and Lo, A. (1984). Examples of an experiment: Evaluating some synthesizers for public announcements. In A. Monk (Ed.), *Fundamental of human-computer interaction* (pp. 89-102). London:Academic Press.

Witten, I. H. (1982). Driving the Vortrax speech synthesizer from a wide phonetic transcription with high- level prosodic markers. *International Journal of Man-Machine Studies*, 16, 393-403.

Witten, I. H. , and Mandams, P. H. C. (1977). The telephone enquiry service: A man-machine system using synthetic speech. *International Journal of Man- Machine Studies*, 9, 449-464.

The purpose of this study is to have potential users assist in the development of the design of a telephone inquiry system. The use of user input in the design process is essential to insure a good design with high acceptability. Your rating of the features in this questionnaire will help determine the attributes that users rate as important and desire in a telephone inquiry system. Your ratings will also be used to determine which features require more research before inclusion in the design of a telephone inquiry system.

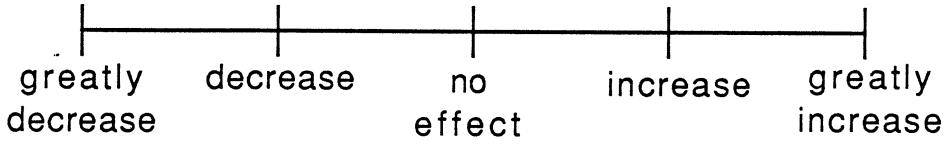
The following is a list of the features in the design of a telephone inquiry system similar to the one just demonstrated. The features include items such as the voice quality, style of text, structure of the data base, and commands. Each feature has a short description including the manner in which it was implemented in the demonstration. You are to rate how you believe the feature will affect your speed, accuracy, and acceptability in using the system.

Speed is the time in which it takes you to find information using the system.

Accuracy is the ability to find information without searching in the wrong part of the data base.

Acceptability is your willingness to use the system.

On each scale you will rate the degree to which you agree that the feature will affect speed, accuracy, and acceptability. Each scale will have 5 options (greatly decrease, decrease, no effect, increase, greatly increase) from which you will select. Select an option by circling the response.

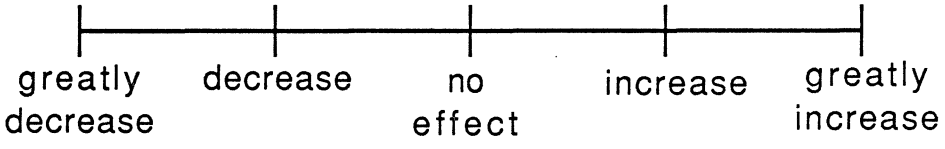


You will also be asked some questions about the features that require a short answer.

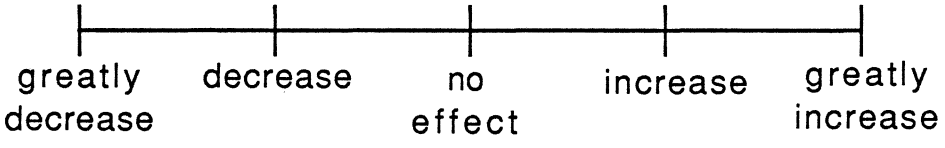
Example

Consider the bogus feature of having the system sing all the information. Normally spoken text would be sung in the opearatic style. If you feel this feature would (1) increase speed, (2) have no effect on your accuracy, but (3) greatly decrease the acceptability of the system, your answer would look like the one below.

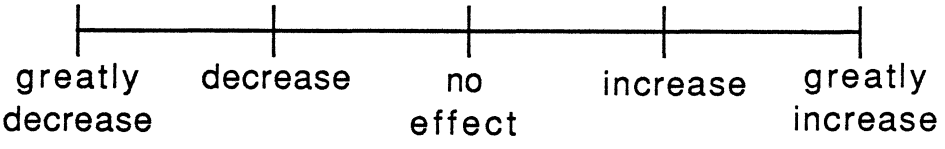
Speed at locating information



Accuracy at locating information



Acceptability of the information system

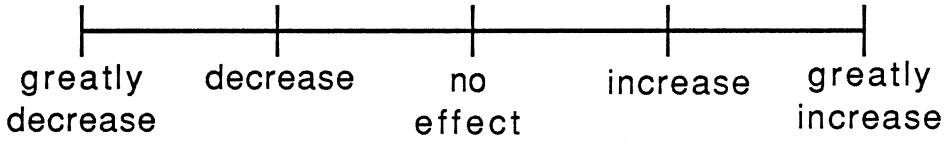


If you have any questions, please ask the experimenter now. If not, please turn the page and begin.

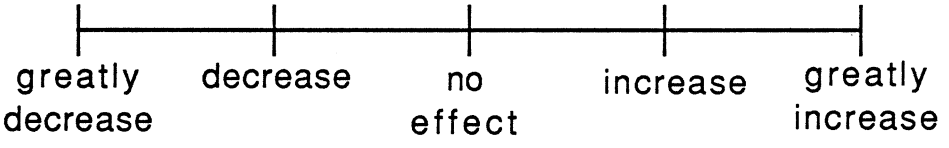
Overall speech rate

Overall speech rate describes how fast the system speaks. This is achieved by changing the length of the pause between words. Currently the speech rate is set at 180 words per minute (wpm). How would increasing overall speech rate affect your:

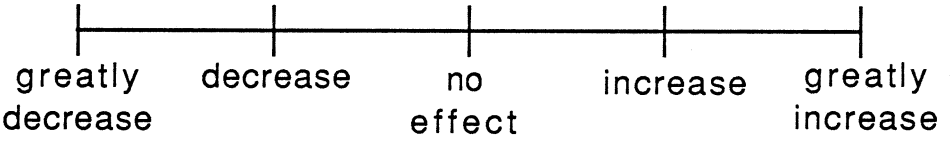
Speed at locating information



Accuracy at locating information



Acceptability of the information system



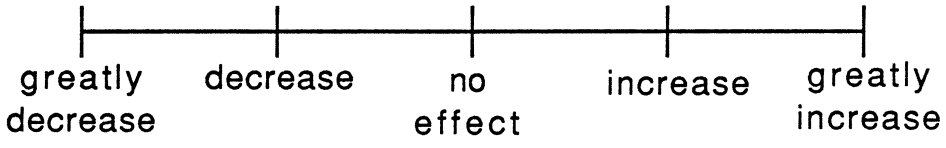
Overall speech rate is presently set at 180 wpm, but can be set from 160 wpm to 250 wpm. Where would you set the overall speech rate?

_____

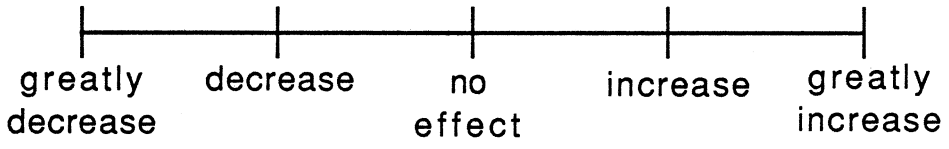
Pauses between phrases

Similar to overall speech rate, the pause between phrases describes the length of time, in seconds, the system waits after a comma. Currently the pause between phrases is set at 0.16 seconds. How would increasing the pause between phrases affect your:

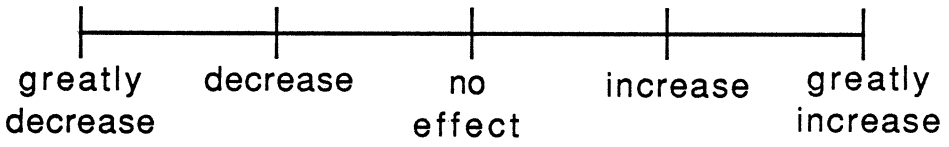
Speed at locating information



Accuracy at locating information



Acceptability of the information system

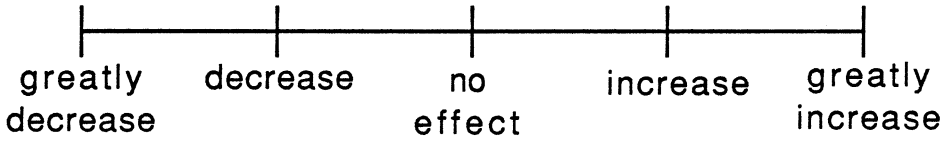


The pause between phrases is presently set at 0.16 seconds and can be increased to 0.41 seconds. Where would you set the pause between phrases? _____

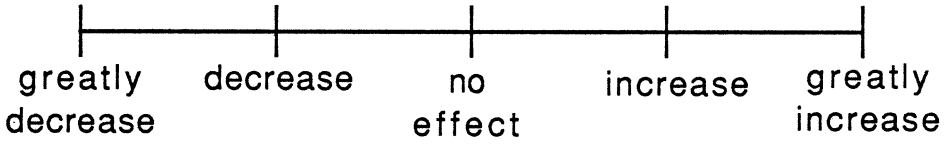
Pauses between sentences

This feature describes the amount of time the system waits after a period at the end of a sentence has been spoken. Currently the pause between sentences is set at 0.64 seconds. How would increasing the pause between sentences affect your:

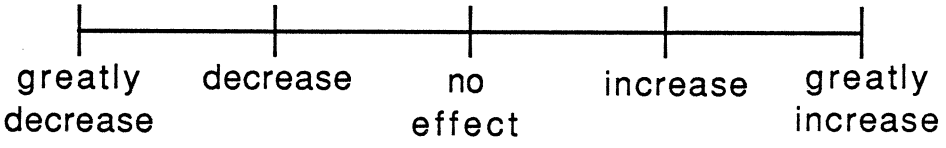
Speed at locating information



Accuracy at locating information



Acceptability of the information system

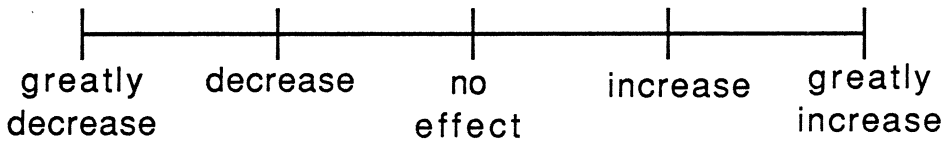


The pause between sentences is presently set at 0 .64 seconds. and can be increased to 2.6 seconds. Where would you set the pause between sentences? _____

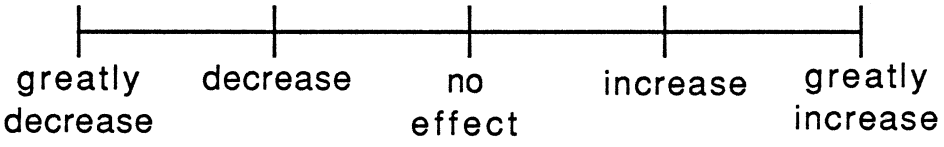
Mean pitch

Mean pitch is the average frequency of the system's speaking voice. This feature determines the average pitch of the speaking voice. The current voice is a male with a mean pitch of 180 Hertz (Hz). How would increasing the mean pitch of the voice affect your:

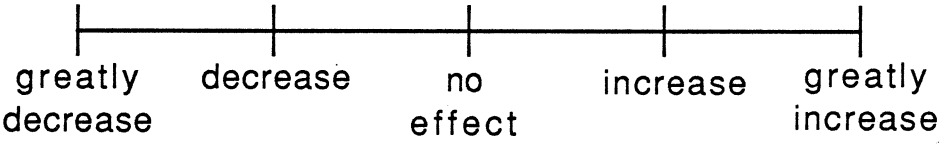
Speed at locating information



Accuracy at locating information



Acceptability of the information system

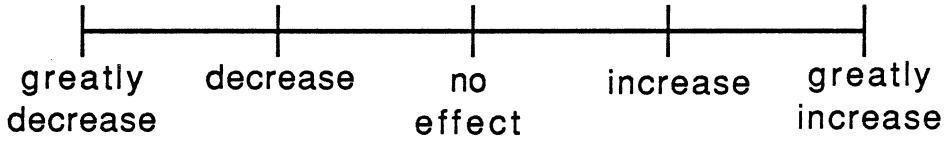


The current voice (Perfect Paul) is a male with a mean pitch of 180 (Hz). The mean pitch can be set from 30 Hz to 300. Where would you set the pitch? _____

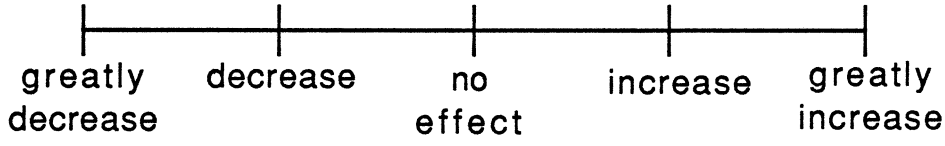
Range of pitch

The range of pitch describes the upper and lower frequencies of the voice. In other words, the highest and lowest pitch of the voice. Currently the range is set at 100%. How would decreasing the range affect your:

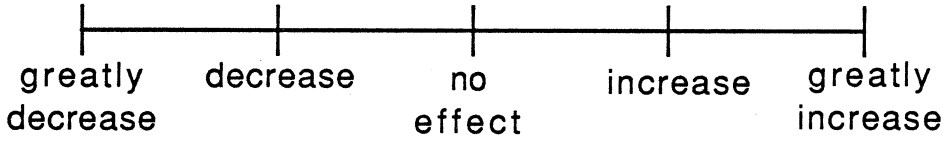
Speed at locating information



Accuracy at locating information



Acceptability of the information system



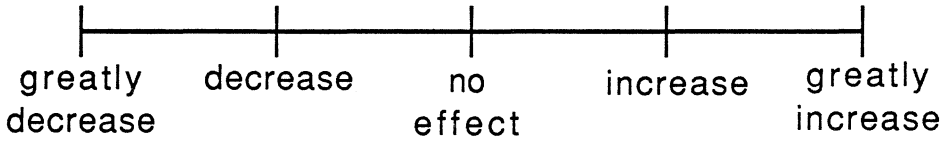
Currently the range of pitch is set at 100%, and can be moved down to 0%. Where would you set the range of pitch?

_____

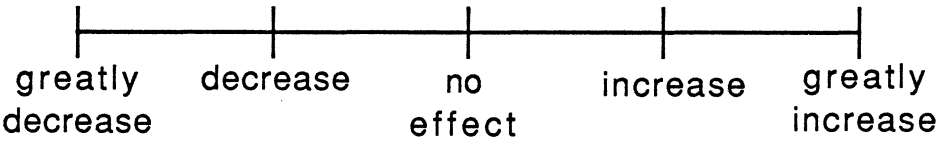
Smoothness

As the name implies, smoothness adjusts the amount of energy a voice appears to have. Increasing smoothness results in a softer, whispering voice, whereas decreasing it results in a more brilliant voice. Currently the voice is set at 34% of possible smoothness. How would increasing the smoothness feature affect your:

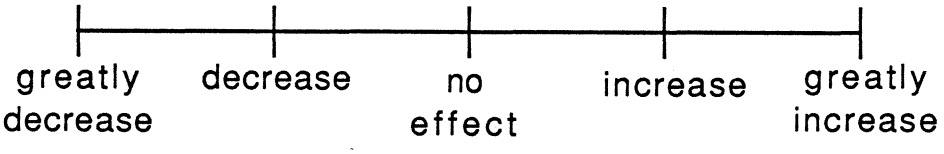
Speed at locating information



Accuracy at locating information



Acceptability of the information system

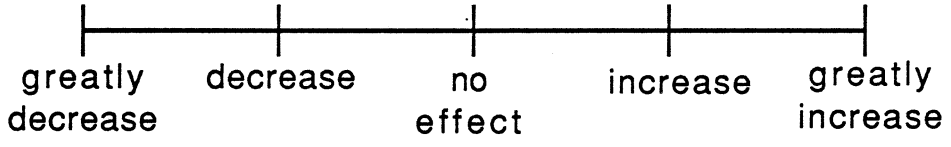


Currently smoothness is set at 34% out of 100%. Where would you set smoothness? _____

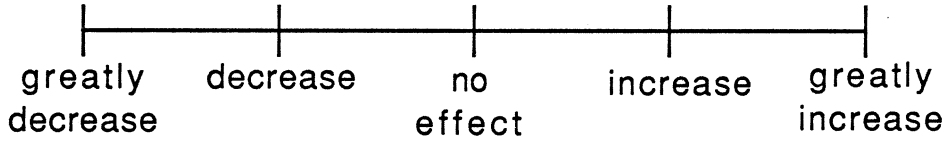
Assertiveness

Assertiveness describes the amount of authority the voice contains. The voice in the current system has 100% of the available assertiveness. How would decreasing the assertiveness affect your:

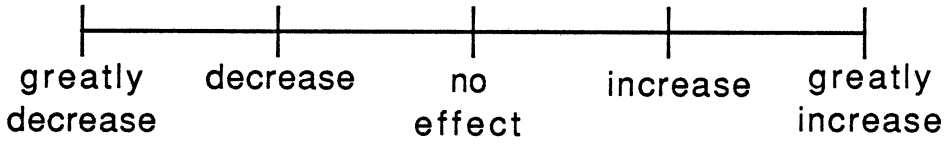
Speed at locating information



Accuracy at locating information



Acceptability of the information system

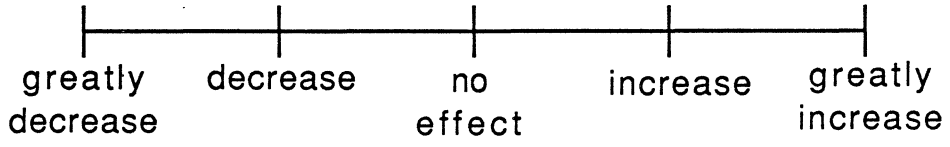


Currently assertiveness is set at 100%. Where would you set assertiveness? _____

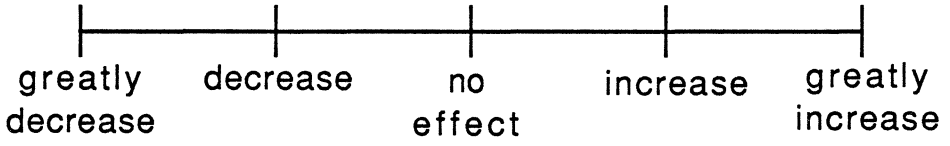
Breathiness

When some people speak, the vocal cords vibrate such that some breath noise escapes. Whispering is an example of much breathiness, and yelling is an example of little. The voice in the current system has no breathiness. How would increasing breathiness affect your:

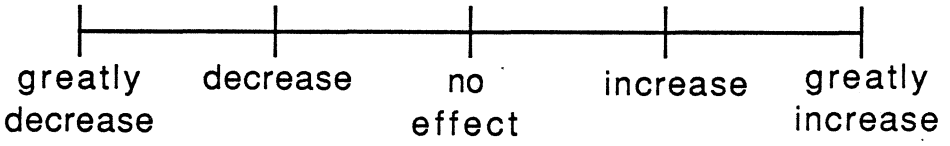
Speed at locating information



Accuracy at locating information



Acceptability of the information system

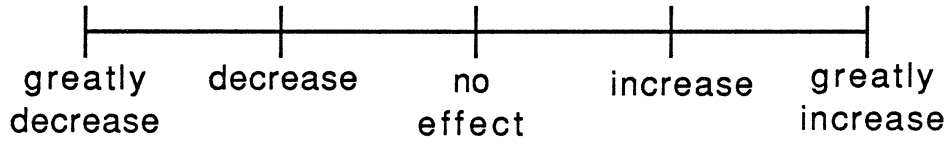


Breathiness is measured in decibels (dB). The current system adds no breathiness, but up to 70 dB could be added. Where would you set the breathiness. _____

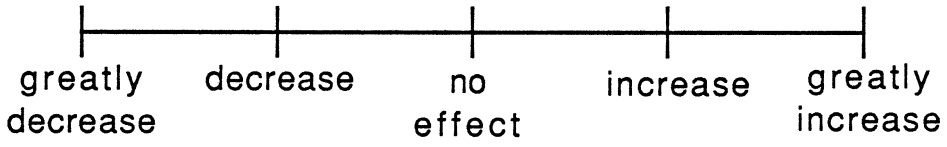
Richness

Richness is similar to brilliance and the opposite of softness. Rich voices carry well and are heard well in noisy places. Soft voices are thought of as friendly. The current voice has 20% of possible richness. How would increasing the richness affect your:

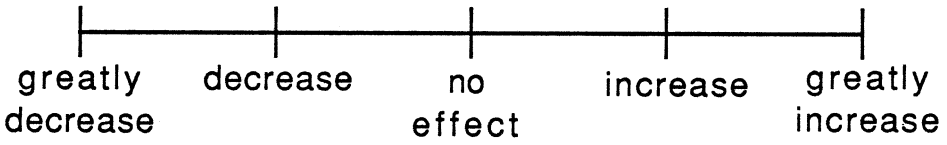
Speed at locating information



Accuracy at locating information



Acceptability of the information system

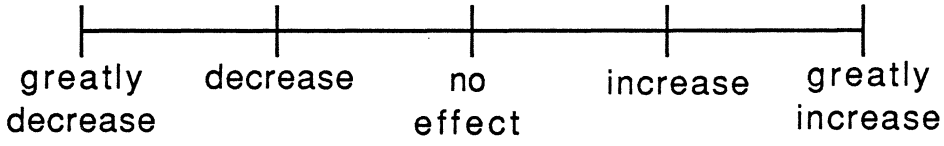


Currently the system is set at 20% out of 100% of possible richness. Where would you set the richness? _____

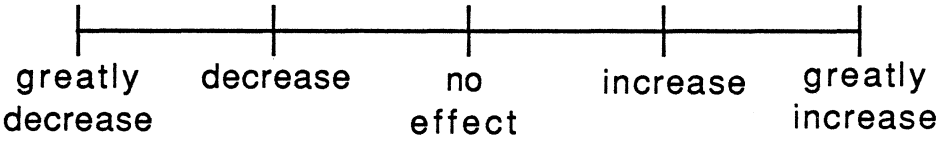
Head size/Resonance

The size of a person's head influences the sound of their voice. Increasing the size of the head increases the resonance thus making the voice sound "bigger" or "fuller". Likewise a smaller head size decreases resonance and makes the voice sound "smaller" or "thinner". Head size is described in the percent larger/smaller than the "normal" size head. The current voice has a normal size head. How would increasing the head size affect your:

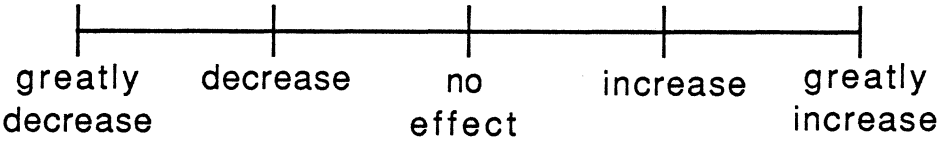
Speed at locating information



Accuracy at locating information



Acceptability of the information system

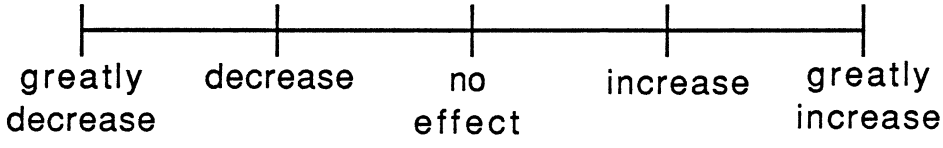


Head size can be varied from 1/2 the size of normal to 2 times normal size. Where would you set the head size? _____

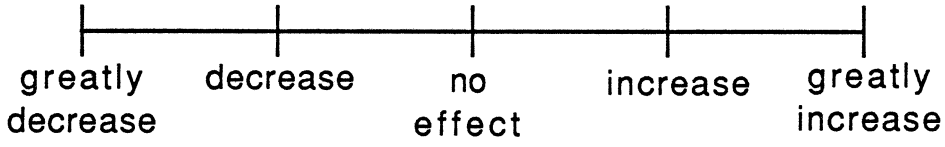
Gain of frication source

The gain of frication source is the degree to which sounds are made by forcing breath through the mouth. The more gain the more force the voice has for sounds made in the mouth. The current voice has 73 dB out of 80 dB. How would increasing the gain of the frication source affect your:

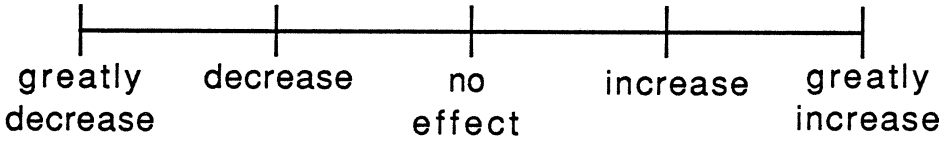
Speed at locating information



Accuracy at locating information



Acceptability of the information system

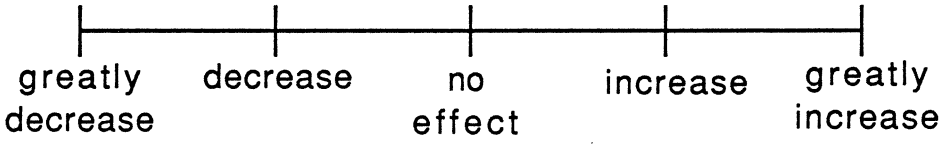


Currently the frication is set at 73 dB. Where would you set the gain of the frication source? _____

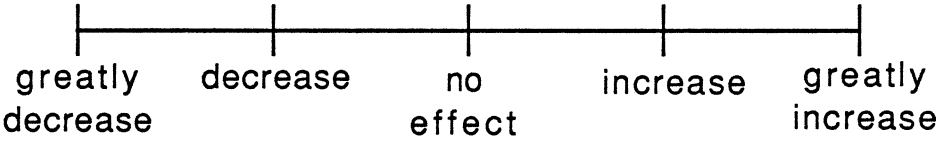
Gain of aspiration source

To aspirate is the breathy emphasis on consonants such as h, p, t, or k. The gain of the aspiration is the force of this component of speech. Currently the gain of aspiration source is set at 70 dB out of 80 dB. How would increasing the gain of the aspiration source affect your:

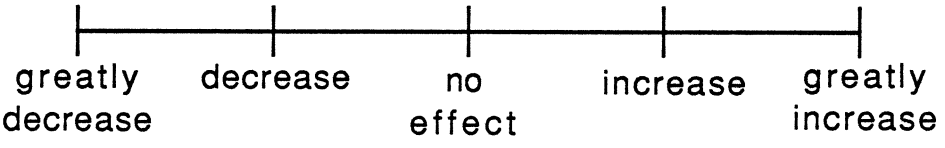
Speed at locating information



Accuracy at locating information



Acceptability of the information system

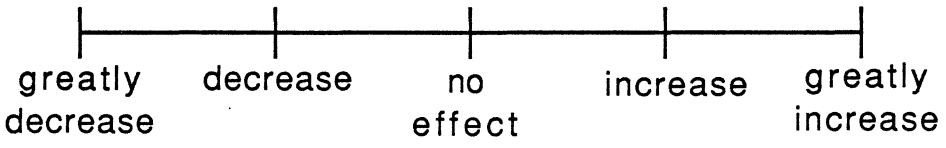


Currently the gain of aspiration source is set at 70 dB. Where would you set the gain of aspiration source? _____

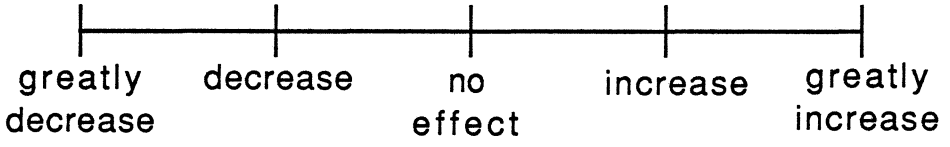
Gain of the nasal resonator

Nasal resonation is the degree to which the voice appears "to speak through the nose." Currently the gain of the nasal resonator is set at 69 dB out of 80 dB. How would increasing the gain of the nasal resonator affect your:

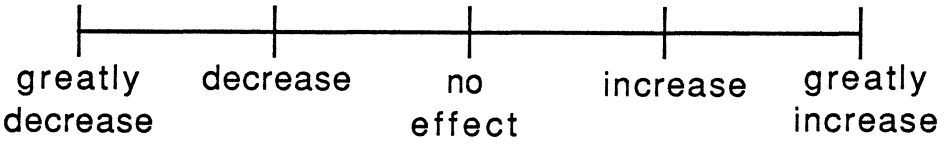
Speed at locating information



Accuracy at locating information



Acceptability of the information system

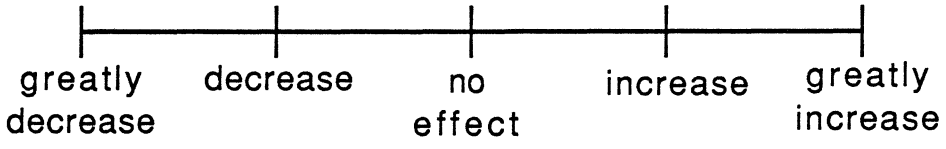


Currently the gain of the nasal resonator is set at 69 dB out of 80 dB. Where would you set the gain of the nasal resonator? _____

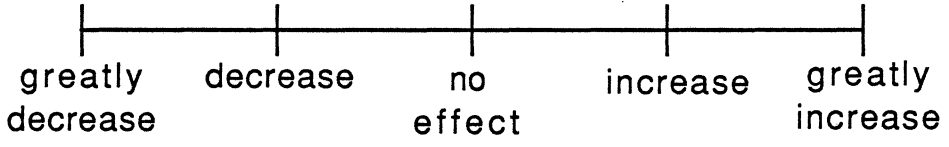
Gain of the voicing source

The force of the speech is referred to as the gain of the voicing source. Currently the gain of the voicing source is set at 71 dB out of 80 dB. How would increasing the gain of the voicing source affect your:

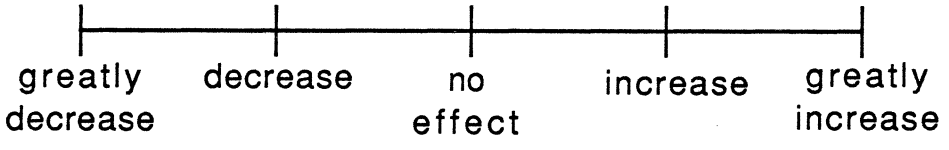
Speed at locating information



Accuracy at locating information



Acceptability of the information system

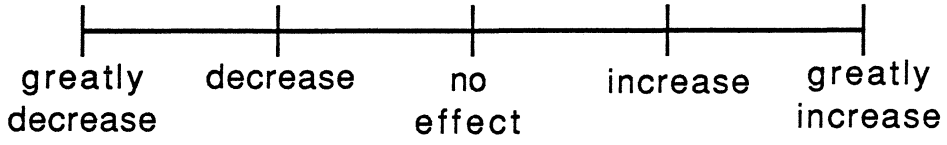


Currently the gain of the voicing source is set at 71 dB. Where would you set the gain of the voicing source? _____

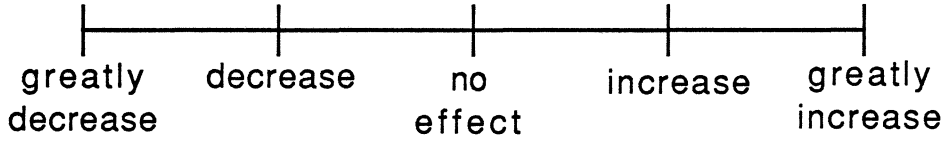
Age

The age of the voice can be varied from a young child to a mature adult. The current voice sounds like an adult. How would using a child's voice affect your:

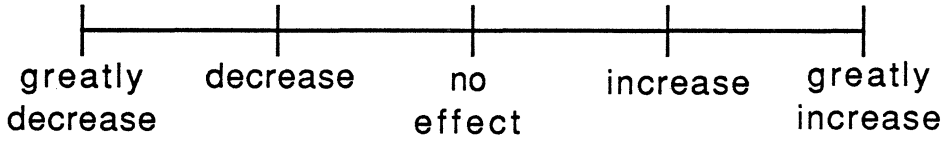
Speed at locating information



Accuracy at locating information



Acceptability of the information system

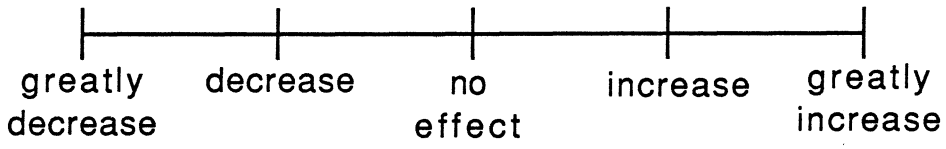


From young child to mature adult, where would you set the age of the voice? _____

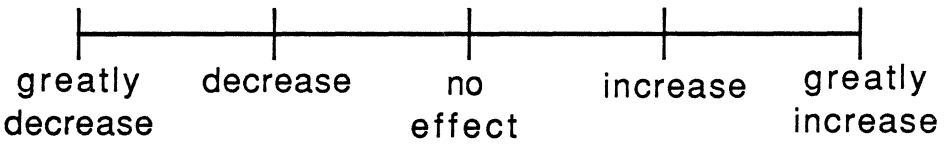
Sex

The voice can be either male or female. How would changing the voice to a female voice affect your:

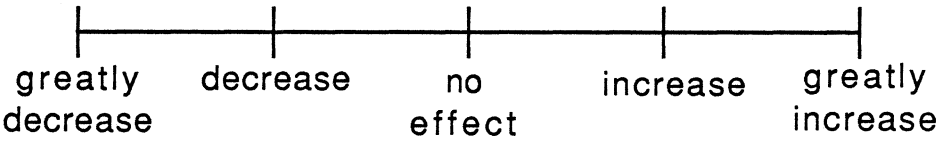
Speed at locating information



Accuracy at locating information



Acceptability of the information system

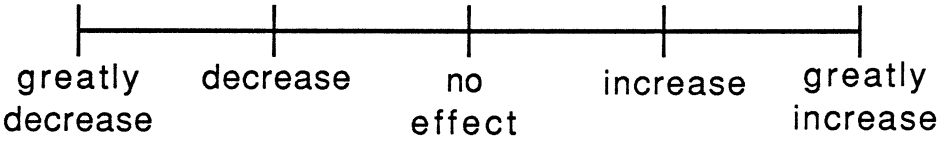


Would you prefer a male or a female voice? _____

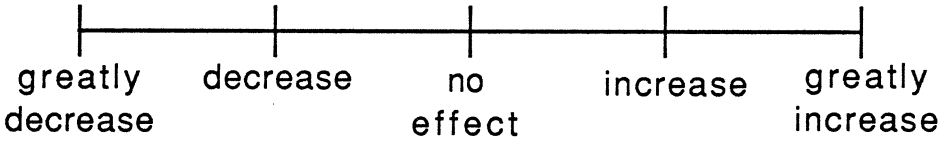
Exception dictionary

The speech synthesizer you heard uses rules to determine how to pronounce words. Sometimes the rules yield drastically mispronounced words. An exception dictionary can be used to provide an alternate pronunciation for each of these mispronounced words. Currently a limited version of this feature is functional. How would inserting more words in the exception dictionary affect your:

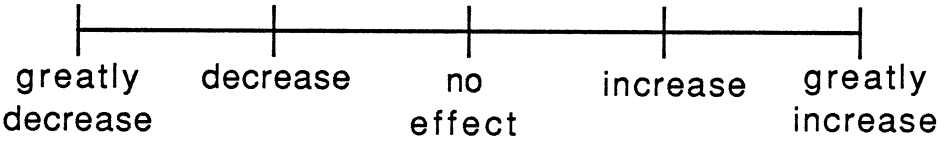
Speed at locating information



Accuracy at locating information



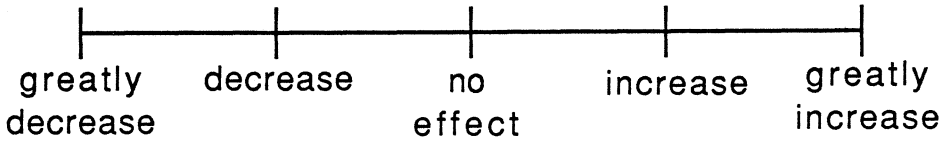
Acceptability of the information system



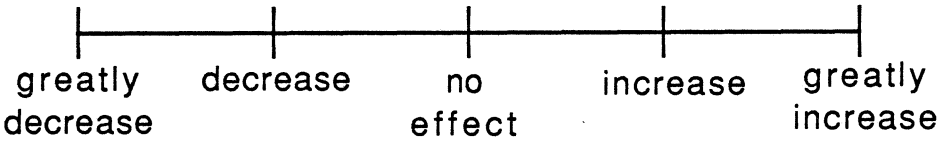
Size of vocabulary

The number of different words spoken by the system can be controlled. Currently the size of the vocabulary is approximately 500 words. How would increasing the size of the vocabulary used in the system affect your:

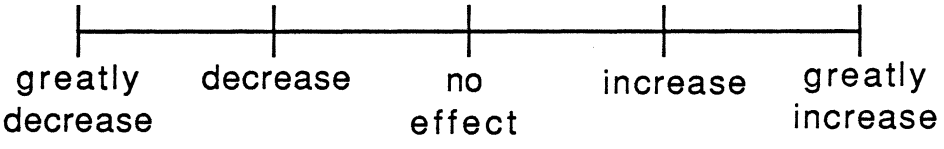
Speed at locating information



Accuracy at locating information



Acceptability of the information system

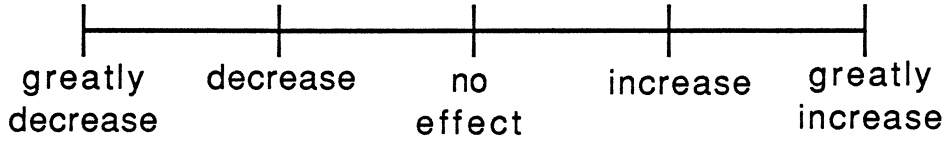


The English language has approximately 10 million words, and the average person uses about 10,000 words. Where would you set the the size of the vocabulary used by the system?

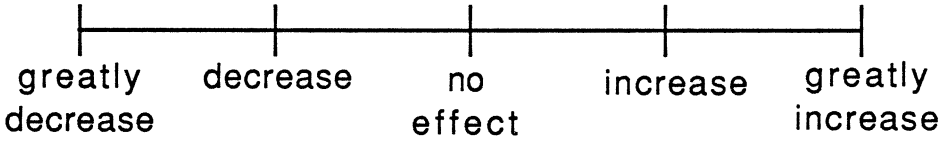
Familiarity of words

Words can be rated as to their familiarity in everyday English. How would increasing the familiarity of the words used in the system affect your:

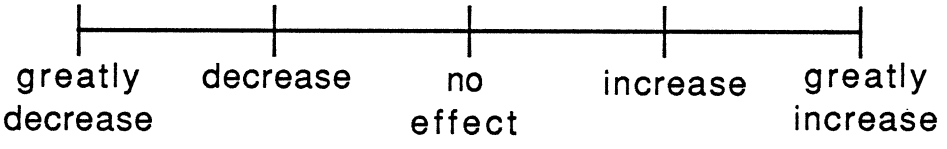
Speed at locating information



Accuracy at locating information



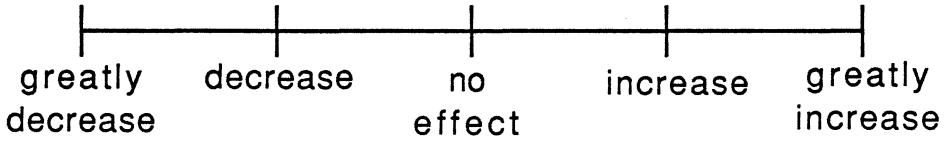
Acceptability of the information system



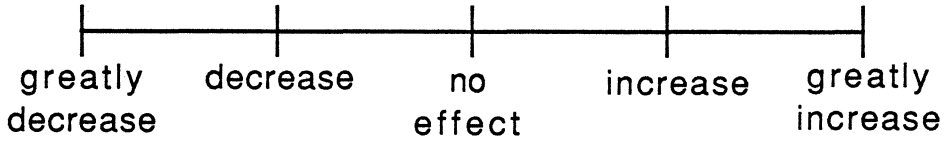
Length of words

The length of a word is measured in number of syllables. Currently there is no restriction on the number of syllables in the words used. How would controlling the length of words used in the system affect your:

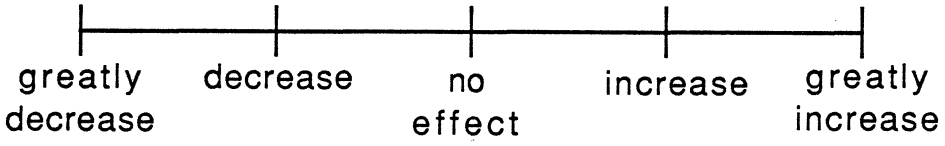
Speed at locating information



Accuracy at locating information



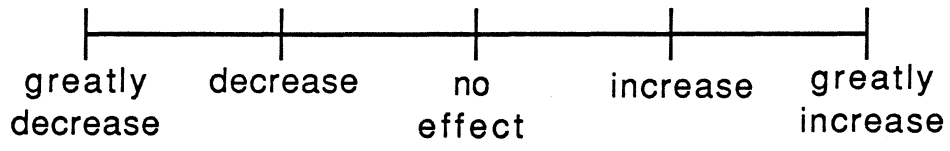
Acceptability of the information system



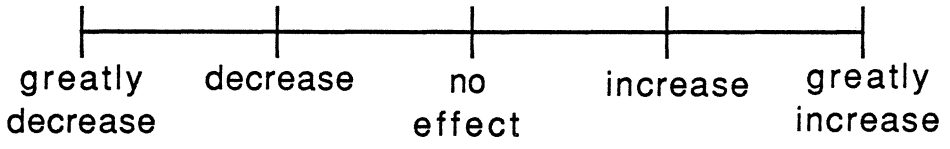
Use of jargon

Some groups of people make extensive use of specialized words known as jargon. These words hold meaning unique to that group alone. How would the inclusion of jargon familiar to you in the system affect your:

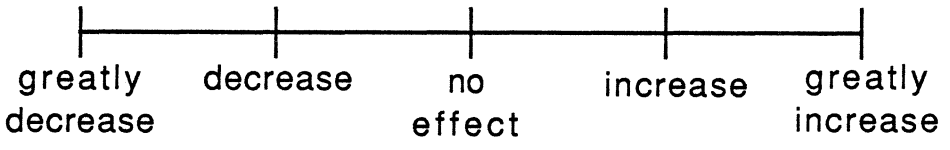
Speed at locating information



Accuracy at locating information



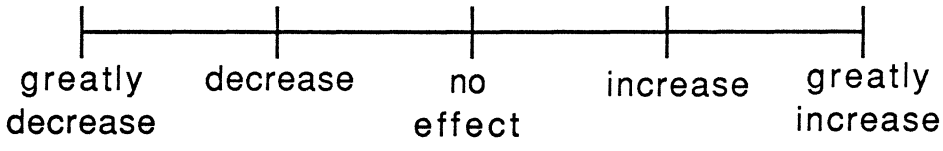
Acceptability of the information system



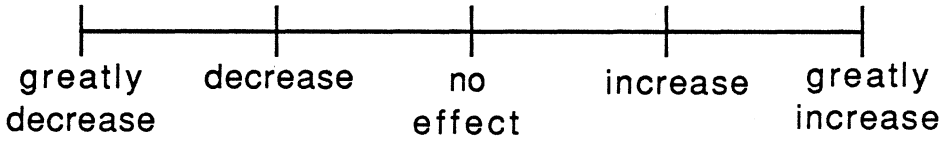
Active -vs- passive voice

In active-voice sentences, the subject takes action. In passive-voice, the subject receives the action of the verb. "Jim hit the ball" is an active sentence, and "The ball was hit by Jim" is the passive-voice counterpart. Active-voice is used in the current system. How would changing to passive-voice affect your:

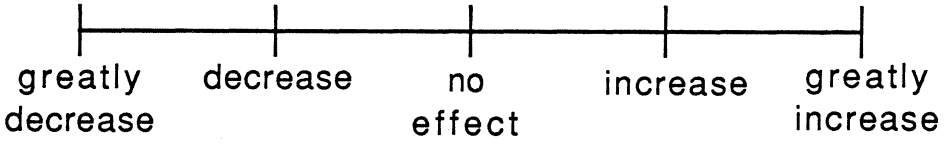
Speed at locating information



Accuracy at locating information



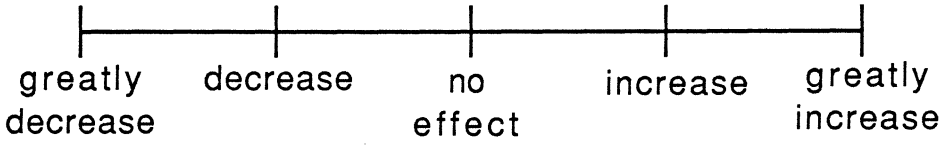
Acceptability of the information system



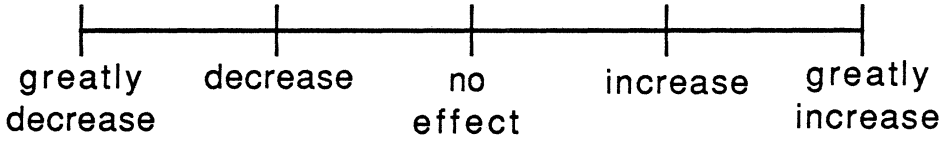
Simple -vs- complex sentences

Simple sentences consist of a subject and verb with very few phrases attached. Complex sentences are composed of multiple phrases and clauses. The current system has a mixture of simple and complex sentences. How would increasing the number of complex sentences affect your:

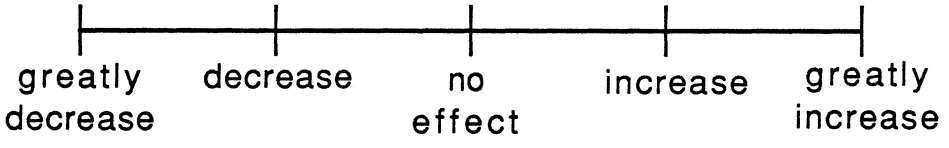
Speed at locating information



Accuracy at locating information



Acceptability of the information system

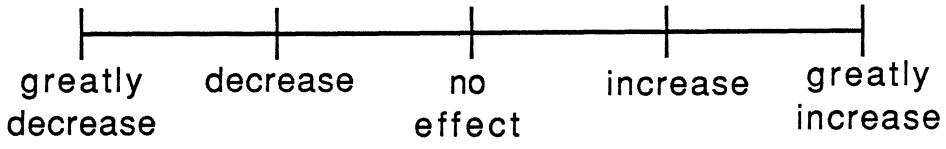


Of simple and complex sentences, which do you believe is more appropriate for this system? _____

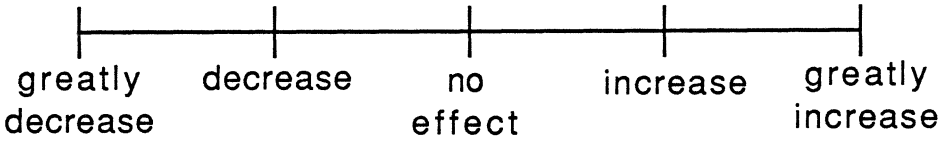
Length of the sentences

The length of sentences can be controlled. Currently there is no control of the length of sentences. How would using only short sentences (8 words or less) affect your:

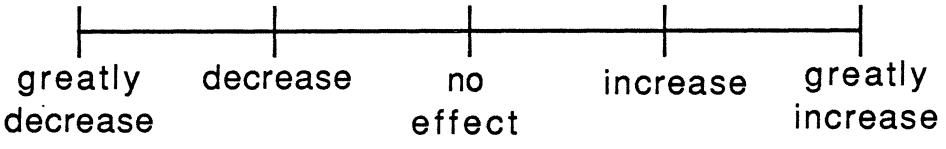
Speed at locating information



Accuracy at locating information



Acceptability of the information system



How long would you make the average sentence? _____ words

Use the components listed
in italics to construct what you believe to be the longest sentence that should be used in the system.

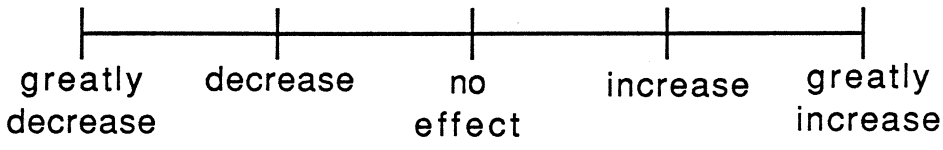
(*subject, verb phrase, object, preposition phrase, clause*)

_____

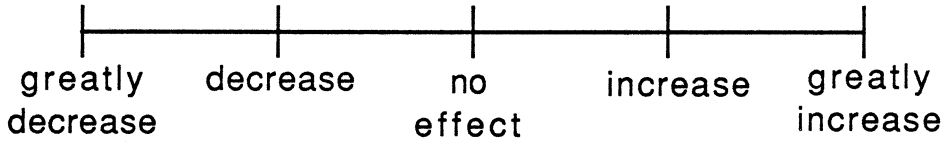
Order of information in a sentence

In longer sentences, the critical information can be ordered such that it appears at the beginning, middle, or end. Currently this is not controlled. How would placing the critical information consistently at the end of the sentence affect your:

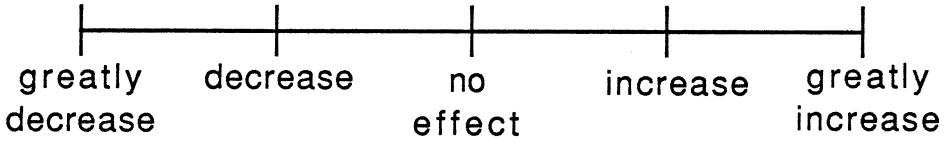
Speed at locating information



Accuracy at locating information



Acceptability of the information system

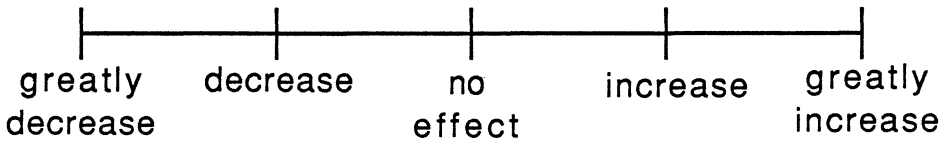


Where would you like to have the critical information placed: beginning, middle, or end? _____

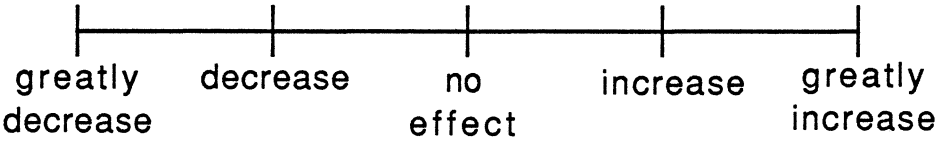
Order of information in a paragraph

The sentences in longer messages can be ordered such that the critical information appears at the beginning, middle, or end of the message. Currently this is not controlled. How would placing the critical information consistently at the end of a message affect your:

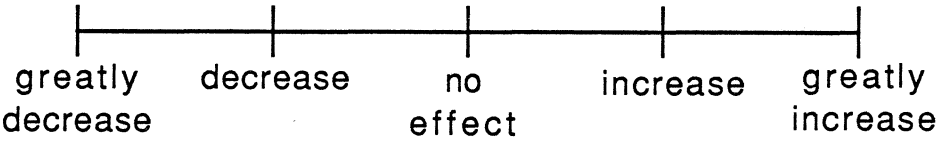
Speed at locating information



Accuracy at locating information



Acceptability of the information system

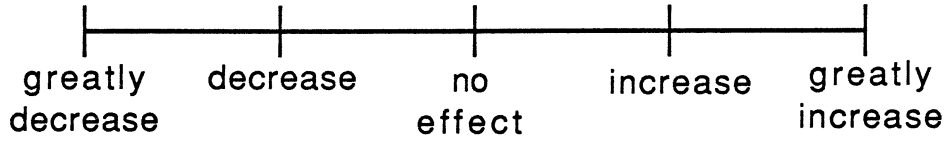


Where would you like to have the critical information placed in a message: beginning, middle, or end? _____

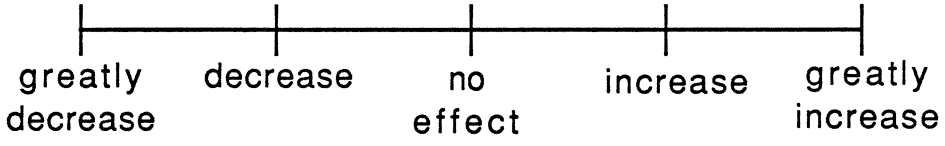
Length of messages

Messages can be made long or short int terms of the sentences. In the current system the length of messages is not controlled. How would increasing the length of messages affect your:

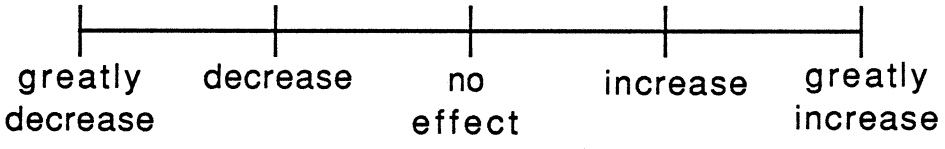
Speed at locating information



Accuracy at locating information



Acceptability of the information system



How long would you make the average message?

_____ sentences

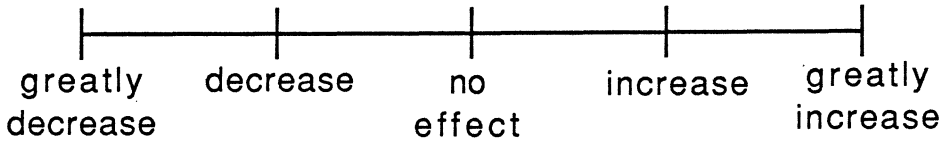
How long would you make the longest message?

_____ sentences

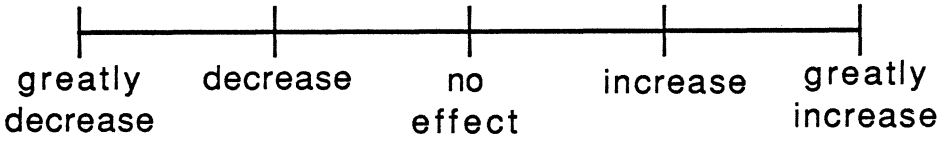
Changing the voice for different types of information

The voice can be changed to denote different types of information in the system. Keywords, messages, help, and errors could all be presented using a unique voice for each. This is not a feature in the current system. How would using unique voices for different types of information affect your:

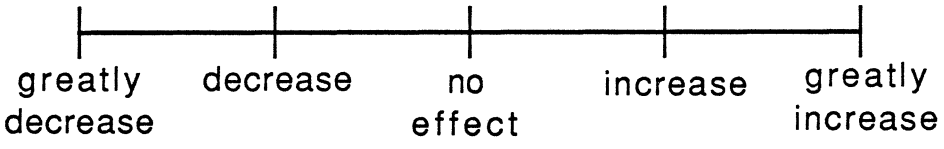
Speed at locating information



Accuracy at locating information



Acceptability of the information system



What type of voice (sex, age, mean pitch, assertiveness) would you select for:

keywords, _____

messages, _____

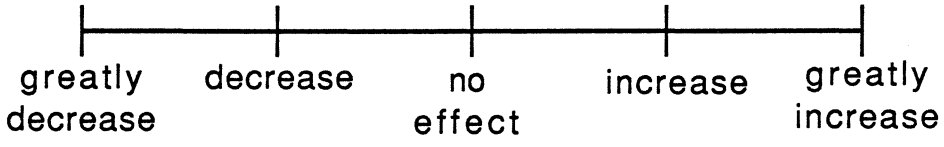
help, _____

and errors? _____

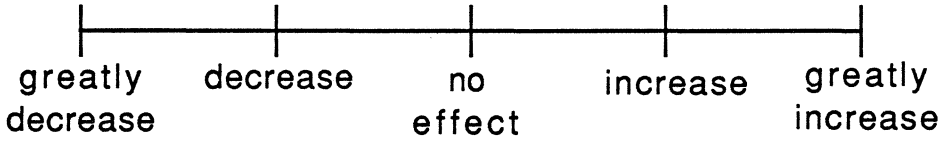
Changing speech rate

The system can be designed to change the speech rate depending on your ease in locating information. This is not a feature in the current system. How would adapting the speech rate affect your:

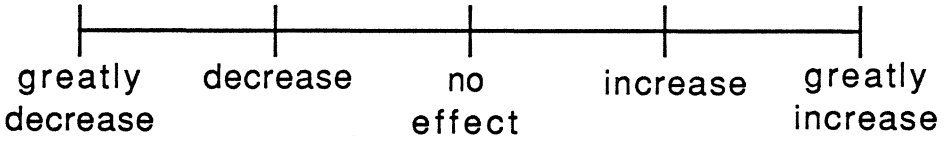
Speed at locating information



Accuracy at locating information



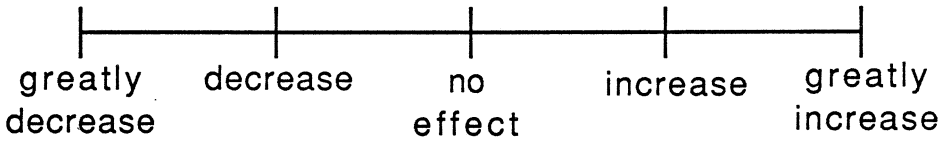
Acceptability of the information system



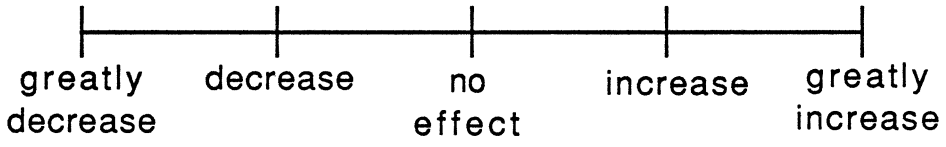
Double keying

In systems where there are more than 10 commands, or where you must type in text using the telephone keypad, two keystrokes are used to differentiate the commands and alphabet. Currently there is no need for double keying. If double keying were required to use the system, how would it affect your:

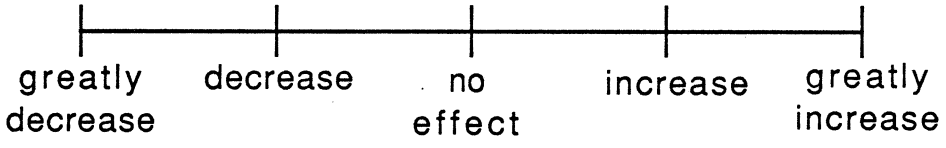
Speed at locating information



Accuracy at locating information



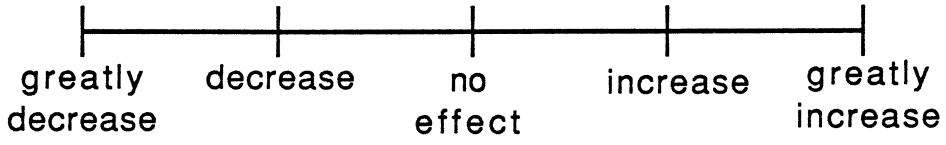
Acceptability of the information system



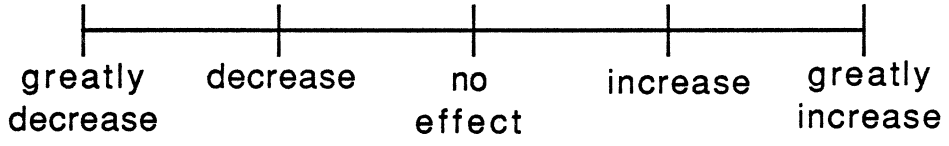
Keypress echoing

If the system speaks your keypresses this is called keypress echoing. The current systems does not use keypress echoing. How would adding keypress echoing affect your:

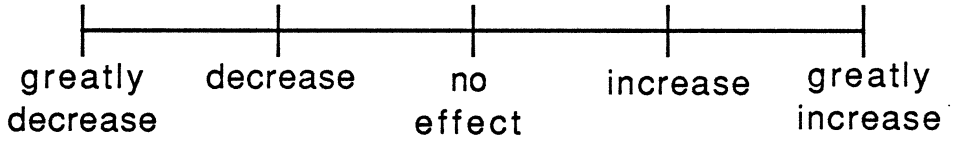
Speed at locating information



Accuracy at locating information



Acceptability of the information system



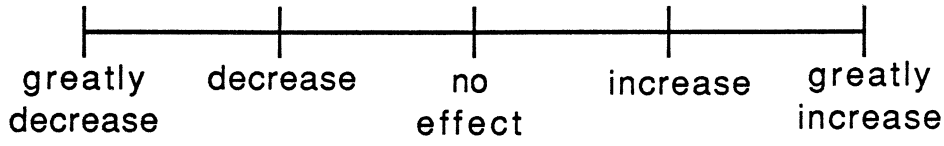
If keypress echoing were implemented, either each keypress or the command names could be spoken. Which would you prefer?

_____

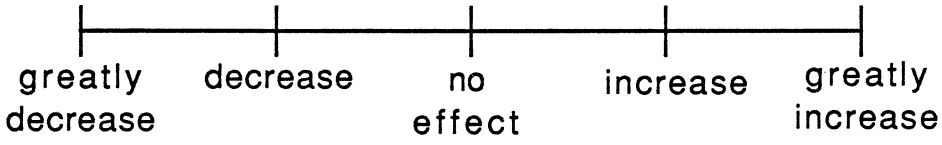
Menu Length

The number of items in a particular menu is called the menu length. The maximum length of the keyword menus in the current system is 9. How would decreasing menu length affect your:

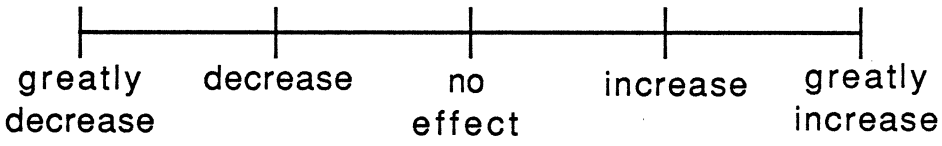
Speed at locating information



Accuracy at locating information



Acceptability of the information system



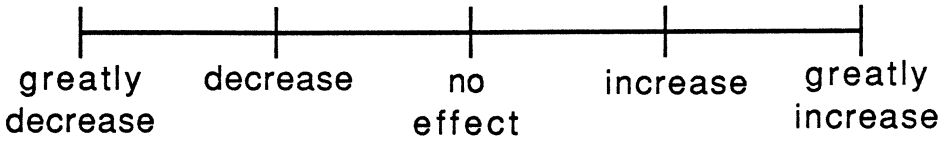
What should be the average length of the menu? _____ items

What should be the maximum length of the menu? _____ items

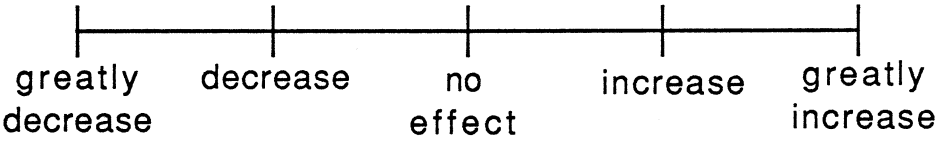
Menu depth

The number of menus you must pass through to hear a message is called the menu depth. The current system has a maximum depth of 4 menus. How increasing menu depth affect your:

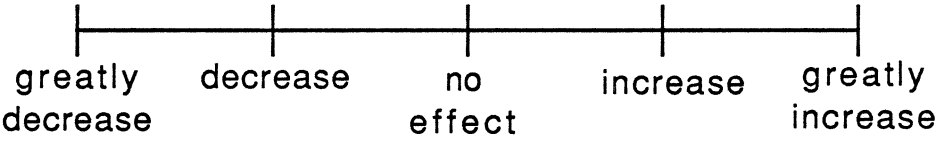
Speed at locating information



Accuracy at locating information



Acceptability of the information system



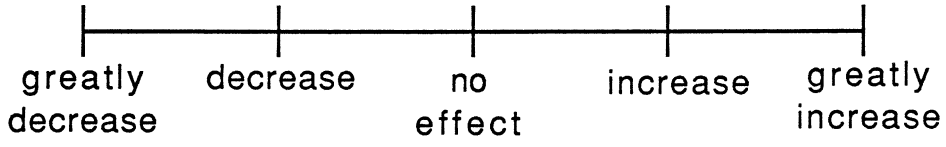
What should be the maximum menu depth? _____

What should be the optimum menu depth? _____

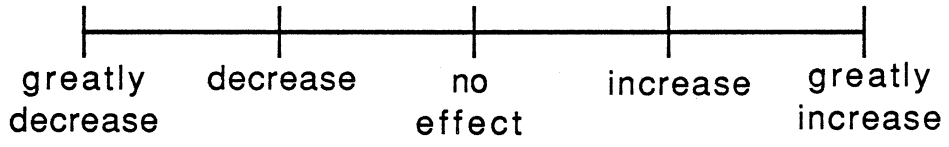
Selection method

In the current system, menu items are selected by pressing the # key during the pause immediately after the item has been spoken. Alternative systems could use numbers or letters to select an item (e.g. press 1 for movies, or press "m" for movies after the entire menu were spoken). How would the use of numbers or letters to select a menu item affect your:

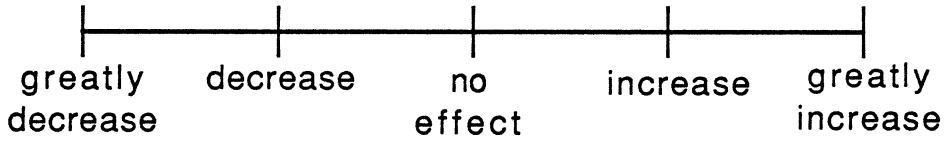
Speed at locating information



Accuracy at locating information



Acceptability of the information system



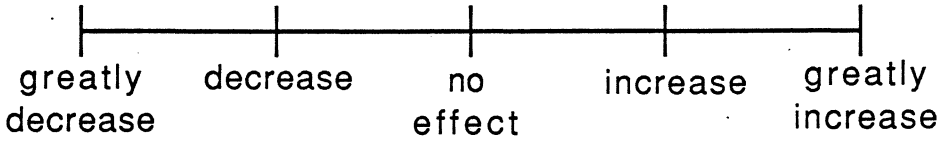
Which selection method (# key, numbers, or letters) would you select?

_____

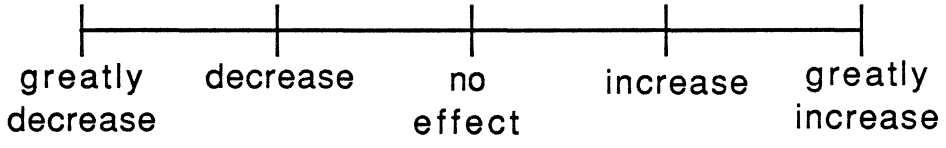
Number of commands

The current system has 4 commands (* - backup, 0 - restart, # - select, and 8 - pause). How would increasing the number of commands affect your:

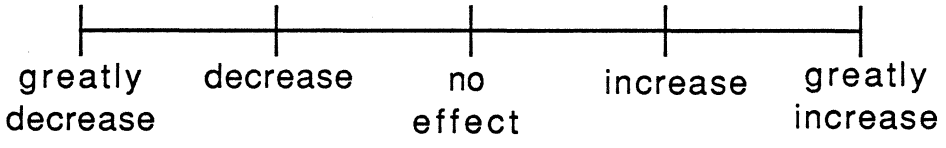
Speed at locating information



Accuracy at locating information



Acceptability of the information system



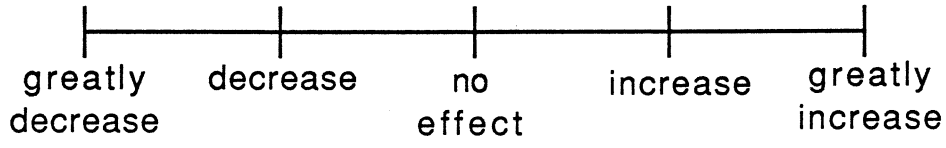
How many commands should be available? _____

What other commands are needed ? _____

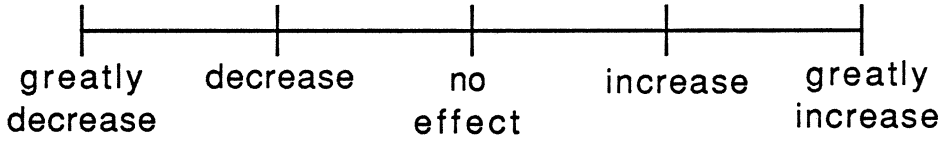
Length of commands

Command length is measured in terms of the number of keystrokes required to enter a command. The current system uses one keystroke to complete a command. How would increasing the number of keystrokes to compose a command affect your:

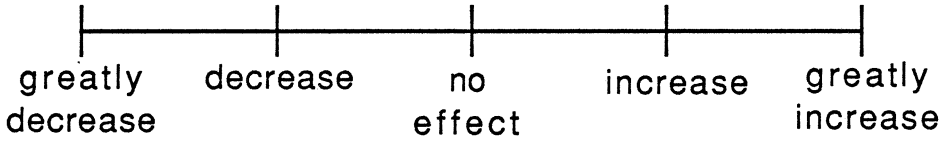
Speed at locating information



Accuracy at locating information



Acceptability of the information system



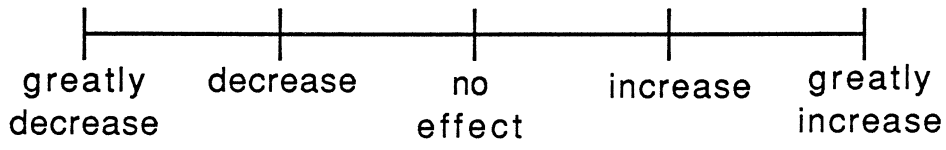
How many keystrokes should be required for each command?

_____

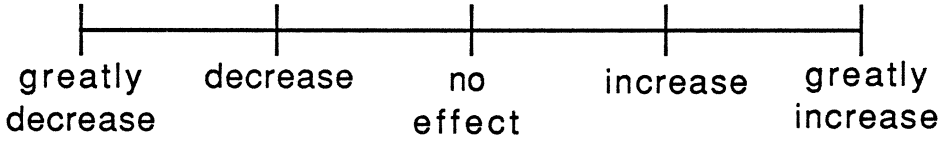
Command abbreviations

If a command name were long, it might be possible to allow for an abbreviation. Since the current system uses single keypress commands, no command abbreviations were created. In a system with longer commands, how would the use of command abbreviations affect your:

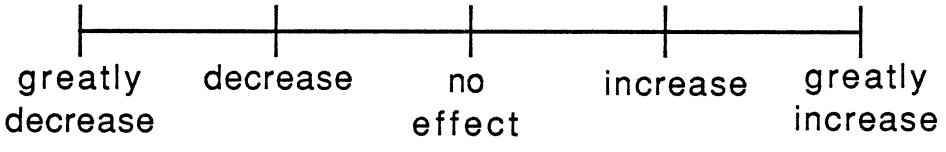
Speed at locating information



Accuracy at locating information



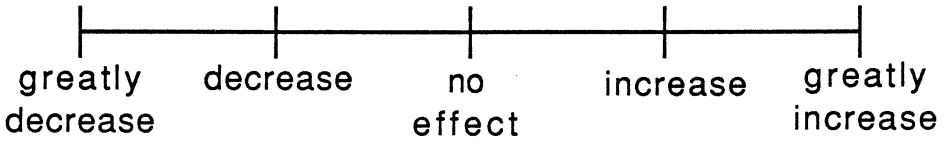
Acceptability of the information system



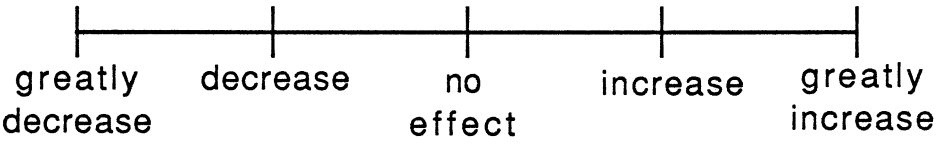
Command synonyms

If a command names were long, it might be possible to allow for a synonym. Since the current system uses single keypress commands, no command synonyms were created. In a system with longer command names how would the use of command synonyms affect your:

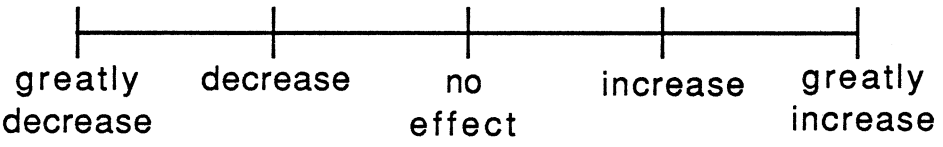
Speed at locating information



Accuracy at locating information



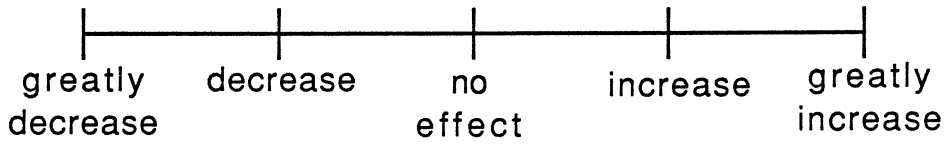
Acceptability of the information system



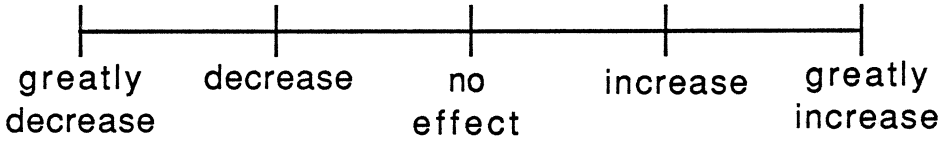
Error detection

It is possible to detect errors in user input. These errors can be detected at either the keypress level or the command level. The current system has no error detection. How would adding an error detection feature affect your:

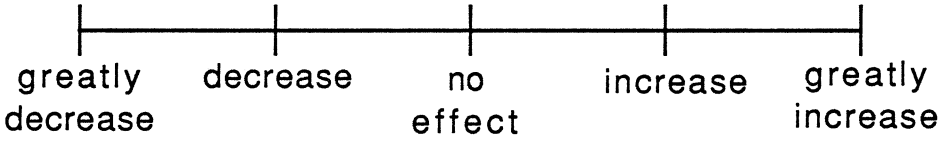
Speed at locating information



Accuracy at locating information



Acceptability of the information system



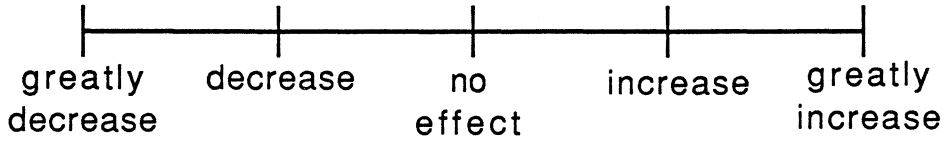
Should error detection occur at the keypress or the command level?

_____

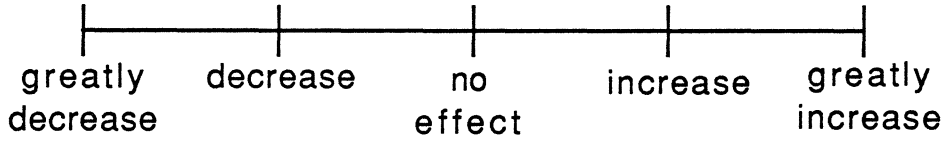
Undoing actions

It is possible to allow you to recover from errors by undoing previous keypresses. If the previous keypress had caused an error you would be able to undo the action. How would allowing you to undo actions affect your:

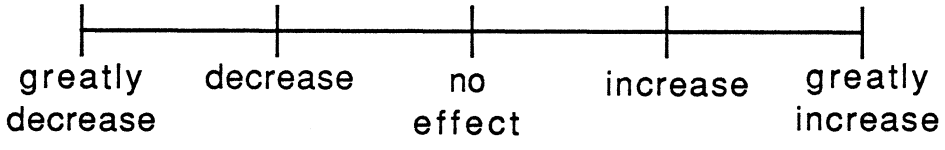
Speed at locating information



Accuracy at locating information



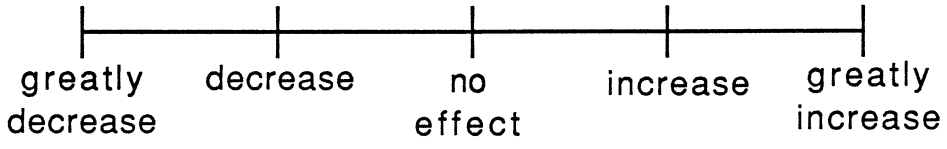
Acceptability of the information system



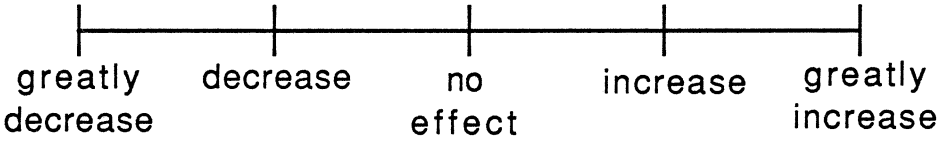
Timeouts

A timeout is the amount of time in seconds the system waits for your response before it continues. Currently the length of time the system waits for your keypress before continuing is 5 seconds. How would increasing the length of time before a timeout affect your:

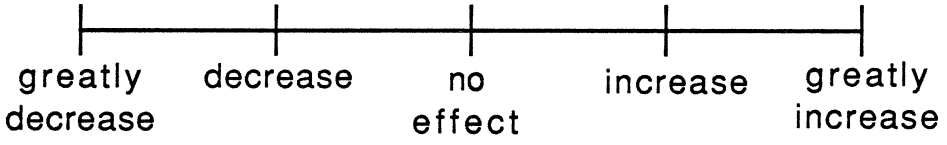
Speed at locating information



Accuracy at locating information



Acceptability of the information system



What is the optimum amount of time a timeout should be?

_____

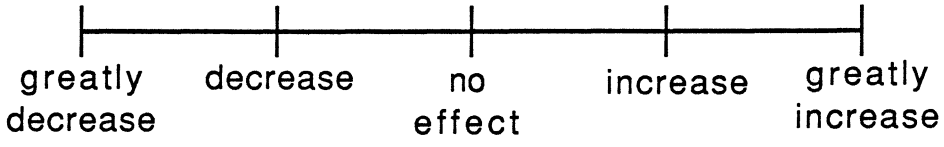
What is the maximum amount of time a timeout should be?

_____

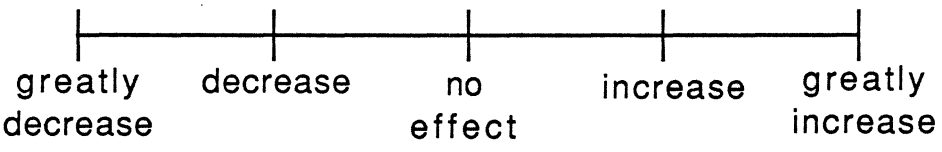
Changing phrasing

It is possible to change the phrasing of a message if you did not understand the message on the first presentation. How would providing an alternative phrasing for the second presentation of a message affect your:

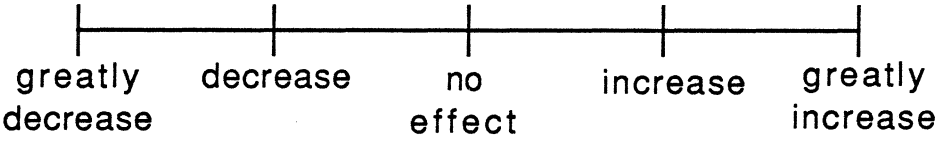
Speed at locating information



Accuracy at locating information



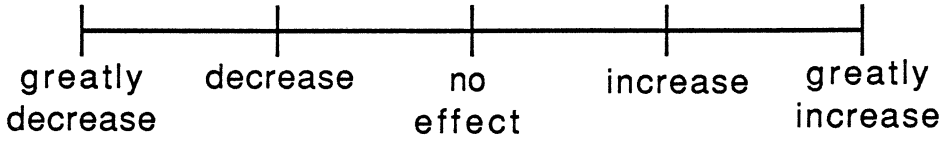
Acceptability of the information system



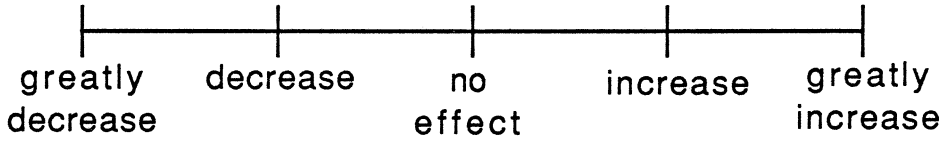
Embedded training

This feature describes the system's ability to teach you how to use the telephone inquiry system. How would the availability of embedded training affect your:

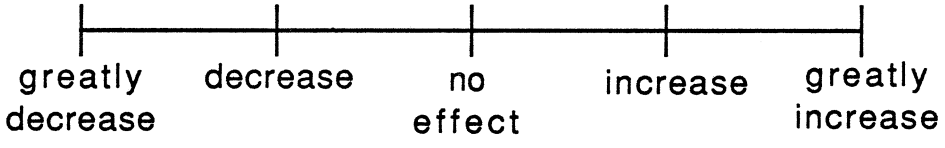
Speed at locating information



Accuracy at locating information



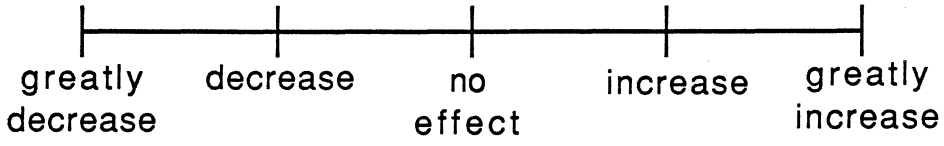
Acceptability of the information system



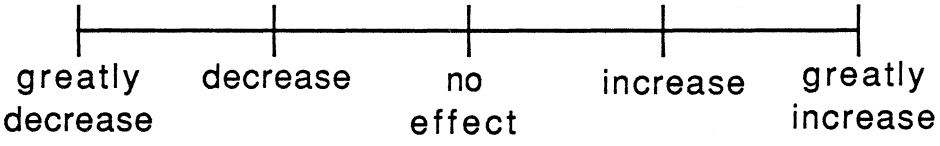
Transaction summaries

It is possible to provide an online list of your actions while using the system. This is not a feature in the current system. How would the availability an online action summary affect your:

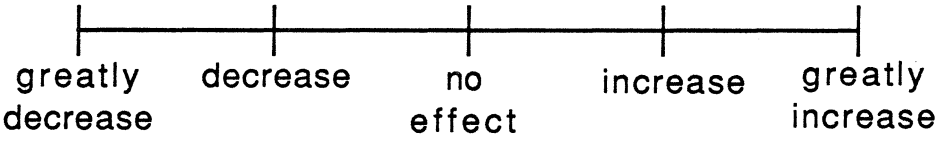
Speed at locating information



Accuracy at locating information



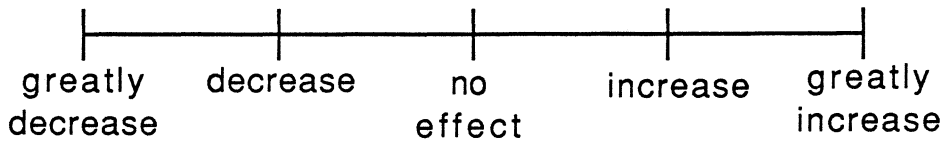
Acceptability of the information system



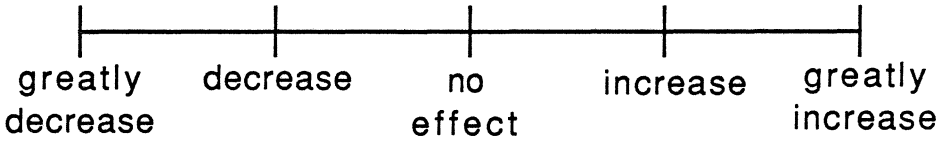
Hardcopy summaries

It is possible to provide a printed listings of your actions on the system. available at a some facility. This is not a feature of the current system. How would the availability of an hardcopy action summary affect your:

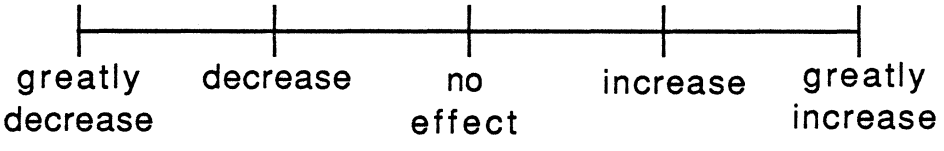
Speed at locating information



Accuracy at locating information



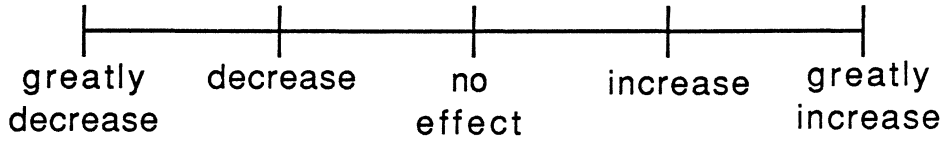
Acceptability of the information system



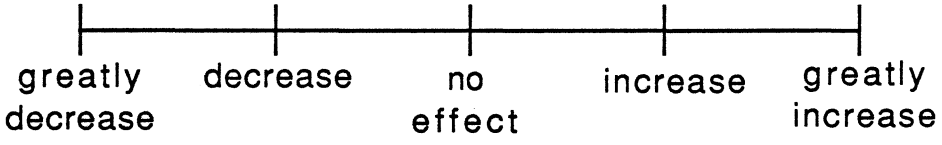
Wallet guide

It is possible to provide a wallet size instruction card to users. Such a card would contain information on available commands and the structure of the data base. A wallet guide is not availble for the current system. How would the availability of a wallet guide affect your:

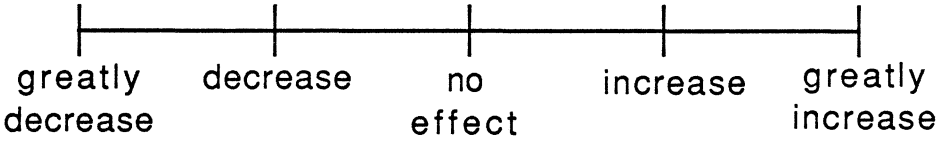
Speed at locating information



Accuracy at locating information



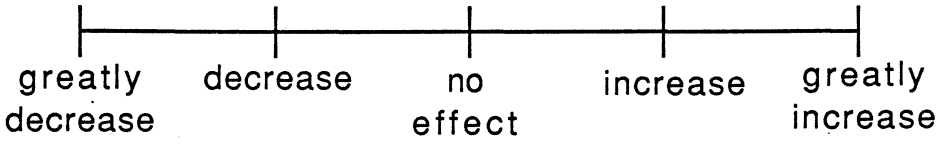
Acceptability of the information system



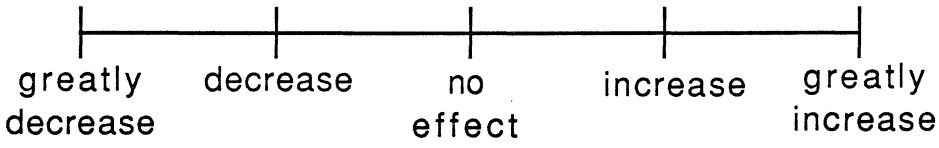
Human assistance

How would the availability of an operator to assist you affect your:

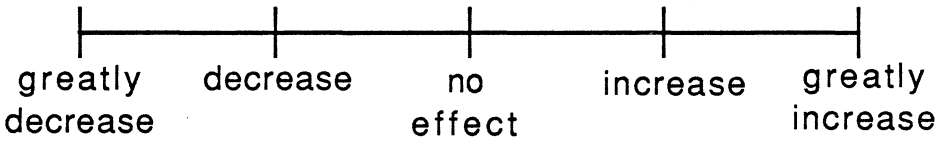
Speed at locating information



Accuracy at locating information



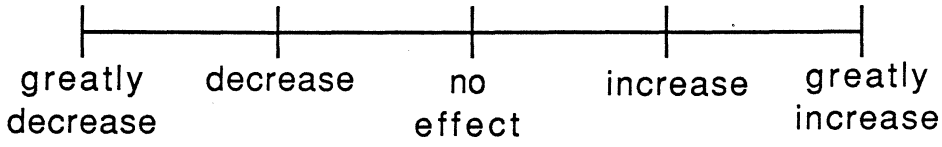
Acceptability of the information system



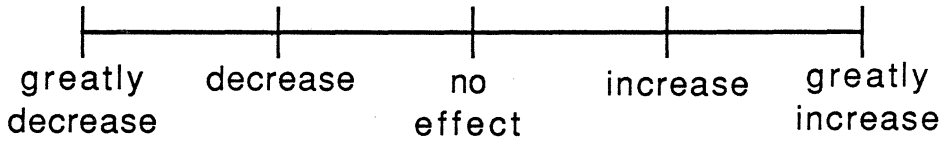
Amplitude control

How would the availability of a volume control on the the telephone affect your:

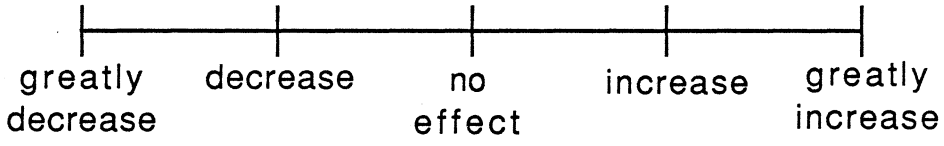
Speed at locating information



Accuracy at locating information



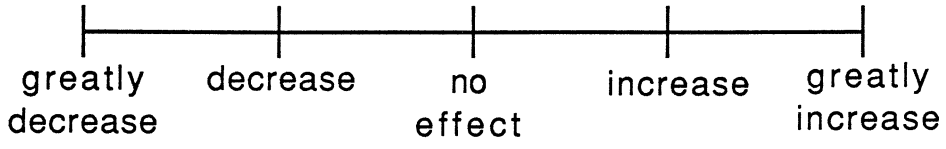
Acceptability of the information system



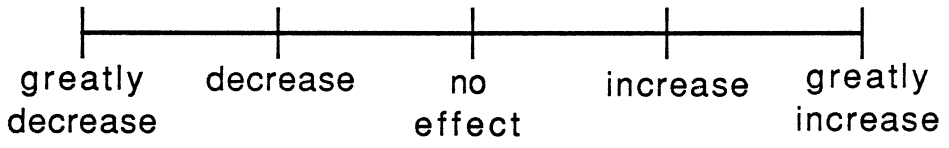
Repeat speech display

The message can be repeated multiple times. How would repeating the message affect your:

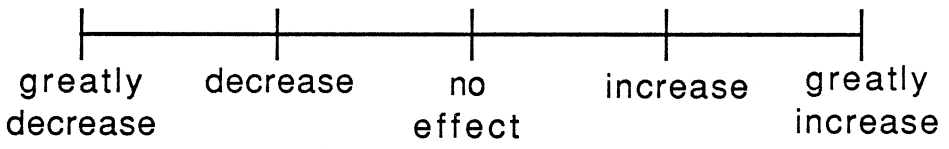
Speed at locating information



Accuracy at locating information



Acceptability of the information system



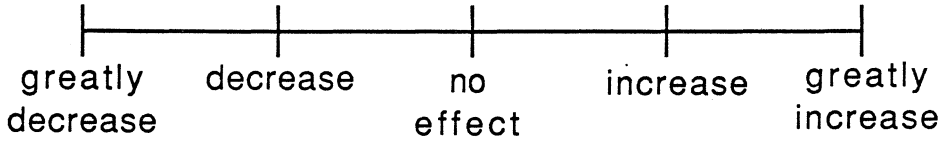
How many times would you want the message to repeated?

_____

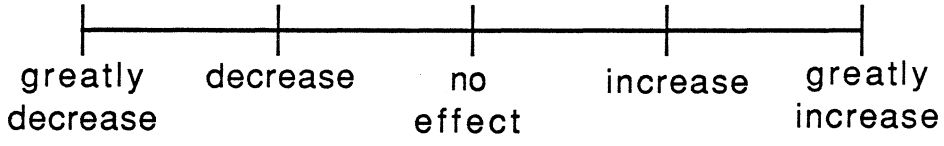
Spell out speech display

It is possible to have the system spell out words you do not understand. How would the availability of spelling affect your:

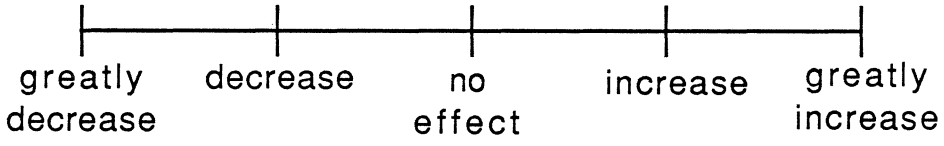
Speed at locating information



Accuracy at locating information



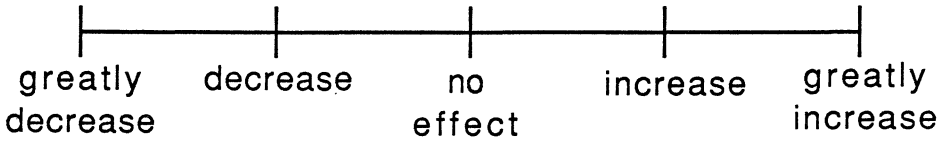
Acceptability of the information system



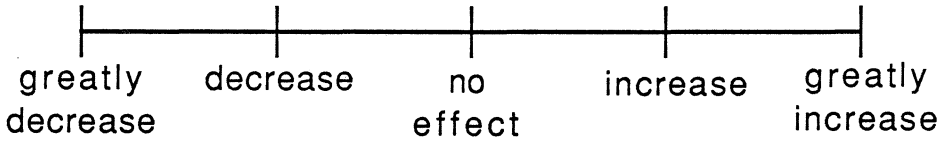
Pause/resume

The current system has a pause/resume function implemented. How would the elimination of this feature affect your:

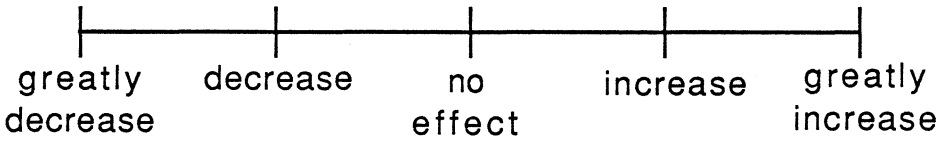
Speed at locating information



Accuracy at locating information



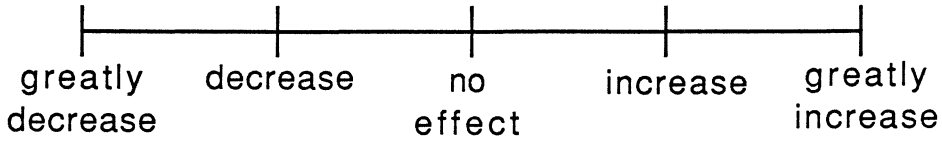
Acceptability of the information system



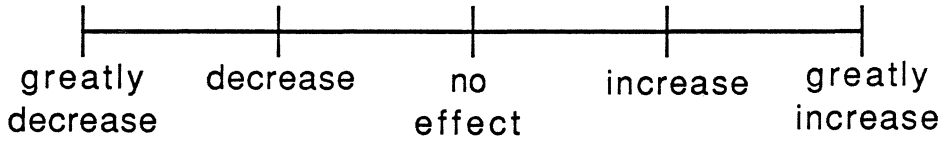
Interrupt

The current system allows you to interrupt messages with another command. How would the elimination of the interrupt feature affect your:

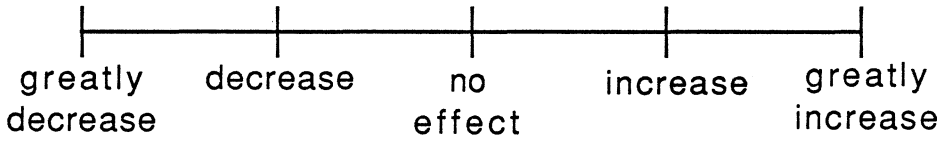
Speed at locating information



Accuracy at locating information



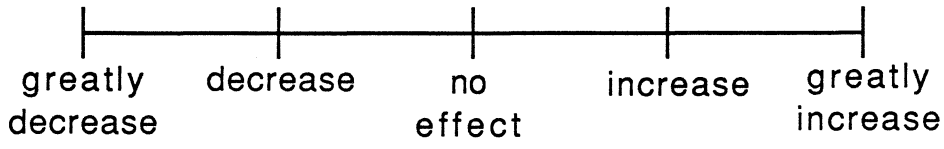
Acceptability of the information system



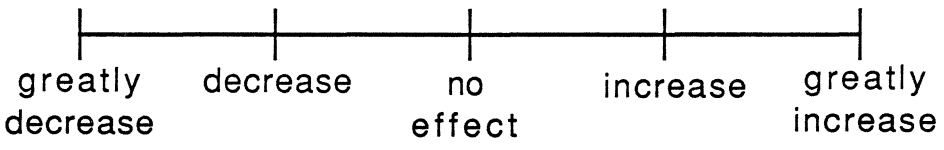
System response time

The current system responds almost instantaneously. How would increasing the system response time affect your:

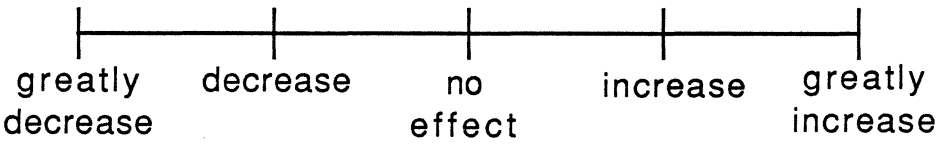
Speed at locating information



Accuracy at locating information



Acceptability of the information system



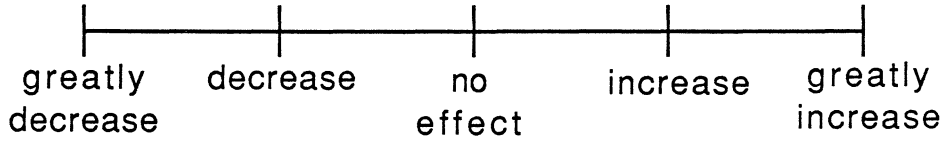
What is the maximum system response delay you would tolerate?

_____

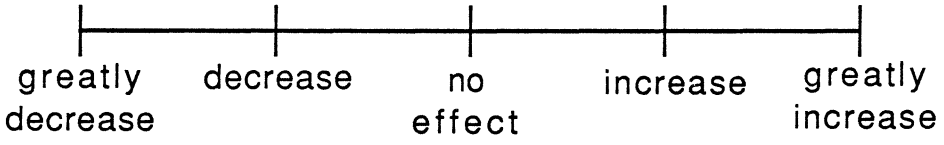
Number of keywords

Currently there are 70 keywords in the system. How would increasing the number of keywords affect your:

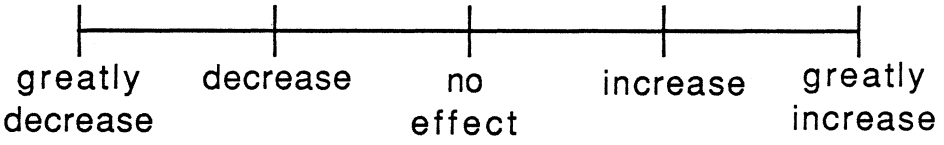
Speed at locating information



Accuracy at locating information



Acceptability of the information system



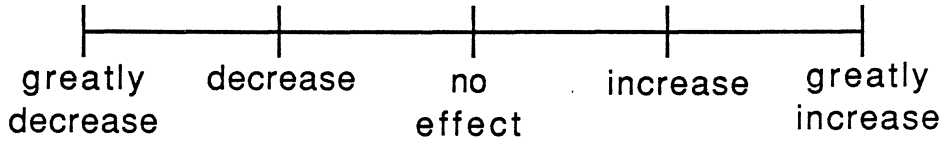
What is the maximum number of keywords you would suggest? _____

What is the optimum number of keywords you would suggest? _____

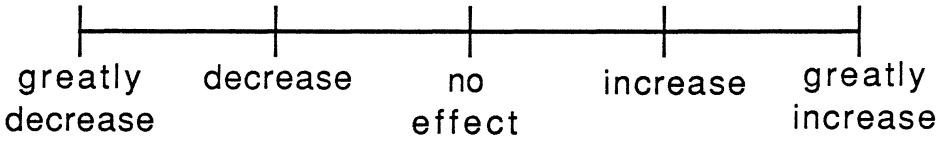
Number of messages

Currently there are 39 messages in the system. How would increasing the number of messages affect your:

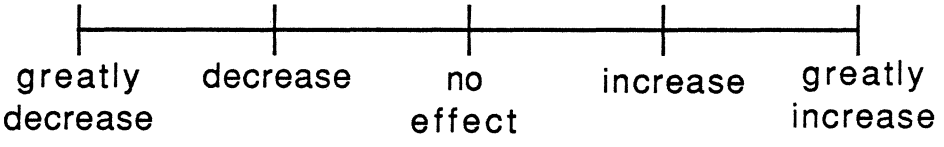
Speed at locating information



Accuracy at locating information



Acceptability of the information system



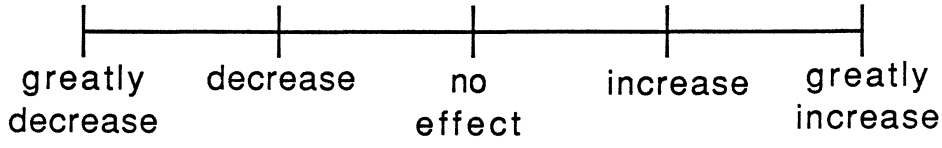
What is the maximum number of messages you would suggest? _____

What is the optimum number of messages you would suggest? _____

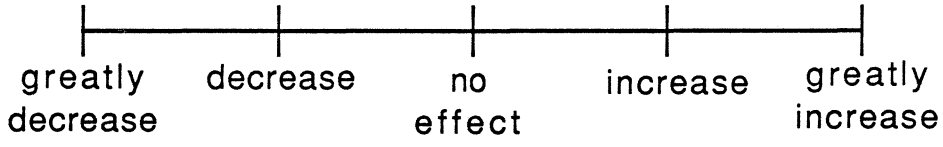
Organization of the data base

Many organization schemes can be used for the data base. How does the organization of the data base affect your:

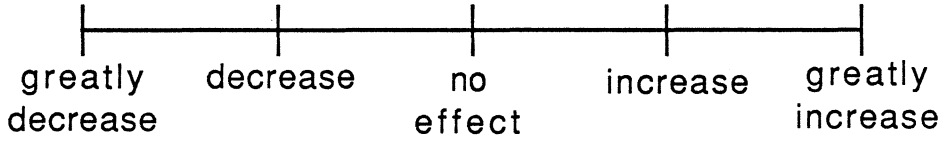
Speed at locating information



Accuracy at locating information



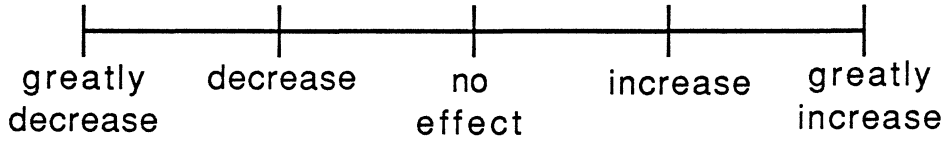
Acceptability of the information system



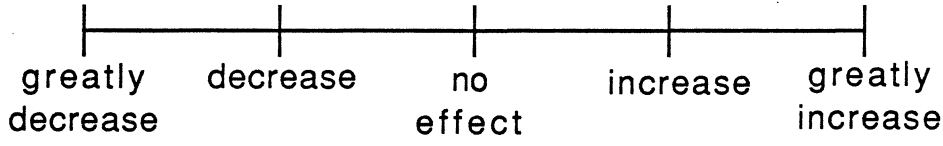
Type of data

Two basic types of data that can be presented within this system: numerical and verbal. How would presenting numerical data rather than verbal data affect your:

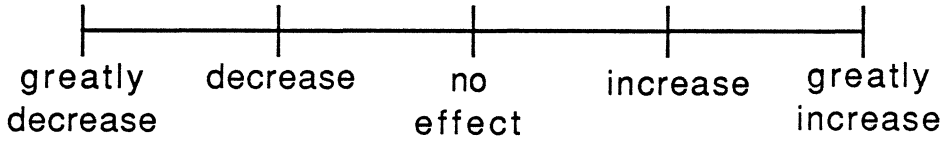
Speed at locating information



Accuracy at locating information



Acceptability of the information system



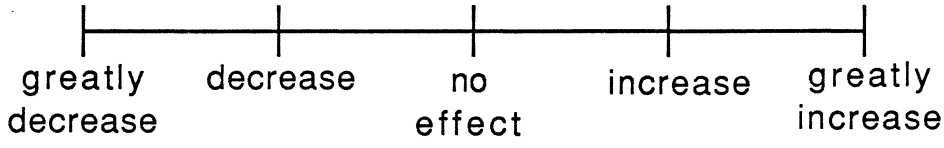
What type(s) of data are appropriate for this system?

_____

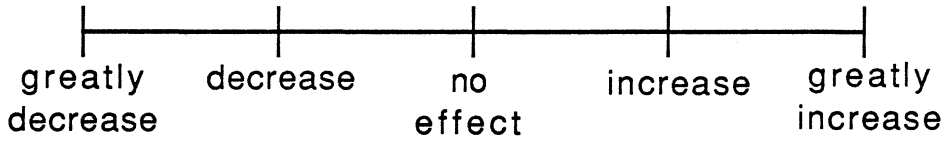
Number of steps in a search

The current system requires between 3 and 11 steps to complete a search. How would increasing the number of steps in a search affect your:

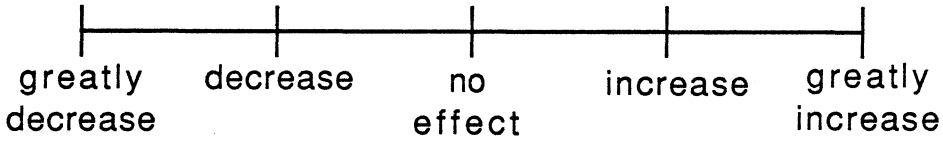
Speed at locating information



Accuracy at locating information



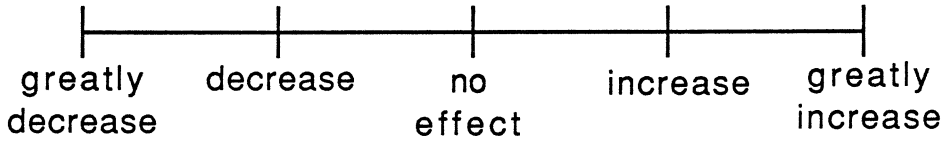
Acceptability of the information system



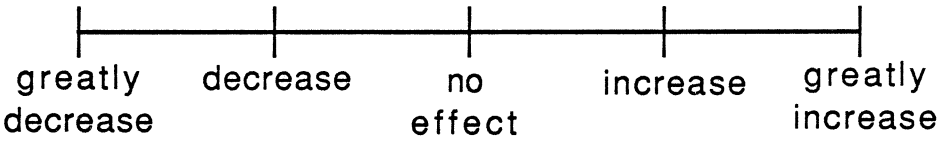
Number of searches per session

You were asked to complete 5 searches. How would increasing the number of searches affect your:

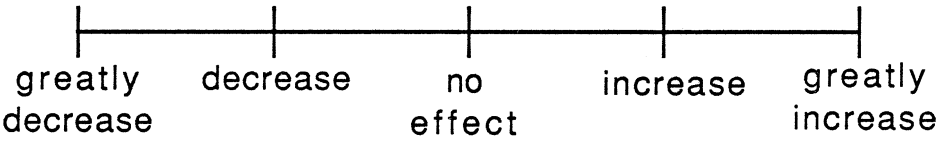
Speed at locating information



Accuracy at locating information



Acceptability of the information system

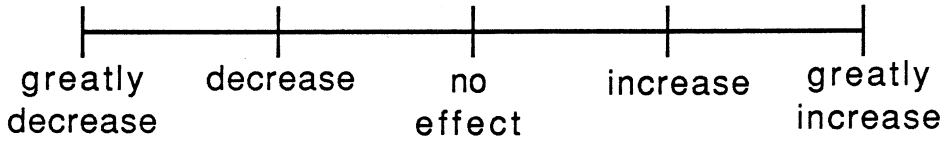


What is the maximum number of searches a subject should be asked to complete? _____

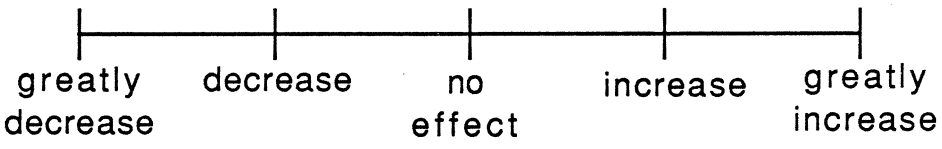
Multiple messages sought to complete a search

Some of your searches required you to locate more than one message. How would increasing the number of messages you must locate to complete a task affect your:

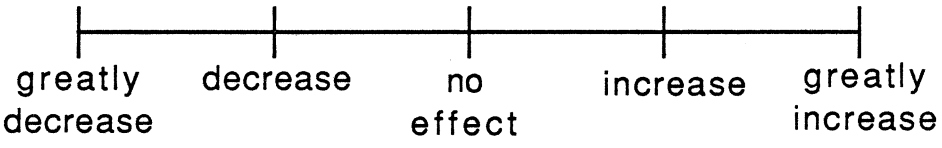
Speed at locating information



Accuracy at locating information



Acceptability of the information system

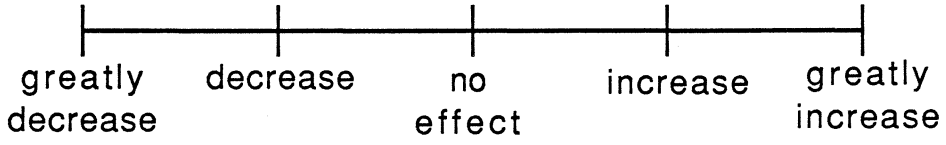


What is the maximum number of messages you would request someone to locate to complete a search? _____

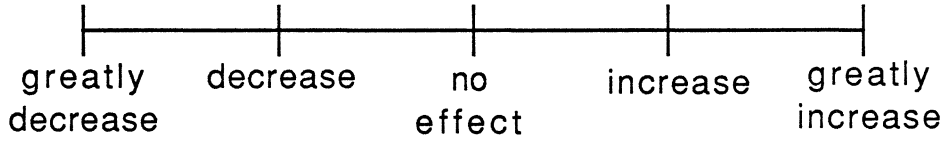
If/then

It is possible to use an if/then structure to complete a search. (e.g. if the movie "Dune" is showing, then what times is it showing) How would the availability of if/then syntax for searching affect your:

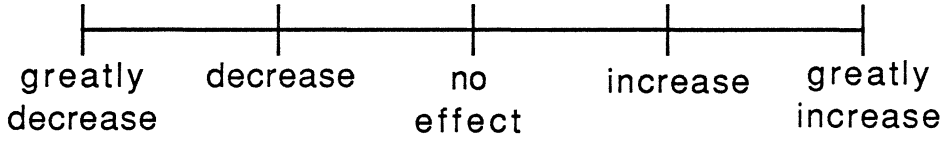
Speed at locating information



Accuracy at locating information



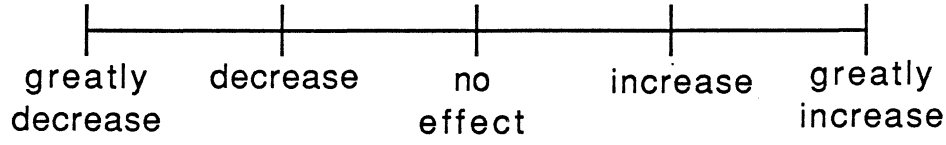
Acceptability of the information system



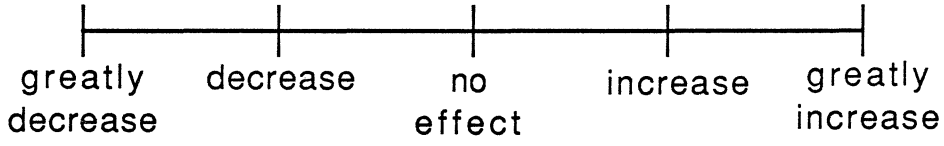
Competing tasks

It is possible to introduce additional activities that subjects must complete while using the system. How would competing tasks affect your:

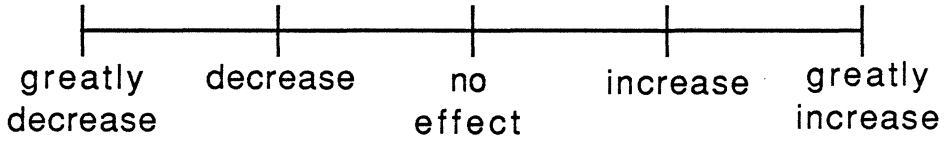
Speed at locating information



Accuracy at locating information



Acceptability of the information system



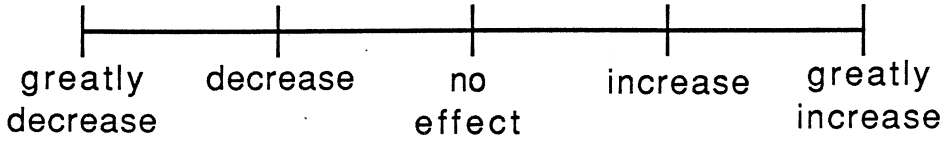
What types of tasks would compete the most with the use of the current system?

_____

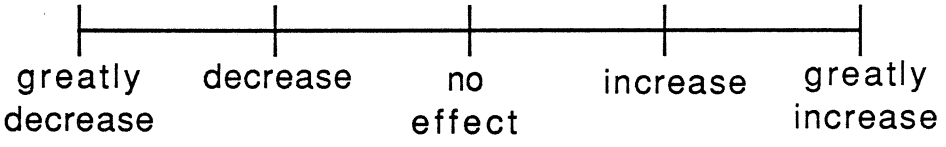
Competing speech

It is possible that you would be asked to use the system while being spoken to or while speaking to someone else. How would competing speech affect your:

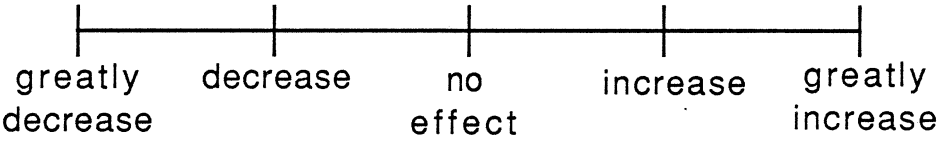
Speed at locating information



Accuracy at locating information



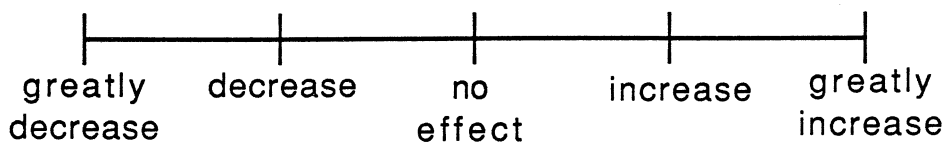
Acceptability of the information system



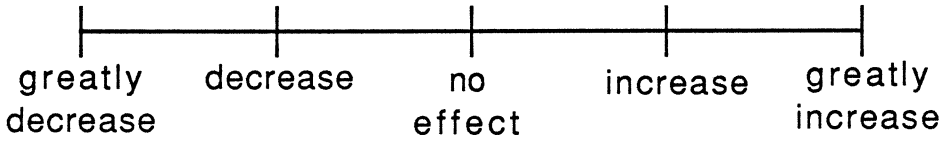
Noise

It is possible that you would be asked to use the system in a noisy (household, office, industrial) environment. How would noise affect your:

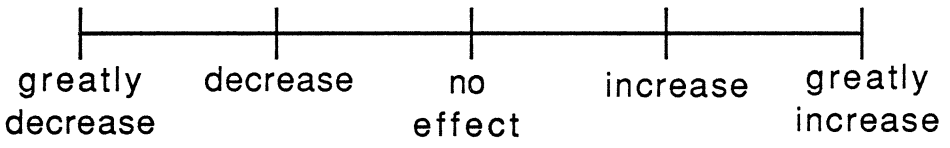
Speed at locating information



Accuracy at locating information



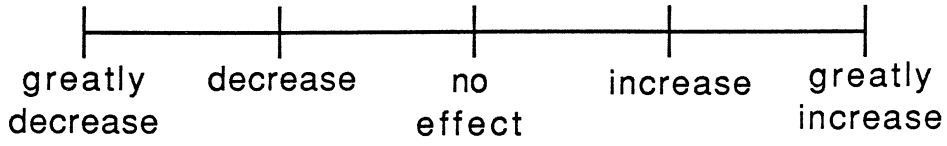
Acceptability of the information system



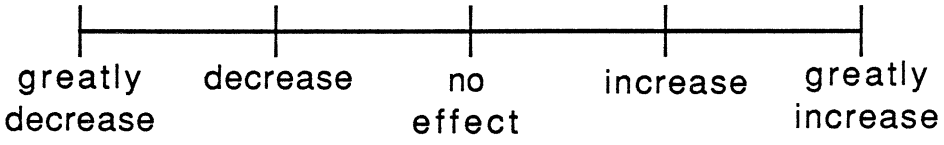
Background music

It is possible that you would be asked to use the system while music was playing in the background. How would background music affect your:

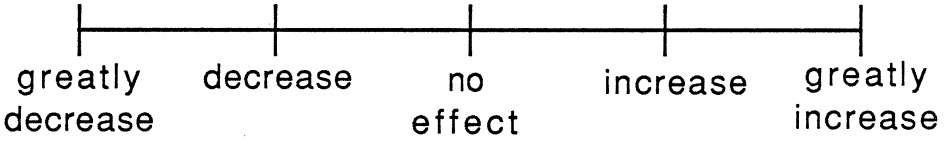
Speed at locating information



Accuracy at locating information



Acceptability of the information system



Other Features

What other features of the system would affect your speed, accuracy, and acceptability. The experimenter will provide you with a page for each feature you wish to list. The page will contain spaces for the name of the feature, a breif description of the feature, and rating scales for each feature.

Name

_____

Description

_____

_____

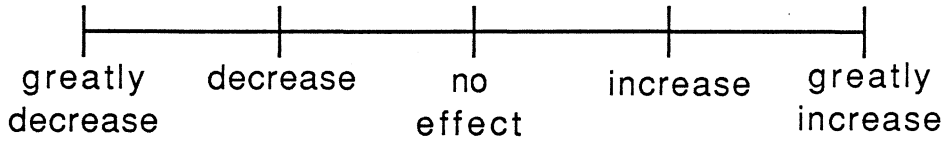
_____

_____

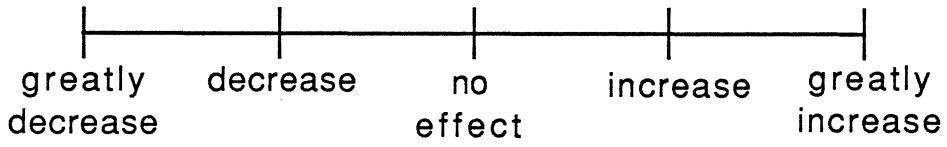
_____

How would this feature affect your:

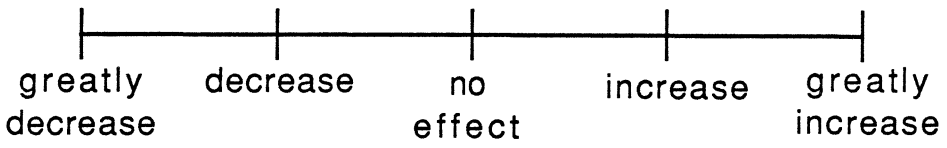
Speed



Accuracy



Acceptability



Appendix C. Participant's Informed Consent Form

The following experiment is a study concerning the evaluation of a telephone-based information system. During the experiment, you will be monitored with a closed-circuit video system. As a participant in this experiment, you have certain rights as explained below. The purpose of this document is to describe these rights and to obtain your written consent to participate in the experiment.

1. You have the right to discontinue your participation in the study at any time for any reason. If you decide to terminate the experiment, inform the researcher and he will pay you for the length of time you have participated.

2. You have the right to inspect your data and withdraw it from the experiment if you feel that you should for any reason. In general, data are processed and analyzed after a subject has completed the experiment. At that time, all identification information will be removed and the data treated with anonymity. Therefore, if you wish to withdraw your data, you must do so immediately after your participation is completed.

3. You have the right to be informed of the overall results of the experiment. If you wish to received a synopsis of the results, include your address with your signature below. If after receiving the synopsis, you would like more in depth information, please contact Virginia Tech's Human Computer Interaction Laboratory and a full report will be made available to you.

This research is funded by a research contract with the National Science Foundation. The co-principal investigators are Dr. Robert Williges, and Ms. Beverly Williges. The researcher is P. Jay Merkle, Jr. Any of these people be contacted at the following address and phone number:

Human Computer Interaction Laboratory
530 Whittemore Hall
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061
(703) 961-4602

Further comments or questions can be addressed to Charles Waring, chairman of the Institutional Review Board for the Use of Human Subjects in research. He can be contacted at the address and the phone number listed below:

Charles Waring
Office of Sponsored Research Programs
301 Burruss Hall
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061
(703) 961-5283

If you have any questions about the experiment or your rights as a participant, please do not hesitate to ask. The researcher will do his best to answer them,subject only to the constraint that he does not pre-bias the experimental results.

Your signature below indicates that you have read and understand your rights as a participant (as stated above), and that you consent to participate.

_____

Participant's Signature

_____

Witness' Signature

_____

_____

_____

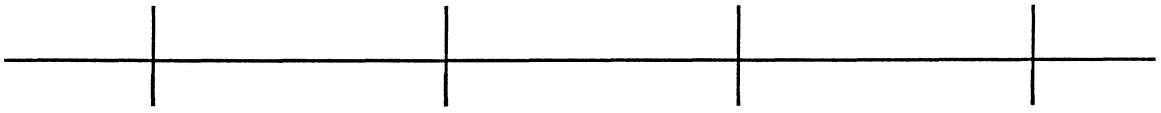
Print your name and address if you
wish to receive a summary of
the experimental results.

Appendix D. Subject Information Questionnaire

Age: _____Sex: _____ Native language: _____

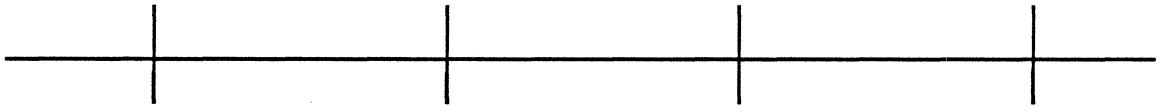
For the following questions, please circle the most accurate response:

1. How experienced are you with using computers?



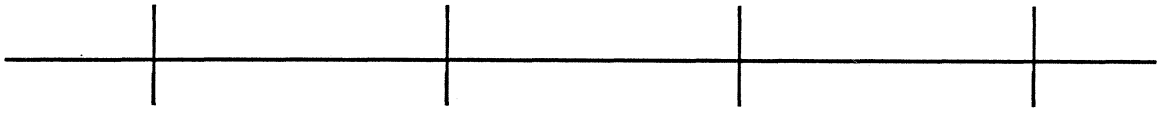
No experience Some experience Experienced Very Experienced

2. How experienced are you with using information systems?



No experience Some experience Experienced Very Experienced

3. How experienced are you with synthesized speech?



No experience Some experience Experienced Very Experienced

Appendix E. Prototype Demonstration Instructions

The object of this task is to locate information using the telephone inquiry system. To do so, turn the speaker-phone on by pushing the orange "sp-phone" button on the lower right hand corner of the telephone. The red indicator above it should come on, and you should hear a dial tone. Once you hear the dial tone, press the buttonmarked "dial" in the upper right hand corner of telephone. The information system will answer, and you can begin yoursearch for the answer to questions that are on the flip cards in front of you. Once you find the answer to the question, record it on the answer sheet provided, and hang up the phone. The experimenter will then cue you to begin searching for the answer to the next question.

Appendix F. Prototype Demonstration Targets

1. What time is the PDQ Bach concert?
2. Where will the movie Dune be playing, and what is its rating?
3. Where will the tractor pull be held?
4. At what theatre(s) will the movies Dune and Gone with the Wind be playing?
5. Who will speak at the next Wildlife Society meeting?

Appendix G. Validation Experiment Instructions

Your task is to search for information on store itemsin the department store's talking database. Store items will be presented as targets on the computer display in front of you. You will find the target by using the telephone keys to move through the talking database.

These are your instructions:

1. Press the on-off key on the telephone key-pad and listen for a dial tone.

2. Press the dial key on the telephone key-pad located inthe upper right corner.

3. The talking computer will answer the telephone andoffer you instructions. Press the # key on the telephone keypad and listen carefully to the instructions for using the telephone keypad.

4. Read the first target on the computer display in front of you.

5. Watch the computer display. It will signal you when the search is about to begin.

6. The talking computer will begin speaking a menu of key-words. Key-words categorize groups of store items. After each key-word is spoken, the computer will pause briefly to allow you to select the item. If you do not select the item, the computer will speak another key-word for that menu.

7. To locate the target, select a keyword from the menu which best categorizes the store item you are searching for. The computer will then speak a new menu of keywords,based on your selection. If you need to hear the key-pad instructions again, select help from any menu.

8. Continue listening to menus and selecting key-words until you reach the desired store item.

9. When you hear the desired store item, select it and listen carefully to the information message

10. Press the 2 key on the telephone key-pad, when you are ready to transcribe the information message.

11. The computer display will prompt you to transcribe what you heard.

12. Type the information message you heard into the computer,and press the return key.

13. Rate the certainty of your transcription being correct on a scale of 1, very uncertain to 7, very certain, and press the return key.

14. Rate the difficulty of understanding the message on a scale of 1, very difficult to 7, very easy, and press the return key.

15. Read the next target on the computer display and getready to start the next search. The computer display will signal you to begin the next search and will speak the first item in the main menu. Locate the next target and transcribe the information message.

16. The experiment will proceed in this fashion. You will search for a total of 16 targets.

17. The computer display will indicate when you have completed the target searches. The computer display will then request that you rate certain characteristicsof the telephone information system. The meaning of eachcharacteristic and how it should be rated will be explained on the computer display.

If you have any questions, please ask the experimenter now.

Appendix H. Experimental Condition Specific Instructions

The video instructions you just watched included a demonstration of how the telephone information system works, and how you should perform the task for this study. The actual telephone information system you will be using today will be similar to the system in the video, but may be different in some ways.

These are the commands that are available to you on thetelephone keypad:

To select an item, press the # key.

To back-up one menu, press the * key.

To select the main menu, press the 0 key.

When you are ready to transcribe the informationmessage, press the 2 key.

To receive a transaction summary, press the 5 key.

Appendix I. Information Message Targets: Main Study

Demonstration targets

1. What is the information message for golf books?
2. What is the information message for women's cotton blouses?

Experiment targets

1. What is the information message for laundry washer?
2. What is the information message for football books?
3. What is the information message for eye mascara?
4. What is the information message for men's blazers?
5. What is the information message for food blenders?
6. What is the information message for guitars?
7. What is the information message for pearl necklaces?
8. What is the information message for hope chests?
9. What is the information message for women's silk blouses?
10. What is the information message for compact disc recordings?
11. What is the information message for women's oriental fragrances?
12. What is the information message for men's sweaters?
13. What is the information message for knit dresses?
14. What is the information message for gold chains?
15. What is the information message for recliner chairs?
16. What is the information message for chicken cookbooks?

Appendix J. Database Information Messages

Message type indicated in parentheses: (I) = Information, (A) = Availability, (P) = Price, and (L) = Location.

1. Washers: Deluxe models are available with green trimming. (A)
2. Football Books:. Faculty discounts are offered to gym teachers. (I)
3. Eye Mascara:. Travel supplies are sold for $17. 50. (P)
4. Men's Blazers:. Garment bags are offered with new purchases. (I)
5. Food Blenders:. Boxes and cartons are in the wrapping center. (L)
6. Guitars:. Carrying cases are reduced by 55 to 63%. (P)
7. Pearl Necklaces:. Sorority clasps are in the school
8. Hope Chests:. Walnut stains are reduced by 34 to 40%. (P)
9. Silk Blouses:. Maternity wear is near ladies lingerie. (L)
10. Compact Discs:. Head cleaneres are on aisle 12. (L)
11. Oriental Fragrances:. Manufacturer's samplers are offered to interested shoppers. (I)
12. Men's Sweaters:. Rugby letters are sold for $11. 60. (P)
13. Knit Dresses:. Designer collections are available in red and ivory. (A)
14. Gold Chains:. Instant financing is available at the central office. (A)
15. Recliner Chairs:. Leather coverings are offered to wholesale buyers. (I)
16. Chicken Cookbooks:. Collector editions are available in limited quantities. (A)

Appendix K. Experimental Debrief

Do you like the idea of an information system like this one?

Would you use an information system like this one?

What applications seem appropriate for an information system such as this one?

What improvements would you suggest?

Overall, did you like (or enjoy) using this system?

What information would you like to add to the instructions?

What would you not include in the instructions?

Did you understand the commands?

if not

Which commands confused you?

What did you understand the command to do?

How did the execution of the command differ from your expectations?

Are there any commands you would like to add?

Are there any commands you would like to eliminate?

Did you understand the use of the transaction summary command?

Did you understand the purpose of the changing speech rate?

What command would you use to restart if you got lost?

What command would you use if you wanted to backup one category?

Do you think you understand the organization of the data base well enough to use the system comfortablely?

Did the keyword categories confuse you?

What would you change about the experimental session?

Was the session length too long?

Was the task interesting or boring?

Appendix L. ANOVA Summary Tables

ANOVA Summary Table for Target Search Time Ratio.

Source	*df*	*SS*	*MS*	*F*	*p*
Between Subjects					
Number Steps (S)	1	0.5746	0.5746	78.92	0.0001 *
Native/ Non-native(N)	1	0.6776	0.6776	9.23	0.0056 *
Adapting Speech Rate(R)	1	0.00003	0.00003	0.01	0.9441
Sex of Voice (V)	1	0.0018	0.0018	0.26	0.6168
Transaction Summaries (T)	1	0.0048	0.0048	0.65	0.4264
S x T	1	0.0049	0.0049	0.67	0.4206
S x N	1	0.0136	0.0136	1.87	0.4206
Subjects/ Treatments	24	0.1747	0.00728		
Total	31	0.84212			

* significant at $p = 0.05$

ANOVA Summary Table for Target Search Efficiency Ratio

Source	*df*	*SS*	*MS*	*F*	*p*
Between Subjects					
Number Steps (S)	1	0.3007	0.3007	47.69	0.0001 *
Native/ Non-native(N)	1	0.05995	0.05995	9.51	0.0051 *
Adapting Speech Rate(R)	1	0.00085	0.00085	0.13	0.7171
Sex of Voice (V)	1	0.00368	0.00368	0.58	0.4523
Transaction Summaries (T)	1	0.00765	0.00765	1.21	0.2817
S x T	1	0.00749	0.00749	1.19	0.2863
S x N	1	0.02079	0.02079	3.30	0.0820
Subjects/ Treatments	24	0.15137	0.0063		
Total	31	0.55257			

* significant at $p = 0.05$

ANOVA Summary Table for Average Invalid Keypresses

Source	*df*	*SS*	*MS*	*F*	*p*
Between Subjects					
Number Steps (S)	1	0.00439	0.00439	0.67	0.4222
Native/ Non-native(N)	1	0.04882	0.04882	7.41	0.0119 *
Adapting Speech Rate(R)	1	0.00195	0.00195	0.30	0.5912
Sex of Voice (V)	1	0.01757	0.01757	2.67	0.1155
Transaction Summaries (T)	1	0.0122	0.0122	1.85	0.1862
S x T	1	0.00048	0.00048	0.07	0.7878
S x N	1	0.001953	0.001953	0.30	0.5912
Subjects/ Treatments	24	0.1582	0.00659		
Total	31	0.2456			

* significant at $p = 0.05$

ANOVA Summary Table for Strict Message Transcription Accuracy

Source	*df*	*SS*	*MS*	*F*	*p*
Between Subjects					
Number Steps (S)	1	0.07629	0.07629	1.07	0.3119
Native/ Non-native(N)	1	1.39758	1.39758	19.55	0.0002 *
Adapting Speech Rate(R)	1	0.10266	0.102663	1.44	0.2425
Sex of Voice (V)	1	0.00305	0.00305	0.04	0.8381
Transaction Summaries (T)	1	0.0048	0.0048	0.65	0.4264
S x T	1	0.00305	0.00305	0.25	0.8381
S x N	1	0.00305	0.00305	1.87	0.8381
Subjects/ Treatments	24	1.71582	0.07149		
Total	31	3.35534			

* significant at $p = 0.05$

Attention Patron:

Page 191 is missing from
all copies

ANOVA Summary Table for Transcription Certainty Rating

Source	*df*	*SS*	*MS*	*F*	*p*
Between Subjects					
Number Steps (S)	1	0.0313	0.0313	0.04	0.8471
Native/ Non-native(N)	1	5.2813	5.2813	6.42	0.0182*
Adapting Speech Rate(R)	1	0.2813	0.2813	0.34	0.5643
Sex of Voice (V)	1	0.0313	0.0313	0.04	0.8471
Transaction Summaries (T)	1	0.7813	0.7813	0.95	0.3396
S x T	1	0.0313	0.0313	0.04	0.8471
S x N	1	0.0313	0.0313	0.04	0.8471
Subjects/ Treatments	24	19.750	0.8229		
Total	31	0.55257			

* significant at $p = 0.05$

ANOVA Summary Table for Ease of Use Rating

Source	*df*	*SS*	*MS*	*F*	*p*
Between Subjects					
Number Steps (S)	1	3.7813	3.7813	4.37	0.0047
Native/ Non-native(N)	1	5.2813	5.2813	6.11	0.0209*
Adapting Speech Rate(R)	1	0.2813	0.2813	0.33	0.5737
Sex of Voice (V)	1	0.7813	0.7813	0.90	0.3513
Transaction Summaries (T)	1	0.7813	0.7813	0.90	0.3513
S x T	1	0.7813	0.7813	0.90	0.3513
S x N	1	0.7813	0.7813	0.90	0.3513
Subjects/ Treatments	24	20.750	0.8646		
Total	31	0.55257			

* significant at $p = 0.05$

ANOVA Summary Table for Speaker Intelligibility Rating

Source	*df*	*SS*	*MS*	*F*	*p*
Between Subjects					
Number Steps (S)	1	0.7813	0.7813	1.06	0.3143
Native/ Non-native(N)	1	3.7813	3.7813	5.11	0.0331*
Adapting Speech Rate(R)	1	3.7813	3.7813	5.11	0.0331*
Sex of Voice (V)	1	0.2813	0.2813	0.38	0.5433
Transaction Summaries (T)	1	0.2813	0.2813	0.38	0.5433
S x T	1	3.7813	3.7813	5.11	0.0331*
S x N	1	0.2813	0.2813	0.38	0.5433
Subjects/ Treatments	24	17.750	0.7396		
Total	31	0.55257			

* significant at $p = 0.05$

The vita has been removed from the scanned document