Spring 1988

# Natural Language Human-Computer Dialogue: Menu-Based Natural Language and Visual Performance

Jeffrey John Hendrickson
*Old Dominion University*

## Recommended Citation

Hendrickson, Jeffrey J.. "Natural Language Human-Computer Dialogue: Menu-Based Natural Language and Visual Performance" (1988). Doctor of Philosophy (PhD), dissertation, Psychology, Old Dominion University, DOI: 10.25777/94hk-pg28
https://digitalcommons.odu.edu/psychology_etds/280

# NATURAL LANGUAGE HUMAN-COMPUTER DIALOGUE:
# MENU-BASED NATURAL LANGUAGE AND VISUAL PERFORMANCE

by

Jeffrey John Hendrickson
B.A. May 1979, Texas Technological University
M.A. May 1984, Stephen F. Austin State University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

INDUSTRIAL/ORGANIZATIONAL PSYCHOLOGY

OLD DOMINION UNIVERSITY
May, 1988

Approved by;

Glynn D. Coates (Director)

# ABSTRACT

## NATURAL LANGUAGE HUMAN-COMPUTER DIALOGUE: MENU-BASED NATURAL LANGUAGE AND VISUAL PERFORMANCE

Jeffrey John Hendrickson
Old Dominion University, 1988
Director: Dr. Glynn D. Coates

The present study was conducted to determine design principles for menu-based natural language (MBNL) interfaces and to provide evidence for the nature of visual search processes with menu-based systems. The effects of window size, window activity, and query length were investigated. Window size was manipulated as a between-subjects variable with three levels representing a sixteen-item window size, an eight-item window size, and a four-item window size. Window activity was manipulated as a within-subjects variable with two levels representing single active and multiple active windows. Query length was manipulated as a within-subjects variable with three levels representing one-, two-, and three-item query lengths. Thirty six subjects randomly assigned to three groups, based on the window size factor, performed queries with the three query lengths in both window activity conditions in counterbalanced order. It was found that two- and three-item queries were performed faster with single active windows. However, subjects rated multiple active windows as more 'natural'. Query times also increased with query length and errors were most likely to occur on the longest query. Longer eye fixation durations were observed with the four-item window size. Fixation frequencies, fixation durations, dwell times, and relative dwell times all varied as a function of query length. Visual behavior also depended on which 'area of interest' subjects were viewing, and this effect interacted with window activity and query length. Finally, it was found that menus were not scanned randomly. However, scanpaths were less deterministic with multiple active windows and tended to become less constrained as query length increased. Eased on the findings, human factors design principles were derived for application to MBNL interfaces.

# ACKNOWLEDGEMENTS

I would like to thank Dr. Glynn Coates for his guidance throughout this project. Without his encouragement and expertise this dissertation would not have have been completed with the same level of professional quality. I would like to thank Dr. Terry Dickinson and Dr. Raymond Kirby for their support and guidance on this project. I would also like to thank Dr. Randall Harris of NASA Langley for his critique of this research.

I am especially indebted to Dr. Gerald Birdwell for providing an internship with the User Systems Engineering Center at Texas Instruments, and for providing the support that made this research possible. I would like to thank Dr. William Muto of Texas Instruments for his critical thinking that contributed to the development of the research concepts represented in this study. I would also like to thank Mr. James Pawlowski of Texas Instruments for his support and encouragement on this project. Without the support of the staff and management at Texas Instruments this project would not have been possible.

Additionally, I owe my sincere thanks to Mr. Ken Stevenson for his dedication and hard work in writing the eyetracking software that was instrumental to the completion of this project.

Finally, I owe my deepest appreciation to my wife, Valette, for her assistance, supportiveness, and understanding over the many long days and nights that were involved in carrying out this project.

## TABLE OF CONTENTS

Page

# LIST OF TABLES

## LIST OF FIGURES

FIGURE PAGE

# INTRODUCTION

Natural language human-computer dialogue has been a subject of much discussion, by advocates as well as opponents (Ballard, 1979; Ballard & Biermann, 1979; Balzer, 1973; Biermann & Ballard, 1980; Biermann, Ballard, & Sigmon, 1983; Dijkstra, 1978; Ford, 1981; Green et al., 1978; Hauptmann & Green, 1983; Heidorn, 1976; Hill, 1972; Kelley, 1981; Ledgard, Whiteside, Singer, & Seymour, 1980; Martin et al., 1974; Morris, 1979; Ogden & Brooks, 1983; Petrick, 1976; Shneiderman, 1980, 1987; Sondheimer, 1978; Tennant, 1980, 1981; Tennant, Ross, Saenz, Thompson, & Miller, 1983; Tennant, Ross, & Thompson, 1983; Thompson, Tennant, Ross, & Saenz, 1983). Proponents of natural language human-computer dialogue claim that it has several advantages over formal command language dialogue in that natural language dialogue is versatile, easy to use, does not require much up front training, and permits the possible use of speech recognizers for input. Furthermore, users do not have to learn a command syntax or new syntactical rules, thereby accommodating the inexperienced user. Shneiderman (1987) argues that natural language human computer dialogue "...can be effective for the user who is knowledgeable about some task domain and computer concepts but who is an intermittent user who cannot retain the syntactic details" (p. 166).

Several applications of restricted scope, such as LUNAR, SOPHIE, ELIZA, CHECKBOOK, BASEBALL, MARGIE, and INTELLECT, have demonstrated that it is possible to design computer programs that will accept natural language instructions to accomplish particular tasks (Bobrow & Collins, 1975; Brown, Burton, & Bell, 1975; Ford, 1981; Green, Wolf, Chomsky, & Laughery, 1963; Pertrick, 1976; Schank, 1975; Schank & Colby, 1973; Suding, 1983; Weizenbaum, 1966; Woods, 1970). Experimental studies of natural language dialogue have included comparisons between natural languages and query languages, laboratory studies of prototype natural query languages, and field studies of prototype systems (Damerau, 1981; Egly & Wescourt, 1981; Hershman, Kelly, & Miller, 1979; Kaplan, 1982; Krause, 1979, 1980; Miller, Hershman, & Kelly, 1978; Shneiderman, 1978; Small & Weldon, 1983; Tennant, 1980; Waltz, 1977). Encouraging results have been reported, but most of the studies also indicate usability problems.

A number of disadvantages and shortcomings of natural language dialogue have been described (Biermann et al., 1983; Hauptmann & Green, 1983; Lowden & DeRoeck, 1985; Ogden & Brooks, 1983; Shneiderman, 1980, 1987; Tennant, Ross, & Thompson, 1983; Weizenbaum, 1966, 1976; Winograd, 1972). Relatively high failure rates, high error rates, ease of use problems, and user frustration have been noted. Some have argued that natural language dialogue leads to ambiguity in the formulation of queries and requests and that natural languages are not only ambiguous but overly verbose. Natural language systems are noted to be mysterious about their coverage and capabilities and natural language dialogue, it has been argued, leads to an overestimation of computer capabilities and intimidation of the user. Features of natural language systems are thus often not used because users are unaware of them or do not trust them.

The desirability of natural language systems for use across the user spectrum and user-system task variety has been questioned. Natural language dialogue is generally considered preferable for inexperienced users. However, a concise command language seems preferable for knowledgeable and frequent users who are thoroughly aware of available functionality. Experts, it is has been noted, generally prefer terse, formal command languages.

From the software development perspective, there are also reservations about natural language systems. The programs must handle relatively large grammars and lexicons, and the code required to parse and translate the natural language input can be extensive and complex. The programs typically require "best guess" algorithms to handle spelling, syntactic, and semantic variations. System resources must consequently be allocated for recognizing the variant syntactical structures and synonymous terms. Resources must also be allocated for error checking and clarification procedures. Conventional natural language systems are thus expensive to build and maintain, and they require large amounts of computer memory.

## Menu-Based Natural Language

Menu-based natural language (MBNL) stands at the middle ground between the restrictive formal command languages and unconstrained free-form natural language. MBNL provides a form of constrained natural language dialogue for human-computer interaction. With a MBNL interface, natural language words and phrases are displayed on a screen as menu items. The user constructs a natural language sentence by selecting the menu items with a pointing device. As the menu items are selected, the natural language sentence is formed in a results window, and when the sentence is complete, it is sent to the underlying application program for execution.

Work in the area of MBNL dialogue has shown promising results (Hendrickson & Williams, 1988; Osga, 1984; Tennant, Ross, Saenz, Thompson & Miller, 1983; Tennant, Ross & Thompson, 1983; Thompson et al., 1983). The coverage and limitations of a MBNL system are made more apparent to the user due to the use of a restricted natural language. The user can thus avoid the frustration of overextending beyond the limits of system functionality. Since MBNL interfaces are closed and manageable, this also allegedly encourages exploration and use of the full range of system resources. Furthermore, MBNL interaction requires only the use of a pointing device such as a mouse, trackball, or lightpen. If a keyboard is used for input, only the cursor keys and enter key are required. Typing is thus eliminated, and the user is guaranteed a semantically and syntactically correct query or command input. MBNL interfaces can also be developed relatively quickly and require fewer memory resources than a conventional natural language system.

## Menu-Based User-Computer Interface Issues

Menu-based human-computer dialogue has proven to be a popular form of human-computer interaction (Brown, 1982; Martin, 1973; Shneiderman, 1980, 1987). Menu-based systems reduce or eliminate the need to learn complex sequences of commands, which simplifies the user-computer interface, and makes these systems particularly appealing for novice and intermittent computer users. With careful design, fast display rates and response times, and provisions for taking shortcuts to deeper menu levels, menu-based systems can also be appealing to knowledgeable and frequent computer users (Shneiderman, 1987).

Menu-driven computer systems can, however, be difficult to use (e.g., see Lee, Whalen, McEwen & Latremouille, 1984). Users may experience difficulties navigating through menu structures. Complex hierarchical menu structures can be rather problematic (Broadbent, Cooper & Broadbent, 1978; Liebelt, McDonald, Stone & Karat, 1982). Difficulties may arise from excessive menu depth (Dray, Ogden & Vesteweig, 1981; Kiger, 1984). Users may lack familiarity with upper level menu items (Somberg & Picardi, 1983), or may have a mental model of the system that differs from that of the designer (Billingsley, 1982; Dumais & Landauer, 1982). Particular training methods or a complete lack of up-front training may also give rise to difficulties in using menu-based systems (Norman, Schwartz & Shneiderman, 1984; Schwartz, Norman & Shneiderman, 1985).

Tombaugh and McEwen (1982) found that users engaged in menu-based information retrieval were very likely to give up on a high proportion of searches. Menu selection response time has also been observed to increase as users traverse deeper into menu tree structures, and when users select an incorrect option they tend to restart at the main menu rather than backtrack to the submenu where the incorrect choice was made (Hagelbarger & Thompson, 1983; Robertson, McCracken & Newell, 1981). Allen (1980) examined the effects of search depth on response times and error rates at each of four levels of a hierarchically structured database. Average response time at each level of the menu was observed to increase and more errors were committed on the deeper searches.

Miller (1981) represented a 64 item menu with four different hierarchical menu structures varied parametrically along breadth and depth dimensions. The menus were configured with two options at each of six levels (2 X 2 X 2 X 2 X 2 X 2), four options at each of three levels (4 X 4 X 4), eight options at each of two levels (8 X 8), and 64 options on one screen. Snowberry, Parkinson and Sisson (1983) also constructed four hierarchical menu structures varied on breadth and depth dimensions as in Miller's study. Subjects in both studies performed simple menu selection tasks. Considering both speed and accuracy, Miller observed the best performances with the menu configuration consisting of eight items at each of two levels, whereas Snowberry et al. observed the best performances with the 64 menu items displayed on a single screen. The results of Snowberry et al., however, held only when the menu items were grouped categorically. A second version of the menu with the 64 items randomly ordered yielded comparable search accuracy but slower search times. Overall, the results of both studies indicate faster search times and fewer errors with decreased depth and increased breadth of menu structures.

Sisson, Parkinson and Snowberry (1983) noted that the menus in their previous study were displayed relatively instantaneously since the menu items were loaded into a display buffer before the display circuitry was enabled. Recognizing that a serial link would be used with most menu-driven implementations, Sisson et al. calculated communication times for the four menu structures at various communication rates and added these times to the search and response times of the subjects to obtain total execution times. It was determined that the menu structure with all 64 items on one screen was optimal at the most rapid communication rates, while the menu structure with four items at each of three levels was optimal at the slowest rates, and the menu structure with eight items at each of two levels was optimal at intermediate rates. The menu structure with two items at each of six levels was never optimal at any communication rate. The conclusion was that one must select menu structures in light of communication rates if the goal is to minimize execution time.

McDonald, Stone and Liebelt (1983) compared five versions of a menu consisting of 64 words from four different categories. In three of the menu configurations the words from the four categories were grouped into four columns. Within each column the words were further grouped into categorical, alphabetical or randomized arrangements. The categorical groupings of words within categories were obtained by grouping words with the highest rated similarity. In the two remaining menu configurations the words from the four categories were intermixed. In one version, all of the words were alphabetized, and in the other version, the words were randomized. Additionally, the nature of the menu search task was manipulated. Some of the subjects were given target items and were instructed to find them, while other subjects were given one line definitions and were instructed to find target items matching the definitions. McDonald et al. found that subjects were able to perform the simple menu item selection task faster than the item-definition matching task. There was also a task type by menu structure interaction with the larger differences in search times observed between the five menu structures in the item-definition matching condition. The most rapid search times were observed with menu items grouped by rated similarity within categories.

## Menu-Based Interfaces and Visual Performance

Card (1982) found that an alphabetically arranged command menu was searched faster than a menu with command options grouped by function. The alphabetically and functionally organized menus both yielded faster search times than a menu with the commands randomized. Card concluded that the alphabetically arranged command menu permitted faster visual searches. Card also asserted that the menus were searched randomly and not by systematic patterns of eye movements. Card theorized that users scan menus in blocks, or 'perceptual chunks', requiring a variable number of eye movements. Depending on menu organization, differences in visual search times will then be observed.

The single study by Card provides the basis for the belief that "...most intra-frame searching is done randomly, not by systematic up-and-down eye movements" (Parton, Huffman, Pridgen, Norman & Shneiderman, 1985, p. 1). However, the study conducted by Card involved only three subjects and, as pointed out by MacGregor and Lee (1987), "Since the study used command menus, and provided high degrees of practice, highly restricted scanning of options may well have occurred, perhaps involving a random search of a subset of options" (p. 63).

MacGregor, Lee and Lam (1986) have argued that Card's evidence for random visual search of command menus is inconclusive, and is as consistent with sequential search as random search. Furthermore, the visual search process for database menus may differ from the search process for command menus, and may be sequential and either self-terminating or exhaustive (see Lee, 1979). Allen (1983), for example, in a study involving a videotex-type task, found significantly faster menu selection times to items at the top of menus than at the bottom, consistent with a sequential, self-terminating, top-down visual search process.

## Experimental Rationale

The overall objectives of the present study were: 1) to determine the usability of a menu-based natural language (MBNL) interface for database query; 2) to establish some design guidelines for MBNL interfaces; and 3) to provide evidence for the nature of visual search processes with menu-based systems. A MBNL system served as an interface to a database. Users performed database retrieval using natural language queries developed by selecting words and phrases from menus embedded within windows. The effects of window size, number of active windows, and query length were investigated.

The interface screen included a total of five windows. The top window on the screen was a results window where the natural language queries were formed in sentence format as users selected menu items. The bottom window was a command window used for displaying available functions (Page Down, Page Up, Back Up, Restart, and Execute). Placed in the center of the screen, between the results and command windows, were three parse windows containing a total of 96 menu items from which selections were made to construct queries. The size of these windows were manipulated so that 4, 8, or 16 of the 32 items within each window were visible at any given time. Users had to page up or down to view any items not currently in view within the windows.

Windows were also manipulated so that only one parse window was active at any given time or all parse windows were active simultaneously. When only one parse window was active at a time, then the cursor could not enter any other parse window until a selection had been made from the currently active window. This manipulation constrained the subjects to selecting items from windows in a particular sequence, effectively reducing the set of selectable items to those contained in the currently active window. When all parse windows were active simultaneously, then the cursor could enter any parse window at any time. This manipulation gave the subjects the freedom to select any item from any window in any sequence, effectively expanding the set of selectable items to the fullest.

Query length was manipulated by requiring users to formulate and enter queries requesting information on one, two, or three database topics. When required to find information on more than one topic, subjects were required to lengthen their query with a recursive sentence. Query length was thus varied parametrically to create three query lengths.

Query performance hypotheses. Performance times and error frequencies were expected to increase with decreasing window size and were expected to increase with query length. Multiple active windows were also expected to yield longer query performance times and higher error frequencies. In terms of subjective evaluations, however, it was expected that multiple active windows would be preferred over single active windows.

It was also predicted that window size and window activity would interact, with the longest query performance times and the highest error frequencies expected in the case of the smallest window size with multiple active windows. Futhermore, it was expected that query length would interact with both window size and window activity, yielding the longest query performance times and the highest error frequencies in the case of the longest query performed with the smallest window size and with multiple active windows.

Visual performance hypotheses. As described in a later section, the visual behavior of subjects was monitored as they performed the database query tasks. In terms of broad categories, the dependent measures of visual performance included global fixation measures and fixation-dwell sequence measures. The global fixation measures included fixation frequencies, fixation durations, and fixation rates. The fixation-dwell sequence measures included fixation frequencies, fixation durations, dwell times, and relative dwell times on each of the windows on the interface screens.

The following hypotheses were formulated with regard to the global fixation measures. Fixation frequencies and ıates were expected to decrease while fixation durations were expected to increase with decreasing window size. Multiple active windows were expected to produce greater fixation frequencies, lower fixation rates, and longer fixation durations. Finally, fixation frequencies and durations were expected to increase while fixation rates were expected to decrease with query length.

For the fixation-dwell sequence measures, it was presumed that window size, window activity, and query length would have functionally similar effects on fixation frequencies and durations as described for the global fixation measures above. Dwell times and relative dwell times were expected to increase with decreasing window size, and were expected to be greater with multiple active windows. Additionally, dwell times were expected to increase with query length, while relative dwell times were expected to be inversely related to query length.

In terms of visual scan patterns, it was expected that random scanning of menus, as described by Card, would not be identified. It was expected, however, that visual scanning of menus could be described reasonably well as a case of stratified random visual sampling. It was also expected that scanpath characteristics would vary with experimental conditions. In particular, scan patterns were expected to decrease in statistical dependency with increasing query length. In other words, scanpaths were expected to become relatively more random as a function of increasing query length. Additionally, scan patterns were expected to exhibit relatively greater statistical independence with multiple active windows. That is, scanpaths were expected to be relatively more random or unconstrained with the use of multiple active windows.

# METHOD

## Experimental Design

The experiment was conducted as a 3 X 2 X 3 mixed three-factor design. Window size was manipulated as a between-subject variable with three levels representing 4, 8, and 16 visible menu items within each of the three parse windows. Window activity was manipulated as a within-subject variable with two levels representing windows active one at a time versus windows simultaneously active. Query length was also manipulated as a within-subject variable with three levels representing queries for information on one, two, and three database items. The experimental design is depicted in Figure 1.

Three independent groups were thus presented with three different window sizes. Within the three groups, subjects were exposed to the two levels of window activity in randomly assigned order. Within each window activity condition, subjects were given eight queries to perform with each of the three query lengths in randomized order. Thus, subjects performed 24 queries with windows active one at a time, and 24 queries with windows simultaneously active, resulting in 48 queries per subject on a total of 96 database items.

Dependent measures of query performance included query performance times and error frequencies. Dependent measures of visual performance included fixation frequencies, fixation durations, fixation rates, dwell times, and relative dwell times. Additionally, conditional information metrics were extracted from first-order conditional transition probability matrices. This metric provided a measure of the amount of statistical dependency in the spatial patterns of fixations represented by the transition matrices. Subjective evaluations were also obtained with a rank-order measure and a set of bipolar rating scales.

| | | Within-Groups | | | | | |
|---|---|---|---|---|---|---|---|
| | | Single | | | Multiple | | |
| | | QL 1 | QL 2 | QL 3 | QL 1 | QL 2 | QL 3 |
| Between-Groups | 16 Item | | | | | | |
| | 8 Item | | | | | | |
| | 4 Item | | | | | | |

Figure 1. Experimental design. 16 Item - Sixteen item window size. 8 Item - Eight item window size. 4 Item - Four item window size. Single - Single active window. Multiple - Multiple active windows. QL 1 - Query length of one. QL 2 - Query length of two. QL 3 - Query length of three.

## Subjects

Thirty six subjects were recruited to participate in the experiment through Manpower Temporary Services, a temporary employment agency. Subjects were paid 20 to 32 dollars for participation, depending on their regular rate of pay from the employment agency. There were 24 females and 12 males ranging in age from 21 to 51, with a median age of 30. Overall, the subjects had an average of 9.3 months of computer-related experience. The most common forms of experience among the subjects were data entry and word processing (72 percent of the subjects), followed by text editing (33 percent), programming (19 percent), and other miscellaneous computer-related experiences (17 percent). Twelve subjects were randomly assigned to each of three groups, based on the window size factor.

The subjects in the four-item window size group included eight females and four males ranging in age from 21 to 44, with a median age of 27. This group had an average of 8 months of computer-related experience, and rated themselves in most cases as having used a computer 'frequently' in the past year.

The subjects in the eight-item window size group included eight females and four males ranging in age from 21 to 41, with a median age of 28.5. This group had an average of 13 months of computer-related experience, and rated themselves in most cases as having used a computer 'occasionally' in the past year.

The subjects in the sixteen-item window size group included eight females and four males ranging in age from 24 to 51, with a median age of 38. This group had an average of 8 months of computer-related experience, and rated themselves in most cases as having used a computer 'occasionally' in the past year. A one-way analysis of variance showed that the three groups did not differ in terms of their average amount of computer-related experience ($F(2,33) = 1.6701$, $p = 0.2037$).

## Materials

Query instructions. Subjects were given written query instructions for retrieving information from a database of information about cars. The instructions were written in the recursive form: 'Find information on the <car> (and the <car> (and the <car>))', where one, two or three specific car makes and models were specified. A conjunctive was always included between the nouns on two- and three-item queries, otherwise query instructions were worded without syntactic or semantic variation from the menu items (e.g., see Table 1, Figures 2, 3, and 4).

There were eight one-item queries, eight two-item queries, and eight three-item queries, with the menu items selected for use in the queries in such a way that total search depth increased linearly with query length. There were twelve randomized orders for the queries assigned to each of the twelve subjects within the three groups defined by the window size factor. Each of the query instructions for a set were placed on a separate page and the entire set was placed in a three-ring binder with dividers separating each instruction page. Full listings of the queries and menu items are available in Appendices A and B. An example of the type of information retrievable from the database is shown in Table 2. The complete database is available in Appendices C through E.

Background questionnaire. Subjects were asked to rate how frequently they had worked with a computer in the past year, and were also asked to indicate types and lengths of computer-related experiences (Appendix F). The subjects' sex and age, obtained verbally, was recorded on a separate vision screening form.

Rating and ranking forms. Subjective evaluations of the interfaces were obtained with five-point bipolar rating scales and a ranking form. The rating scale anchor points were: simple - complex, weak - powerful, fatiguing - relaxing, pleasing - irritating, easy to use - hard to use, natural - unnatural, confusing - clear, predictable - unpredictable, meaningless - meaningful, and good - bad (Appendix

TABLE 1

Example Queries for the Three Query Lengths

Query Length One

Find information on the Buick Century.

Find Information on the Dodge Lancer.

Query Length Two

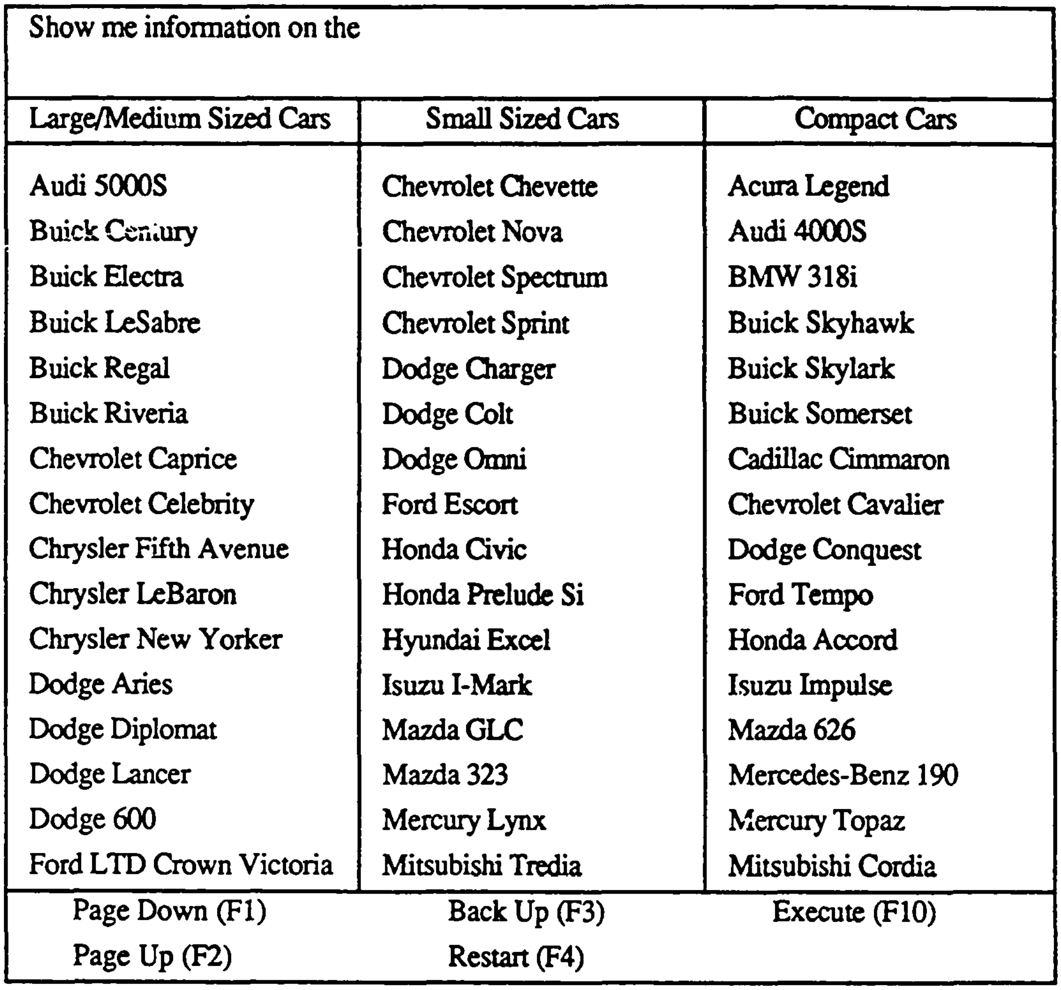
Find information on the Mazda 323 and the Buick Electra.

Find information on the Pontiac 1000 and the Dodge 600.

Query Length Three

Find information on the Mercedes-Benz 300 and the Renault Encore and the Audi 4000S

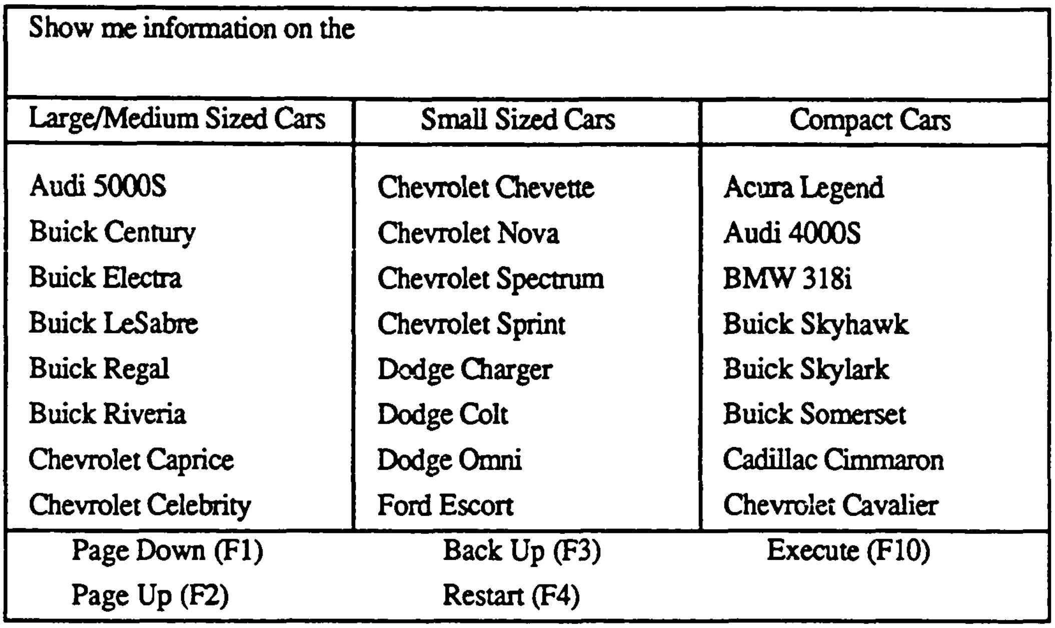
Find information on the Mercury Grand Marquis and the Chevrolet Spectrum and the Pontiac Sunbird.



Show me information on the

| Large/Medium Sized Cars | Small Sized Cars | Compact Cars |
|---|---|---|
| Audi 5000S | Chevrolet Chevette | Acura Legend |
| Buick Century | Chevrolet Nova | Audi 4000S |
| Buick Electra | Chevrolet Spectrum | BMW 318i |
| Buick LeSabre | Chevrolet Sprint | Buick Skyhawk |
| Buick Regal | Dodge Charger | Buick Skylark |
| Buick Riveria | Dodge Colt | Buick Somerset |
| Chevrolet Caprice | Dodge Omni | Cadillac Cimmaron |
| Chevrolet Celebrity | Ford Escort | Chevrolet Cavalier |
| Chrysler Fifth Avenue | Honda Civic | Dodge Conquest |
| Chrysler LeBaron | Honda Prelude Si | Ford Tempo |
| Chrysler New Yorker | Hyundai Excel | Honda Accord |
| Dodge Aries | Isuzu I-Mark | Isuzu Impulse |
| Dodge Diplomat | Mazda GLC | Mazda 626 |
| Dodge Lancer | Mazda 323 | Mercedes-Benz 190 |
| Dodge 600 | Mercury Lynx | Mercury Topaz |
| Ford LTD Crown Victoria | Mitsubishi Tredia | Mitsubishi Cordia |
| Page Down (F1) | Back Up (F3) | Execute (F10) |
| Page Up (F2) | Restart (F4) | |

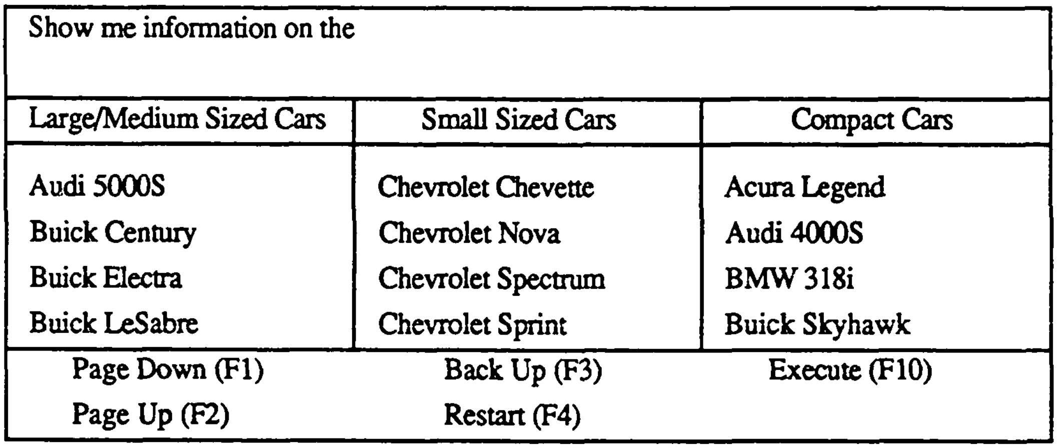
Figure 2. Sixteen item window size interface. Menus are paged up to the top level leaving sixteen items out of view within each window.



Show me information on the

| Large/Medium Sized Cars | Small Sized Cars | Compact Cars |
|---|---|---|
| Audi 5000S | Chevrolet Chevette | Acura Legend |
| Buick Century | Chevrolet Nova | Audi 4000S |
| Buick Electra | Chevrolet Spectrum | BMW 318i |
| Buick LeSabre | Chevrolet Sprint | Buick Skyhawk |
| Buick Regal | Dodge Charger | Buick Skylark |
| Buick Riveria | Dodge Colt | Buick Somerset |
| Chevrolet Caprice | Dodge Omni | Cadillac Cimmaron |
| Chevrolet Celebrity | Ford Escort | Chevrolet Cavalier |
| Page Down (F1) | Back Up (F3) | Execute (F10) |
| Page Up (F2) | Restart (F4) | |

Figure 3. Eight item window size interface. Menus are paged up to the top level leaving twenty four items out of view within each window.



Show me information on the

| Large/Medium Sized Cars | Small Sized Cars | Compact Cars |
|---|---|---|
| Audi 5000S | Chevrolet Chevette | Acura Legend |
| Buick Century | Chevrolet Nova | Audi 4000S |
| Buick Electra | Chevrolet Spectrum | BMW 318i |
| Buick LeSabre | Chevrolet Sprint | Buick Skyhawk |
| Page Down (F1) | Back Up (F3) | Execute (F10) |
| Page Up (F2) | Restart (F4) | |

Figure 4. Four item window size interface. Menus are paged up to the top level leaving twenty eight items out of view within each window.

TABLE 2

Example of Database Information

Audi 5000S

Predicted Reliability - Average. Repair costs are high.

Fuel Economy - Mpg with non-turbo engine and automatic transmission: city, 14; expressway, 28. Gallons used in 15,000 miles, 745. Cruising range, 475 miles.

Comments - The Audi 5000S performs as a European sports sedan should. Seating and ride comfort are good also. Be sure that all the factory recalls relating to the "sudden acceleration runaway" have been performed. Bumper test damage: none.

G). The ranking form simply asked subjects to give a rank of "1" to the system they liked the most and a rank of "2" to the system they liked the least (Appendix H).

Software. The menu-based natural language interfaces were developed using NaturalLink™ (Texas Instruments, 1985a, 1985b, 1985c). NaturalLink™ combines an interactive menu-based system with a semantic grammar analysis approach to natural language processing (where sentences are parsed according to semantic rather than syntactic categories). NaturalLink™ includes interface building utilities, runtime, and linkable object code for creating the MBNL software interface.

The interaction between the user and application software is handled by a window manager, a parser, a translator and a sessioner (driver). The window manager runtime controls the screen displays and returns inputs from the windows when menu items are selected. The parser receives the inputs from menu selections, consults grammar and lexicon files, and builds a parse tree. The parse tree is then passed to the translator when the user completes and enters the query. The translator receives the parse tree, maps it to the elements of the underlying application program, and passes it to the sessioner. As the user builds and executes the queries, the sessioner coordinates the interaction among the parser, translator, and window manager, passing control among these software components and the application. The application ultimately calls the window manager to display the results of the query.

A high level language program, compiled and linked with the NaturalLink™ libraries, is used to make the calls to the NaturalLink runtime software. Microsoft FORTRAN version 4.0 was used for this purpose. In addition to calling the NaturalLink™ software, the FORTRAN program received key codes returned from the window manager and performed DOS time calls with each return. The time-stamped keystrokes were written to a memory buffer, and were then written to disk whenever queries were executed. The resolution of the time measurements was one one-hundredths of a second.

## Equipment

Computer. The computer used for testing was a Texas Instruments Business Pro operating in IBM compatible mode. The system runs at an 8 MHz clock speed and was configured with 640K of memory, a 20 megabyte hard disk, a Wyse monochrome graphics adaptor, and a Wyse 700 monochrome monitor.

Eyetracker. An Applied Science Laboratory Model 1998 Eye View Monitor was used to collect eye point-of-regard data. The system illuminates the left eye of a subject with a collimated, near-infrared light source beamed coaxially with a TV camera (the pupil camera), which produces a backlighted bright pupil and a corneal reflection. The Eye View Monitor identifies the centers of these features and computes direction of gaze, independently of eye translation, by computing the x-y vector distance between the center of the pupil and the center of the corneal reflection. The eye position data are collected at a 60 Hz sampling rate and are accurate to within a one degree radius. A magnetic head tracking system, coupled with automatic pupil camera focusing capability and a servo-controlled tracking mirror, enables the system to track the eye during head movements within one cubic foot of space.

## Procedure

Subjects were told that the purpose of the study was to evaluate the usability of database retrieval systems. A global description of the database was given. Subjects were informed that their eye movements would be monitored as they worked with the systems. The simple explanation was given that the eyetracker was to be used so the experimenter could know where the subject was looking while he or she was working with the systems. A description of the eyetracker was given in general terms.

After the introductory formalities, the subjects were given an informed consent form, followed by the background questionnaire. Tests of acuity, phoria, fusion, stereopsis, and color perception were given with a Keystone VS II vision screener. The subject was then seated at the workstation and a standard calibration procedure was performed with the eyetracker.

Subjects were informed that they were to find information on different cars and then decide if they would consider purchasing the particular cars they read about. Subjects were instructed to say out loud what they thought about each of the cars as they read about them.

Subjects were then given three practice trials with the first system they were to use, with either single or multiple active windows. After the practice trials, subjects were given a set of query instructions. Subjects were instructed to find the information about the cars as quickly and accurately as possible. At the end of the first trial block, subjects were given the rating scales for evaluating the interface, followed by a 20 minute break.

Before the second trial block, subjects were reminded that they were to find information on different cars and then decide if they would consider purchasing the particular cars they read about. Subjects were reminded to say out loud what they thought about each of the cars as they read about them.

Subjects were then given three practice trials with the alternate system. Following the practice trials, subjects were given a set of query instructions. Subject were reminded to find the information about the cars as quickly and accurately as possible. After completing the second trial block, subjects were then given the rating scales for evaluating the second interface. Finally, subjects were asked to rank the two systems they had used in terms of overall preference.

# RESULTS

The effects of window size, window activity, and query length on user performance with the MBNL interface were measured in terms of query performance times and error frequencies. Window activity preferences were assessed with the rank-order preference measure. The subjective evaluations of the interfaces, based on the bipolar rating scale data, were broken down and evaluated by window size and by window activity.

## Query Performance Time

The time-stamped keystroke data were processed by an algorithm which determined the absolute time for each keystroke and then computed the relative start time for each query by reference to eyetrack timing data. Keystroke times for each query were then sequentially differenced and the resulting relative times were then summed. Times for 'back ups' or 'restarts' were then subtracted from the total query construction times to yield the performance times for each query (equal to the sum of the times for each menu selection plus the time for query execution). Query performance times were subsequently analyzed with a 3 X 2 X 3 mixed three-factor analysis of variance (ANOVA), with the between-groups factor representing window size and the within-groups factors representing window activity and query length.

The main effect of window activity was found to be significant (F(1,33) = 5.17, p = 0.0296; Table 3). Queries performed with single active windows were performed faster (20.49 s) than queries with multiple active windows (23.76 s; Figure 5). The main effect of query length was also significant (F(2,66) = 322.73, p = 0.0001; Table 3, Figure 6). Newman-Keuls tests showed that one-item queries were performed faster (10.83 s) than two-item queries (21.17 s) or three-item queries

TABLE 3

ANOVA Summary Table for Query Performance Time

| SOURCE | MS | df | F | P |
|---|---|---|---|---|
| WINDOW SIZE | 32.860 | 2 | 0.23 | 0.7954 |
| SUBJECT(WSIZE) | 142.554 | 33 | . | . |
| | | | | |
| WINDOW ACTIVITY | 578.108 | 1 | 5.17 | 0.0296 * |
| WSIZE x WACT | 66.702 | 2 | 0.60 | 0.5565 |
| WACT x SUBJ(WSIZE) | 111.816 | 33 | . | . |
| | | | | |
| QUERY LENGTH | 10010.123 | 2 | 322.73 | 0.0001 * |
| WSIZE x QL | 25.994 | 4 | 0.84 | 0.5060 |
| QL x SUBJ(WSIZE) | 31.017 | 66 | . | . |
| | | | | |
| WACT x QL | 102.257 | 2 | 3.64 | 0.0317 * |
| WSIZE x WACT x QL | 15.384 | 4 | 0.55 | 0.7016 |
| WACT x QL x SUBJ(WSIZE) | 28.111 | 66 | . | . |

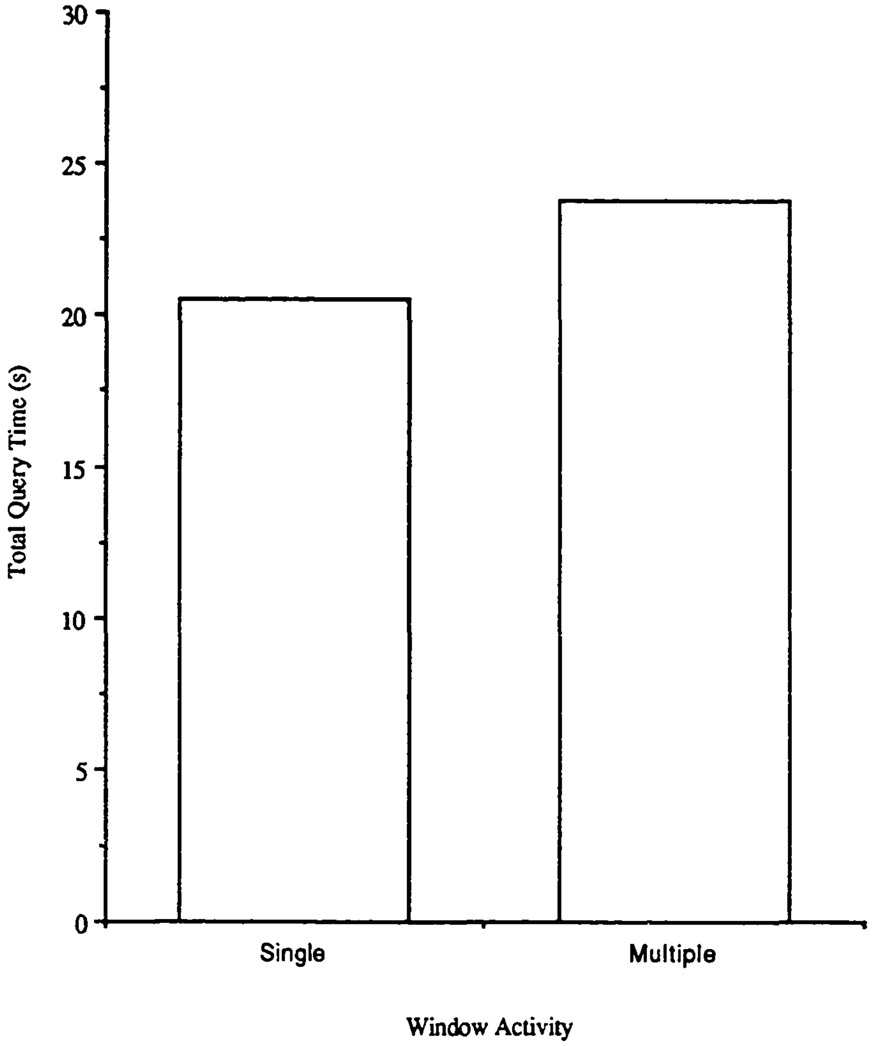


Figure 5. Total query performance time for window activity conditions.

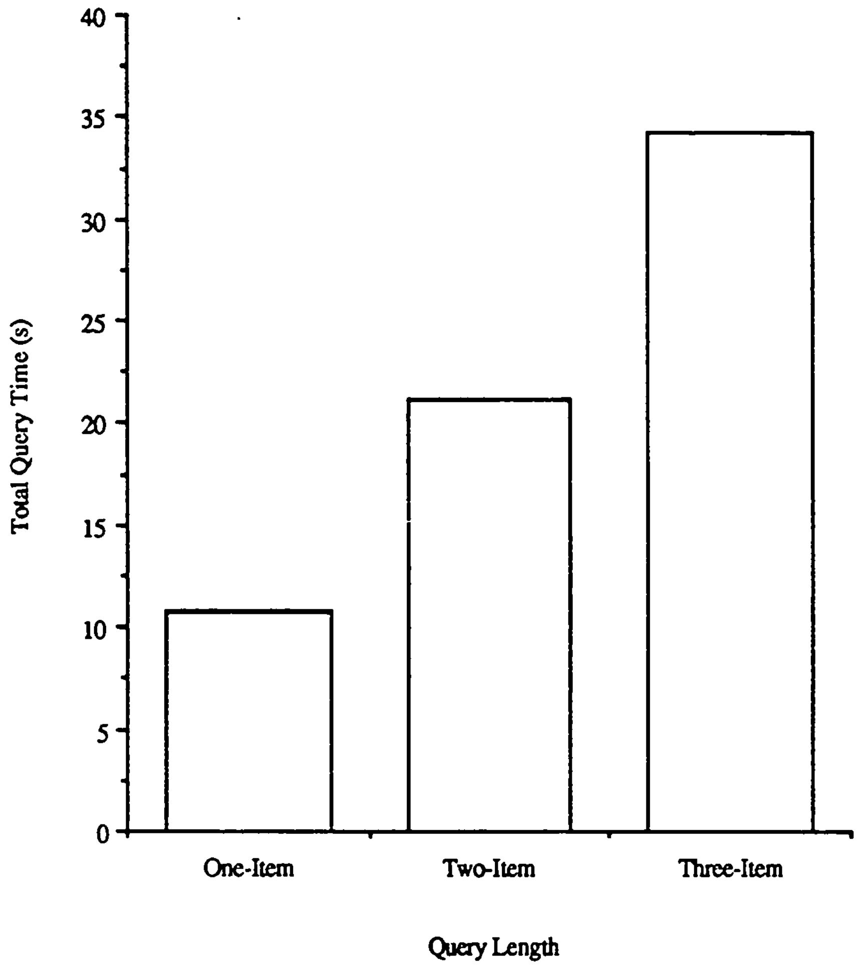


Figure 6. Total query performance time by query length.

(34.36 s), and two-item queries, in turn, were performed faster thaır three-item queries (Table 4).

Finally, there was a significant interaction between window activity and query length (F(2,66) = 3.64, p = 0.0317; Table 3, Figure 7). Newman-Keuls tests showed that mean query performance times for one-item queries did not differ across window activity conditions. However, two-item queries were performed faster with single active windows (19.25 s) than with multiple active windows (23.10 s). Similarly, three-item queries were performed faster with single active windows (31.70 s) than with multiple active windows (37.02 s; Table 5).

## Error Frequencies

There were three types of errors identified from keystroke recordings. Subjects either made an incorrect menu selection, selected too many items, or failed to make a required selection (too few selections). The frequencies of each type of error were collapsed for analysis (Table 6).

Hierarchical log-linear analysis was then used with a stepwise backward elimination method to identify a model that would best fit the error frequency data (Benedetti & Brown, 1978; Marascuilo & Levin, 1983). The three-way interaction between window size, window activity, and query length was excluded from the model, as were all two-way interactions, and the first-order effects for window size and window activity. The 'best' fitting model consequently included only query length as the generating class (Partial Chi-Square = 24.114, df = 2, p < 0.0001; Likelihood Ratio Chi-Square = 16.895, df = 15, p = 0.325; Pearson Chi-Square = 16.286, df = 15, p = 0.363).

Errors frequencies for the three query lengths were subsequently tested for goodness of fit to a uniform distribution. As expected, based on the log-linear analysis, the total Chi-Square was significant (Chi-Square = 25.551, df = 2, p <

TABLE 4

Newman-Keuls Tests on Mean Query Performance Times by Query Length

| Query Length | Mean |
|---|---|
| 1 | 10.835 (A) |
| 2 | 21.171 (B) |
| 3 | 34.359 (C) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Times given in seconds.

TABLE 5

Newman-Keuls Tests on Mean Query Performance Times for Window Activity by Query Length Interaction

| Single Active Windows | | Multiple Active Windows | |
|---|---|---|---|
| Query Length | Mean | Query Length | Mean |
| 1 | 10.508 (A) | 1 | 11.161 (A) |
| 2 | 19.246 | 2 | 23.096 |
| 3 | 31.703 | 3 | 37.016 |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Times given in seconds.

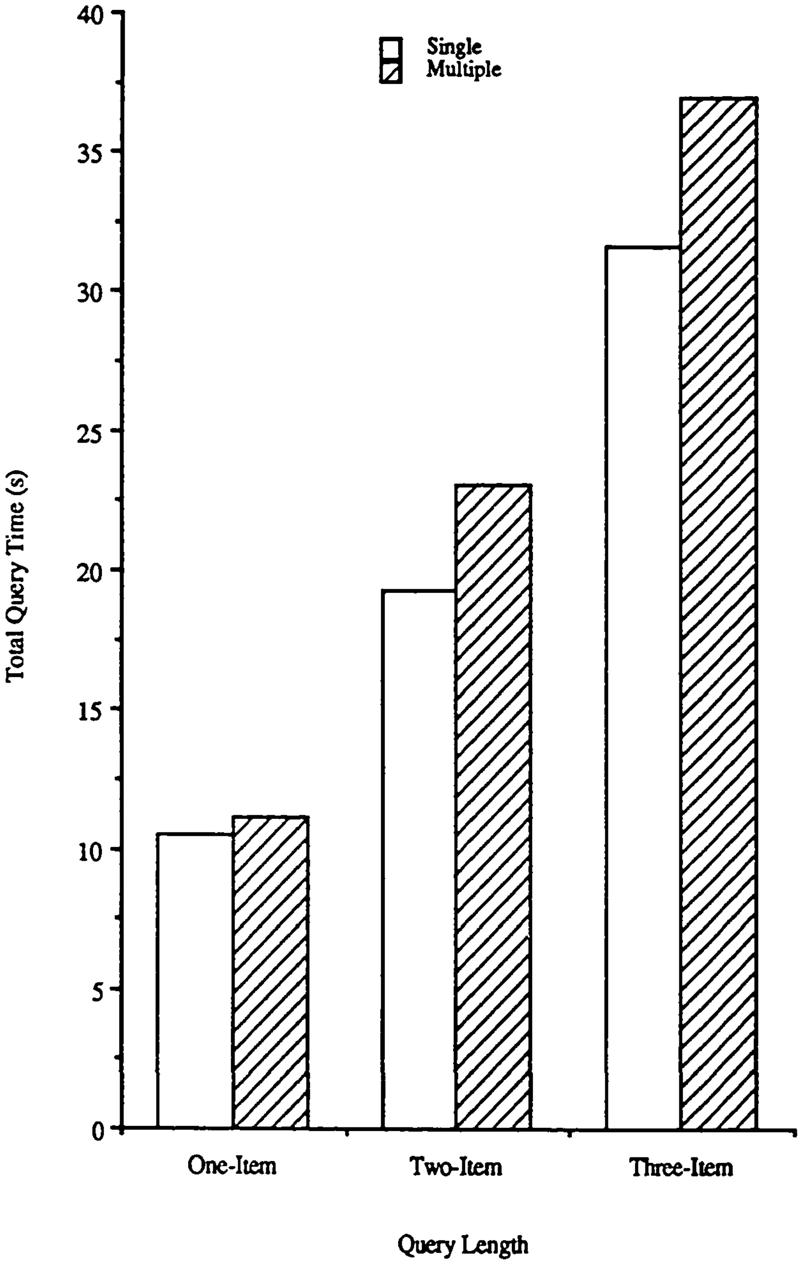


Figure 7. Total query performance time for window activity by query length interaction.

0.001). A Chi-Square goodness-of-fit test for query lengths one and two indicated uniform error frequencies (Chi-Square = 0.857, df = 1, p = 0.355). However, error frequencies for query lengths one and three were not distributed uniformly (Chi-Square = 19.514, df = 1, p < 0.001), nor were error frequencies for query lengths two and three (Chi-Square = 12.800, df = 1, p < 0.001). Thus, error frequencies were found to depend only on query length, and errors were most likely to occur in the longest (three-item) query (Figure 8).

Error Correction Frequencies. There were two means of error correction available to subjects, namely, 'backing up' and deleting the last menu selection from the query under construction and 'restarting' the query altogether. The frequencies for each type of correction were collapsed for analysis (Table 7).

Hierarchical log-linear analysis was then used with a stepwise backward elimination method to identify a model that would best fit the error correction frequency data. In similar fashion to the error frequency data, the three-way interaction between window size, window activity, and query length was excluded from the model, as were all two-way interactions, and the first-order effects for window size and window activity. The 'best' fitting model consequently included only query length as the generating class (Partial Chi-Square = 11.674, df = 2, p = 0.0029; Likelihood Ratio Chi-Square = 18.653, df = 15, p = 0.230; Pearson Chi-Square = 15.948, df = 15, p = 0.386).

Error correction frequencies for each of the three query lengths were subsequently tested for goodness-of-fit to a uniform distribution. The total Chi-Square, as expected from the log-linear analysis, was significant (Chi-Square = 12.531, df = 2, p = 0.002). A Chi-Square test for query lengths one and two indicated uniform error correction frequencies (Chi-Square = 0.048, df = 1, p = 0.827). However, error correction frequencies for query lengths one and three were

TABLE 6

Error Frequencies for Each Cell in the Experimental Design

| | | Single | Multiple |
|---|---|---|---|
| | QL 1 | 4 | 1 |
| Sixteen-Item Window | QL 2 | 5 | 4 |
| | QL 3 | 8 | 7 |
| | QL 1 | 7 | 1 |
| Eight-Item Window | QL 2 | 6 | 5 |
| | QL 3 | 10 | 9 |
| | QL 1 | 1 | 4 |
| Four-Item Window | QL 2 | 1 | 3 |
| | QL 3 | 9 | 13 |

TABLE 7

Error Correction Frequencies for Each Cell in the Experimental Design

| | | Single | Multiple |
|---|---|---|---|
| | QL 1 | 3 | 0 |
| Sixteen-Item Window | QL 2 | 2 | 2 |
| | QL 3 | 3 | 3 |
| | QL 1 | 4 | 0 |
| Eight-Item Window | QL 2 | 2 | 1 |
| | QL 3 | 7 | 4 |
| | QL 1 | 1 | 3 |
| Four-Item Window | QL 2 | 1 | 1 |
| | QL 3 | 4 | 8 |

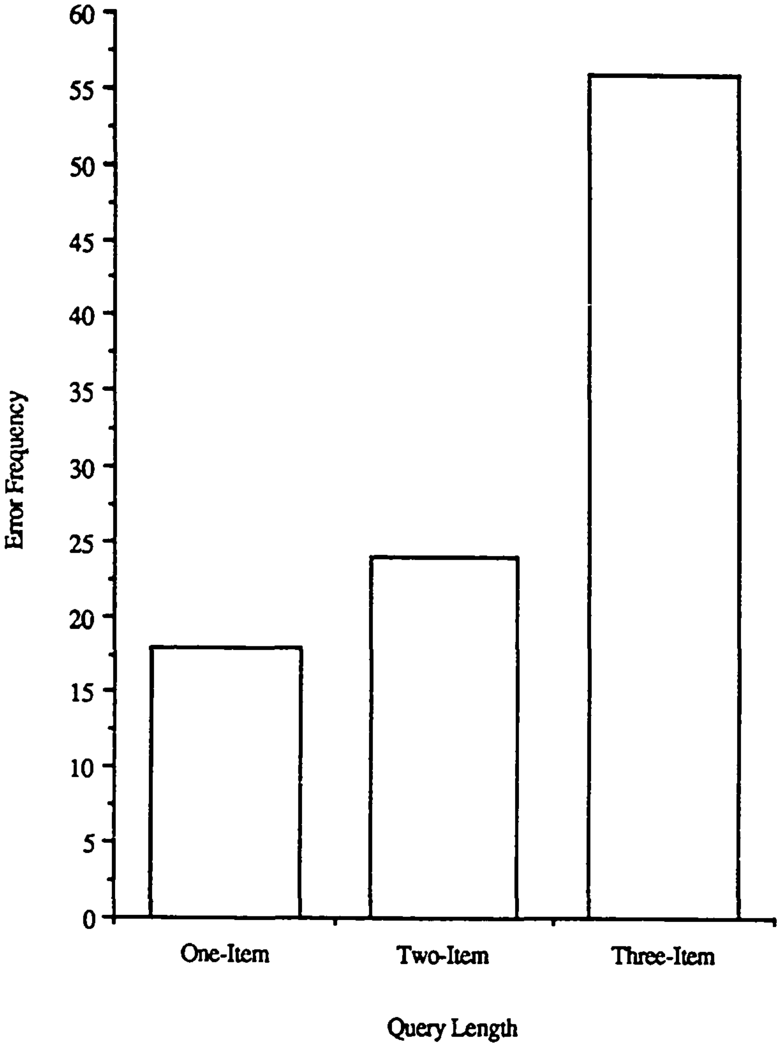


Figure 8. Error frequencies by query length.

not distributed uniformly (Chi-Square = 8.526, df = 1, p = 0.004), nor were error correction frequencies for query lengths two and three (Chi-Square = 7.140, df = 1, p = 0.006). Thus, even though errors were more likely to occur on three-item queries, they also tended to be detected and corrected (Figure 9).

## Subjective Evaluations

Rank-order preferences. The rank-order preference data for single versus multiple active windowing were analyzed with a binomial test. The binomial test result across all 36 subjects failed to reach statistical significance (p = 0.1336). Sixty four percent of the subjects preferred multiple active windows.

Seventy five percent of the subjects who used the system with the sixteen-item window size preferred multiple active windows (p = 0.1460). Fifty eight percent of the subjects who used the system with the eight-item window size preferred multiple active windows (p = 0.7744). Likewise, fifty eight percent of the subjects who used the system with the four-item window size preferred multiple active windows (p = 0.7744).

Overall, then, there was a tendency for subjects to prefer multiple active windows, particularly for subjects with the larger sixteen-item window size. However, the proportion of subjects preferring multiple active windows was not statistically significant in any case.

Rating scales. The bipolar rating scale data were broken down by window size (between groups) and window activity (within groups). The rating scale data were analyzed by window size for each scale dimension with a Kruskal-Wallis One-Way Analysis of Variance by Ranks. No significant differences due to window size were found for any scale dimension (Table 8).

The rating scale data were then analyzed by window activity level for each scale dimension with a Friedman One-Way Analysis of Variance by Ranks (Table 9). A

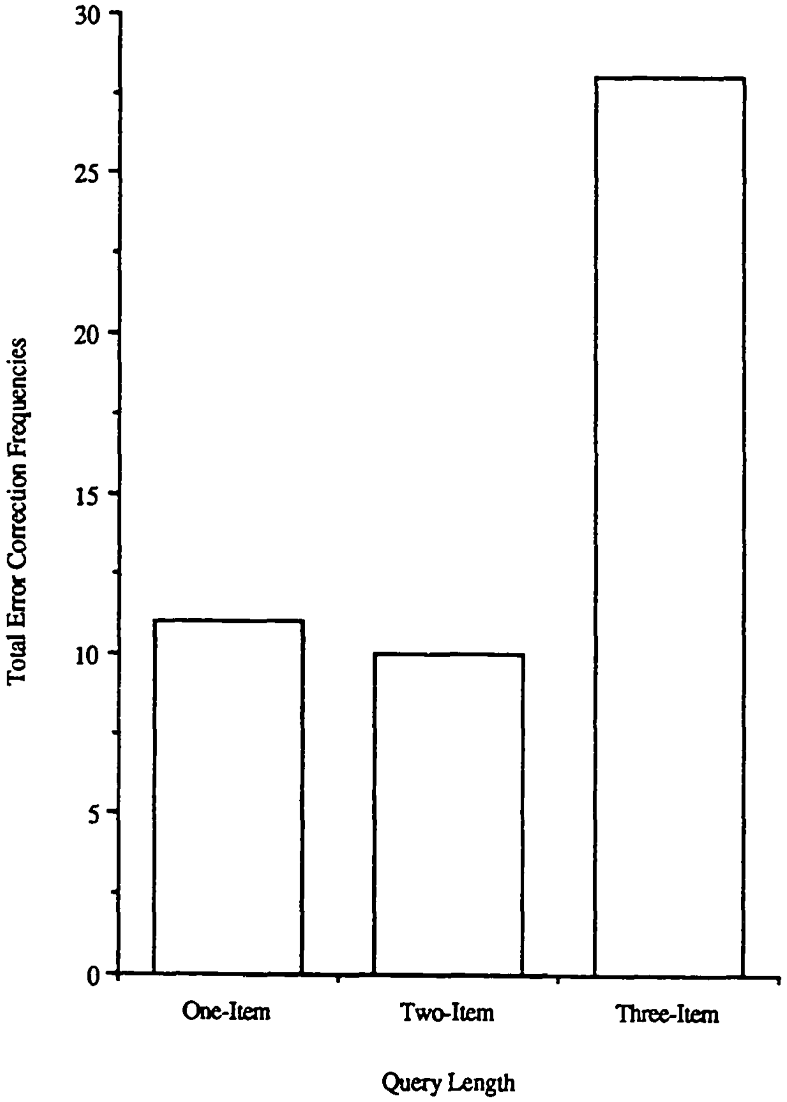


Figure 9. Total error correction frequencies by query length.

TABLE 8

Kruskal-Wallis Analyses of Variance by Window Size for Rating Scale Dimensions

| | Mean Ranks | | | | |
|---|---|---|---|---|---|
| Dimension | 16 Item | 8 Item | 4 Item | Chi-Square | P |
| simple | 41.00 | 33.23 | 35.27 | 2.0714 | 0.3550 |
| powerful | 34.31 | 38.44 | 36.75 | 0.5252 | 0.7691 |
| relaxing | 35.06 | 37.40 | 37.04 | 0.1944 | 0.9074 |
| pleasing | 36.54 | 34.81 | 38.15 | 0.3243 | 0.8503 |
| easy to use | 36.63 | 35.08 | 37.79 | 0.2709 | 0.8733 |
| natural | 38.10 | 37.60 | 33.79 | 0.6617 | 0.7813 |
| clear | 38.46 | 36.42 | 34.63 | 0.4922 | 0.7819 |
| predictable | 37.83 | 37.77 | 33.90 | 0.6167 | 0.7347 |
| meaningful | 34.40 | 38.77 | 36.63 | 0.5900 | 0.7445 |
| good | 38.10 | 35.31 | 36.08 | 0.2605 | 0.9779 |

NOTE: The positive scale anchor point is shown for each scale dimension.

TABLE 9

Friedman Analyses of Variance by Window Activity for Rating Scale Dimensions

| | Mean Ranks | | | |
|---|---|---|---|---|
| Dimension | Single Active | Multiple Active | Chi-Square | P |
| simple | 1.49 | 1.51 | 0.0278 | 0.8676 |
| powerful | 1.44 | 1.56 | 0.4444 | 0.5050 |
| relaxing | 1.43 | 1.57 | 0.6944 | 0.4047 |
| pleasing | 1.47 | 1.53 | 0.1111 | 0.7389 |
| easy to use | 1.56 | 1.44 | 0.4444 | 0.5050 |
| natural | 1.33 | 1.67 | 4.0000 | 0.0455 * |
| clear | 1.50 | 1.50 | 0.0000 | 0.9988 |
| predictable | 1.44 | 1.56 | 0.4444 | 0.5050 |
| meaningful | 1.46 | 1.54 | 0.2500 | 0.6171 |
| good | 1.44 | 1.56 | 0.4444 | 0.5050 |

NOTE: The positive scale anchor point is shown for each scale dimension.

rank order difference was found for the natural-unnatural dimension. The multiple active window interface ranked as more natural than the single active window interface (Chi-Square = 4.00, df = 1, $\underline{p}$ = 0.0455).

## Visual Performance

Data smoothing algorithm. The raw x-y eye coordinate data of the subjects were smoothed by fitting a third degree least-squares polynomial through a sliding window of five data points by means of an integer convolution technique (Evans & Gutmann, 1978; Hershey, Zakin & Simha, 1967; Savitsky & Golay, 1964; Steiner, Termonia & Deltour, 1972).

Fixation algorithm. Eye fixations were then identified from the filtered x-y eye coordinate data using a one degree-100 msec operational definition for a fixation. That is, a fixation was identified if the point-of-gaze remained within a one degree by one degree area for at least one hundred milliseconds.

More specifically, a six point sliding window technique with three criteria was used to identify fixations. The algorithm first computed the standard deviations of the x and y cooordinates for the first six data samples. If the standard deviations were less than 0.5 deg, then the means of these points were used as temporary fixation coordinates. If the standard deviations were greater than 0.5 deg, then the six point window was moved up one sample and the calculations were repeated, until six samples passed the 0.5 deg criterion.

Once a fixation start point was identified, the x and y distances of the next data sample from the temporary means were computed. If the distances were less than 1.0 deg, the sample was included in the fixation. If the distances were greater than 1.0 deg, then the next sample was tested. This process was then continued until a measurement sample passed the 1.0 deg criterion or until six sequential samples had been tested.

If one of the measurements did fall within the 1.0 degree criterion, previous samples that did not were tested against a 1.5 deg criterion. All of the samples that passed the 1.0 deg or 1.5 deg criteria were included in the fixation and were used in the final calculation of the fixation coordinates.

If the x and y distances of six sequential samples exceeded the 1.0 deg criterion, then the x and y means of these samples were computed. If these means did not differ from the temporary means by more than 1.0 deg, they were included in the fixation. Otherwise, the fixation was closed at the last acceptable sample.

Blinks, defined as pupil losses of 200 msec or less, were ignored and did not terminate a fixation. Pupil losses for longer than 200 msec did close a fixation at the last acceptable sample.

Visual performance measures. The effects of window size, window activity, and query length on visual performance were assessed in terms of global fixation measures and fixation-dwell sequence measures. The global fixation measures included fixation frequencies, fixation durations, and fixation rates. Fixation frequencies represent a count of the number of eye fixations performed by the subjects in a given experimental condition. Fixation durations represent the length of the fixations in milliseconds, and fixation rates represent the number of fixations performed over a unit of time, in this case, one second. The global fixation measures were computed from fixation data including off-screen fixations (on the keyboard, on the query instructions, etc.).

To derive fixation-dwell sequence measures, the five windows for each interface screen plus an 'off-area' were defined as 'areas of interest'. Then, based on their coordinates, fixations were identified as falling within one of the five windows (or areas of interest), or in the off-area, not within one of the defined areas. A frequency count of the number of fixations falling within each area was then computed along with mean fixation durations, dwell times, and relative dwell times for each area of interest. A dwell time represents the sum of the durations of a contiguous sequence of fixations falling within the same area of interest. A relative dwell time represents the ratio of the dwell time for an area to the total dwell time for all areas.

First-order joint transition probabilities were computed for each subject across the window size by window activity by query length conditions. To provide a measure of the amount of statistical dependency in scanning, the joint probability matrices were then transformed into conditional probability matrices and the total conditional "information" in them was determined. The information metric provided a measure of the amount of statistical dependency in the spatial patterns of fixations represented by the transition matrices (Brillouin, 1962; Ellis & Stark, 1986).

Global fixation measures. To analyze the effects of the experimental factors on general fixational behavior, a 3 X 2 X 3 mixed three-factor multivariate analysis of variance (MANOVA) was performed. The between-groups factor again represented window size and the within-groups factors represented window activity and query length. The dependent measures were the global fixation frequencies, fixation durations, and fixation rates.

The overall effect of the window size factor only bordered on significance (Wilk's Lambda = 0.6864; F(6,62) = 2.14, $p < 0.0613$; Hotelling-Lawley Trace = 0.4480; F(6,60) = 2.24, $p < 0.0513$; Table 10). However, because the effects of window size on visual performance were of theoretical significance, 3 X 2 X 3 mixed three-factor ANOVA's were performed for each of the global fixation measures. The ANOVA for the mean fixation duration measure alone showed a reliable difference across window sizes (F(2,33) = 4.18, $p < 0.0242$; Table 11, Figure 10). Newman-Keuls tests showed that mean fixation durations tended to be longer with the four-item window (401 ms) than with the sixteen-item window (335 ms) or

TABLE 10

MANOVA Summary Table for Global Fixation Measures

| SOURCE | df | F | P |
|---|---|---|---|
| WINDOW SIZE | 6,62 | 2.14 | 0.0613 |
| WINDOW ACTIVITY | 3,31 | 1.37 | 0.2711 |
| WSIZE x WACT | 6,62 | 0.69 | 0.6608 |
| QUERY LENGTH | 6,128 | 42.50 | 0.0001 * |
| WSIZE x QL | 12,169.62 | 0.53 | 0.8903 |
| WACT x QL | 6,128 | 0.50 | 0.8108 |
| WSIZE x WACT x QL | 12,169.62 | 0.84 | 0.6133 |

NOTE: All results are exact F-tests based on Wilk's criterion, except in the cases where df's are approximate, which indicates F approximation based on Wilk's criterion.

TABLE 11

ANOVA Summary Table for Global Fixation Duration Measure

| SOURCE | MS | df | F | P |
|---|---|---|---|---|
| WINDOW SIZE | 114369.950 | 2 | 4.18 | 0.0242 * |
| SUBJECT(WSIZE) | 27388.368 | 33 | . | . |
| | | | | |
| WINDOW ACTIVITY | 593.352 | 1 | 0.35 | 0.5591 |
| WSIZE x WACT | 1267.580 | 2 | 0.75 | 0.4806 |
| WACT x SUBJ(WSIZE) | 1703.898 | 33 | . | . |
| | | | | |
| QUERY LENGTH | 8546.100 | 2 | 7.83 | 0.0001 * |
| WSIZE x QL | 341.401 | 4 | 0.31 | 0.8683 |
| QL x SUBJ(WSIZE) | 1090.858 | 66 | . | . |
| | | | | |
| WACT x QL | 63.598 | 2 | 0.09 | 0.9147 |
| WSIZE x WACT x QL | 144.375 | 4 | 0.20 | 0.9360 |
| WACT x QL x SUBJ(WSIZE) | 712.112 | 66 | . | . |

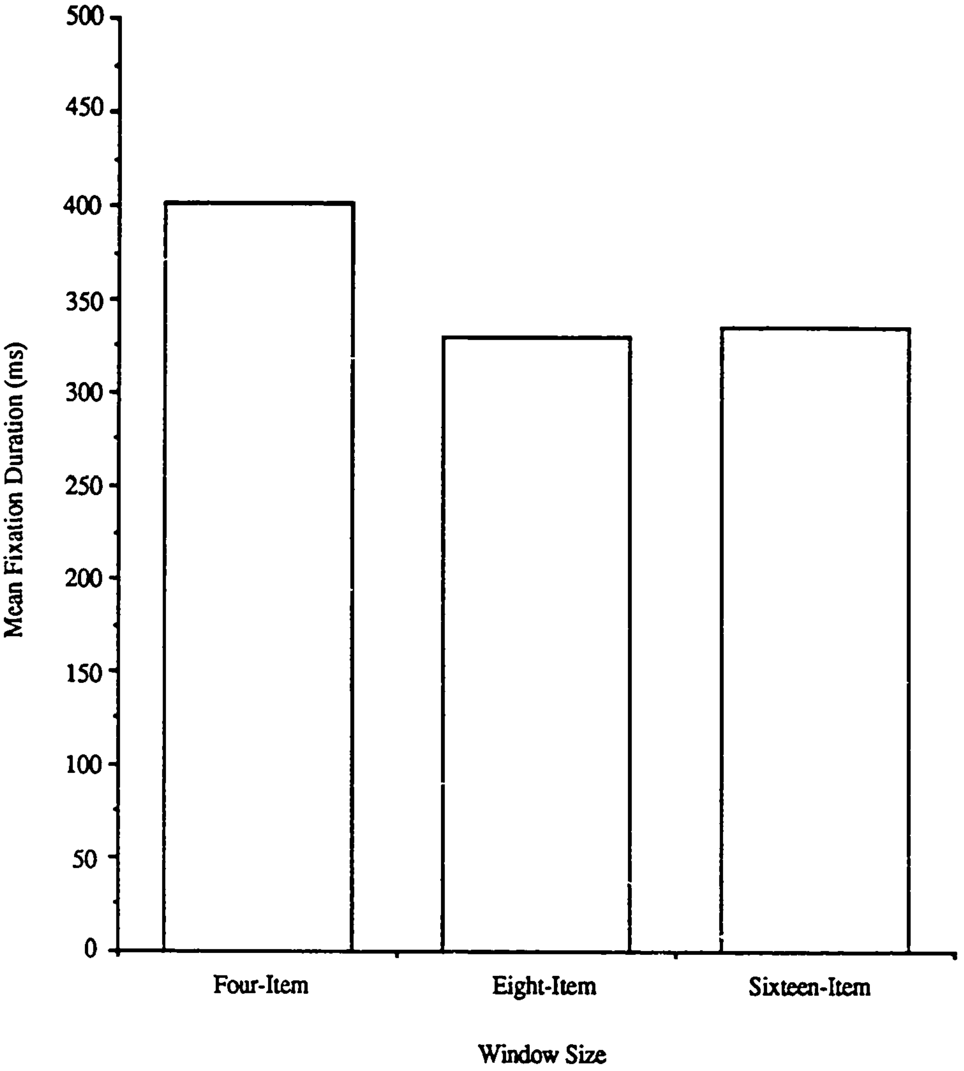


Figure 10. Mean fixation duration by query length.

the eight-item window (330 ms), but mean fixation durations for the sixteen- and eight-item windows did not differ (Table 12).

The overall effect of query length was also significant (Wilk's Lambda = 0.1117; F(6,128) = 42.5, p < 0.0001; Table 10). A 3 X 2 X 3 mixed three-factor ANOVA showed that mean fixation frequencies differed across query lengths (F(2,66) = 233.64, p < 0.0001; Table 13, Figure 11). Newman-Keuls tests showed that there were fewer fixations on one-item queries (n = 17) than on two-item queries (n = 35) or three-item queries (n = 59), and there were fewer fixations, in turn, on two-item queries than on three-item queries (Table 14).

Mean fixation durations also differed reliably across query lengths (F(2,66) = 7.83, p = 0.0009; Table 11, Figure 12). Newman-Keuls tests showed that mean fixation durations were significantly shorter on one-item queries (343 ms) than on two-item queries (358 ms) or three-item queries (364 ms), but mean fixation durations for two- and three-item queries did not differ (Table 15).

Fixation-dwell sequence measures. The fixation-dwell sequence measures were analyzed by means of a 3 X 2 X 3 X 6 mixed four-factor MANOVA, where the between-groups factor was window size, and the within-group factors were window activity, query length, and area of interest. The dependent measures included the fixation-dwell sequence fixation frequencies, fixation durations, dwell times, and relative dwell times. Significant main effects and interactions were found for all of the fixation-dwell sequence measures.

The main effect of window size was found to be significant (Wilk's Lambda = 0.5531; F(8,60) = 2.58, p = 0.0170; Table 16). A 3 X 2 X 3 X 6 mixed four-factor ANOVA revealed that mean fixation durations differed across window sizes (F(2,33) = 4.52, p = 0.0184; Table 17, Figure 13). Newman-Keuls tests showed that the mean fixation duration on the four-item window (344 ms) was significantly greater than

TABLE 12

Newman-Keuls Tests on Global Mean Fixation Durations by Window Size

| Window Size | Mean |
|---|---|
| Eight-Item | 330 (A) |
| Sixteen-Item | 335 (A) |
| Four-Item | 401 |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.05$). Times given in milliseconds.

TABLE 13

ANOVA Summary Table for Global Fixation Frequency Measure

| SOURCE | MS | df | F | P |
|---|---|---|---|---|
| WINDOW SIZE | 369.063 | 2 | 1.00 | 0.3803 |
| SUBJECT(WSIZE) | 370.699 | 33 | . | . |
| WINDOW ACTIVITY | 128.025 | 1 | 0.29 | 0.5926 |
| WSIZE x WACT | 109.943 | 2 | 0.25 | 0.7798 |
| WACT x SUBJ(WSIZE) | 438.601 | 33 | . | . |
| QUERY LENGTH | 30835.191 | 2 | 233.64 | 0.0001 * |
| WSIZE x QL | 91.986 | 4 | 0.70 | 0.5967 |
| QL x SUBJ(WSIZE) | 131.975 | 66 | . | . |
| WACT x QL | 7.514 | 2 | 0.08 | 0.9209 |
| WSIZE x WACT x QL | 57.051 | 4 | 0.63 | 0.6455 |
| WACT x QL x SUBJ(WSIZE) | 91.102 | 66 | . | . |

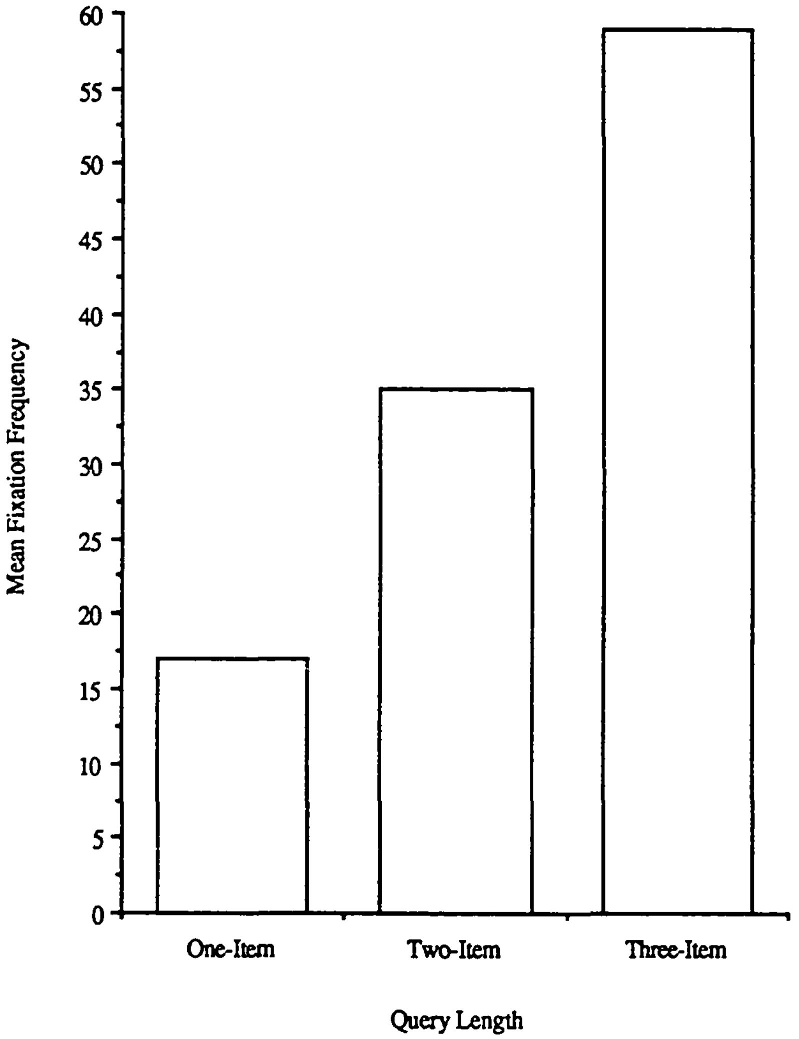


Figure 11. Mean fixation frequency by query length.

TABLE 14

Newman-Keuls Tests on Global Mean Fixation Frequencies by Query Length

| Query Length | Mean |
|---|---|
| 1 | 17 (A) |
| 2 | 35 (B) |
| 3 | 59 (C) |

NOTE: Means sharing a common letter in parantheses were not significantly different ($p > 0.01$).

TABLE 15

Newman-Keuls Tests on Global Mean Fixation Durations by Query Length

| Query Length | Mean |
|---|---|
| 1 | 343 |
| 2 | 358 (A) |
| 3 | 364 (A) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Times given in milliseconds.

TABLE 16

MANOVA Summary Table for Fixation-Dwell Sequence Measures

| SOURCE | df | F | P |
|---|---|---|---|
| WINDOW SIZE | 8,60 | 2.58 | 0.0170 * |
| WINDOW ACTIVITY | 4,30 | 2.58 | 0.0572 |
| WSIZE x WACT | 8,60 | 0.98 | 0.4580 |
| QUERY LENGTH | 8,126 | 45.16 | 0.0001 * |
| WSIZE x QL | 16,193.11 | 1.38 | 0.1553 |
| AREA | 20,538.24 | 26.03 | 0.0001 * |
| WSIZE x AREA | 40,616.14 | 1.29 | 0.1125 |
| WACT x QL | 8,126 | 0.85 | 0.5642 |
| WSIZE x WACT x QL | 16,193.11 | 1.00 | 0.4599 |
| WACT x AREA | 20,538.24 | 2.04 | 0.0051 * |
| WSIZE x WACT x AREA | 40,616.14 | 0.89 | 0.6595 |
| QL x AREA | 40,1241.80 | 31.38 | 0.0001 * |
| WSIZE x QL x AREA | 80,1292.40 | 1.22 | 0.0965 |
| WACT x QL x AREA | 40,1241.80 | 1.22 | 0.1697 |
| WSIZE x WACT x QL x AREA | 80,1292.40 | 1.13 | 0.2071 |

NOTE: All results are exact F-tests based on Wilk's criterion, except in the cases where df's are approximate, which indicates F approximations based on Wilk's criterion.

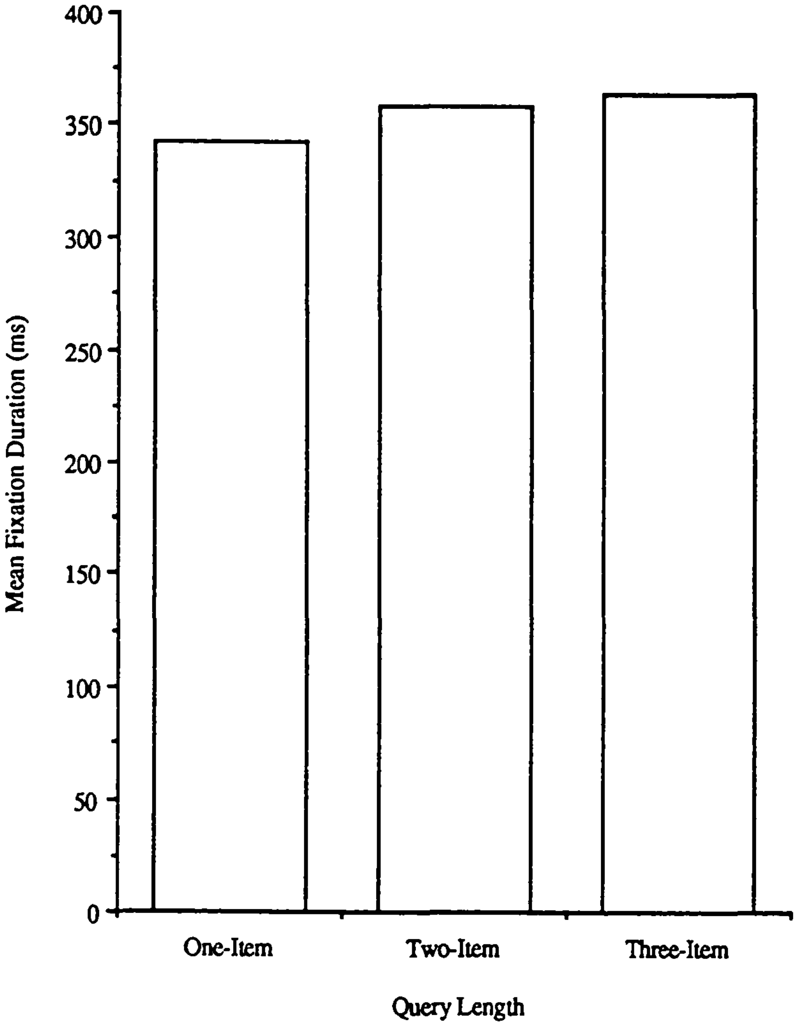


Figure 12. Mean fixation duration by query length.

TABLE 17

ANOVA Summary Table for Fixation-Dwell Sequence Fixation Duration Measure

| SOURCE | MS | df | F | P |
|---|---|---|---|---|
| WINDOW SIZE | 529733.450 | 2 | 4.52 | 0.0148 * |
| SUBJECT(WSIZE) | 117135.130 | 33 | . | . |
| | | | | |
| WINDOW ACTIVITY | 14221.866 | 1 | 0.40 | 0.5319 |
| WSIZE x WACT | 6586.278 | 2 | 0.18 | 0.8321 |
| WACT x SUBJ(WSIZE) | 35639.048 | 33 | . | . |
| | | | | |
| QUERY LENGTH | 970854.900 | 2 | 49.48 | 0.0001 * |
| WSIZE x QL | 5270.241 | 4 | 0.27 | 0.8971 |
| QL x SUBJ(WSIZE) | 19619.297 | 66 | . | . |
| | | | | |
| AREA | 772169.580 | 5 | 17.48 | 0.0001 * |
| WSIZE x AREA | 24863.417 | 10 | 0.56 | 0.8425 |
| AREA x SUBJ(WSIZE) | 44185.156 | 165 | . | . |
| | | | | |
| WACT x QL | 3365.428 | 2 | 0.30 | 0.7451 |
| WSIZE x WACT x QL | 35620.988 | 4 | 3.13 | 0.0203 * |
| WACT x QL x SUBJ(WSIZE) | 11387.306 | 66 | . | . |
| | | | | |
| WACT x AREA | 12822.180 | 5 | 0.72 | 0.6074 |
| WSIZE x WACT x AREA | 12663.772 | 10 | 0.71 | 0.7109 |
| WACT x AREA x SUBJ(WSIZE) | 17746.281 | 165 | . | . |
| | | | | |
| QL x AREA | 326694.220 | 10 | 19.46 | 0.0001 * |
| WSIZE x QL x AREA | 17497.199 | 20 | 1.04 | 0.4114 |
| QL x AREA x SUBJ(WIZE) | 16787.022 | 330 | . | . |
| | | | | |
| WACT x QL x AREA | 10040.619 | 10 | 0.84 | 0.5908 |
| WSIZE x WACT x QL x AREA | 23991.571 | 20 | 2.01 | 0.0069 * |
| WACT x QL x AREA x SUBJ(WSIZE) | 11960.531 | 330 | . | . |

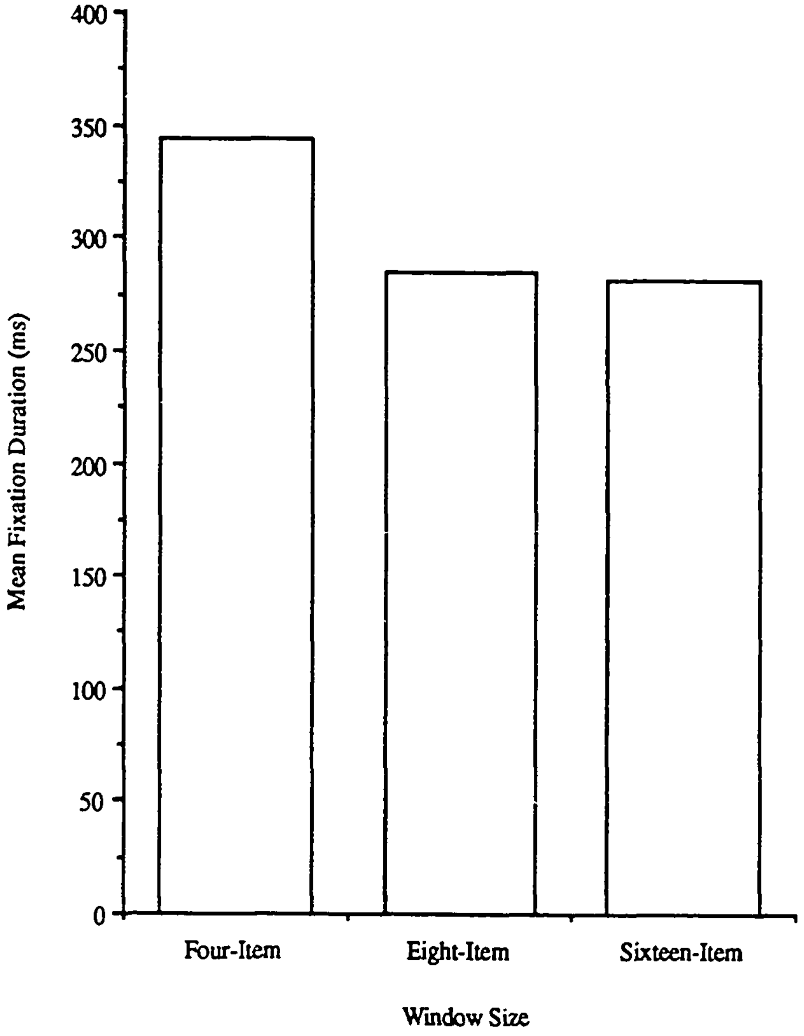


Figure 13. Mean fixation duration by window size.

with the eight-item window (285 ms) or the sixteen-item window (281 ms), but mean fixation durations for the eight- and sixteen-item windows did not differ (Table 18).

The main effect of query length was significant (Wilk's Lambda = 0.0669; F(8,126) = 45.16, p = 0.0001; Table 16). Mean dwell times differed across query lengths (F(2,66) = 173.11, p = 0.0001; Table 19, Figure 14). Mean relative dwell times differed across query lengths (F(2,66) = 57.62, p = 0.0001; Table 20, Figure 15). Mean fixation frequencies differed across query lengths (F(2,66) = 184.53, p = 0.0001; Table 21, Figure 16). Mean fixation durations differed across query lengths (F(2,66) = 49.48, p = 0.0001; Table 17, Figure 17).

Newman-Keuls tests showed that the mean dwell time on a one-item query was shorter (1.188 s) than on a two-item query (2.266 s) or three-item query (3.623 s), and the mean dwell time on a two-item query, in turn, was shorter than on a three-item query (Table 22). Conversely, the mean relative dwell time on a one-item query was greater (19.33 %) than on a two-item query (18.05 %) or three-item query (17.21 %), and the mean relative dwell time on a two-item query, in turn, was greater than on a three-item query (Table 23).

Newman-Keuls tests further showed that the mean number of fixations on a one-item query was less (n = 3.6) than on a two-item query (n = 6.5) or three-item query (n = 10.1), and the mean number of fixations on a two-item query, in turn, was less than on a three-item query (Table 24). Finally, the mean fixation duration on a one-item query was shorter (250 ms) than on a two-item query (319 ms) or three-item query (341 ms), but mean fixation durations on two- and three-item queries did not differ (Table 25).

The main effect for the area of interest factor was significant (Wilk's Lambda = 0.1060; F(20,538.24) = 26.03, p = 0.0001; Table 16). Mean dwell times differed across areas (F(5,165) = 92.83, p = 0.0001; Table 19, Figure 18). Mean relative dwell times differed across areas (F(5,165) = 174.18, p = 0.0001; Table 20, Figure

TABLE 18

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Fixation Durations by Window Size

| Window Size | Mean |
|---|---|
| Sixteen-Item | 281 (A) |
| Eight-Item | 285 (A) |
| Four-Item | 344 |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.05$). Times given in milliseconds.

TABLE 19

ANOVA Summary Table for Fixation-Dwell Sequence Dwell Time Measure

| SOURCE | MS | df | F | P |
|---|---|---|---|---|
| WINDOW SIZE | 9.683 | 2 | 0.69 | 0.5108 |
| SUBJECT(WSIZE) | 14.125 | 33 | . | . |
| WINDOW ACTIVITY | 12.310 | 1 | 1.14 | 0.2925 |
| WSIZE x WACT | 5.719 | 2 | 0.53 | 0.5926 |
| WACT x SUBJ(WSIZE) | 10.759 | 33 | . | . |
| QUERY LENGTH | 642.970 | 2 | 173.11 | 0.0001 * |
| WSIZE x QL | 1.835 | 4 | 0.49 | 0.7400 |
| QL x SUBJ(WSIZE) | 3.714 | 66 | . | . |
| AREA | 563.850 | 5 | 92.83 | 0.0001 * |
| WSIZE x AREA | 4.710 | 10 | 0.78 | 0.6523 |
| AREA x SUBJ(WSIZE) | 6.074 | 165 | . | . |
| WACT x QL | 0.102 | 2 | 0.05 | 0.9534 |
| WSIZE x WACT x QL | 0.929 | 4 | 0.43 | 0.7833 |
| WACT x QL x SUBJ(WSIZE) | 2.139 | 66 | . | . |
| WACT x AREA | 5.022 | 5 | 2.56 | 0.0294 |
| WSIZE x WACT x AREA | 1.466 | 10 | 0.75 | 0.6798 |
| WACT x AREA x SUBJ(WSIZE) | 1.963 | 165 | . | . |
| QL x AREA | 57.546 | 10 | 43.61 | 0.0001 * |
| WSIZE x QL x AREA | 1.641 | 20 | 1.24 | 0.2158 |
| QL x AREA x SUBJ(WIZE) | 1.319 | 330 | . | . |
| WACT x QL x AREA | 0.751 | 10 | 1.08 | 0.3762 |
| WSIZE x WACT x QL x AREA | 0.897 | 20 | 1.29 | 0.1821 |
| WACT x QL x AREA x SUBJ(WSIZE) | 0.695 | 330 | . | . |

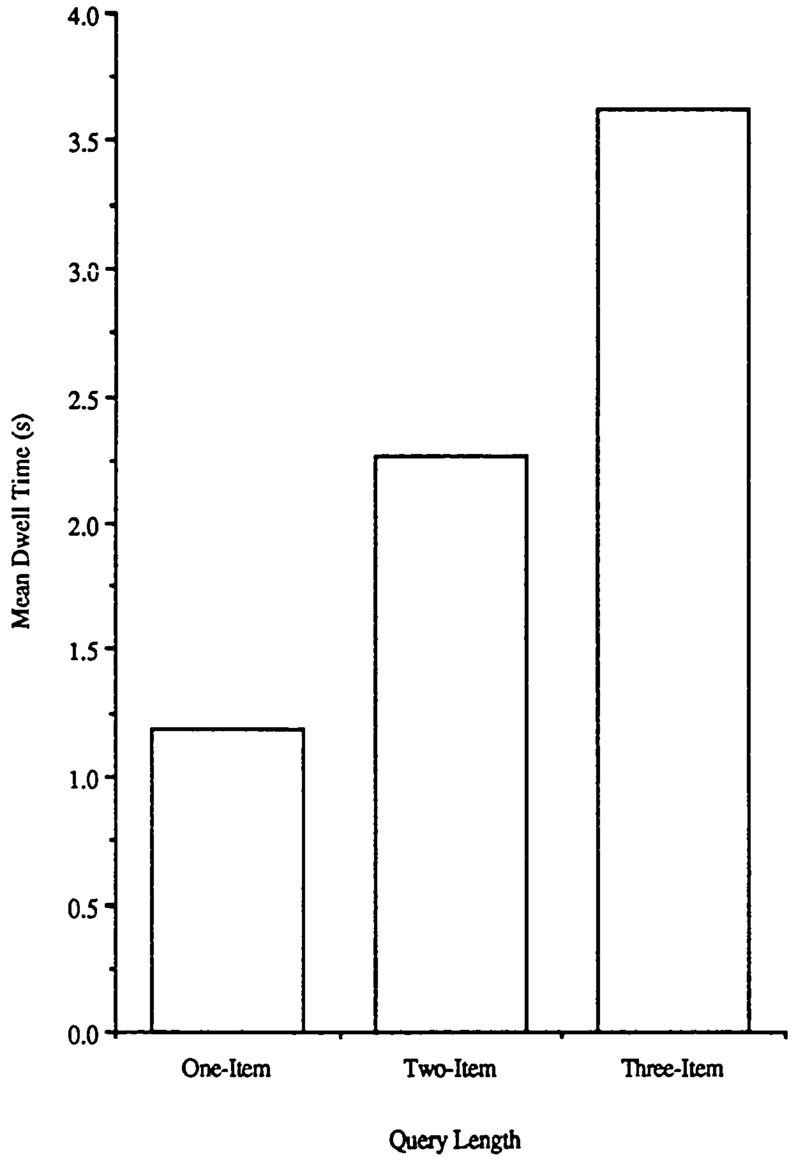


Figure 14. Mean dwell time by query length.

TABLE 20

ANOVA Summary Table for Fixation-Dwell Sequence Relative Dwell Time Measure

| SOURCE | MS | df | F | P |
|---|---|---|---|---|
| WINDOW SIZE | 18.612 | 2 | 0.85 | 0.4353 |
| SUBJECT(WSIZE) | 21.818 | 33 | . | . |
| WINDOW ACTIVITY | 13.894 | 1 | 2.52 | 0.1222 |
| WSIZE x WACT | 15.037 | 2 | 2.72 | 0.0804 |
| WACT x SUBJ(WSIZE) | 5.522 | 33 | . | . |
| QUERY LENGTH | 491.618 | 2 | 57.62 | 0.0001 * |
| WSIZE x QL | 1.535 | 4 | 0.18 | 0.9484 |
| QL x SUBJ(WSIZE) | 8.532 | 66 | . | . |
| AREA | 44912.072 | 5 | 174.18 | 0.0001 * |
| WSIZE x AREA | 249.045 | 10 | 0.97 | 0.4752 |
| AREA x SUBJ(WSIZE) | 257.848 | 165 | . | . |
| WACT x QL | 16.796 | 2 | 2.53 | 0.0871 |
| WSIZE x WACT x QL | 11.360 | 4 | 1.71 | 0.1576 |
| WACT x QL x SUBJ(WSIZE) | 6.631 | 66 | . | . |
| WACT x AREA | 195.789 | 5 | 3.77 | 0.0030 * |
| WSIZE x WACT x AREA | 69.232 | 10 | 1.33 | 0.2172 |
| WACT x AREA x SUBJ(WSIZE) | 51.972 | 165 | . | . |
| QL x AREA | 3819.964 | 10 | 105.32 | 0.0001 * |
| WSIZE x QL x AREA | 23.578 | 20 | 0.65 | 0.8732 |
| QL x AREA x SUBJ(WIZE) | 36.269 | 330 | . | . |
| WACT x QL x AREA | 52.226 | 10 | 2.13 | 0.0219 * |
| WSIZE x WACT x QL x AREA | 25.195 | 20 | 1.03 | 0.4290 |
| WACT x QL x AREA x SUBJ(WSIZE) | 24.526 | 330 | . | . |

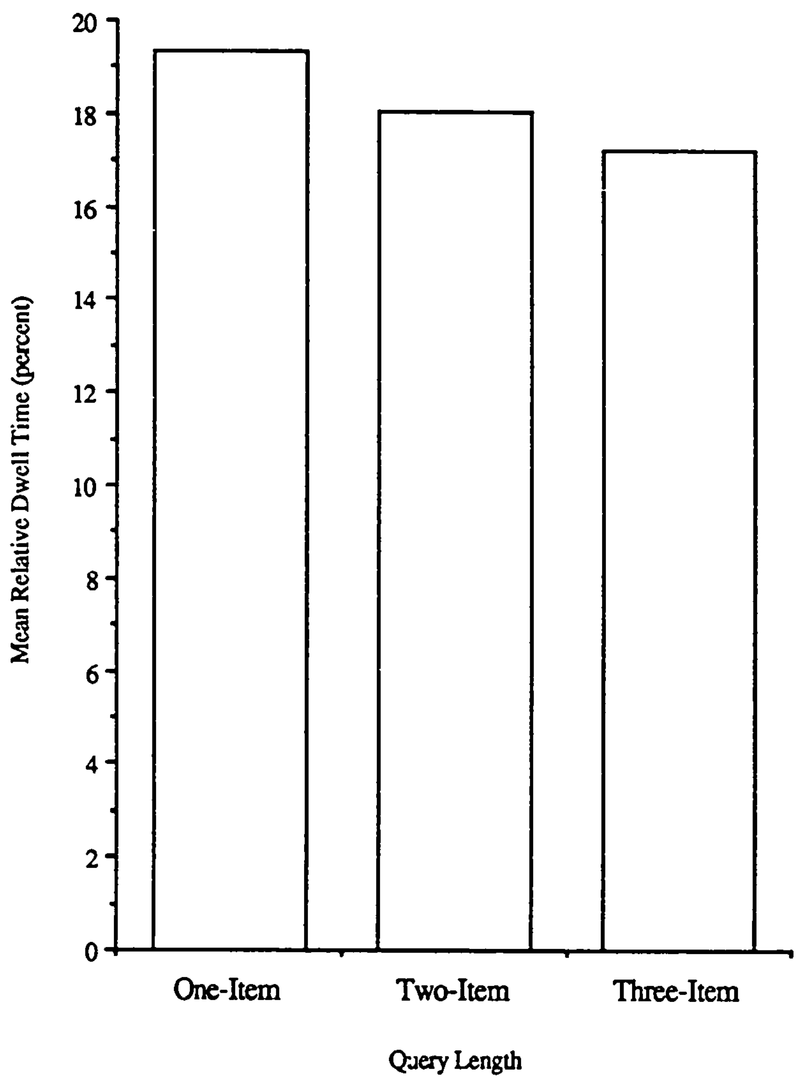


Figure 15. Mean relative dwell time by query length.

TABLE 21

ANOVA Summary Table for Fixation-Dwell Sequence Fixation Frequency Measure

| SOURCE | MS | df | F | P |
|---|---|---|---|---|
| WINDOW SIZE | 48.253 | 2 | 0.72 | 0.4951 |
| SUBJECT(WSIZE) | 67.196 | 33 | . | . |
| WINDOW ACTIVITY | 33.824 | 1 | 0.41 | 0.5267 |
| WSIZE x WACT | 43.676 | 2 | 0.53 | 0.5944 |
| WACT x SUBJ(WSIZE) | 82.632 | 33 | . | . |
| QUERY LENGTH | 4609.091 | 2 | 184.53 | 0.0001 * |
| WSIZE x QL | 23.416 | 4 | 0.94 | 0.4479 |
| QL x SUBJ(WSIZE) | 24.978 | 66 | . | . |
| AREA | 4169.363 | 5 | 119.79 | 0.0001 * |
| WSIZE x AREA | 58.909 | 10 | 1.69 | 0.0862 |
| AREA x SUBJ(WSIZE) | 34.805 | 165 | . | . |
| WACT x QL | 0.359 | 2 | 0.02 | 0.9782 |
| WSIZE x WACT x QL | 8.717 | 4 | 0.54 | 0.7102 |
| WACT x QL x SUBJ(WSIZE) | 16.283 | 66 | . | . |
| WACT x AREA | 22.182 | 5 | 1.73 | 0.1315 |
| WSIZE x WACT x AREA | 4.584 | 10 | 0.36 | 0.9632 |
| WACT x AREA x SUBJ(WSIZE) | 12.859 | 165 | . | . |
| QL x AREA | 377.723 | 10 | 47.76 | 0.0001 * |
| WSIZE x QL x AREA | 11.213 | 20 | 1.42 | 0.1109 |
| QL x AREA x SUBJ(WIZE) | 7.909 | 330 | . | . |
| WACT x QL x AREA | 5.991 | 10 | 1.39 | 0.1842 |
| WSIZE x WACT x QL x AREA | 4.251 | 20 | 0.98 | 0.4802 |
| WACT x QL x AREA x SUBJ(WSIZE) | 4.316 | 330 | . | . |

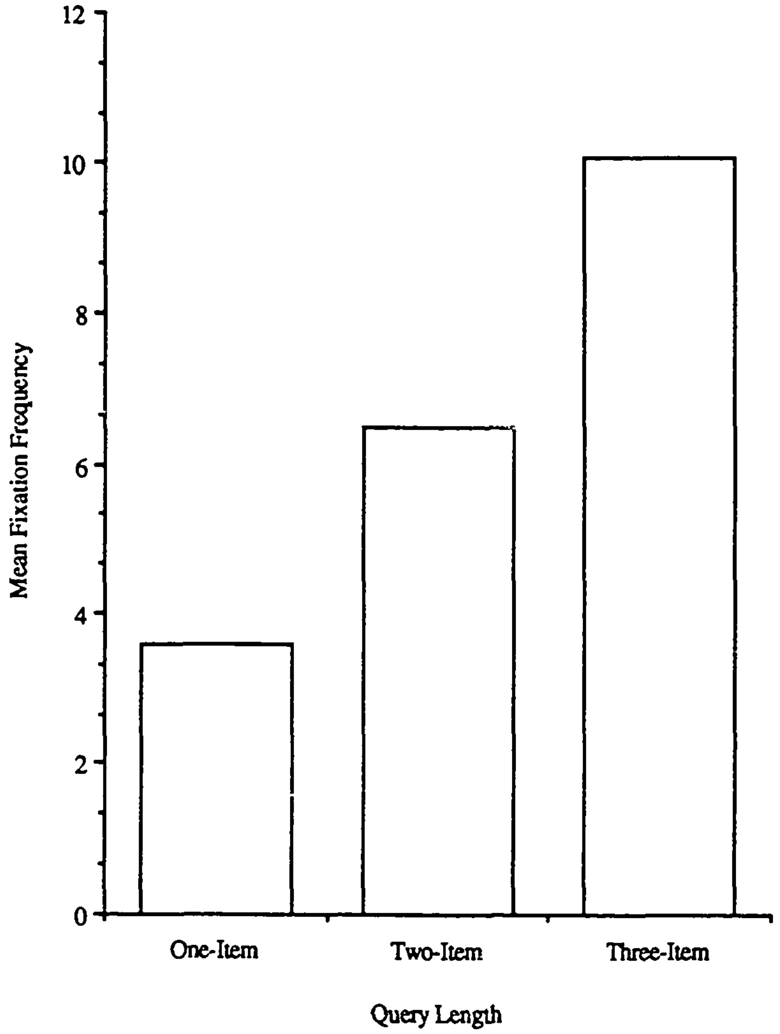


Figure 16. Mean fixation frequency by query length.

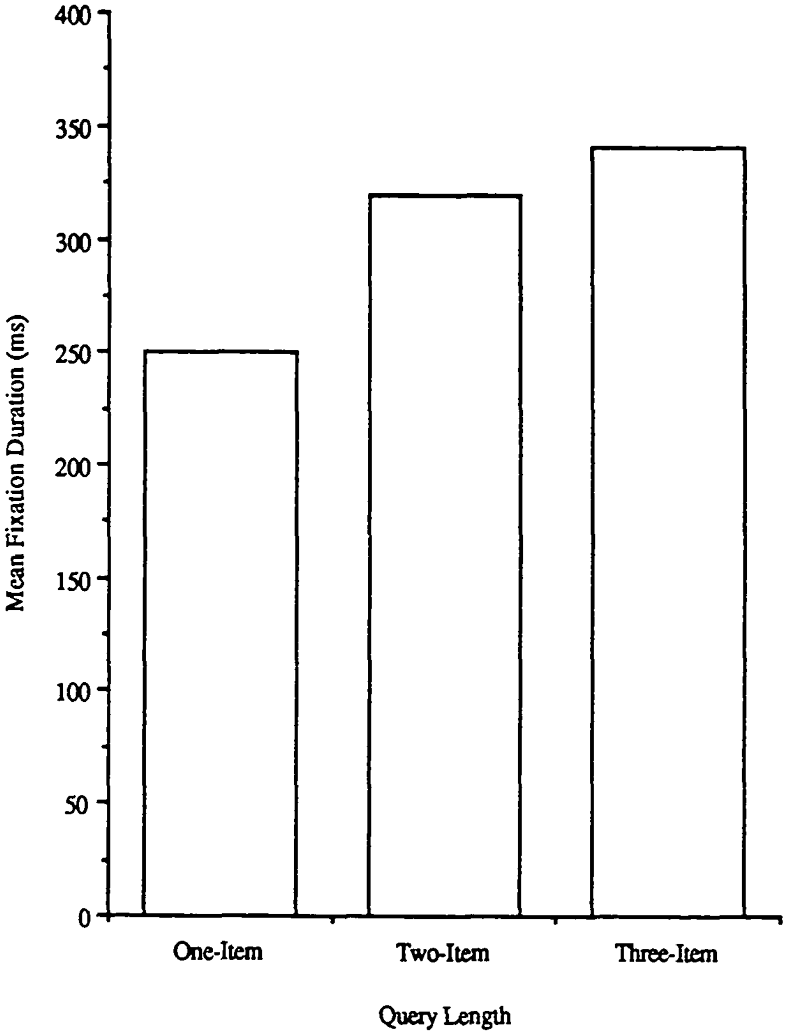


Figure 17. Mean fixation duration by query length.

TABLE 22

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Dwell Times by Query Length

| Query Length | Mean |
|---|---|
| 1 | 1.188 (A) |
| 2 | 2.266 (B) |
| 3 | 3.623 (C) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Times given in seconds.

TABLE 23

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Relative Dwell Times by Query Length

| Query Length | Mean |
|---|---|
| 1 | 19.33 (A) |
| 2 | 18.05 (B) |
| 3 | 17.21 (C) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Relative times given in percentages.

TABLE 24

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Fixation Frequencies by Query Length

| Query Length | Mean |
|---|---|
| 1 | 3.6 (A) |
| 2 | 6.5 (B) |
| 3 | 10.1 (C) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$).

TABLE 25

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Fixation Durations by Query Length

| Query Length | Mean |
|---|---|
| 1 | 250 |
| 2 | 319 (A) |
| 3 | 341 (A) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Times given in milliseconds.

19). Mean fixation frequencies differed across areas (F(5,165) = 119.79, p = 0.0001; Table 21, Figure 20). Mean fixation durations differed across areas (F(5,165) = 17.48, p = 0.0001; Table 17, Figure 21).

Newman-Keuls tests showed that mean dwell times on the results and command windows did not differ. However, all other mean dwell times were significantly different (Table 26). The greatest mean dwell time was in the off-area (4.57 s), followed by the left parse window (3.69 s), middle parse window (2.92 s), right parse window (1.55 s), and then results window (0.72 s) and command window (0.70 s).

Newman-Keuls tests showed that mean relative dwell times for the results window, command window, and right parse window did not differ. However, all other mean relative dwell times were significantly different (Table 27). The greatest mean relative dwell time was in the off-area (39.27 %), followed by the left parse window (31.38 %), middle parse window (19.27 %), and then right parse window (8.15 %), results window (5.75 %), and command window (5.36 %).

Newman-Keuls tests further showed that mean fixation frequencies for the results window, command window, and right parse window did not differ. However, all other mean fixation frequencies were significantly different (Table 28). The greatest mean fixation frequency was in the off-area (n = 12.83), followed by the left parse window (n = 10.47), middle parse window (n = 8.09), and then right parse window (n = 3.96), results window (n = 2.71) and command window (n = 2.30).

Finally, Newman-Keuls tests showed that mean fixation durations for the results window (221 ms) and command window (247 ms) did not differ. Mean fixation durations for the right parse window (301 ms), middle parse window (325 ms), and left parse window (353 ms) also did not differ. Nor did the mean fixation durations for the middle parse window, left parse window and off-area (374 ms; Table 29). In other words, the mean fixation duration for the off-area was significantly greater than for the right parse window, results window, or command window, while mean

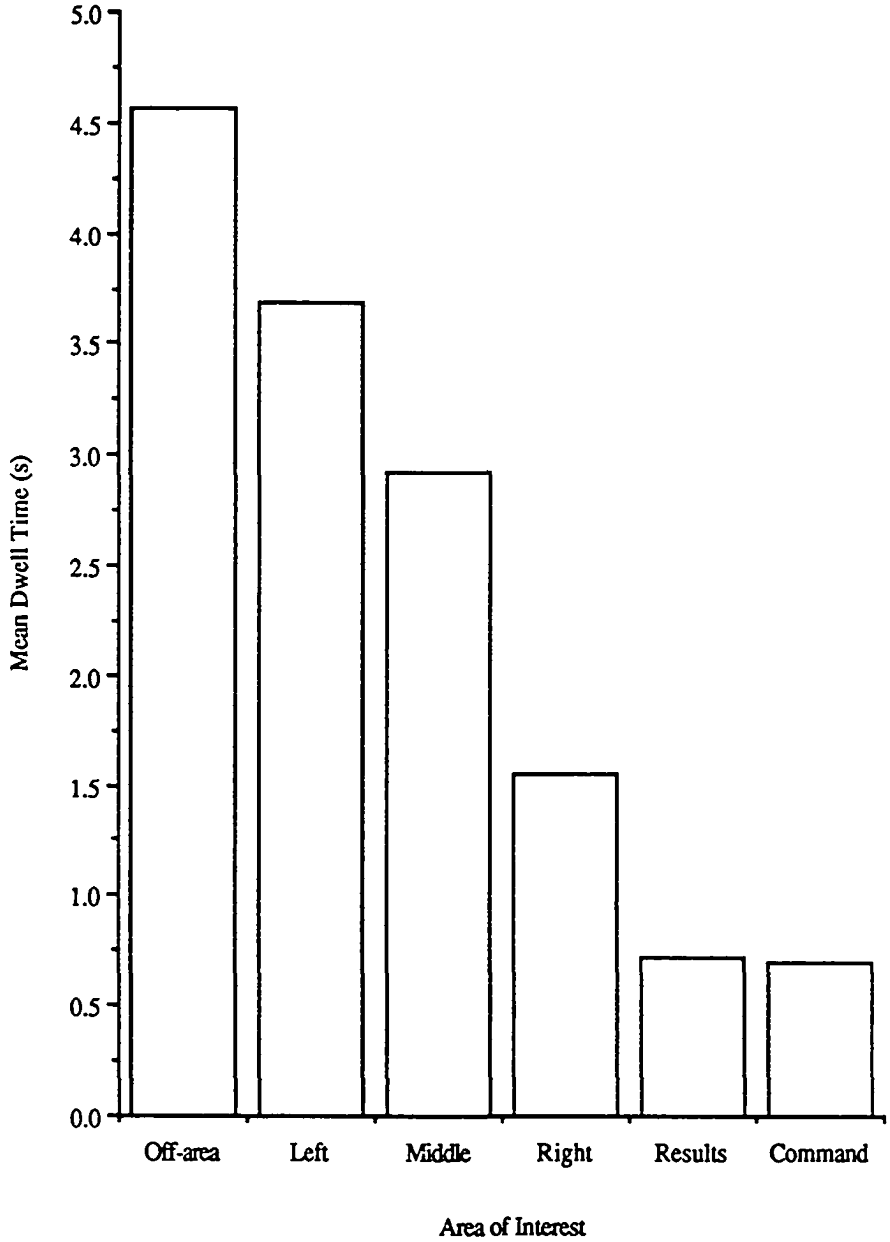


Figure 18. Mean dwell time by area of interest.

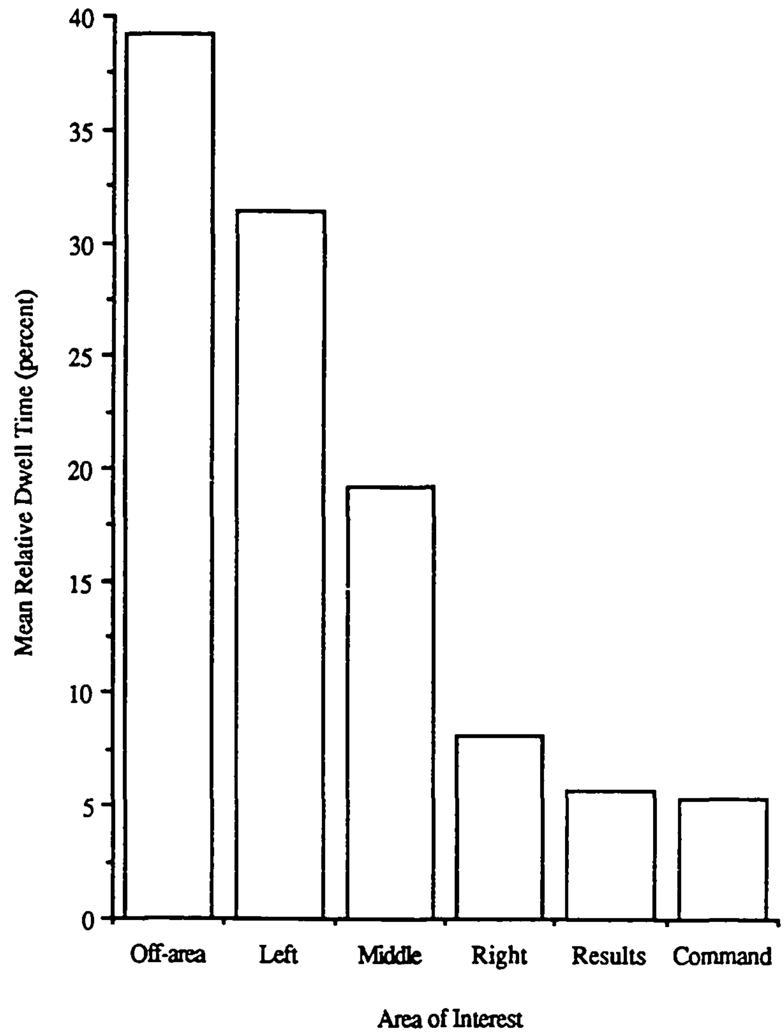


Figure 19. Mean relative dwell time by area of interest.

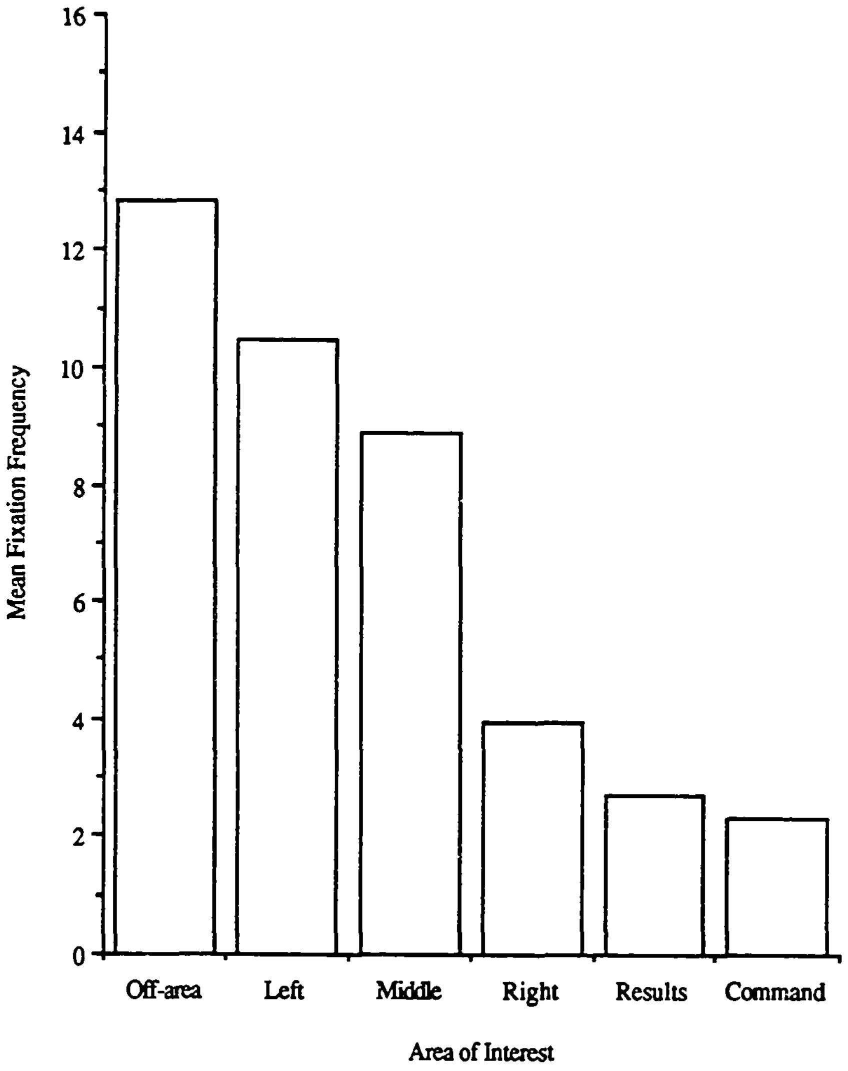


Figure 20. Mean fixation frequency by area of interest.

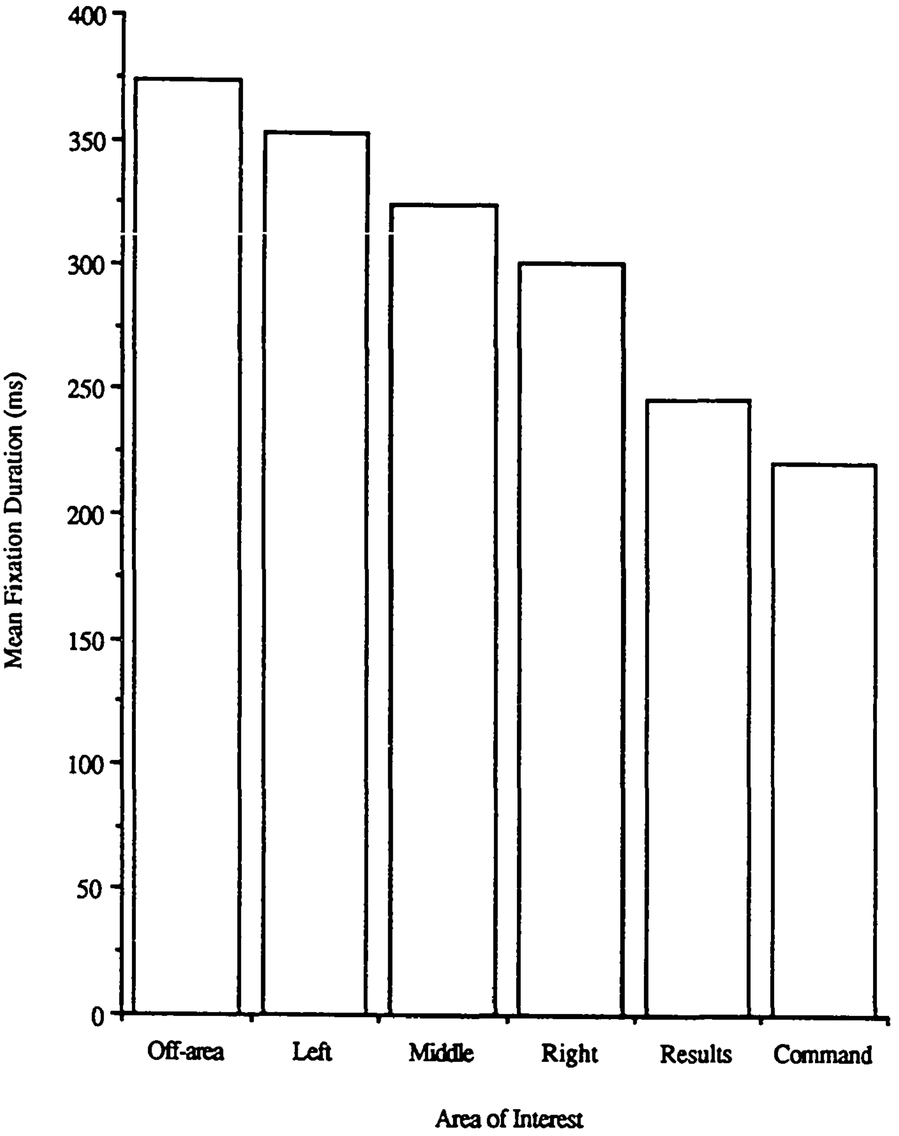


Figure 21. Mean fixation duration by area of interest.

TABLE 26

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Dwell Times by Area of Interest

| Area of Interest | Mean |
|---|---|
| Off-Area | 4.57 |
| Left Parse Window | 3.69 |
| Middle Parse Window | 2.92 |
| Right Parse Window | 1.55 |
| Results Window | 0.72 (A) |
| Command Window | 0.70 (A) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Times given in seconds.

TABLE 27

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Relative Dwell Times by Area of Interest

| Area of Interest | Mean |
|---|---|
| Off-Area | 39.27 |
| Left Parse Window | 31.38 |
| Middle Parse Window | 19.27 |
| Right Parse Window | 8.15 (A) |
| Results Window | 5.75 (A) |
| Command Window | 5.26 (A) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Relative times given in percentages.

TABLE 28

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Fixation Frequencies by Area of Interest

| Area of Interest | Mean |
|---|---|
| Off-Area | 12.83 |
| Left Parse Window | 10.47 |
| Middle Parse Window | 8.09 |
| Right Parse Window | 3.96 (A) |
| Results Window | 2.71 (A) |
| Command Window | 2.30 (A) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$).

TABLE 29

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Fixation Durations by Area of Interest

| Area of Interest | Mean |
|---|---|
| Off-Area | 374.20 (A) |
| Left Parse Window | 352.67 (AB) |
| Middle Parse Window | 324.61 (AB) |
| Right Parse Window | 301.38 (B) |
| Results Window | 246.47 (C) |
| Command Window | 221.30 (C) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Times given in milliseconds.

fixation durations for the results and command windows were significantly shorter than for any other area.

The two-way interaction between window activity and area of interest was also significant (Wilk's Lambda = 0.7852; F(20,538.24) = 2.04, $p$ = 0.0051; Table 16). Results of a 3 X 2 X 3 X 6 mixed-design ANOVA showed that mean dwell times differed as a function of the window activity by area of interest interaction (F(5,165) = 25.109, $p$ < 0.0294; Table 19, Figure 22). Mean relative dwell times also differed as a function of this interaction (F(5,165) = 3.77, $p$ < 0.0030; Table 20, Figure 23).

The most notable result of Newman-Keuls tests on mean dwell times was that mean dwell time on the middle parse window was significantly longer with multiple active windows (3.25 s) than with single active windows (2.60 s; Table 30). With regard to mean relative dwell times, there were two principal results from Newman-Keuls testing (Table 31). First, subjects dwelled relatively longer in the off-area in the single active window condition (40.83 %) than in the multiple active window condition (37.72 %). Second, in the multiple active window condition subjects dwelled relatively longer on the right parse window (8.96 %) than on the results window (5.86 %) or command window (5.62 %), but in the single active window condition this result did hold true (7.35 % for right parse window; 5.88 % for results window; 4.86 % for command window).

Finally, the two-way interaction between query length and area of interest was significant (Wilk's Lambda = 0.0707; F(40,1241.80) = 31.38, $p$ < 0.0001; Table 16). Mean dwell times differed as a function of the interaction between query length and area of interest (F(10,330) = 43.61, $p$ = 0.0001; Table 19, Figure 24), as did mean relative dwell times (F(10,330) = 105.32, $p$ = 0.0001; Table 20, Figure 25), mean fixation frequencies (F(10,330) = 47.76, $p$ = 0.0001; Table 21, Figure 26), and mean fixation durations (F(10,330) = 19.46, $p$ = 0.0001; Table 17, Figure 27).

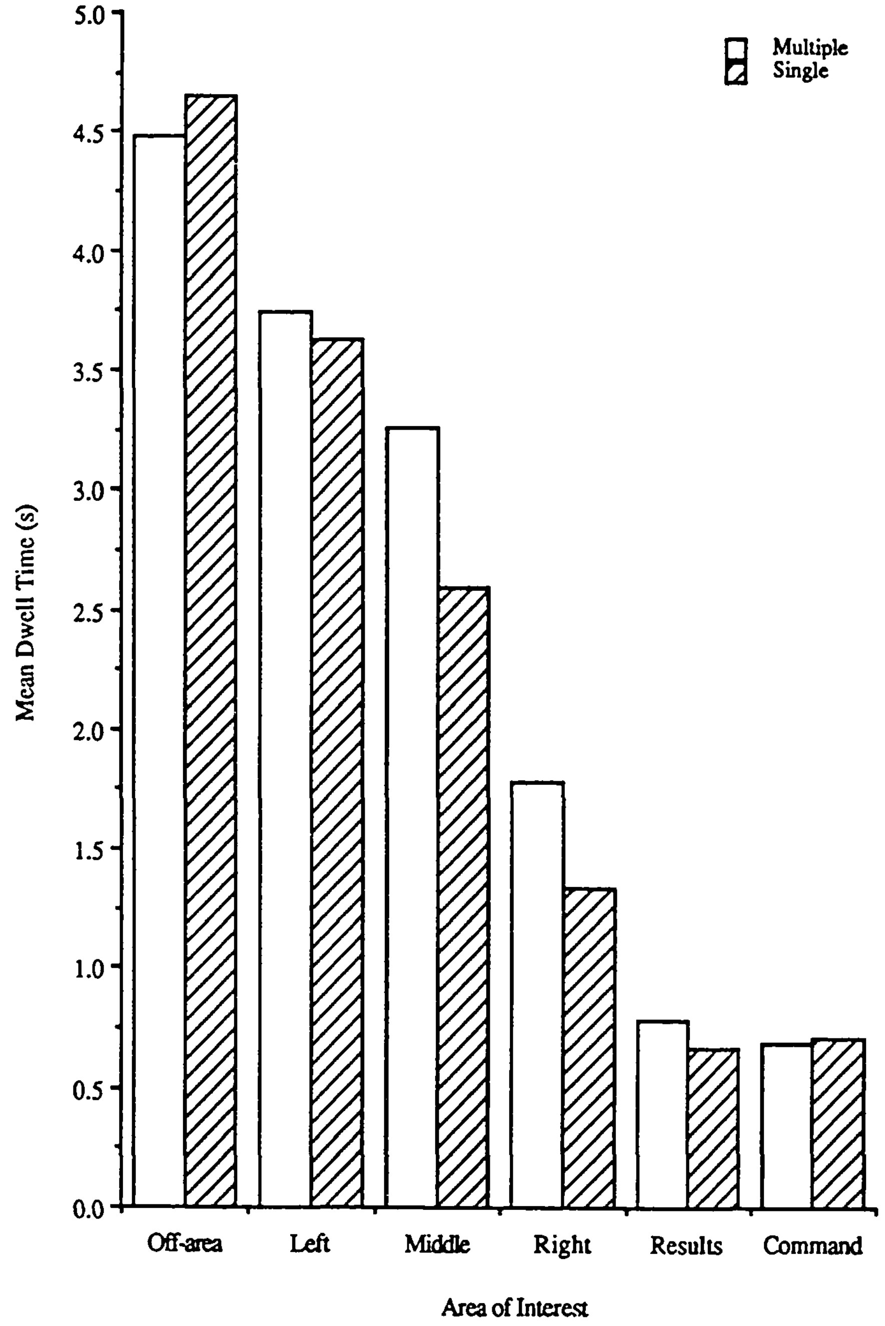


Figure 22. Mean dwell time for window activity by area of interest interaction.

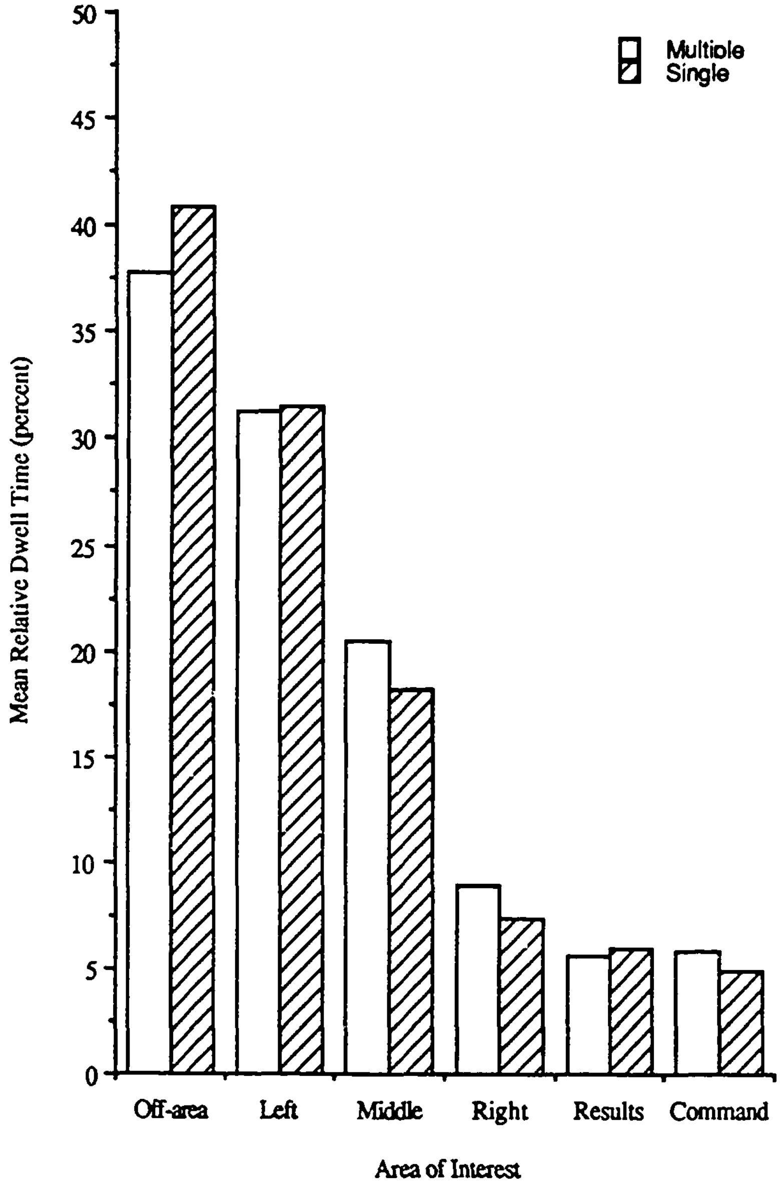


Figure 23. Mean relative dwell time for window activity by area of interest interaction.

TABLE 30

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Dwell Times for Window Activity by Area of Interest Interaction

| Single Active Windows | | Multiple Active Windows | |
|---|---|---|---|
| Area of Interest | Mean | Area of Interest | Mean |
| Off-Area | 4.653 (A) | Off-Area | 4.486 (A) |
| Left Parse Window | 3.628 (B) | Left Parse Window | 3.746 (B) |
| Middle Parse Window | 2.595 | Middle Parse Window | 3.251 (B) |
| Right Parse Window | 1.328 (C) | Right Parse Window | 1.780 (C) |
| Command Window | 0.705 (D) | Results Window | 0.787 (D) |
| Results Window | 0.661 (D) | Command Window | 0.689 (D) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Times given in seconds.

TABLE 31

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Relative Dwell Times for Window Activity by Area of Interest Interaction

| Single Active Windows | | Multiple Active Windows | |
|---|---|---|---|
| Area of Interest | Mean | Area of Interest | Mean |
| Off-Area | 40.83 | Off-Area | 37.72 |
| Left Parse Window | 31.27 (A) | Left Parse Window | 31.48 (A) |
| Middle Parse Window | 18.16 (B) | Middle Parse Window | 20.38 (B) |
| Right Parse Window | 7.35 (CD) | Right Parse Window | 8.96 (C) |
| Results Window | 5.88 (D) | Command Window | 5.86 (D) |
| Command Window | 4.86 (D) | Results Window | 5.62 (D) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Relative times given in percentages.

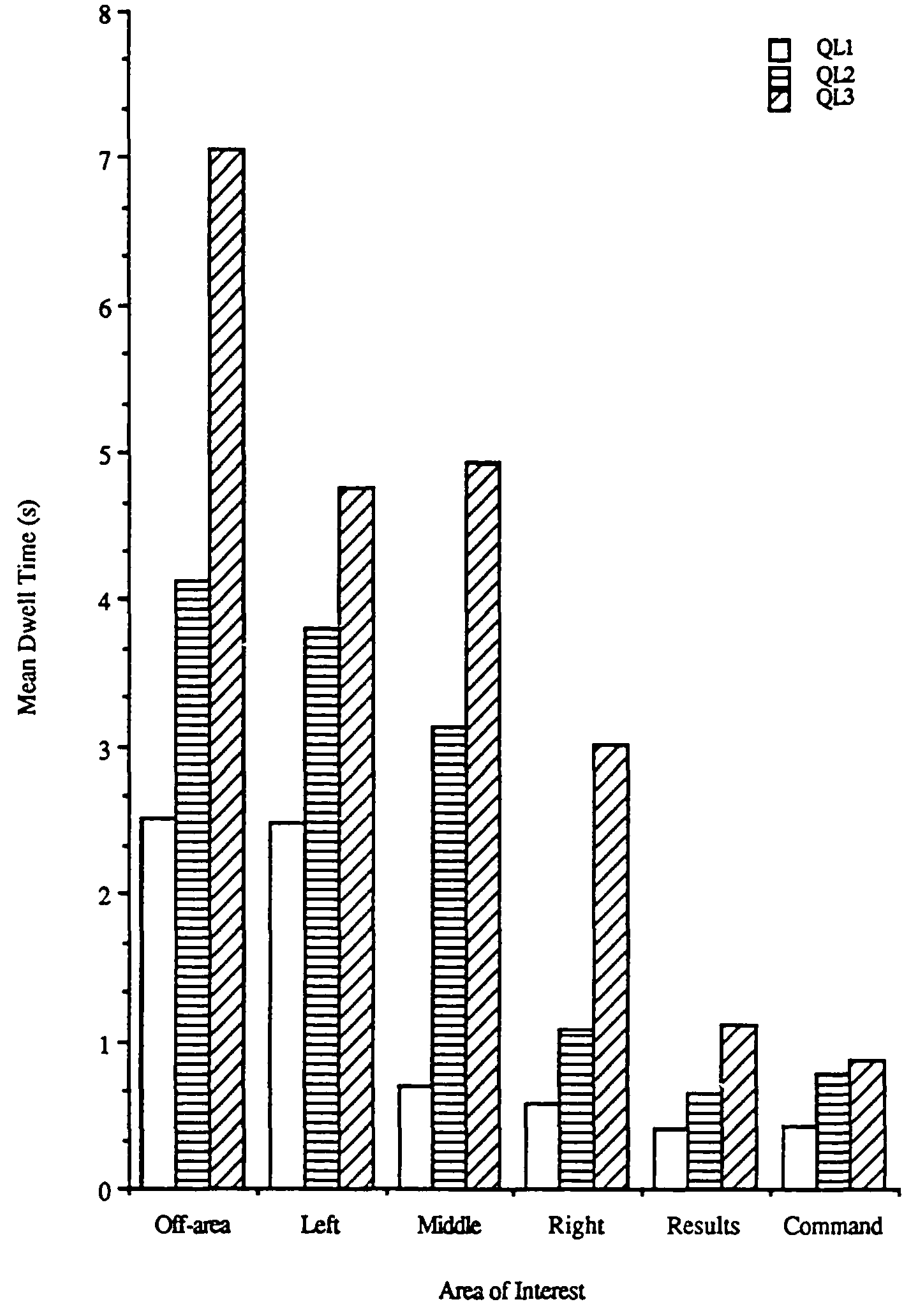


Figure 24. Mean dwell time for query length by area of interest interaction.

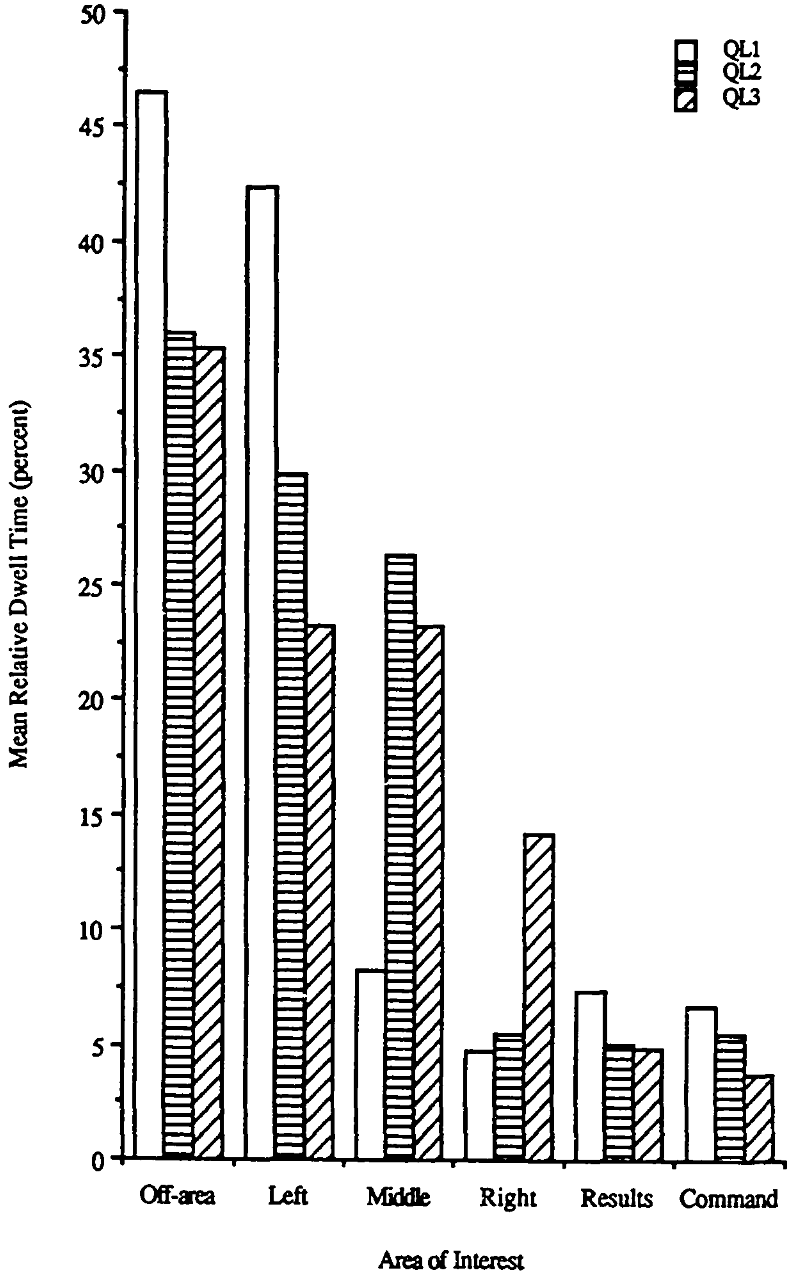


Figure 25. Mean relative dwell time for query length by area of interest interaction.

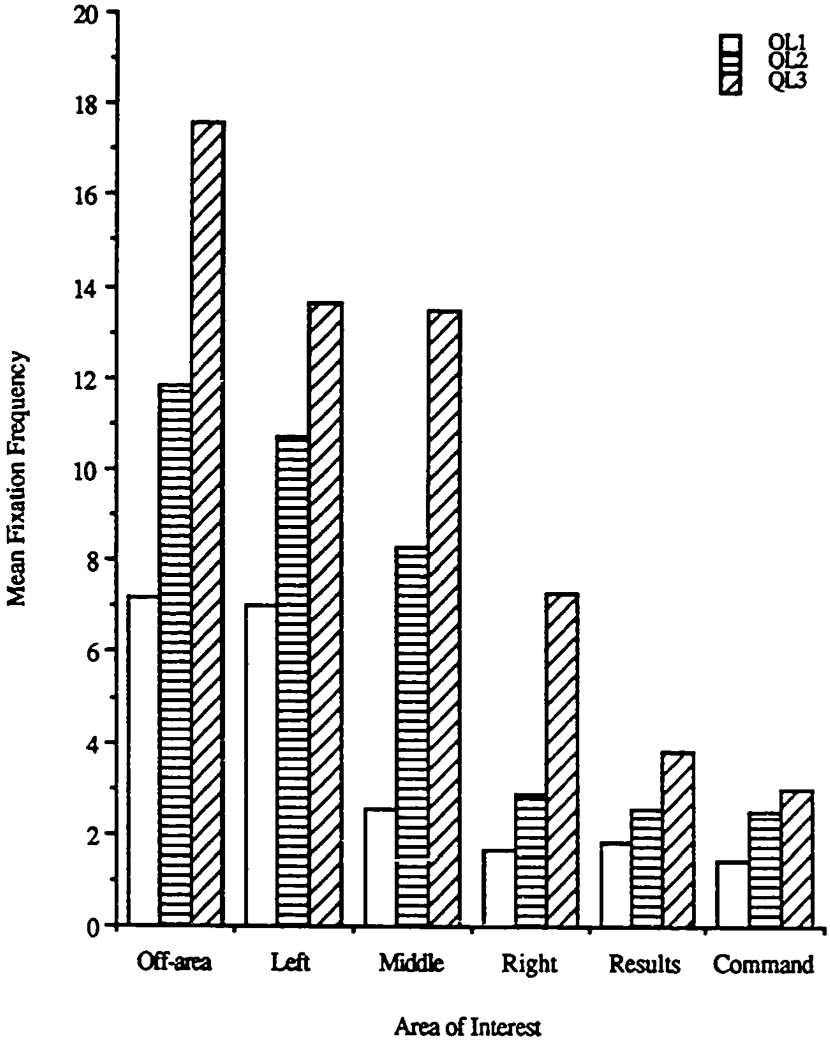


Figure 26. Mean fixation frequencies for query length by area of interest interaction.

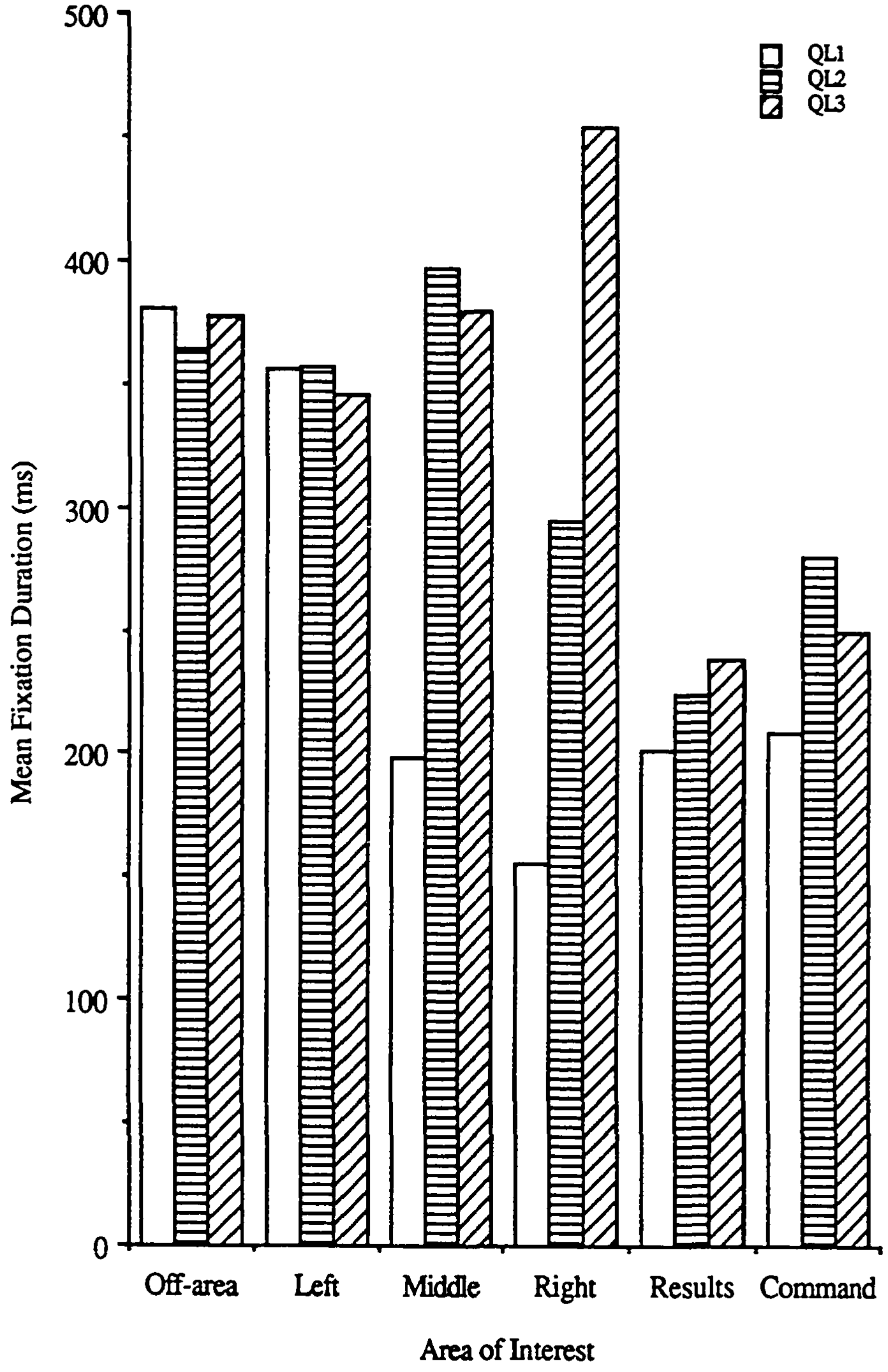


Figure 27. Mean fixation duration for query length by area of interest interaction.

Newman-Keuls tests for mean dwell time differences showed that mean dwell times on the off-area, on the left parse window, and on the middle parse window increased with increasing query length (Table 32). Mean dwell time on the right parse window was also longer on three-item queries than on one- or two-item queries, and mean dwell time on the results window was longer on three-item queries than on one-item queries.

Newman-Keuls tests for mean relative dwell time differences indicated that mean relative dwell times for the off-area and the left parse window increased with decreasing query length (Table 33). Mean relative dwell time for the middle parse window was also greater for two-item queries than for one- or three-item queries, and was greater for three-item queries than for one-item queries. Furthermore, mean relative dwell time for the right parse window was greater for three-item queries than for one- or two-item queries.

Newman-Keuls tests for mean fixation frequency differences revealed that mean fixation frequencies on the off-area, on the left parse window, and on the middle parse window increased with increasing query length (Table 34). Mean fixation frequencies on the right parse window and on the results window were also greater for three-item queries than for one- or two-item queries.

Finally, Newman-Keuls tests for mean fixation duration differences showed that the mean duration of a fixation on the middle parse window was longer on two- and three-item queries than on one-item queries, while the mean duration of a fixation on the right parse window increased with increasing length (Table 35).

Conditional information metrics. If there are statistical dependencies in scanning, areas of interest are not viewed with equal probability (zero-order fixation probabilities are unequal) and transition probabilities between pairs of areas are not equal (first-order joint transition probabilities are unequal). In the most statistically dependent case, only one type of transition occurs from each area of interest. If the

TABLE 32

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Dwell Times for Query Length by Area of Interest Interaction

| Query Length | | | | | |
|---|---|---|---|---|---|
| One-Item | | Two-Item | | Three-Item | |
| Areas | Means | Areas | Means | Areas | Means |
| Off-Area | 2.526 (A) | Off-Area | 4.121 (D) | Off-Area | 7.061 |
| L. Parse | 2.492 (A) | L. Parse | 3.811 (D) | M. Parse | 4.928 (F) |
| M. Parse | 0.699 (BC) | M. Parse | 3.143 (E) | L. Parse | 4.759 (F) |
| R. Parse | 0.583 (BC) | R. Parse | 1.071 (BC) | R. Parse | 3.008 (AE) |
| Cmnd | 0.429 (BC) | Cmnd | 0.792 (BC) | Results | 1.112 (B) |
| Results | 0.403 (C) | Results | 0.657 (BC) | Cmnd | 0.870 (BC) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Times given in seconds.

TABLE 33

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Relative Dwell Times for Query Length by Area of Interest Interaction

| Query Length | | | | | |
|---|---|---|---|---|---|
| One-Item | | Two-Item | | Three-Item | |
| Areas | Means | Areas | Means | Areas | Means |
| Off-Area | 46.51 | Off-Area | 36.00 (C) | Off-Area | 35.31 (C) |
| L. Parse | 42.41 | L. Parse | 29.84 | M. Parse | 23.24 (D) |
| M. Parse | 8.24 (A) | M. Parse | 26.32 | L. Parse | 21.88 (D) |
| Results | 7.36 (AB) | Cmnd | 5.54 (AB) | R. Parse | 14.22 |
| Cmnd | 6.77 (AB) | R. Parse | 5.54 (AB) | Results | 4.85 (AB) |
| R. Parse | 4.70 (AB) | Results | 5.05 (AB) | Cmnd | 3.77 (B) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Relative times given in percentages.

TABLE 34

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Fixation Frequencies for Query Length by Area of Interest Interaction

| Query Length | | | | | |
|---|---|---|---|---|---|
| One-Item | | Two-Item | | Three-Item | |
| Areas | Means | Areas | Means | Areas | Means |
| Off-Area | 7.16 (A) | Off-Area | 11.84 (D) | Off-Area | 19.49 |
| L. Parse | 7.01 (A) | L. Parse | 10.74 (D) | L. Parse | 13.66 (E) |
| M. Parse | 2.54 (BC) | M. Parse | 8.27 (A) | M. Parse | 13.48 (E) |
| Results | 1.80 (C) | R. Parse | 2.88 (BC) | R. Parse | 7.32 (A) |
| R. Parse | 1.67 (C) | Results | 2.53 (BC) | Results | 3.80 (B) |
| Cmnd | 1.45 (C) | Cmnd | 2.48 (BC) | Cmnd | 2.97 (BC) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$).

TABLE 35

Newman-Keuls Tests on Fixation-Dwell Sequence Mean Fixation Durations for Query Length by Area of Interest Interaction

| Query Length | | | | | |
|---|---|---|---|---|---|
| One-Item | | Two-Item | | Three-Item | |
| Areas | Means | Areas | Means | Areas | Means |
| Off-Area | 380.79 (A) | M. Parse | 396.67 (A) | R. Parse | 454.49 |
| L. Parse | 355.49 (ABD) | Off-Area | 364.46 (A) | M. Parse | 378.93 (A) |
| Cmnd | 208.91 (CD) | L. Parse | 356.93 (A) | Off-Area | 377.34 (A) |
| Results | 201.55 (CD) | R. Parse | 293.98 (BF) | L. Parse | 345.59 (AB) |
| M. Parse | 198.23 (CE) | Cmnd | 280.58 (CF) | Cmnd | 249.91 (CF) |
| R. Parse | 155.68 (E) | Results | 223.86 (DF) | Results | 238.50 (EF) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Times given in milliseconds.

transitions from each area all uniquely go to a single area, the information metric will then have a minimum value. Thus, the smaller the information metric, the greater the statistical dependency in scanning. Such statistical dependencies are indicative of a first-order or higher Markov process.

If visual scanning occurs in a stratified random fashion, areas of interest are viewed with unequal probabilities and the transition probabilities between pairs of areas are equal. Transitions between areas of interest are likely to be due in this case simply to the zero-order probability of viewing the respective areas. There are then by chance more transitions between the areas with higher zero-order fixation probabilities. In this case, the probability of fixating on any area of interest is statistically independent of fixation on the preceding area, characteristic of a zero-order Markov process.

If visual scanning occurs in completely random fashion, areas of interest are viewed with equal probability and all transition probabilities between pairs of areas are equal. Scanpaths are then completely unconstrained (and unpredictable). If transitions from each area of interest are equally distributed to all other areas, the information metric will then have a maximum value. Therefore, the larger the information metric, the less the statistical dependency in scanning.

The information metrics extracted from the conditional transition probability matrices of the subjects were used as dependent measures in a 3 X 2 X 3 mixed-design ANOVA. The between-groups factor again represented window size, and the within-groups factors represented window activity and query length.

The results of the ANOVA indicated a significant main effect due to query length ($F(2,66) = 309.19$, $p < 0.0001$; Table 36, Figure 28). The results of Newman-Keuls tests indicated that scan patterns were relatively more statistically dependent on one-item queries (0.915 bits) than on two-item queries (1.149 bits) or three-item queries (1.340 bits). Scanpaths, in turn, were relatively more statistically dependent

TABLE 36

ANOVA Summary Table for Conditional Information Measure

| SOURCE | MS | df | F | P |
|---|---|---|---|---|
| WINDOW SIZE | 0.0446 | 2 | 0.80 | 0.4563 |
| SUBJECT(WSIZE) | 0.0555 | 33 | . | . |
| | | | | |
| WINDOW ACTIVITY | 0.0506 | 1 | 2.78 | 0.1047 |
| WSIZE x WACT | 0.0117 | 2 | 0.64 | 0.5316 |
| WACT x SUBJ(WSIZE) | 0.0182 | 33 | . | . |
| | | | | |
| QUERY LENGTH | 3.2636 | 2 | 309.19 | 0.0001 * |
| WSIZE x QL | 0.0110 | 4 | 1.04 | 0.3919 |
| QL x SUBJ(WSIZE) | 0.0106 | 66 | . | . |
| | | | | |
| WACT x QL | 0.0318 | 2 | 4.83 | 0.0111 * |
| WSIZE x WACT x QL | 0.0070 | 4 | 1.06 | 0.3857 |
| WACT x QL x SUBJ(WSIZE) | 0.0066 | 66 | . | . |

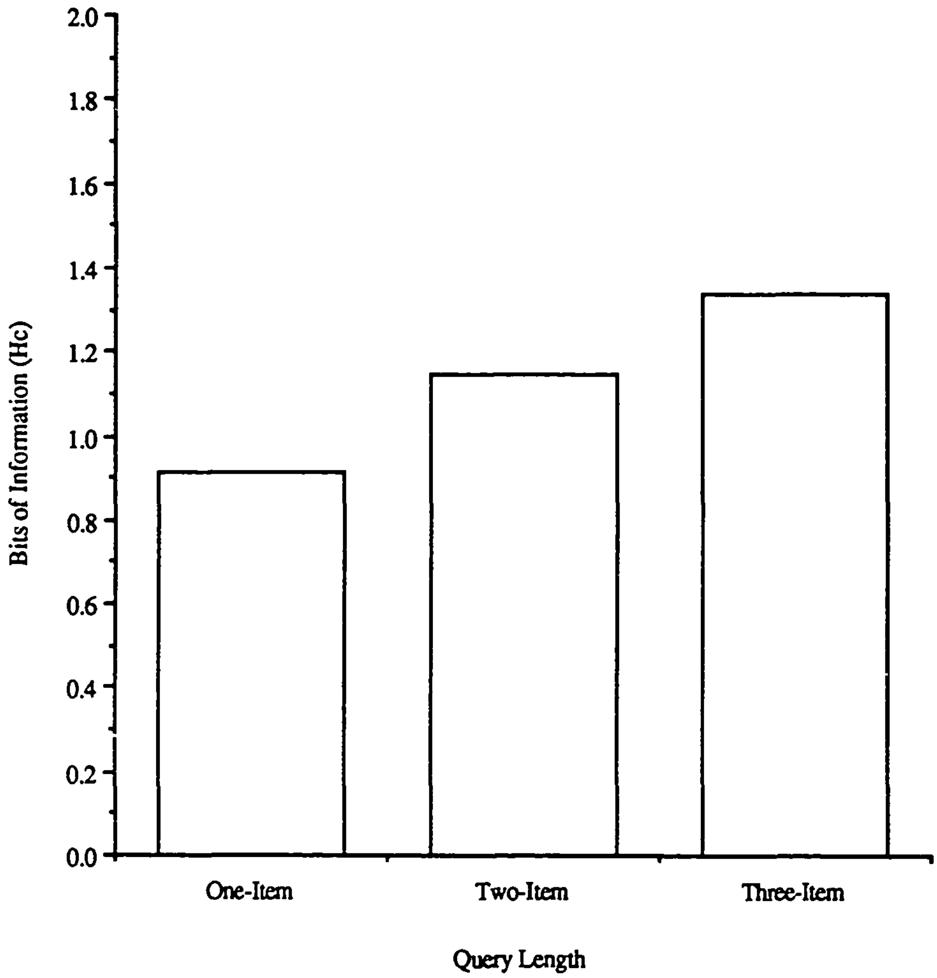


Figure 28. Number of bits of conditional information by query length.

on two-item queries than on three-item queries (Table 37). Thus, the shorter the query the greater the statistical dependency in scanning.

For reference, Table 38 shows the observed conditional information for each query length along with the expected conditional information assuming stratified random sampling. If scanning were completely random, the conditional information metric would have a value of 2.585 bits, which represents the maximum possible conditional information. The relative amounts of statistical independence in the scan patterns for the three query lengths are shown in Table 39. In all cases, the observed information metrics were consistently smaller than the corresponding expected values. Overall, scan patterns were thus more statistically dependent than predicted by either stratified or completely random sampling.

The results of the ANOVA for the information metrics also indicated a significant interaction between window activity and query length (F(2,66) = 4.83, $\underline{p} < 0.0111$; Table 36, Figure 29). Newman-Keuls tests showed that scanpaths were more statistically dependent with single active windows than with multiple active windows on two-item queries (1.124 vs 1.175 bits) and on three-item queries (1.311 vs 1.370 bits; Table 40). There was no difference in scanpath dependency across window activity conditions in the case of one-item queries. Thus, on the two- and three-item queries scanpaths were more statistically dependent with single active windows than with multiple active windows.

For reference, Table 41 shows the observed conditional information for each combination of window activity and query length along with the expected conditional information assuming stratified random sampling. If scanning were completely random, the conditional information metric would again have a maximum value of 2.585 bits. Table 42 shows the relative amount of statistical independence in the scan patterns for each combination of window activity and

TABLE 37

Newman-Keuls Tests on Mean Conditional Information Metrics by Query Length

| Query Length | Mean |
|---|---|
| 1 | 0.915 (A) |
| 2 | 1.149 (B) |
| 3 | 1.340 (C) |

NOTE: Means sharing a common letter in parentheses were not significantly different ($p > 0.01$). Metrics given in bits.

TABLE 38

Observed and Expected Conditional Information Metrics by Query Length

| Query Length | Observed | Expected |
|---|---|---|
| 1 | 0.915 | 1.361 |
| 2 | 1.149 | 1.733 |
| 3 | 1.340 | 1.971 |

NOTE: Expected information was computed on the assumption of stratified random sampling. Metrics given in bits.

TABLE 39

Relative Statistical Independence of Scanpaths by Query Length

| Query Length | Stratified | Random |
|---|---|---|
| 1 | 0.675 | 0.354 |
| 2 | 0.663 | 0.445 |
| 3 | 0.680 | 0.518 |

NOTE: The relative measures represent ratios of observed information to expected information assuming, respectively, stratified random sampling and pure random sampling.

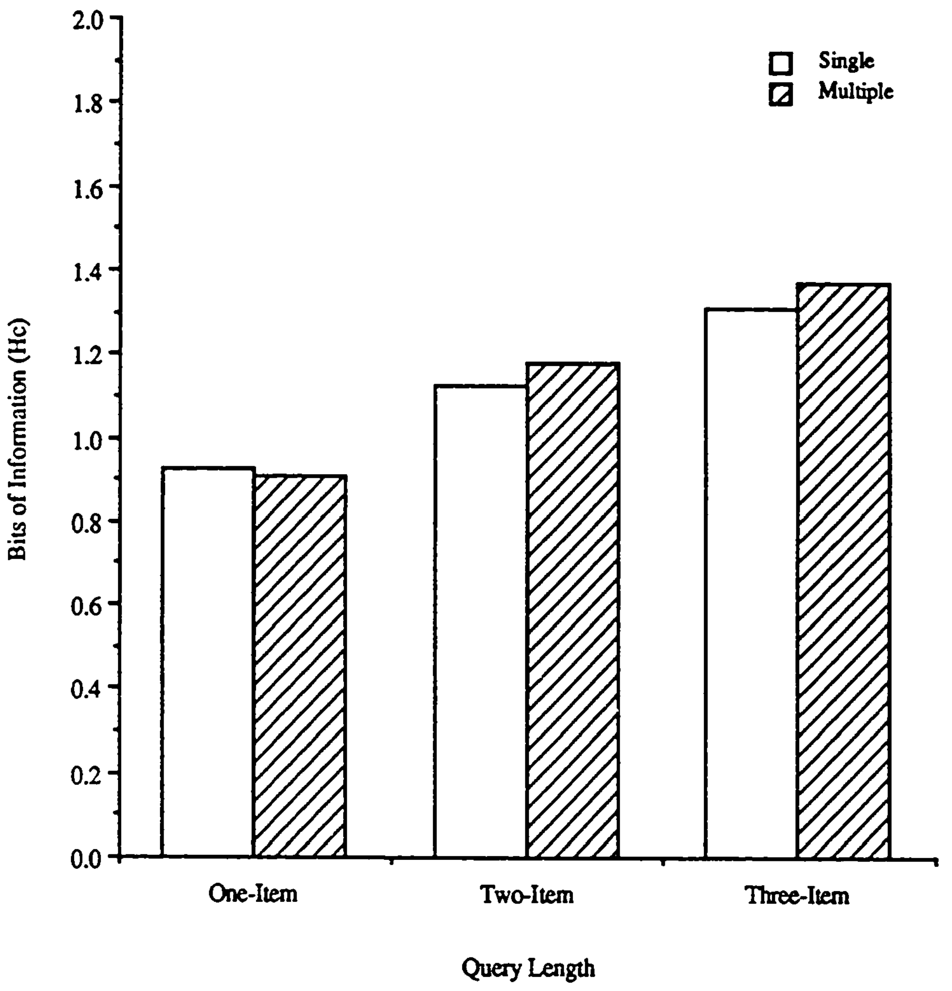


Figure 29. Number of bits of conditional information for window activity by query length interaction.

TABLE 40

Newman-Keuls Tests on Mean Conditional Information Metrics for Window Activity by Query Length Interaction

| Singie Active Windows | | Multiple Active Windows | |
|---|---|---|---|
| Query Length | Mean | Query Length | Mean |
| 1 | 0.924 (A) | 1 | 0.906 (A) |
| 2 | 1.124 | 2 | 1.175 |
| 3 | 1.311 | 3 | 1.369 |

NOTE: Means sharing a common letter in parentheses were not significantly different (p > 0.01). Metrics given in bits.

TABLE 41

Observed and Expected Conditional Information Metrics for Window Activity by Query Length Interaction

| | Single Active Windows | | Multiple Active Windows | |
|---|---|---|---|---|
| Query Length | Observed | Expected | Observed | Expected |
| 1 | 0.924 | 1.378 | 0.906 | 1.345 |
| 2 | 1.124 | 1.705 | 1.175 | 1.762 |
| 3 | 1.311 | 1.940 | 1.369 | 2.002 |

NOTE: Expected information was computed on the assumption of stratified random sampling. Metrics given in bits.

TABLE 42

Relative Statistical Independence of Scanpaths for Window Activity by Query Length Interaction

| | Single Active Windows | | Multiple Active Windows | |
|---|---|---|---|---|
| Query Length | Stratified | Random | Stratified | Random |
| 1 | 0.671 | 0.357 | 0.674 | 0.351 |
| 2 | 0.659 | 0.435 | 0.667 | 0.455 |
| 3 | 0.676 | 0.507 | 0.684 | 0.530 |

NOTE: The relative measures represent ratios of observed information to expected information assuming, respectively, stratified random sampling and pure random sampling.

query length. In all cases, scan patterns were again observed to be more statistically dependent than predicted by either stratified or completely random sampling.

# DISCUSSION AND CONCLUSIONS

In the past, natural language processing systems have generally required so much memory and processing time that only very limited applications were possible, even on mainframes. However, with the development of high performance microcomputers and LISP machines, the possibilities for natural language processing appear more promising. Current work in this field can be divided into the major areas of language translation, text scanning and intelligent indexing, text generation, speech recognition and processing, development tools and shells, and database interfaces. There is, of course, overlap between the latter two areas since natural language shells can be used to develop natural language interfaces to databases.

Most of the efforts in the field of natural language processing are devoted to the development of natural language interfaces to databases. The likely reason for this is that natural language systems designed for other purposes have generally not been successful as commercial products. Overall, then, the current largest area of application for natural language processing is in the area of database interfaces. An example of this line of effort, is the work of the Naval Ocean Systems Center (NOSC) which is attempting to develop a large scale command and control database system with a MBNL interface (Hendrickson & Williams, 1988; Osga, 1984).

In the present study, a system more modest in scope was used in an attempt to derive fundamental human factors principles for the design of MBNL interfaces to databases. Anecdotally, the database system used in the study was generally very well accepted. With user acceptance as the criterion, then the kind of MBNL system used in this study can be considered an acceptable system for use by the type of novice and beginning computer user represented by the sample.

It is interesting, though not too suprising, that multiple active windows were considered more 'natural' than single active windows. Furthermore, although the rank-order preference for multiple active windows failed to reach statistical significance, the probability of users preferring multiple active windows ($p = 0.13$) may be considered to be of practical significance. Additionally, it appears that users particularly preferred greater degrees of freedom for menu selection with the broader menu structure (the sixteen-item window size).

Given that a natural language system should be natural to the user, and assuming that "people won't use it if they don't like it", then the conclusion follows that multiple active windows should be used with a MBNL interface. Multiple active windows should be used, that is, to the extent allowed by the constraints of the grammar and vocabulary required by the particular domain of interest. In any case, multiple active windows appear to be preferable, and this type of windowing would appear to facilitate user acceptance of a MBNL interface.

The above conclusion is interesting in light of the nature of MBNL. The semantic grammar approach, together with a predefined grammar and lexicon and a menu-based interface yields a constrained form of natural language dialogue. Additionally, MBNL is grounded on a context-free grammar, which is less complex but more constrained than a context-sensitive grammar, or a grammar that does not follow any set patterns or requirements (requiring a Turing machine). Thus, MBNL is by definition a highly constrained form of natural language dialogue. The use of single active windows represents fundamental agreement with this definition insofar as their use entails greater constraint. On the other hand, the use of multiple active windows represents a fundamental discrepancy from the definition of MBNL insofar as their use entails a certain lack of constraint. Nevertheless, on the basis of the subjective data obtained in this study, the use of multiple active windows should be encouraged.

The subjective data, unfortunately, do not agree with the performance data, and thus the conclusion that multiple active windows should be used must be qualified.

Query performance was, in fact, slower with multiple active windows than with single active windows, and this difference was exaggerated on the longer two- and three-item queries. There was also an effect related to the spatial arrangement of the menus where dwell times on the middle parse window were longer with multiple active windows. Subjects were apparently exhibiting a tendency to focus attention on the center of the screen, perhaps reflecting greater uncertainty in the search for target items. On the other hand, subjects dwelled relatively longer in the off-area in the single active window condition. Presumably, subjects were looking at the query instructions longer. If so, then in the single active window condition, subjects were probably better prepared for searches, to the extent that target items were more thoroughly rehearsed in preparation for the search task.

If subjects were in fact more uncertain in their searches with multiple active windows, then fixation frequencies could be expected to be greater in that case, which did not hold true, or dwell times could be expected to be longer, which did hold true in the case of two- and three-item queries. Interestingly, in the multiple active window condition on the longer two- and three-item queries, subjects also tended to engage in less statistically dependent search patterns, indicating relatively greater uncertainty in the spatial patterns of scanning. In any case, search strategies were apparently less efficient with multiple active windows.

Consequently, whether single or multiple active windows should be used presents a human factors tradeoff which must be addressed in context. If a MBNL system is designed for general use, then multiple active windows should be used, as concluded above, to facilitate user acceptance. Fortunately, errors are not more likely to occur with the use of multiple active windows. However, if a MBNL system is to be used under time-critical circumstances, as would likely be the case with the NOSC system, for example, then constraining the user to single active windows would be advisable. In fact, under time-critical conditions the use of multiple active windows could be predicted to result in performance decrements, particularly as the complexity of the grammar, the size of the vocabulary, and the corresponding lengths of queries increased. Thus, for a system to be used under this type of circumstance, the use of multiple active windows should probably be minimized.

There appear to be similarities between MBNL information retrieval and menu-based information retrieval in general. As has been observed with menu-based retrieval systems in general, with the MBNL system subjects occasionally gave up on searches, if failure to make a selection can be taken as an indication of giving up a search. However, subjects by no means gave up on a high proportion of searches, as has been observed with general menu-based information retrieval. Additionally, when subjects made an incorrect selection they would sometimes restart the query construction process rather than simply backing up to correct the error, which would have been more efficient. This is similar to the observation made with general menu-based systems, that users tend to restart at the main menu when correcting an error rather than backtracking to the submenu where the incorrect choice was made.

It was assumed a priori that the smaller window size was similar to a deeper menu structure, since fewer items were visible and subjects were required to page down more to search for target items. While the larger window size was assumed to be similar to a broader menu structure, since more items were visible and subjects were required to page down less to find target items. It was consequently predicted, based on the menu breadth\depth tradeoff literature, that the smaller window size would produce slower query performance times and a higher error rate. However, contrary to predictions, no window size displayed any relative disadvantage in terms of query performance times or error frequencies. In terms of performance, it thus appears that there are somewhat tenuous similarities between MBNL and menu-based systems in general. Consequently, generalizations about performance with

MBNL do not necessarily hold for menu-based systems in general and any such generalizations must be qualified and stated with caution.

There was only one significant effect due to the window size factor. The mean duration of a fixation was longer with the four-item window size than with any other window size. It would appear that subjects were taking longer to reach decisions about matches between target items and menu items in this case. Perhaps having fewer visible items lead subjects to adopt a strategy of comparing more than one target item with each menu item on which they were fixating, thereby producing the longer observed fixation durations. However, the cause of this effect is not clear and, in fact, it may be an artifact. With the smaller four-item window size saccades tended to be smaller in magnitude, owing to the limited range of possible fixation points. With frequent smaller saccades, against a measurement resolution of one degree, there would be a greater probability of two contiguous fixations being identified as one, resulting in spuriously longer fixation durations. Consequently, no firm design recommendations regarding window size can be made.

The finding that three-item queries produced more errors than one- and two-item queries is not too suprising. However, error frequencies might be expected to increase with query length. Inspection of the errors that occurred on three-item queries revealed a reason for this apparent anomaly. On one particular three-item query subjects frequently confused the Buick Skyhawk with the Buick Skylark. Consequently, the higher error frequency associated with the three-item query was due partially to confusions between these two menu items.

Overall, most errors were detected and corrected, and given a possible 3456 incorrect menu selections across all subjects, the error frequencies that were observed were relatively low. Of course, one of the claimed benefits of MBNL is that users are guaranteed semantically and syntactically correct queries, which is true, assuming that the system developer correctly specifies the grammar. However, as shown by the present study, task-specific errors can be expected to occur, albeit infrequently, with a MBNL interface.

It is not too suprising that query performance times and fixation frequencies increased with query length. That mean fixation durations were shorter on one-item queries than on two- or three-item queries simply reflects the fact that on one-item queries there was only one target item to match with a menu item. That is, on one-item queries subjects were required to make only a binary decision. On two- and three-item queries subjects fixated for statistically equal lengths of time on menu items, which would appear to reflect something other than a series of binary decisions. It would also appear to rule out the possibility that subjects were making full exhaustive comparisons of target items and menu items.

It might be assumed that for two- and three-item queries, subjects started out by comparing more than one target item with each menu item they were viewing. Once a match was determined and a selection made, the target set was reduced. On two-item queries one binary decision would then remain. On three-item queries two target items would remain, and perhaps the search strategy then became similar to that used on two-item queries. Once two items had finally been selected on the three-item query, one binary decision would then remain. However, if these strategies were employed, then average fixation durations might be expected to increase in nearly linear fashion with query length. It can only be surmised that subjects employed mixed strategies, including possibly an exhaustive comparison strategy, or a stepwise ruling out process, and binary decisions. Furthermore, the menu selection process may well have changed over time. The decision-making literature might supply hypotheses for interesting eyetracking research in this area.

The fact that query performance times increased with query length mirrors the fact that dwell times increased with query length. Descriptively, the increase in dwell times with query length reflects the fact that fixation frequencies increased with query length and these fixations tended to be spatially and temporally contiguous. As intra-menu searching increased with query length, intra-menu dwell times increased, and query performance times increased. The inverse relationship between relative dwell time and query length can be said to reflect the fact that as query length increased, visual attention tended to be more evenly distributed across menus in the search for target items, which consequently lowered the relative dwell time for any given area.

More interesting is the finding that as query length increased, the spatial patterns of fixations became less statistically dependent and became instead distributed in more of a stratified random fashion. So as query length increased, 'uncertainty' in the spatial patterns of scanning increased. Presumably, then, the increase in the stochastic nature of scanning with increasing query length reflects increased uncertainty related to the search for a greater number of target items. As discussed earlier, these effects were apparently exaggerated to an even greater extent with multiple active windows.

Design tradeoffs are implied by these findings. Where feasible, query lengths should be decreased, particularly if multiple active windows will be used. One design strategy would be to first determine the frequency with which menu items are included in queries of different lengths, perhaps by user testing with a prototype system. Then, where feasible and meaningful, one could join menu items which are frequently included in longer queries. Of course, longer menu items would result and there may be some redundancy created in the content of the menus, but there would be a decrease in the number of selections required to build longer queries.

As a rule, however, it should be kept in mind that menu items should not be placed in the same window that are semantically dissimilar or that do not serve a similar syntactic function, mainly because writing a parsable grammar would thereby become more difficult. Decisions regarding the classification of menu items for inclusion in the same menu can usually be based on knowledge of semantic similarity and syntactic functionality. Menu item similarity or relatedness can also be determined psychometrically.

Dwell times, as noted above, increased as a function of query length. However, they also decreased from left to right across the visual workspace. Similarly, relative dwell times decreased from left to right across the visual workspace. However, relative dwell times on the left side of the workspace were longer on shorter queries, while relative dwell times on the right side of the workspace were longer on longer queries. In other words, the longer the query the more likely it was that subjects would distribute visual attention across the visual workspace towards the right side of the screen. Fixation frequencies and fixation durations also generally decreased from left to right across the visual workspace.

Overall, then, the majority of visual attention was apparently allocated to preparation for query construction (off-window viewing). However, it appears that during actual query construction, the spatial distribution of attention was allocated primarily to viewing the left parse window, followed by the middle parse window, and then the right parse window. These patterns of visual behavior mainly reflect task constraints since most of the target items were, in fact, located in the left parse window, followed by the middle parse window, and then the right parse window. More importantly, it appears that the majority of the time was spent in formulating the query in preparation for performing the selection task.

Subjects viewed the results window very little and most often viewed it only with quick glances, indicating that subjects felt little need to verify each menu selection or query under construction. Presumably, the size of the results window should be large enough to contain the longest possible query, however the findings of this study indicate that this rule need not be strictly followed. It appears that screen space could be safely conserved by reducing the size of the results window, since subjects often did not view the results window to verify menu selections. Of course, the detrimental effect would be that if users wanted to verify menu selections, then they would have to scroll the results window to see any part of the query not within view. The consequence would be an increase in query performance time. Decisions about reducing the size of the results window should be tempered by a consideration of the complexity of possible queries and the corresponding need to see the full query under construction.

Subjects also viewed the command window very little, implying that they needed little reminding about available command functions. Thus, it also appears that the size of the command window could be reduced to conserve screen space for parse windows. However, decisions about reducing the size of the command window should be tempered by a consideration of the frequency with which certain command functions are executed, and by a consideration of the need for visual feedback (by highlighting) that a command function has been selected.

A final note, of theoretical interest, is that scan patterns could not be described as purely random. Visual scan patterns were also more statistically dependent than predicted by stratified random sampling, although the stratified random sampling model better approximated menu scanning behavior than did the pure random sampling model. Therefore, though menu scanning was not purely deterministic, menus were apparently searched to a large extent by systematic patterns of eye movements. It would thus appear that menu scanning can be represented reasonably well as a first-order Markov process. However, note that the identified statistical dependencies in scanning were observed to vary as a function of task requirements (query length) and as a function of menu navigation constraints (window activity). Therefore, descriptions of menu scanning behavior should be stated in the context of task requirements and menu navigation constraints.

# REFERENCES

Allen, R. B. (1980). Usage of menus and tress. Technical Memorandum 80-1359-3, Bell Laboratories.

Allen, R. B. (1983). Cognitive factors in the use of menus and trees: An Experiment. IEEE Journal on Selected Areas in Communications, SAC-1, 333-336.

Ballard, B. W. (1979). Semantic processing for a natural language programming system. Doctoral Dissertation, Duke University.

Ballard, B. W., & Biermann, A. W. (1979). Programming in natural language: NLC as prototype. Proceedings of the 1979 Annual Conference of the ACM, 228-237.

Balzer, R. M. (1973). A global view of automatic programming. Proceedings of the 3rd Joint Conference on Artificial Intelligence.

Benedetti, J. K., & Brown, M. B. (1978). Strategies for the selection of log-linear models. Biometrics, 34, 680-686.

Biermann, A. W., & Ballard, B. W. (1980). Toward natural language computation. American Journal of Computational Linguistics, 6, 71.

Biermann, A. W., Ballard, B. W., & Sigmon, A. M. (1983). An experimental study of natural language programming. International Journal of Man-Machine Studies, 18, 71-87.

Billingsley, P. (1982). Navigation through hierarchical menu structures: Does it help to have a map? Proceedings of the 26th Annual Meeting of the Human Factors Society, 103-107.

Bobrow, D. G., & Collins, A. (1975). Representation and understanding. New York: Academic Press.

Brillouin, L. (1962). Science and information theory. New York: Academic Press.

Broadbent, D. E., Cooper, P. J., & Broadbent, M. H. P. (1978). A comparison of hierarchical and matrix retrieval schemes in recall. Journal of Experimental Psychology, 4, 486-497.

Brown, J. (1982). Controlling the complexity of menu networks. Communications of the ACM, 25, 412-418.

Brown, J. S., Burton, R. R., & Bell, A. G. (1975). SOPHIE: A step towards creating a reactive learning environment. International Journal of Man-Machine Studies, 7, 675-696.

Card, S. K. (1982). User perceptual mechanisms in the search of computer command menus. Proceedings of CHI '82 Human Factors in Computer Systems, 25-31.

Damerau, F. J. (1981). Operating statistics for the Transformational Question Answering system. American Journal of Computational Linguistics, 7, 30.

Dijkstra, E. W. (1978). On the foolishness of natural language programming. Technical Note.

Dray, S. M., Ogden, W. G., & Vesteweig, R. E. (1981). Measuring performance with a menu-selection human-computer interface. Proceedings of the 25th Annual Meeting of the Human Factors Society, 746-748.

Dumais, S. T., & Landauer, T. K. (1982). Naming categories: Describing objects for menu and keyword systems. Paper presented at the meeting of the American Psychological Assoociation, Washington, DC.

Egly, D., & Wescourt, K. (1981). Cognitive style categorization, and vocational effects on performance of REL database users. Joint Conference on Easier and More Productive Use of Computing Systems. Ann Arbor: University of Michigan.

Ellis, S. R., & Stark, L. (1986). Statistical dependency in visual scanning. Human Factors, 28, 421-438.

Evans, J. E., III, & Gutmann, J. C. (1978). Minicomputer processing of dual Purkinje image eye-tracker data. Behavior Research Methods and Instrumentation, 10, 701-704.

Ford, W. R. (1981). Natural language processing by computer: A new approach. Doctoral Dissertation, Johns Hopkins University, Baltimore, MD.

Green, B. F., Wolf, A. D., Chomsky, C., & Laughery, K. (1963). BASEBALL: An automatic question answerer. In E. A. Feigenbaum & J. Feldman (Eds.), Computers and thought. New York: McGraw-Hill.

Green, C. C., Gabriel, R. P., Ginsparg, J. M., Kant, E., Ludlow, J. J., McCune, B. P., Phillips, J. V., Steinberg, L. I., Tappel, S. T., & Westfold, S. J. (1978). Progress report on knowledge based programming. Technical Report. Palo Alto, CA: Systems Control, Inc.

Hagelbarger, D., & Thompson, R. (1983). Experiments in teleterminal design. IEEE Spectrum, 20, 40-45.

Hauptmann, A. G., & Green, B. F. (1983). A comparison of command, menu-selection and natural-language computer programs. Behavior and Information Technology, 2, 163-178.

Heidorn, G. E. (1976). Automatic programming through natural language dialogue: A survey. IBM Journal of Research and Development, 20, 302.

Hendrickson, J. J., & Williams, R. D. (1988). The effect of input device on user performance with a menu-based natural language interface (TR-1224). San Diego: Naval Ocean Systems Center.

Hershey, H. C., Zakin, J. L., & Simha, R. (1967). Numerical differentiation of equally spaced and not equally spaced experimental data. Industrial and Engineering Chemistry Fundamentals, 6, 413-420.

Hershmann, R. L., Kelly, R. T., & Miller, H. G. (1979). User performance with actual language query system for command control (NPRDC-TR-79-7). San Diego, CA: Navy Personnel Research and Development Center.

Hill, I. D. (1972). Wouldn't it be nice if we could write programs in ordinary English-or would it? The Computer Bulletin, 16, 306-312.

Kaplan, S. J. (1982). Special section on natural language processing. SIGART Newsletter, 79, 29.

Kelley, J. F. (1981). Calendar access language: Natural language input for computers. Doctoral Dissertation, Johns Hopkins University, Baltimore, MD.

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

Krause, J. (1979). Results of a user study with the User Specialty Language system and consequences for the architecture of natural language interfaces. Technical Report 79.04.003. IBM Heidelberg Scientific Center.

Ledgard, H., Whiteside, J., Singer, A., & Seymour, W. (1980). The natural language of interactive systems. Communications of the ACM, 23, 556.

Lee, E. (1979). The optimum number of alternatives to display on an index page in an interactive Telidon database. In D. Phillips (Ed.), Telidon Behavioural Research I. Ottawa: Department of Communications.

Lee, E., Whalen, T., McEwen, S. A., & Latremouille, S. (1984). Optimizing the design of menu pages for information retrieval. Ergonomics, 27, 1051-1069.

Liebelt, L. S., McDonald, J. E., Stone, J. D., & Karat, J. (1982). The effects of organization on learning menu access. Proceedings of the 26th Annual Meeting of the Human Factors Society, 546-550.

Lowden, B. G. T., & DeRoeck, A. (1985). Generating English paraphrases from relational query expressions. Behaviour and Information Technology, 4, 337-348.

MacGregor, J. N., & Lee, E. S. (1987). Performance and preference in videotex information retrieval: A review of the empirical literature. Behaviour and Information Technology, 6, 43-68.

MacGregor, J. N., Lee, E. S., & Lam, N. (1986). Optimizing the structure of menu indexes: A decision model of menu search. In D.J. Osborne (Ed.), Contemporary ergonomics 1986. London: Taylor & Francis.

Marascuilo, L. A., & Levin, J. R. (1983). Multivariate statistics in the social sciences. Monterey, CA: Brooks-Cole.

Martin, J. (1973). Design of man-computer dialogues. Englewood Cliffs: Prentice-Hall.

Martin, W. A., Ginzberg, M. T., Krumland, R., Mark, B., Morgenstern, M., Niamir, B., & Sunguroff, A. (1974). Internal Memos. M.I.T., Cambridge, MA: Automatic Programming Group.

McDonald, J., Stone, J., & Liebelt, L. (1983). Searching for items in menus: The effects of organization and type of target. Proceedings of the 27th Annual Meeting of the Human Factors Society, 834-837.

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

Miller, H. G., Hershman, R. L., & Kelly, R. T. (1978). Performance of a natural language query system in a simulated command control environment. Technical Report. Advanced Command and Control Architectural Testbed Facility, U.S. Navy.

Morris, J. M. (1979). Natural language systems: A user's view. SQ/BTN Newsletter, 2, 16.

Norman, K. L., Schwartz, J. P., & Shneiderman, B. (1984). Memory for menus: Effects of study mode (CAR-TR-69). College Park, MD: University of

Maryland, Center for Automation Research and the Department of Computer Science.

Ogden, W. C., & Brooks, S. R. (1983). Query languages for the casual user: Exploring the middle ground between formal and natural languages. Proceedings of CHI '83 Human Factors in Computing Systems, 161-165.

Osga, G. A. (1984). Menu-based natural language query for naval command and control (TR-1006). San Diego: Naval Ocean Systems Center.

Parton, D., Huffman, K., Pridgen, P., Norman, K., & Shneiderman, B. (1985). Learning a menu selection tree: Training methods compared. Behaviour and Information Technology, 4, 81-91.

Petrick, S. R. (1976). On natural language based computer systems. IBM Journal of Research and Development, 20, 314-325.

Robertson, G., McCracken, D., & Newell, A. (1981). The ZOG approach to man-machine communication. International Journal of Man-Machine Studies, 14, 461-488.

Savitsky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36, 1627-1639.

Schank, R. C. (1975). Conceptual information processing. New York: North-Holland.

Schank, R. C., & Colby, K. (1973). Computer models of thought and language. San Francisco: Freedman.

Schwartz, J. P., Norman, K. L., & Shneiderman, B. (1985). Performance on content-free menus as function of study method (CAR-TR-110). College Park, MD: University of Maryland, Center for Automation Research and the Department of Computer Science.

Shneiderman, B. (1978). Improving the human factors aspect of database interactions. ACM Transactions on Database Systems, 3, 417-439.

Shneiderman, B. (1980). Software psychology. Cambridge, MA: Winthrop.

Shneiderman, B. (1987). Designing the user interface. Reading, MA: Addison-Wesley.

Sisson, N., Parkinson, S., & Snowberry, K. (1983). Proceedings of the Second Annual Phoenix Conference on Computers and Communication, Phoenix, AZ, 14-16.

Small, D. W., & Weldon, L. J. (1983). An experimental comparison of natural and structured query languages. Human Factors, 25, 253-263.

Snowberry, K., Parkinson, S., & Sisson, N. (1983). Computer menu displays. Ergonomics, 26, 699-712.

Somberg, B. L., & Picardi, M. C. (1983). Locus of the information familiarity effect in the search of computer menus. Proceedings of the 27th Annual Meeting of the Human Factors Society, 826-830.

Sondheimer, N. K. (1978). Why don't we - how can we - when will we - have practical English language interfaces? Technical Report H00002. Bluebell, PA: Sperry Univac.

Steiner, J., Termonia, Y., & Deltour, J. (1972). Comments on smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 44, 1906-1909.

Suding, A. (1983). The users perspective. In Tutorial on Natural Language Interfaces. Symposium presented at the conference on Applied Natural Language Processing, Santa Monica, CA.

Tennant, H. R. (1980). Evaluation of natural language processors (Report T-103). Urbana, IL: University of Illinois, Coordinated Science Laboratory.

Tennant, H. (1981). Natural language processing. New York: Petrocelli.

Tennant, H. R., Ross, K. M., Saenz, R. M., Thompson, C. W., & Miller. J. R. (1983). Menu-based natural language understanding. 21st Annual Meeting of the Association for Computational Linguistics, MIT.

Tennant, H. R., Ross, K. M., & Thompson, C. W. (1983). Usable natural language interfaces through menu-based natural language understanding. Proceedings of CHI '83 Human Factors in Computing Systems, 154-160.

Texas Instruments. (1985a). NaturalLink Toolkit (No. 2240316-0001).

Texas Instruments. (1985b). NaturalLink Window Manager (No. 2240317-001).

Texas Instruments. (1985c). NaturalLink Technical Report (No. 2249795-001).

Thompson, C. W., Tennant, H. R., Ross, K. M., & Saenz, R. M. (1983). Building usable menu-based natural language interfaces to databases. Ninth International Conference on Very Large Databases, Florence, Italy.

Tombaugh, J., & McEwen, S. (1982). Comparison of two information retrieval methods of videotex: Tree-structure versus alphabetical directory. Proceedings of CHI '82 Human Factors in Computing Systems, 106-110.

Waltz, D. L. (Ed.) (1977). Natural language interfaces. SIGART Newsletter.

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. Communications of the ACM, 10, 474-480.

Weizenbaum, J. (1976). Computer power and human reason. San Francisco: Freeman.

Winograd, T. (1972). Understanding natural language. New York: Academic Press.

Woods, W. A. (1970). Transition network grammars for natural language analysis. Communications of the ACM, 13, 591-606.

## Appendix A: Menu Items

Large/Medium Sized Cars

Audi 5000S
Buick Century
Buick Electra
Buick LeSabre
Buick Regal
Buick Riveria
Chevrolet Caprice
Chevrolet Celebrity
Chrysler Fifth Avenue
ChryslerLeBaron
Chrysler New Yorker
Dodge Aries
Dodge Diplomat
Dodge Lancer
Dodge 600
Ford LTD Crown Victoria
Ford Thunderbird
Lincoln Mark VI
Mercedes-Benz 300
Mercury Cougar
Mercury Grand Marquis
Mercury Sable
Olds Cutlass Ciera
Olds Cutlass Supreme
Olds Delta 88 Royale
Olds Toronado
Plymouth Caravelle
Plymouth Gran Fury
Plymouth Reliant
Pontiac Bonneville
Pontiac Parisienne
Pontiac 6000

Small Sized Cars

Chevrolet Chevette
Chevrolet Nova
Chevrolet Spectrum
Chevrolet Sprint
Dodge Charger
Dodge Colt
Dodge Omni
Ford Escort
Honda Civic
Honda Prelude Si
Hyundai Excel
Isuzu I-Mark
Mazda GLC
Mazda 323
Mercury Lynx
Mitsubishi Tredia
Nissan Sentra
Nissan 200SX
Plymouth Colt
Plymouth Horizon
Plymouth Turismo
Pontiac Fiero
Pontiac 1000
Porsche 944
Renault Alliance
Renault Encore
Subaru
Toyota Corolla
Toyota Tercel
Volkswagen Golf
Volkswagen Jetta
Yugo GV

Compact Cars

Acura Legend
Audi 4000S
BMW 318i
Buick Skyhawk
Buick Skylark
Buick Somerset
Cadillac Cimmaron
Chevrolet Cavalier
Dodge Conquest
Ford Tempo
Honda Accord
Isuzu Impulse
Mazda 626
Mercedes-Benz 190
Mercury Topaz
Mitsubishi Cordia
Mitsubishi Galant
Nissan Maxima
Nissan Stanza
Oldsmobile Calais
Oldsmobile Firenza
Peugot 505
Plymouth Conquest
Pontiac Grand AM
Pontiac Sunbird
Saab 900
Toyota Camry
Toyota Cressida
Toyota MR2
Volkswagen Quantum
Volvo DL
Volvo 240

## Appendix B: Queries

1. Find information on the Buick Skylark and the Audi 5000S and the Dodge Colt.
2. Find information on the Buick Century.
3. Find information on the Mazda 323 and the Buick Electra.
4. Find information on the Buick Regal.
5. Find information on the Volkswagen Jetta and the Buick Riveria.
6. Find information on the Chevrolet Caprice and the Toyota Tercel and the Mitsubishi Galant.
7. Find information on the Chrysler Fifth Avenue and the Renault Alliance.
8. Find information on the Toyota Corolla and the Mercedes-Benz 190 and the Chrysler LeBaron.
9. Find information on the Chrysler New Yorker.
10. Find information on the Dodge Diplomat and the Pontiac Fiero and the Honda Accord.
11. Find information on the Dodge Lancer.
12. Find information on the Pontiac 1000 and the Dodge 600.
13. Find information on the Ford Thunderbird.
14. Find information on the Nissan Sentra and the Lincoln Mark VII.
15. Find information on the Mercedes-Benz 300 and the Renault Encore and the Audi 4000S.
16. Find information on the Mercury Grand Marquis and the Chevrolet Spectrum and the Pontiac Sunbird.
17. Find information on the Mercury Sable and the Hyundai Excel.
18. Find information on the Olds Cutlass Ciera.
19. Find information on the Olds Delta 88 Royale.
20. Find information on the Olds Toronado and the Chevrolet Nova.

21. Find information on the Plymouth Conquest and the Plymouth Caravelle and the Honda Civic.
22. Find information on the Plymouth Reliant and the Dodge Charger.
23. Find information on the Mercury Lynx and the Volvo DL and the Pontiac Bonneville.
24. Find information on the Pontiac Parisienne.

## Appendix C: Database for Menu Items in the Left Parse Window

### Audi 5000S

Predicted Reliability - Average. Repair costs are high.

Fuel Economy - Mpg with non-turbo engine and automatic transmission: city, 14; expressway, 28. Gallons used in 15,000 miles, 745. Cruising range, 475 miles.

Comments - The Audi 5000S performs as a European sports sedan should. Seating and ride comfort are good also. Be sure that all the factory recalls relating to the "sudden acceleration runaway" have been performed. Bumper test damage: none.

### Buick Century

Predicted Reliability - For V4, average. For V6, worse than average.

Fuel Economy - Mpg for sedan with V4 and automatic transmission: city, 16; expressway, 37. Gallons used in 15,000 miles, 610. Cruising range, 455 miles. Mpg for wagon with V6 engine: city, 14; expressway, 33. Gallons used in 15,000 miles, 710. Cruising range, 365 miles.

This GM A-body model appears to be improving in quality and reliability. The 4-cylinder engine is recommended; it provides adequate acceleration and gives good mileage. The sedan version will benefit from any of the heavy duty or performance suspension options. Bumper test damage: moderate with sedans; none with wagons.

**Buick Electra**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 3.8-liter V6: city, 14; expressway, 33. Gallons used in 15,000 miles, 700. Cruising range, 440 miles.

Comments - This recently downsized model is considerably lighter and shorter than the previous rear wheel drive model, but is roomy inside and has comfortable seating. Its ride and handling, however, are not as good as they should be, especially on more challenging roads or road surfaces. Bumper test damage: moderate.

**Buick LeSabre**

Predicted Reliability - No data.

Fuel Economy - Mpg with 3-liter V6: city, 14; expressway, 33. Gallons used in 15,000 miles, 700. Cruising range, 440 miles.

Comments - This model is essentially the same car as the more expensive Buick Electra, and thus is a better value because of its lower first cost. The power drivers seat and tilt steering column are good options to look for. Bumper test damage: moderate.

**Buick Regal**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 3.8-liter V6 and automatic transmission: city, 13; expressway, 28. Gallons used in 15,000 miles, 780. Cruising range, 385 miles.

Comments - This aging GM model has little going for it anymore, and its reliability has been poor.

**Buick Riveria**

Predicted Reliability - No data.

Fuel Economy - Mpg with V6 and automatic transmission: city, 13; expressway, 34. Gallons used in 15,000 miles, 750. Cruising range, 385 miles.

Comments - The Riveria appears overpriced and overcomplicated. Getting repairs for a Riveria may be worrisome. The T-Type version is considerably more competent in ride and handling than the standard version. Bumper test damage: moderate.

### Chevrolet Caprice

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with V8: city, 14; expressway, 31. Gallons used in 15,000 miles, 735. Cruising range, 545 miles.

Comments - This model is a traditional large GM rear wheel drive car. It is not as comfortable, especially in the rear seat, as the new front wheel drive GM models, but offers the familiar feel and the heavier trailer towing capability of older models.

### Chevrolet Celebrity

Predicted Reliability - For V4, average. For V6, worse than average.

Fuel Economy - Mpg for sedan with V4 and automatic transmission: city, 16; expressway, 37. Gallons used in 15,000 miles, 610. Cruising range, 455 miles. Mpg for wagon with V6 engine: city, 14; expressway, 33. Gallons used in 15,000 miles, 710. Cruising range, 365 miles.

This GM A-body model appears to be improving in quality and reliability. The 4-cylinder engine is recommended; it provides adequate acceleration and gives good mileage. The sedan version will benefit from any of the heavy duty or performance suspension options. Bumper test damage: moderate with sedans; none with wagons.

**Chrysler Fifth Avenue**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg: city, 11; expressway, 26. Gallons used in 15,000 miles, 885. Cruising range, 350 miles.

Comments - This older design Chrysler product is sold in the large car market class but is really medium in size. The Fifth Avenue has been successful in sales.

**Chrysler LeBaron**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 2.2-liter V4: city, 16; expressway, 34. Gallons used in 15,000 miles, 640. Cruising range, 360 miles.

Comments - The Chrysler LeBaron could benefit from a heavy duty suspension. The standard 2.2-liter V4 is adequate for this model. The optional 2.5-liter V4 and turbocharged V4 provide extra punch.

### Chrysler New Yorker

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with 2.2-liter V4 and automatic transmission: city, 16; expressway, 34. Gallons used in 15,000 miles, 640. Cruising range, 360 miles.

Comments - This model is a stretched K-car. It has a slightly longer wheel base and more fore-and-aft room in the rear seat. As with the K-cars, the suspension is overly soft and can benefit from a heavy duty option. Bumper test damage: moderate.

### Dodge Aries

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with 2.2-liter V4 and automatic transmission: city, 16; expressway, 34. Gallons used in 15,000 miles, 640. Cruising range, 360 miles.

Comments - This is a basic K-car. Low first cost is its primary advantage. The standard 2.2-liter V4 is recommended for this model, as is the heavy duty suspension. Bumper test damage: moderate.

### Dodge Diplomat

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg: city, 11; expressway, 26. Gallons used in 15,000 miles, 885. Cruising range, 350 miles.

Comments - This older design Chrysler product is sold in the large car market class but is really medium in size. The Diplomat is popular primarily as a police and fleet car.

### Dodge Lancer

Predicted Reliability - For nonturbo model, average. For turbo model, better than average.

Fuel Economy - Mpg with 2.5-liter V4 and automatic transmission: city, 16; expressway, 31. Gallons used in 15,000 miles, 665. Cruising range, 335 miles. Mpg with turbocharged V4 and automatic transmission: city, 16; expressway, 29. Gallons used in 15,000 miles, 675. Cruising range, 335 miles.

Comments - This Chrysler product is a model of choice in the family-sized line. Its hatchbacked body gives added versatility and its suspension, even in the standard version, is more competent than the suspension in the K-cars. The turbo V4 with the performance suspension is even more capable. Bumper test damage: minor.

**Dodge 600**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 2.2-liter V4: city, 16; expressway, 34. Gallons used in 15,000 miles, 640. Cruising range, 360 miles.

Comments - The Dodge 600 could benefit from a heavy duty suspension. The standard 2.2-liter V4 is adequate for this model. The optional 2.5-liter V4 and turbocharged V4 provide extra punch.

**Ford LTD Crown Victoria**

Predicted Reliability - Better than average.

Fuel Economy - Mpg: city, 11; expressway, 27. Gallons used in 15,000 miles, 880. Cruising range, 350 miles.

Comments - This large Ford model has consistently had good overall repair records, and for that reason is a model of choice in this group. It is comfortable and smooth riding, but is not as fuel efficient as its front wheel drive competitors. It does have a heavy trailer towing ability. Bumper test damage: moderate.

## Ford Thunderbird

Predicted Reliability - Average.

Fuel Economy - Mpg 3.8-liter V6 and automatic transmission: city, 16; expressway, 29. Gallons used in 15,000 miles, 675. Cruising range, 505 miles.

Comments - This rear wheel drive Ford product is a model of choice in the domestic speciality coupe field, primarily because of its relatively good overall repair record. Bumper test damage: none.

## Lincoln Mark VII

Predicted Reliability - Average.

Fuel Economy - Mpg: city, 12; expressway, 35. Gallons used in 15,000 miles, 770. Cruising range, 500 miles.

Comments - The Mark VII is an older design car, but it delivers the luxury and smoothness that one expects from a car of this type. Bumper test damage: moderate.

### Mercedes-Benz 300

Predicted Reliability - Much better than average.

Fuel Economy - Mpg with automatic transmission: city, 16; expressway, 28. Gallons used in 15,000 miles, 700. Cruising range, 415 miles.

Comments - This model is a very expensive car, but it is likely to give good service to those that can afford it. Bumper test damage: extensive.

### Mercury Cougar

Predicted Reliability - Average.

Fuel Economy - Mpg 3.8-liter V6 and automatic transmission: city, 16; expressway, 29. Gallons used in 15,000 miles, 675. Cruising range, 505 miles.

Comments - This rear wheel drive Ford product is a model of choice in the domestic speciality coupe field, primarily because of its relatively good overall repair record. Bumper test damage: none.

### Mercury Grand Marquis

Predicted Reliability - Better than average.

Fuel Economy - Mpg: city, 11; expressway, 27. Gallons used in 15,000 miles, 880. Cruising range, 350 miles.

Comments - This large Ford model has consistently had good overall repair records, and for that reason is a model of choice in this group. It is comfortable and smooth riding, but is not as fuel efficient as its front wheel drive competitors. It does have a heavy trailer towing ability. Bumper test damage: moderate.

### Mercury Sable

Predicted Reliability - No data.

Fuel Economy - Mpg for sedan with V6: city, 15; expressway, 35. Gallons used in 15,000 miles, 660. Cruising range, 400 miles. Mpg for wagon: city, 13; expressway, 35. Gallons used in 15,000 miles, 730. Cruising range, 435 miles. Mpg with 2.5-liter V4 and automatic transmission: city, 14; expressway, 33. Gallons used in 15,000 miles, 700. Cruising range, 385 miles.

Comments - This Ford model combines the comfort expected in a domestic car with the competent handling expected in European models. The V6 combined with the overdrive automatic transmission is preferred over the 4 cylinder version. Bumper test damage: none.

### Olds Cutlass Ciera

Predicted Reliability - For V4, average. For V6, worse than average.

Fuel Economy - Mpg for sedan with V4 and automatic transmission: city, 16; expressway, 37. Gallons used in 15,000 miles, 610. Cruising range, 455 miles. Mpg for wagon with V6 engine: city, 14; expressway, 33. Gallons used in 15,000 miles, 710. Cruising range, 365 miles.

This GM A-body model appears to be improving in quality and reliability. The 4-cylinder engine is recommended; it provides adequate acceleration and gives good mileage. The sedan version will benefit from any of the heavy duty or performance suspension options. Bumper test damage: moderate with sedans; none with wagons.

### Olds Cutlass Supreme

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 3.8-liter V6 and automatic transmission: city, 13; expressway, 28. Gallons used in 15,000 miles, 780. Cruising range, 385 miles.

Comments - This aging GM model has little going for it anymore, and its reliability has been poor.

**Olds Delta 88 Royale**

Predicted Reliability - No data.

Fuel Economy - Mpg with 3-liter V6: city, 14; expressway, 33. Gallons used in 15,000 miles, 700. Cruising range, 440 miles.

Comments - This model is essentially the same car as the more expensive Olds 98, and thus is a better value because of its lower first cost.
The power drivers seat and tilt steering column are good options to look for. Bumper test damage: moderate.

**Olds Toronado**

Predicted Reliability - No data.

Fuel Economy - Mpg: city, 13; expressway, 35. Gallons used in 15,000 miles, 735. Cruising range, 400 miles.

Comments - The standard suspension is too soft for best handling; the performance option would be a better compromise. Bumper test damage: moderate.

**Plymouth Caravelle**

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with 2.2-liter V4 and automatic transmission: city, 16; expressway, 34. Gallons used in 15,000 miles, 640. Cruising range, 360 miles.

Comments - This model is a stretched K-car. It has a slightly longer wheel base and more fore-and-aft room in the rear seat. As with the K-cars, the suspension is overly soft and can benefit from a heavy duty option. Bumper test damage: moderate.

**Plymouth Gran Fury**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg: city, 11; expressway, 26. Gallons used in 15,000 miles, 885. Cruising range, 350 miles.

Comments - This older design Chrysler product is sold in the large car market class but is really medium in size. The Gran Fury is popular primarily as a police and fleet car.

**Plymouth Reliant**

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with 2.2-liter V4 and automatic transmission: city, 16; expressway, 34. Gallons used in 15,000 miles, 640. Cruising range, 360 miles.

Comments - This is a basic K-car. Low first cost is its primary advantage. The standard 2.2-liter V4 is recommended for this model, as is the heavy duty suspension. Bumper test damage: moderate.

**Pontiac Bonneville**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 3.8-liter V6 and automatic transmission: city, 13; expressway, 28. Gallons used in 15,000 miles, 780. Cruising range, 385 miles.

Comments - This aging GM model has little going for it anymore, and its reliability has been poor.

## Pontiac Parisienne

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with V8: city, 14; expressway, 31. Gallons used in 15,000 miles, 735. Cruising range, 545 miles.

Comments - This model is a traditional large GM rear wheel drive car. It is not as comfortable, especially in the rear seat, as the new front wheel drive GM models, but offers the familiar feel and the heavier trailer towing capability of older models.

## Pontiac 6000

Predicted Reliability - For V4, average. For V6, worse than average.

Fuel Economy - Mpg for sedan with V4 and automatic transmission: city, 16; expressway, 37. Gallons used in 15,000 miles, 610. Cruising range, 455 miles. Mpg for wagon with V6 engine: city, 14; expressway, 33. Gallons used in 15,000 miles, 710. Cruising range, 365 miles.

Comments - This GM A-body model appears to be improving in quality and reliability. The 4-cylinder engine is recommended; it provides adequate acceleration and gives good mileage. The sedan version will benefit from any of the heavy duty or performance suspension options. Bumper test damage: moderate with sedans; none with wagons.

## Appendix D: Database for Menu Items in the Middle Parse Window

### Chevrolet Chevette

Predicted Reliability - For gasoline model, much worse than average.

Fuel Economy - Mpg with gasoline V4 and automatic transmission: city, 19; expressway, 37. Gallons used in 15,000 miles, 555. Cruising range, 370 miles.

Comments - Passenger space is tight, especially width. The 4 door has better seating than the 2 door. Reliability has been poor, though major components have not been particularly troublesome. Service and parts should be readily available at low cost. Chevettes are widely used as a fleet car.

### Chevrolet Nova

Predicted Reliability - No data.

Fuel Economy - Mpg with manual transmission: city, 24; expressway, 46. Gallons used in 15,000 miles, 435. Cruising range, 530 miles.

Comments - The Nova is a top-rated high quality car. It should be a good buy and service would certainly be widely available. Bumper test damage: none.

## Chevrolet Spectrum

Predicted Re!:ability - No data.

Fuel Economy - Mpg with manual transmission: city, 25; expressway, 48. Gallons used in 15,000 miles, 420. Cruising range, 445 miles.

Comments - The Spectrum does not compete well with other small cars. Its mechanical reliability is yet unknown. It might be a serviceable car at a favorable price. Bumper test damage: extensive.

## Chevrolet Sprint

Predicted Reliability - Average.

Fuel Economy - Mpg with manual transmission: city, 37; expressway, 59. Gallons used in 15,000 miles, 310. Cruising range, 445 miles.

Comments - The Sprint is a suprisingly competent small car. It would make a good town car because it is small and gives superlative fuel economy, but its ride is punishing at times and its rear seat is only habitable, at best. Bumper test damage: moderate.

## Dodge Charger

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 1.6-liter V4 and manual transmission: city, 22; expressway, 45. Gallons used in 15,000 miles, 475. Cruising range, 450 miles. Mpg with 2.2-liter V4 and automatic transmission: city, 19; expressway, 33. Gallons used in 15,000 miles, 595. Cruising range, 335 miles.

Comments - The repair record for this car remains much worse than average up through the latest data. Buy one only if the price and mileage are so low that coping with potential problems seems worth it.

## Dodge Colt

Predicted Reliability - Much better than average, and repair costs are low.

Fuel Economy - Mpg with manual transmission: city, 23; expressway, 45. Gallons used in 15,000 miles, 465. Cruising range, 410 miles.

Comments - The Colt is among the better small cars in overall quality, although its acceleration, ride and heating system are a little below par. Colts are likely to be priced a bit below most top selling small cars.

**Dodge Omni**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 1.6-liter V4 and manual transmission: city, 22; expressway, 45. Gallons used in 15,000 miles, 475. Cruising range, 450 miles. Mpg with 2.2-liter V4 and automatic transmission: city, 19; expressway, 33. Gallons used in 15,000 miles, 595. Cruising range, 335 miles.

Comments - The repair record for this car remains much worse than average up through the latest data. Buy one only if the price and mileage are so low that coping with potential problems seems worth it.

**Ford Escort**

Predicted Reliability - For gasoline V4, average. For diesel, no data.

Fuel Economy - Mpg with gasoline engine and manual transmission: city, 21; expressway, 41. Gallons used in 15,000 miles, 505. Cruising range, 410 miles.

Comments - The Escort is a better small car choice than the traditional U.S. built small cars. Its repair history is better, and it performs somewhat better. However, it does not have the passenger room and ride comfort of the high rated small cars, and suffers from some poor design of its controls. Service and parts should be widely available and repair costs relatively low. Bumper test damage: none.

### Honda Civic

Predicted Reliability - Much better than average.

Fuel Economy - Mpg with au.omatic transmission: city, 22; expressway, 39. Gallons used in 15,000 miles, 505. Cruising range 380 miles.

Comments - This is one of the better small cars, but its not up to the level of the Toyota Corolla in overall quality. The Civic is very reliable and at its best with a 5 speed manual transmission.

### Honda Prelude Si

Predicted Reliability - Much better than average.

Fuel Economy - Mpg with 2-liter V4 and manual transmission: city, 21; expressway, 35. Gallons used in 15,000 miles, 535. Cruising range, 475 miles.

Comments - The Prelude is responsive and peppy, and it handles well also. The standard version has plenty of acceleration, but the Si has more power. Bumper test damage: none.

**Hyundai Excel**

Predicted Reliability - No data.

Fuel Economy - Mpg with manual transmission: city, 23; expressway, 45. Gallons used in 15,000 miles, 465. Cruising range, 360.

Comments - The Hyundai has an unusually low price. However, it is not up to the quality level of the better Japanese cars. Its engine hesitates at times and its acceleration, handling, braking and rear seat comfort are below par. The reliability of the car is uncertain, but if its price stays below market, it could be a good buy. Bumper test damage: none.

**Isuzu I-Mark**

Predicted Reliability - Better than average.

Fuel Economy - Mpg with 1.8-liter engine and automatic transmission: city, 19; expressway, 35. Gallons used in 15,000 miles, 575. Cruising range, 380 miles.

Comments - The I-Mark is an older rear wheel drive design. Its reliability is likely to be above average.

### Mazda GLC

Predicted Reliability - Better than average.

Fuel Economy - Mpg with manual transmission: city, 24; expressway, 49. Gallons used in 15,000 miles, 430. Cruising range, 430 miles. Mpg with 1.5-liter engine and automatic transmission: city, 21; expressway, 33. Gallons used in 15,000 miles, 550. Cruising range, 315 miles.

Comments - While not among the best small cars in some respects, such as ride or noise level, the GLC is reliable, handles well and gives good gas mileage.

### Mazda 323

Predicted Reliability - No data.

Fuel Economy - Mpg with manual transmission: city, 22; expressway, 42. Gallons used in 15,000 miles, 480. Cruising range, 420 miles.

Comments - This front wheel drive model gives both good acceleration and good fuel economy. The 323 is a good car buy when available at competitive prices. Bumper test damage: none.

### Mercury Lynx

Predicted Reliability - For gasoline V4, average. For diesel, no data.

Fuel Economy - Mpg with gasoline engine and manual transmission: city, 21; expressway, 41. Gallons used in 15,000 miles, 505. Cruising range, 410 miles.

Comments - The Lynx is a better small car choice than the traditional U.S. built small cars. Its repair history is better, and it performs somewhat better. However, it does not have the passenger room and ride comfort of the high rated small cars, and suffers from some poor design of its controls. Service and parts should be widely available and repair costs relatively low. Bumper test damage: none.

### Mitsubishi Tredia

Predicted Reliability - Better than average.

Fuel Economy - Mpg with manual transmission: city, 21; expressway, 40. Gallons used in 15,000 miles, 510. Cruising range, 405 miles.

Comments - The Tredia is a satisfactory but not outstanding car. Its reliability should be good, but is below par in accomodations and handling. Bumper test damage: none.

**Nissan Sentra**

Predicted Reliability - For gasoline model, better than average. For diesel model, no data.

Fuel Economy - Mpg with gasoline V4 and manual transmission: city, 24; expressway, 45. Gallons used in 15,000 miles, 440. Cruising range, 510 miles. Mpg with gasoline V4 and automatic transmission: city, 20; expressway, 36. Gallons used in 15,000 miles, 545. Cruising range, 390 miles.

Comments - The Sentra is a competent car worth considering. It gives good fuel economy and rides more stiffly than most small cars.

Nissan 200SX

Predicted Reliability - Better than average.

Fuel Economy - Mpg with turbocharged V4 and manual transmission: city, 20; expressway, 39. Gallons used in 15,000 miles, 545. Cruising range, 405 miles.

Comments - The 200SX is a good all-around performer. The turbo version has especially strong acceleration.

### Plymouth Colt

Predicted Reliability - Much better than average, and repair costs are low.

Fuel Economy - Mpg with manual transmission: city, 23; expressway, 45. Gallons used in 15,000 miles, 465. Cruising range, 410 miles.

Comments - The Colt is among the better small cars in overall quality, although its acceleration, ride and heating system are a little below par. Colts are likely to be priced a bit below most top selling small cars.

### Plymouth Horizon

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 1.6-liter V4 and manual transmission: city, 22; expressway, 45. Gallons used in 15,000 miles, 475. Cruising range, 450 miles. Mpg with 2.2-liter V4 and automatic transmission: city, 19; expressway, 33. Gallons used in 15,000 miles, 595. Cruising range, 335 miles.

Comments - The repair record for this car remains much worse than average up through the latest data. Buy one only if the price and mileage are so low that coping with potential problems seems worth it.

**Plymouth Turismo**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 1.6-liter V4 and manual transmission: city, 22; expressway, 45. Gallons used in 15,000 miles, 475. Cruising range, 450 miles. Mpg with 2.2-liter V4 and automatic transmission: city, 19; expressway, 33. Gallons used in 15,000 miles, 595. Cruising range, 335 miles.

Comments - The repair record for this car remains much worse than average up through the latest data. Buy one only if the price and mileage are so low that coping with potential problems seems worth it.

**Pontiac Fiero**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 2.5-liter V4 and 4 speed manual transmission: city, 20; expressway, 38. Gallons used in 15,000 miles, 535. Cruising range, 305 miles. Mpg with V6 and 4 speed manual transmission: city, 17; expressway, 31. Gallons used in 15,000 miles, 635. Cruising range, 255 miles.

Comments - The Fiero has improved over previous years. The performance suspension is desirable. Bumper test damage: none.

**Pontiac 1000**

Predicted Reliability - For gasoline version, much worse than average.

Fuel Economy - Mpg with gasoline V4 and automatic transmission: city, 19; expressway, 37. Gallons used in 15,000 miles, 555. Cruising range, 370 miles.

Comments - Passenger space is tight, especially width. The 4 door has better seating than the 2 door. Reliability has been poor, though major components have not been particularly troublesome.

**Porsche 944**

Predicted Reliability - Data over the past three years shows few trouble spots, but costs of maintenance and repair have been very high.

Fuel Economy - Mpg with manual transmission: city, 16; expressway, 35. Gallons used in 15,000 miles, 635. Cruising range, 560 miles.

Comments - The 944 accelerates, handles and stops extremely well. Bumper test damage: none.

**Renault Alliance**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 1.4-liter V4 and 5 speed manual transmission: city, 21; expressway, 45. Gallons used in 15,000 miles, 495. Cruising range, 395 miles. Mpg with 1.4-liter V4 and automatic transmission: city, 22; expressway, 37. Gallons used in 15,000 miles, 520. Cruising range, 370 miles. Mpg with 1.7-liter V4 and manual transmission: city, 25; expressway, 44. Gallons used in 15,000 miles, 445. Cruising range, 445 miles.

Comments - The Alliance is a reasonably good car, but the driving position, controls and rear seating are not good. Because its repair record is much worse than average, the Alliance is not apt to be a good buy. Bumper test damage: minor.

**Renault Encore**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with 1.4-liter V4 and 5 speed manual transmission: city, 21; expressway, 45. Gallons used in 15,000 miles, 495. Cruising range, 395 miles. Mpg with 1.4-liter V4 and automatic transmission: city, 22; expressway, 37. Gallons used in 15,000 miles, 520. Cruising range, 370 miles. Mpg with 1.7-liter V4 and manual transmission: city, 25; expressway, 44. Gallons used in 15,000 miles, 445. Cruising range, 445 miles.

Comments - The Encore is a reasonably good car, but the driving position, controls and rear seating are not good. Because its repair record is much worse than average, the Encore is not apt to be a good buy. Bumper test damage: minor.

**Subaru**

Predicted Reliability - Much better than average.

Fuel Economy - Mpg with manual transmission: city, 21; expressway, 45. Gallons used in 15,000 miles, 485. Cruising range, 555 miles.

Comments - The Subaru sedan and wagon are worthy of consideration. They provide good seating, very good handling and good gas mileage. On the bad side, fresh air ventilation with air conditioning is poor, and the engine growls noticeably when accelerating. Bumper test damage: extensive.

**Toyota Corolla**

Predicted Reliability - Much better than average. Low maintenance and repair costs.

Fuel Economy - MPg with manual transmission: city, 23; expressway, 48. Gallons used in 15,000 miles, 445. Cruising range, 500 miles. Mpg with 3 speed automatic transmission: city, 21; expressway, 43. Gallons used in 15,000 miles, 495. Cruising range, 445 miles.

Comments - The Corolla is one of the best all-around small cars. It has excellent reliability and low repair costs. Bumper test damage: none.

**Toyota Tercel**

Predicted Reliability - Much better than average.

Fuel Economy - Mpg with 5 speed manual transmission: city, 23; expressway, 46. Gallons used in 15,000 miles, 460. Cruising range, 420 miles. Mpg 6 speed four wheel drive transmission: city, 22; expressway, 38. Gallons used in 15,000 miles, 520. Cruising range, 380 miles. Mpg with automatic transmission: city, 23; expressway, 42. Gallons used in 15,000 miles, 475. Cruising range, 445 miles.

Comments - The Tercei is a top choice in the small car market. It is a very reliable car and has low operating costs. The Tercel station wagon model is quite roomy and, with four wheel drive, is an excellent performer.

**Volkswagen Golf**

Predicted Reliability - Average.

Fuel Economy - Mpg with gasoline V4 and manual transmission: city, 21; expressway, 40. Gallons used in 15,000 miles, 500. Cruising range, 490 miles.

Comments - The most recent data indicates that the Golf's repair record is average. Bumper test damage: extensive.

**Volkswagen Jetta**

Predicted Reliability - Much better than average.

Fuel Economy - Mpg with gasoline V4 and manual transmission: city, 20; expressway, 40. Gallons used in 15,000 miles, 520. Cruising range, 480 miles.

Comments - The European-built Jetta has an improved repair record over previous years. Expect this model to be expensive. Bumper test damage: extensive.

**Yugo GV**

Predicted Reliability - No data.

Fuel Economy - Mpg: city, 24; expressway, 42. Gallons used in 15,000 miles, 465. Cruising range, 255 miles.

Comments - The Yugo's low price cannot make up for its shortcomings. Bumper test damage: extensive.

## Appendix E: Database for Menu Items in the Right Parse Window

**Acura Legend**

Predicted Reliability - No data.

Fuel Economy - Mpg with automatic transmission: city, 15; expressway, 33. Gallons used in 15,000 miles, 695. Cruising range, 400 miles.

Comments - The Legend is the senior model in a new line of cars made by Honda and marketed by Acura dealers. It compares very favorably to the better European sports sedans in ride and handling, and is roomy inside and quite comfortable. Bumper test damage: none.

**Audi 4000S**

Predicted Reliability - Better than average. Repair costs above average.

Fuel Economy - Mpg with 2.2-liter V5 manual transmission: city, 18; expressway, 31. Gallons used in 15,000 miles, 625. Cruising range, 470 miles.

Comments - The basic Audi 4000 uses the VW 1.8-liter V4. The Quattro and Coupe GT use the 5 cylinder engine from the Audi 5000S. The Quattro is an excellent performer. Bumper test damage: extensive.

**BMW 318i**

Predicted Reliability - Better than average. However, costs of maintenance and repair have been very high.

Fuel Economy - Mpg with manual transmission: city, 20; expressway, 34. Gallons used in 15,000 miles, 565. Cruising range, 430 miles.

Comments - The 318i does not have the snappy response that one expects from a BMW sports sedan, but it does deliver good fuel economy.

**Buick Skyhawk**

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with 2-liter V4 and automatic transmission: city, 16; expressway, 35. Gallons used in 15,000 miles, 640. Cruising range, 335 miles. Mpg with 1.8-liter V4 and 5 speed manual transmission: city, 20; expressway, 42. Gallons used in 15,000 miles, 515. Cruising range, 445 miles.

Comments - This GM J-car has improved, but is still not up to the level of its Japanese competitors.

**Buick Skylark**

Predicted Reliability - For V4, average. For V6, much worse than average.

Fuel Economy - Mpg with V4 and automatic transmission: city, 18; expressway, 38. Gallons used in 15,000 miles, 565. Cruising range, 405 miles. Mpg with V6 and automatic transmission: city, 16; expressway, 32. Gallons used in 15,000 miles, 645. Cruising range, 350 miles.

Comments - This N-car from GM performs well. Bumper test damage: moderate.

**Buick Somerset**

Predicted Reliability - For V4, average. For V6, much worse than average.

Fuel Economy - Mpg with V4 and automatic transmission: city, 18; expressway, 38. Gallons used in 15,000 miles, 565. Cruising range, 405 miles. Mpg with V6 and automatic transmission: city, 16; expressway, 32. Gallons used in 15,000 miles, 645. Cruising range, 350 miles.

Comments - This N-car from GM performs well. Bumper test damage: moderate.

### Cadillac Cimmaron

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with V6 and automatic transmission: city, 15; expressway, 28. Gallons used in 15,000 miles, 720. Cruising range, 295 miles.

Comments - The Cimmaron has considerably more up-market content than the other J-cars from GM. Handling and comfort are quite good. Bumper test damage: moderate.

### Chevrolet Cavalier

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with 2-liter V4 and automatic transmission: city, 16; expressway, 35. Gallons used in 15,000 miles, 640. Cruising range, 335 miles. Mpg with 1.8-liter V4 and 5 speed manual transmission: city, 20; expressway, 42. Gallons used in 15,000 miles, 515. Cruising range, 445 miles.

Comments - This GM J-car has improved, but is still not up to the level of its Japanese competitors.

**Dodge Conquest**

Predicted Reliability - No data.

Fuel Economy - Mpg with manual transmission: city, 17; expressway, 33. Gallons used in 15,000 miles, 625. Cruising range, 525 miles.

Comments - The Conquest is a good all-around performer. It has strong turbocharged acceleration.

**Ford Tempo**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with gasoline engine and 5 speed manual transmission: city, 19; expressway, 41. Gallons used in 15,000 miles, 530. Cruising range 445 miles. Mpg with gasoline engine and automatic transmission: city, 18; expressway, 34. Gallons used in 15,000 miles, 590. Cruising range, 425 miles.

Comments - The Tempo is equivalent in performance to GM's J-cars. Bumper test damage: none.

**Honda Accord**

Predicted Reliability - Much better than average.

Fuel Economy - Mpg with automatic transmission: city, 19; expressway, 40. Gallons used in 15,000 miles, 550. Cruising range, 460 miles.

Comments - The Accord is one of the better compact cars and a good dollar value. Bumper test damage: none.

**Isuzu Impulse**

Predicted Reliability - Average.

Fuel Economy - Mpg with manual transmission: city, 17; expressway, 37. Gallons used in 15,000 miles. 590. Cruising range, 440 miles.

Comments - The Impulse does not perform up to the level of the better compact cars.

**Mazda 626**

Predicted Reliability - Much better than average.

Fuel Economy - Mpg with 5 speed manual transmission: city, 20; expressway, 38. Gallons used in 15,000 miles, 530. Cruising range, 505 miles. Mpg with automatic transmission: city, 18; expressway, 33. Gallons used in 15,000 miles, 595. Cruising range, 445 miles.

Comments - The 626 is a good compact car. The manual transmission version will offer considerably better overall performance than the automatic version. Bumper test damage: none.

**Mercedes-Benz 190**

Predicted Reliability - For gasoline model, much better than average. For diesel model, no data.

Fuel Economy - Mpg with gasoline V4 and automatic transmission: city, 23; expressway, 33. Gallons used in 15,000 miles, 530. Cruising range, 420 miles.

Comments - This smaller Mercedes is an excellent performer but commands a high price. Rear seating comfort is not what it should be. The diesel model has a more powerful engine.

**Mercury Topaz**

Predicted Reliability - Much worse than average.

Fuel Economy - Mpg with gasoline engine and 5 speed manual transmission: city, 19; expressway, 41. Gallons used in 15,000 miles, 530. Cruising range 445 miles. Mpg with gasoline engine and automatic transmission: city, 18; expressway, 34. Gallons used in 15,000 miles, 590. Cruising range, 425 miles.

Comments - The Topaz is equivalent in performance to GM's J-cars. Bumper test damage: none.

**Mitsubishi Cordia**

Predicted Reliability - Better than average.

Fuel Economy - Mpg with manual transmission: city, 21; expressway, 40. Gallons used in 15,000 miles, 510. Cruising range, 405 miles.

Comments - The Cordia is the sporty version of the Mitsubishi Tredia sedan.

**Mitsubishi Galant**

Predicted Reliability - Average.

Fuel Economy - Mpg: city, 17; expressway, 38. Gallons use in 15,000 miles, 590. Cruising range, 445 miles.

Comments - The Galant is an excellent "high tech" compact with a roomy and comfortable rear seat. Bumper test damage: none.

**Nissan Maxima**

Predicted Reliability - Much better than average.

Fuel Economy - Mpg: city, 15; expressway, 32. Gallons used in 15,000 miles, 680. Cruising range, 385 miles.

Comments - The Maxima is the top of the line for Nissan. It is a powerful performer, but its seating package is not the best. Bumper test damage: none.

**Nissan Stanza**

Predicted Reliability - Much better than average.

Fuel Economy - Mpg for wagon with automatic transmission: city, 15; expressway, 32. Gallons used in 15,000 miles, 675. Cruising range, 400 miles.

Comments - The sedan versions of the Stanza have been redesigned. The Stanza wagon is tall with sliding doors on each side and lots of head room and cargo volume. A four wheel drive version is also available. Bumper test damage: none.

**Oldsmobile Calais**

Predicted Reliability - For V4, average. For V6, much worse than average.

Fuel Economy - Mpg with V4 and automatic transmission: city, 18; expressway, 38. Gallons used in 15,000 miles, 565. Cruising range, 405 miles. Mpg with V6 and automatic transmission: city, 16; expressway, 32. Gallons used in 15,000 miles, 645. Cruising range, 350 miles.

Comments - This N-car from GM performs well. Bumper test damage: moderate.

**Oldsmobile Firenza**

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with 2-liter V4 and automatic transmission: city, 16; expressway, 35. Gallons used in 15,000 miles, 640. Cruising range, 335 miles. Mpg with 1.8-liter V4 and 5 speed manual transmission: city, 20; expressway, 42. Gallons used in 15,000 miles, 515. Cruising range, 445 miles.

Comments - This GM J-car has improved, but is still not up to the level of its Japanese competitors.

**Peugot 505**

Predicted Reliability - Average to worse than average. Costs of maintenance and repairs have been fairly high.

Fuel Economy - Mpg for wagon with automatic transmission: city, 15; expressway, 28. Gallons used in 15,000 miles, 720. Cruising range, 400 miles.

Comments - The 505 wagon is a real cargo hauler with a very large and useful volume. The acceleration is fairly sluggish, but a turbocharged version is available. Bumper test damage: none.

## Plymouth Conquest

Predicted Reliability - No data.

Fuel Economy - Mpg with manual transmission: city, 17; expressway, 33. Gallons used in 15,000 miles, 625. Cruising range, 525 miles.

Comments - The Conquest is a good all-around performer. It has strong turbocharged acceleration.

## Pontiac Grand AM

Predicted Reliability - For V4, average. For V6, much worse than average.

Fuel Economy - Mpg with V4 and automatic transmission: city, 18; expressway, 38. Gallons used in 15,000 miles, 565. Cruising range, 405 miles. Mpg with V6 and automatic transmission: city, 16; expressway, 32. Gallons used in 15,000 miles, 645. Cruising range, 350 miles.

Comments - This N-car from GM performs well. Bumper test damage: moderate.

**Pontiac Sunbird**

Predicted Reliability - Worse than average.

Fuel Economy - Mpg with 2-liter V4 and automatic transmission: city, 16; expressway, 35. Gallons used in 15,000 miles, 640. Cruising range, 335 miles. Mpg with 1.8-liter V4 and 5 speed manual transmission: city, 20; expressway, 42. Gallons used in 15,000 miles, 515. Cruising range, 445 miles.

Comments - This GM J-car has improved, but is still not up to the level of its Japanese competitors.

**Saab 900**

Predicted Reliability - For turbocharged model, average. For non-turbo model, better than average.

Fuel Economy - Mpg with turbo engine and manual transmission: city, 18; expressway, 32. Gallons used in 15,000 miles, 620. Cruising range, 420 miles.

Comments - The Saab turbo is a powerful performer, but the standard model is peppy as well. Both models command high prices. Repair costs will be high. Bumper test damage: none.

## Toyota Camry

Predicted Reliability - Much better than average.

Fuel Economy - Mpg with gasoline V4 and 5 speed manual transmission: city, 23; expressway, 46. Gallons used in 15,000 miles, 455. Cruising range, 510 miles. Mpg with gasoline V4 and automatic transmission: city, 19; expressway, 44. Gallons used in 15,000 miles, 515. Cruising range, 490 miles.

Comments - The Camry is a good all-around performer, gives good gas mileage and has a comfortable rear seat. Bumper test damage: none.

## Toyota Cressida

Predicted Reliability - Much better than average.

Fuel Economy - Mpg: city, 15; expressway, 29. Gallons used in 15,000 miles, 700. Cruising range, 450 miles.

Comments - The Cressida is a very powerful performer. Bumper test damage: none.

**Toyota MR2**

Predicted Reiiability - Much better than average.

Fuel Economy - Mpg with manual transmission: city, 25; expressway, 45. Gallons used in 15,000 miles, 435. Cruising range, 415 miles.

Comments - The MR2 is an agiie and responsive car. There is very little room inside. Bumper test damage: none.

**Volkswagen Quantum**

Predicted Reliability - No data.

Fuel Economy - Mpg with automatic transmission: city, 14; expressway, 26. Gallons used in 15,000 miles, 765. Cruising range, 335 miles.

Comments - The Quantum is typical of European sports sedans in its ride and handling qualities. It is basically an Audi 4000 at a lower price. Bumper test damage: moderate.

**Volvo DL**

Predicted Reliability - Better than average.

Fuel Economy - Mpg with automatic transmission: city, 18; expressway, 32. Gallons used in 15,000 miles, 620. Cruising range, 395 miles.

Comments - The Volvo DL is an old design but it still performs well.

**Volvo 240**

Predicted Reliability - Average.

Fuel Economy - Mpg with automatic transmission: city, 18; expressway, 32. Gallons used in 15,000 miles, 620. Cruising range, 395 miles.

Comments - The compact Volvo 240 is a durable and competent model.

## Appendix F: Background Questionnaire

Please indicate your computer-related experience below.

In the past year I worked with a computer:

___ frequently

___ occasionally

___ rarely

| | |
|---|---|
| Types of experience: | Length of experience (months): |
| ___ Data Entry | ____________________ |
| ___ Word Processing | ____________________ |
| ___ Text Editing | ____________________ |
| ___ Programming | ____________________ |
| ___ Other (describe) | ____________________ |

## Appendix G: Bipolar Rating Scales

Please rate the system that you just used by marking the scales below. Overall, the system was:

simple I_____I_____I_____I_____I_____I complex
very slightly neutral slightly very

weak I_____I_____I_____I_____I_____I powerful
very slightly neutral slightly very

fatiguing I_____I_____I_____I_____I_____I relaxing
very slightly neutral slightly very

pleasing I_____I_____I_____I_____I_____I irritating
very slightly neutral slightly very

easy to use I_____I_____I_____I_____I_____I hard to use
very slightly neutral slightly very

natural I_____I_____I_____I_____I_____I unnatural
very slightly neutral slightly very

confusing I_____I_____I_____I_____I_____I clear
very slightly neutral slightly very

predictable I_____I_____I_____I_____I_____I unpredictable
very slightly neutral slightly very

meaningless I_____I_____I_____I_____I_____I meaningful
very slightly neutral slightly very

good I_____I_____I_____I_____I_____I bad
very slightly neutral slightly very

## Appendix H: Ranking Form

Please rank the two systems that you have just used by giving a rank of "1" to the system you liked the most and a rank of "2" to the system you liked the least.

First system _____ Second system _____

# VITA

Jeff Hendrickson was born in Lubbock, Texas, on April 26, 1957. In May of 1979 he received the Bachelor of Arts degree in Psychology from Texas Tech University. In May of 1984 he received the Master of Arts degree in Experimental Psychology from Stephen F. Austin State University. In the Fall of 1984 he entered the Graduate School of Old Dominion University, and subsequently received the Doctor of Philosophy degree in Industrial/Organizational Psychology in May, 1988. Jeff performed his doctoral internship at the User Systems Engineering Center of Texas Instruments in Dallas, Texas. In May of 1988 he joined the technical staff at Texas Instruments as a Human Factors Engineer.