CONCURRENT VERSUS RETROSPECTIVE VERBAL PROTOCOL FOR COMPARING WINDOW USABILITY

by

Victoria A. Bowers

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Industrial Engineering and Operations Research

APPROVED:

H L Snyder

Harry L. Snyder, Chairman

Daniel T. Donohoo

Brian W. Epps

Paul T. Kemmerling

John W. Roach

Robert C. Williges

January, 1990

Blacksburg, Virginia

Concurrent Versus Retrospective Verbal Protocol For
Comparing Window Usability

Victoria A. Bowers

(ABSTRACT)

The measurement of software usability has become an important issue in recent years. Metrics of usability include time, errors, questionnaires, ratings, and results of verbal protocols. Concurrent verbal protocol, a method in which the user "thinks aloud" while completing given tasks, has been heavily employed by software usability researchers who want to know the reason a user is having difficulties. Possible problems associated with using concurrent verbal protocol are (1) that verbalization may interfere with the processing required to complete the task, and (2) that subjects may not be able to monitor and express the information of interest to the researcher. A relatively new approach which may avoid these problems is heavily cued retrospective verbal protocol in which the user is presented subsequently with a representation (a video tape, for example) which helps him recall his thoughts during the task without interfering with task completion.

This research compared the performance of subjects while completing tasks using both methods of verbal protocol. The verbal data collected by the two protocol techniques was compared to assess any information differences due to the methods of collection. No performance differences were found between the two protocol methods. Reasons for this lack of degradation due to concurrent verbalization are discussed. The kinds of information gathered were quite different for the two methods, with concurrent protocol subjects giving procedural information and retrospective protocol subjects giving explanations and design statements. Implications for usability testing are discussed.

The two methods of protocol were employed in a comparison of two different size monitors, a 30.48 cm diagonal and a 53.34 cm diagonal. The subjects' performance, as measured by steps to completion, task completion time, and errors committed, was compared across the two monitors. Subjects were required to complete 12 tasks which varied in the difficulty of the windowing required. Subjective data were also collected in the form of task difficulty ratings, as well as a global measure of user satisfaction. These performance measures and subjective measures were compared across protocol methods as well as monitors. Performance data, as well as subjective data, indicate that on tasks that do not require extensive windowing, there are no difference between the two monitor sizes. As windowing difficulty increases, however, the large monitor's advantages become apparent. Tasks with a high level of windowing difficulty are judged to be easier and require fewer steps on the large monitor than on the small monitor.

Acknowledgments

I would first like to acknowledge Dr. Brian Epps, Dr. Daniel Donohoo, Professor Paul Kemmerling, Dr. John Roach, and Dr. Robert Williges for their insights and contributions to this research. I am truly grateful to Dr. Harry Snyder; without his guidance, encouragement, and friendship, this research would have never been completed. Dr. Snyder has provided me not only with a professional role model but with a personal one. Willard W. Farley is acknowledged for his efforts and patience during the more frustrating times of this effort. I thank Deborah Cox for her skills and patience under the stress of deadlines. I wish to acknowledge Texas Instruments Corporation for their graduate internship and for allowing me to pursue my own interests.

I wish to thank my family for their support; special appreciation is given to my parents, John and Carol Bowers, for their complete faith in me throughout the years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Developing software that is "user friendly" is a major goal of designers. Usability has become a major economic issue in the last 10 years. Early in software development, the majority of software was written for programmers and computer scientists who were willing to cope with the challenge offered by software that was not easy to use. Today's software designer must realize that if his/her software application is not easy to use, the potential purchaser of the software has the option to choose another, equally effective, software package that is easier to use. While designers are aware of the need to create software that is easy to use, they may not be achieving their goal because they are unaware of the process by which ease of use may be accomplished (Gould and Lewis, 1983).

Gould and Lewis (1983) interviewed software system designers and developers to discover whether they employed several principles that should be used in software design. Gould and Lewis contended that a software interface system should be designed with four goals in mind: it should be (1) easy to learn, (2) useful, (3) easy to use, and (4) pleasant to use. They further suggest that to attain these goals four principles must be followed. First, designers must understand who the users will be. Second, the expected users should work in conjunction with the design team during the formative stages of development. Third, early prototypes should be tested with users on representative tasks. Fourth, when problems are found in testing, they should be fixed. Lewis and Gould conducted a survey which asked system designers, planners, and developers to write down the five major steps in developing and evaluating a computer system for end-users. They found that of those polled, only 2% included all four principles, while 26% did not mention any of them. Closer examination revealed that the goals mentioned by the designers were to make the system "easy to use," "user friendly," and "simple." These goals, however, are difficult to reach without an explicit process which includes the four design principles.

Defining Usability

Usability must be carefully defined before it can be included in design goals. "Usability is concerned with the process of use, the steps the user must take: *how* the user achieves results" (Bennett, 1984, p. 162). It is related to ease of learning and ease of use, but these must be considered in the context of the user, task, tool, and environment (Schackel, 1984). Not only must ease of use be considered in usability but the performance of the user on the system must be considered as well. According to Schackel, four goals which need to be met for user acceptance are: (1) Learnability - a certain level of performance is acquired by a sample of intended users in an allotted time for training, (2) Throughput - the required task can be accomplished by a percentage of intended users in an allotted time span with less than a given number of errors, (3) Flexibility - the system must be adaptable to a range of environment and skill levels of intended users, and (4) Attitude - the user must want to continue to use the system. The usability goals must be quantifiable and measurable in order to evaluate usability. Thus, usability must be operationally defined.

USABILITY METRICS

Many measurements of usability are presently employed in usability testing. These metrics are often used in combination to evaluate whether usability goals have been met. These metrics include time to learn, time to complete a task, errors committed, time spent recovering from errors, knowledge of the system after a given amount of training, verbal protocol data, and subjective data. The current literature suggests that different metrics of usability may be employed to answer different types of usability questions. Measuring usability can be very molecular, as in the keystroke-level method, to very molar, as in assessing general knowledge about the system from a picture probe method. Each metric reveals its own type of information and may pertain to specific usability goals. While the suggestion might be made to develop a unified usability metric, using one metric of usability would be an oversimplification. Perhaps an integrated optimal combination of usability metrics should be used to glean the most information about usability. To compare the possible advantages and disadvantages of specific metrics, a review of available metrics seems appropriate.

Time and Error Measurements

Perhaps one of the most often used methods of measuring the usability of software is the time the user takes to learn to perform prescribed tasks and the time to perform these tasks once they have been learned. These measurements are direct measures of the first two usability goals, learnability and throughput. One method for estimating throughput is the keystroke level-model of user performance, which analytically predicts the amount of time an expert on a system would take to perform a specified task with no errors (Card, Moran, and Newell, 1980). This model is based on the assumption that the amount of time it takes an expert to do a task is determined by the time taken to do keystrokes as well as other tasks which are determined at the keystroke-level such as homing, pointing, mental operations, response, and drawing. The given task is broken down into the keystroke-level operations which are required to perform that task. The time of each of these operations is summed to result in the total task time. Card et al. (1980) empirically validated the keystroke-level model of user performance time. Twenty-eight expert users were tested across 14 tasks on 10 different systems. The execution times were calculated using the model and then compared to empirical execution times of the expert users. The empirical times could be well predicted from the model. Given this task time predictor from analytic data, the authors suggest that the model could be used to estimate the time of benchmark tasks on systems; this approach would enable the researcher to compare the systems without empirical research.

Another standard method for comparing usability across systems was developed by Roberts and Moran (1983). This method measures two of Schackel's usability goals, learnability and throughput. The method specifically compares text editors by comparing those editors on a benchmark task which is made up of core tasks that all text editors must perform. While the methodology was specifically developed for use with text editors, the basic idea of comparing systems on a benchmark task could be employed with any software applications.

The benchmark task in Roberts and Moran's comparison consisted of 53 editing tasks. This evaluation involved novice users as well as experts, and included measurements of time, errors, learning, and functionality. Expert subjects were required to do these tasks on each of nine editors. During the session, an observer measured the overall time spent on the task and the amount of time the subject spent correcting errors. The expert's score, then, was a combination of the error time and error free time. The four experts' scores were combined for each editor. Four novices were also trained for a given amount of time on each editor to arrive at a score for the learning dimension. This learning score was the average time it took a subject to learn a task. The score was reached by dividing the total session time by the number of core tasks the subject learned. Finally, functionality for each editor was determined by the percentage of all possible tasks that the editor could perform.

Roberts and Moran (1983) not only compared the nine text editors, but also compared the resulting times with those predicted by the Card et al. keystroke-level model. The researchers found a very high correlation between the predicted and actual times (r = .9), thus lending further validity to the keystroke-level model. The authors stated that this method proved successful in comparing and evaluating text editors.

The Roberts and Moran method was replicated in a study by Borenstein (1985) which compared three text editors, one which had been previously studied in the original Roberts and Moran study and two that had not. In general, the Roberts and Moran results were replicated and their benchmarking method validated. While Borenstein followed the same general procedure as Roberts and Moran, several additional measurements were made. Borenstein, as did Roberts and Moran, timed the subjects with a stop watch; in addition, however, he videotaped the sessions. Through watching the tapes, Borenstein discovered that errors under 15 seconds in length, which were ignored by Roberts and Moran, made up nearly 50% of the error time. Borenstein suggested that videotaping might be a better means of timing if more careful time measurements are required.

Another study that used time and error measures to ascertain whether acceptable usability had been attained was conducted by Butler (1985) in evaluating a financial analysis package. The subjects were given a manual for the software package and a problem which they had to solve. An accepted time criterion was set at 180 minutes; the subjects had to solve the problem without an "unrecoverable error" in this time. All the subjects were able to meet the criterion. This requirement was felt to meet the learnability goal of usability.

Usability measures employing time and error data were employed by Brewley, Roberts, Schroit, and Verplank (1983) in the development of the Xerox 8010 "Star" office workstation. In one study, the usability of several text selection schemes by use of a mouse was compared. Six groups of subjects performed text editing tasks with each of six mouse button configurations. Mean selection times and numbers of errors were compared to measure the usability goals of learnability and throughput. These results were used to create another mouse configuration which was evaluated using the same method. This new configuration had the shortest time needed for selection of text compared to all the others and was eventually incorporated into the "Star" design.

The "Star" work station developers also employed time and error measurements to help select the best icons. Subjects were first given pictures of all the icons to discover their ability to name them. In the first "round," subjects were given each icon one at a time and asked to describe what they thought it was. The subjects were then given the entire set and the opportunity to change any descriptions. Next, the subjects were given names and descriptions and had to choose the icon which best fit the description; finally, the subjects were given names for all the icons and asked to match them. Mislabeling or mismatching of the icons was considered an error. To take time measures, subjects were presented with the all icons randomly displayed on a VDT. A particular icon was named and the subject had to point as quickly as possible to that icon. The times taken to locate the given icons were combined graphically with the errors in labeling and matching the icons. These combined time and error data were employed to select the best icon.

Whiteside, Jones, Levy, and Wixon (1985) also used time scores to compare usability of different interactive dialog systems. Subjects were required to complete a standardized benchmark task on each of the dialog systems. Whiteside et al. noted that the use of a benchmark task was similar to the Roberts and Moran approach. Further, the use of benchmark tasks to compare dialog systems shows that benchmarking can be extended past use with text editors to compare other interfaces as well. Again, time was used as a measure of the learnability and throughput goals. The benchmark task was a file manipulation task which included tasks such as displaying, merging, and sending files. A user performance score which reflects rate of task completion was computed by the equation

$$S = PC / T \qquad (1)$$

where T = time spent on task in minutes, P = percentage of task completed, C = arbitrary time unit constant of five minutes, and S = user's performance score

The time constant is the estimated value of the shortest possible time in which an expert could complete the task. Thus, the user performance score is the percentage of the task that could be completed by the user per 5-minute interval. The usability of the interactive systems was compared across the user performance measures on the benchmark task. This holistic comparison of systems on benchmark tasks seems to be the most realistic way to compare usability across systems because critical information about the way separate elements of that system interact may be lost if each part of the system is studied individually rather than studying the system as a whole.

The Whiteside et al. (1985) user performance measure was employed by Good, Spine, Whiteside, and George (1986) to compare usability of an improved version of software for a windowing environment to that of an old version of the software. The benchmark task included creating windows, moving, pushing, and popping windows. A goal of a 20% improvement in user performance was set for users on the benchmark task; users exceeded this goal with a 37% increase in the user performance scores. This study again demonstrates the value of a benchmark task in comparing usability of interfaces.

Questionnaires and Ratings

While time and error measurements are the most frequently used metrics in usability, subjective ratings and questionnaires are probably the next most frequently used measures of usability. Often they are used to assess Schackel's usability goal of attitude, to discover if the user wants to continue to use the system. The ratings and questionnaires assess the subject's feelings about ease of use and ease of learning. Ratings have been found more successful, in general, than open-ended questions.

Several problems that exist with the questionnaire's use as a usability metric have been suggested (T. M. Spine, personal communication, September 15,1988). First, the questionnaire may not, in fact, measure the dimension that the experimenter believes is being measured. This problem of validity must be addressed with any measure that uses the self-report questionnaire format. Secondly, if not properly constructed, the data generated by questionnaires may not be readily interpreted with standard statistical tests. A third problem is that the global evaluation often provided by questionnaires does not include the detailed information necessary for correcting problems in the system. Finally, the questionnaire may not elicit an honest evaluation from the user; the user may not want to reveal that he found the system difficult as he might be perceived as inadequate or may not want to offend the experimenter. While these are real problems with the questionnaire format, a properly constructed questionnaire may circumvent them and be a useful tool for measuring user attitude.

Root and Draper (1983) assessed questionnaires as a software evaluation tool and found that the information gleaned from questionnaires could be categorized as (1) identifying problem areas and (2) evaluating specific features on the existing system. Root and Draper had subjects with various levels of computer experience evaluate a screen editor by completing a questionnaire. Two basic kinds of questions were used, checklists and open-ended questions. The checklists required subjects to rate the usefulness of existing commands on a scale of one to three. Subjects also rated the usefulness of commands that might be implemented. The ratings of existing commands were very consistent. In order to discover how well people predicted the usefulness of potential commands, those potential commands that had been rated were implemented; subjects then rated these commands after using them. The ratings of usefulness of potential commands correlated poorly with the actual ratings of those commands. This result indicates that subjects may not be good at predicting their opinions about features that they have not used. The open-ended questionnaire did result in many useful suggestions, but was more useful for more experienced users. Root and Draper suggested that the checklist ratings give consistent data about problems that already exist in the system. An open-ended question, however, may be useful when completed by experienced users to find features that should be implemented in the system.

The use of specific ratings has been found to be a successful measure of usability. The developers of the "Star" workstation (Brewley et al., 1983) used ratings in combination with their time and error data to help choose the best icon. After the subjects had completed the naming and recognition tasks discussed earlier, they were asked to choose any icons that were especially difficult or easy to recognize. They were also asked to rank sets of the icons.

The subjects' rankings of the icons corresponded to their performance, a result which seems to validate the rating and ranking methods.

Whiteside et al. (1985) also augmented performance measures of usability with subjective ratings of the interactive systems. After the subject had completed the benchmark task on a system, he/she gave an evaluative rating of the overall system which ranged from extremely positive to extremely negative. These semantic differential scales included pleasant-unpleasant, good-bad, nice-awful, helpful-unhelpful, and friendly-unfriendly. While the authors did find significant differences on preference ratings, these results were not fully discussed.

One problem with the Whiteside et al. (1985) rating scales for measuring preference is that they may be too global. Coleman, Williges, and Wixon (1985) suggested that rating specific parts on a system may be a better reflection of the attitude of the user than rating the overall system. To build a questionnaire for measuring the attitude of the user, Coleman et al. (1985) had users suggest adjectives that described user interfaces. The subjects' adjectives were added to a previous list of adjectives to create a new list on which subjects were asked to rate the importance of each adjective; a list of 17 adjectives resulted. Each of the final 17 adjectives was paired with its antonym to form 17 bipolar adjective rating scales. In the next phase of the study, subjects were asked to rate their overall satisfaction of a text editor on a 7-point scale ranging from Extremely Satisfied to Extremely Unsatisfied. Then they rated 16 functions of the editor on two 7-point scales ranging from Good to Bad and from Unimportant to Important. Finally, the subjects rated the 16 functions on the 17 adjective scales. The Good/Bad scales were collapsed across the

16 functions to yield a "function goodness" rating. The "goodness" rating was not significantly correlated to the overall satisfaction rating. All 17 bipolar scales were significantly correlated (r ranging from .153 to .455) with the "goodness function." Coleman et al. (1985) suggested that having subjects rate parts of a system is better for accurately assessing the users feelings about the system because the user is presented with cues to remind him of characteristics of the system.

Another global questionnaire for measuring user attitude, called the System Usability (SU) questionnaire, was suggested by Spine (T. M. Spine, personal communication, September 15,1988). This questionnaire includes 10 items that are positive or negative statements about the usability of the system. The user rates each of the items on a 5-point Likert-type scale which ranges from Strongly Agree to Strongly Disagree. The general attitude of the user towards the system may be determined by the System Usability questionnaire.

In an interesting implementation of the questionnaire method, Cordes (1984) used magnitude estimation for evaluating ease of use of software for a space management product involving several tasks. After the subject had completed each task he/she was asked to draw a line representing the task difficulty; the subject was given a reference line that represented the difficulty of logging on to the system. Using the information from the ratings of task difficulties, changes were incorporated in the space management system. The new system was tested in the same manner as the first. Because magnitude estimation results in ratio data, a direct comparison of the old and new management system could be made; Cordes concluded that the second system was twice as easy to use as the first.

In another unique implementation of the questionnaire format, del Galdo, Williges, Williges, and Wixon (1986) described a critical incident tool. This tool was used as an on-line questionnaire which required subjects to record any interface problems, "critical incidents," which they encounter during a benchmark task. When a critical incident occurred, the subject classified the incident as a success or failure and briefly described what happened. The subject rated the incident on a 7-point scale from extremely critical to extremely non-critical. These success and failure frequencies and ratings were used to compare different systems on benchmark tasks, as well as to pinpoint problems in order to correct them. del Galdo et al. (1986) suggested that this method is better than post-session questionnaires because the subjects may not be able to verbalize the interface problems in retrospect. This argument is also a justification for the use of verbal protocol.

Perhaps the questionnaire which has received the most vigorous testing to date is the Questionnaire for User Interface Satisfaction (QUIS), which is being developed at the Human-Computer Interaction Laboratory at the University of Maryland (Chin, Diehl, and Norman, 1988; Chin, Norman, and Shneiderman, 1987). The original questionnaire consisted of three sections. The first two sections ascertained the user's background, the type of system being evaluated, and the amount of time the user had spent using the system. The third section contained 103 nine-point bipolar scales on which to rate the system being evaluated. The structure is of special interest as each scale had the negative anchor on the left side and the questions were organized into 20 groups of scales related to certain aspects of the evaluated system. This structure may lead to artifacts in the data which will be discussed later.

The QUIS was administered to an introductory computer science class which was split into two groups, each of which learned one of two systems for six weeks and then learned the other system. QUIS was administered to each group after six weeks on the first system and again after six weeks on the other system. QUIS was, however, modified after the first administration by dropping 33 items that had low beta weights when a multiple regression was conducted on each of the 20 groups of questions. While differences in the systems were indeed found, the interest here is in the high reliability achieved by QUIS as measured by Cronbach's alpha. This measure of inter-correlation among questionnaire items was found to be high for both versions of QUIS (.89 and .94). In an effort to measure validity of the questionnaire, QUIS scores were compared with class performance; no significant relationship was found.

Another test of the reliability of QUIS was conducted with a further modified version, QUIS 5.0. The QUIS was administered to members of a PC users group who were asked to rate four software packages including one they liked and one they disliked. Again, an internal consistency measure of reliability yielded high reliability. The liked and disliked software packages were reliably discriminated by QUIS, a result which gives some measure of its validity. While the ability to differentiate liked and disliked software does lend validity to QUIS, it gives no indication whether QUIS would be sensitive enough to discriminate software in the mid-range between liked and disliked, or between software versions in either the liked of disliked categories.

Verbal Protocol

A frequently used method in usability testing is verbal protocol. In the most commonly employed method of verbal protocol, concurrent verbal protocol, the subject is required to "think aloud" while completing specified tasks on a given system. In other words, the subject is asked to verbalize what he is doing and why he is doing it while he is actually performing the task. A description of the methodology was presented by Lewis (1982) who suggested that the instructions be kept simple and that the subject be told to tell the experimenter what he is thinking while working through the tasks. The subject might be given examples of what types of thoughts to verbalize such as why he is confused and how he is making his decisions. Also, Lewis recommended that the subject be told that the experimenter is not interested in his secret thoughts and that the experimenter has no stake in the system and will not be upset if the subject has difficulties or does not like the system. Finally, perhaps most importantly, the experimenter should stress that the system is being evaluated, not the subject. While running the subject, the experimenter should prompt the subject to continue speaking, but must be careful not to interview the subject, but instead to remain an observer. Lewis recommended video taping for data collection as well as key-stroke capture. Finally, the analysis is more qualitative than traditional analysis due to the nature of the data collected. The statements made by the subjects should be summarized and categorized into episodes of a problem in the interface to determine the problems and why they occurred.

Lewis stated that concurrent verbal protocol is useful for pinpointing problems, discovering why a problem occurs, catching problems when they occur, and discovering how a subject feels about the system. Lewis also discussed disadvantages associated with concurrent verbal protocol such as its possible interference with the processing required for subjects to perform tasks, possible interference with time data, and the fact that the collected data do not lend themselves to traditional statistical analysis.

While arguments about the accuracy and usefulness of verbal protocol data date back to the introspectionist movement in psychology, perhaps two papers that best discuss the issues involved are those of Nisbett and Wilson (1977) and Ericsson and Simon (1980). The first issue is concerned with whether the subjects have the ability to monitor the processes of interest to the researcher. The second issue, and perhaps more important in usability testing, is whether the act of verbalizing interferes with the processes that are of interest to the experimenter. Because of the frequent use of verbal protocol in usability testing, it is crucial that these issues are understood and that the limitations of verbal protocol are recognized.

Nisbett and Wilson (1977) argued that it may be impossible for subjects to accurately monitor and express their higher order cognitive processes. They posited that (1) subjects cannot monitor the processes in order to report them; (2) when subjects do report the processes, their reports are not based on observations but on their own implicit theories about these processes; and (3) while reports may at times be correct, this is not a result of the subject actually being directly aware of the process. The conclusion that all information learned from verbal protocol is not accurate seems extreme; certainly, if a subject reports that he is confused during a task this would seem to be accurate information, regardless of the processes involved.

Perhaps the important question the experimenter might ask is whether the act of verbalizing his thoughts has changed the process that the subject employed to perform the task, thereby contributing to the subject's confusion. Ericsson and Simon (1980) proposed that verbalization which is unrelated to the task will interfere with task performance. They further proposed that verbalization about processes which are directly attended to during the task add negligible cognitive load. These processes which the subject describes must be processes that must be attended to during the task; if they do require further processing, the experimenter can expect interference. Ericsson and Simon posited that the distinction between these processes is whether the subject must query short-term memory or long-term memory in order to verbalize the information. They suggested that the decision to use verbal protocol should depend on the task and on the information desired. According to Ericsson and Simon, the information of interest may be retrieved in retrospective protocols if the protocols are taken immediately after the tasks are completed. They suggested, however, that the information taken in retrospective protocol is subject to elaboration or deterioration during the retrieval process. These problems during retrieval may be particularly damaging to the accuracy of the information if the subject completes similar tasks over a short period of time as is the case with most usability benchmark tasks. According to Ericsson and Simon, then, if the experimenter is interested in cognitive processing which must take place in order to complete the task, then concurrent verbal protocol not only does not interfere with the task performance but is the most appropriate means of measuring these processes. This point, however, has been argued by others.

Verbal protocol has been used a great deal in usability testing with most studies employing concurrent verbal protocol. Lund (1985) employed concurrent verbal protocol for assessing usability during the development of an interactive graphics system. Lund pointed out that one advantage of concurrent verbal protocol is that it calls attention to a problem the first time that problem is encountered. This issue is important because often as the test continues, the user adapts to an initially confusing item in the interface so that by the end of the session, the problem may not be cited by the user as a problem. Lund used both direct observation and review of videotapes to collect the verbal data. Reviewing video tapes may be time consuming, but use of this method as a backup to direct observation is important because the experimenter may not record all the critical information. One problem with concurrent verbal protocol that Lund discussed is that subjects may be stressed by the experimenter and video equipment observing and recording the interaction.

In order to avoid the problem of subjects feeling uncomfortable, O'Malley, Draper, and Riley (1984) used the constructive interaction verbal protocol method. In this verbal protocol method, two subjects work through a problem together, talking to each other. This avoids the stress of the subject having to verbalize directly to the experimenter. Another advantage of having subjects verbalize to each other is that if the ideas are not well communicated, the subjects will query each other to clarify the ideas. If the experimenter does not understand the subject in a traditional verbal protocol scenario, he must decide whether to query the subject, perhaps interrupting the subject's train of thought. Videotaping is important for data collection in this method.

Perhaps the most damaging problem associated with concurrent verbal protocol is that verbalizing during the task completion may in fact distort or disrupt the cognitive processes that are required to complete the task. In order to try to compensate for this, Bailey, Knox, and Lynch (1988) trained subjects in verbal protocol before the actual testing of the system. The verbal protocol training required the subjects to "think aloud" while they performed a construction task with an Erector® set. During the training task they verbalized their problem solving process; a beep was sounded if subjects were silent for more than three seconds. The training session lasted approximately 20 minutes. After training, the subjects performed the experimental task, a task which required subjects to make measurements with an oscilloscope, under the same conditions. The study categorized the statements of subjects to compare two oscilloscopes. A major problem with this study is that the intrusive beep in both conditions would likely be very distracting to the subject and therefore likely to change the processes involved in the tasks.

In another effort to avoid the possibility of concurrent verbalization changing normal cognitive processes, Barnard, Wilson, and MacLean (1986) collected verbal data *after* the experimental sessions were completed in a method called prompted knowledge elicitation (PKE), a kind of retrospective verbal protocol. In order to test the learnability of a work station interface, the subjects were trained on the interface and then presented with retrieval cues which were photographs of screens of the interface in different states. When shown these photographs the subjects were asked how they had arrived at that screen, what they might do while at that screen, and what they would do next. Specific questions were also asked about the screens. Two photographs of each screen were shown to the subjects; the first contained minimal contextual information while the second had more information and some data in the screen to give some clues as to the nature of the screen. This method has the advantage that the collection of the verbal protocol data does not interfere with performing the task. Two disadvantages are that (1) the subject may not recall what problems he had and (2) the data are collected after the subject has adapted to the interface so he may not mention initially confusing items. A method that may correct these two deficiencies was presented by Deffner (1988).

Deffner (1988) suggested the use of a heavily cued retrospective verbal protocol to avoid the possible interference of concurrent verbal protocol with cognitive processing. Deffner (1987) found that subjects required to verbalize during problem solving may not employ efficient strategies as quickly as subjects not required to verbalize. Deffner had subjects solve tasks which could be solved with two distinct strategies; one strategy was more efficient than the other. The task problems were n-term series problems in which the subject is given information about the spatial arrangement of a group of people sitting on a bench. They are asked a particular question about the arrangement given that information. Two strategies are available to solve this type of problem, the method of series formation and the method of elimination. In the more efficient strategy, the method of elimination, the subject searches for and compares pieces of the given information, while in the method of series formation the subject builds a spatial image of the group. Because eye track data result in two distinct normative sequences for the two strategies, eye track data were used to discover which strategy the subjects employed. Half the subjects were required to verbalize their problem solving while half were allowed to solve the problem in silence.

Deffner (1987) found that subjects required to verbalize their problem solving did not differ from the non-verbalizing subjects in their overall use of the two strategies. If, however, only the first half of the tasks are examined, subjects using verbal protocol are found to use the less efficient method for problem solving significantly more than those subjects not verbalizing. Subjects who discovered the more efficient strategy did so more quickly in the silent condition. Deffner concluded that this quicker discovery for the silent subjects is due to a more flexible response, while verbalizing subjects are impaired in flexibility. Again, Deffner suggested a heavily cued retrospective verbal protocol which presents detailed cues of the interaction to subjects after the interaction. He recommended eye track data be recorded and played back to subjects to help them remember the interaction. The advantages offered by this method are many. First, the collection of data does not interfere with the cognitive processes involved in the task. Second, the subjects are given such detailed cues of their interactions that they are reminded of the initial problems to which they may later adapt. Some might argue, however, that the information discovered in retrospective verbal protocol is not the information of interest.

Integration of the Literature

As demonstrated in the review of the metrics of usability, each of the measures provides different kinds of data. These data may be related to performance (time and error measures), to subjective attitudes (subjective ratings), or to a more general knowledge (verbal protocol data collection). Obviously, in choosing the correct metric or metrics, the researcher should consider the specific questions that he wishes to address and the kind of information which is most important. One major deficiency in many usability metrics that has plagued researchers is the inability of those metrics to yield specific information regarding why problems were encountered. Due to their dissatisfaction with these metrics, researchers have begun to rely heavily on concurrent verbal protocol which they feel allows the subject to express this "why" information as he encounter problems.

According to Ericsson and Simon (1980), if the subject is verbalizing only those processes to which he directly attends during task performance, the interference of verbalization should not be significant; the cognitive load incurred in this case would be minimal. The current study does not try to disprove this inherently untestable hypothesis because empirically defining the processes to which the subject "directly attends" during the task would be impossible. Instead, this study explores possible interference of verbal protocol on tasks of varying difficulty. As discussed earlier, some researchers believe that there will be no interference of verbalization if the subjects report processes to which they have direct access during the task performance. Others believe that there may be some alteration of the cognitive processing which may change task performance. If there are, in fact, cognitive changes, then the use of concurrent verbal protocol for measuring usability may be a problem. If the subjects do change their processing, results taken during concurrent verbal protocol may not be generalizable to real situations in which users are not required to verbalize cognitive processes.

The information gained through the different procedures of verbal protocol collection, concurrent and retrospective, is also of interest. Some researchers propose that the information gained during the actual task performance may not be gathered at any other time. This study explores the kind of information found by both verbal protocol collection procedures. While some have indicated that subjects who give retrospective verbal protocol may have forgotten information that is available during task performance, Deffner (1988) argued that if given extensive cues during the retrospective verbal protocol subjects will be able to remember that information. If retrospective verbal protocol can provide relevant information, then the experimenter should consider heavily cued retrospective verbal protocol rather than concurrent verbal protocol collection. The advantage with the retrospective protocol would be that verbalization has not had the opportunity to interfere with the task at hand; therefore, the researcher can generalize to more realistic settings where a user would not be required to "think aloud." Retrospective verbal protocol can also be used with benchmarks to compare systems; benchmark tasks are not recommended with the use of concurrent verbal protocol due to the problems with time data.

While the issues have been discussed in the literature, many researchers remain unaware of the potential problems with concurrent verbal protocol. Further, because much of the research dealing with possible problems with concurrent verbal protocol has been done in areas outside of applied human-computer interaction, human-computer interaction researchers have remained relatively unaware of these studies. The lack of attention given to the possible deficiencies of concurrent verbal protocol indicates that these possible problems should be directly addressed in an applied human-computer interaction setting.

Applied Setting

In order to select a realistic applied setting, a real human-computer interaction research question is necessary. One question which has been intuitively addressed, but not empirically supported, was raised by Shneiderman (1987) who suggested that windows become more useful on a larger screen and are not as effective on a smaller screen. Consumers purchasing large screens for their Macintoshes would certainly be interested in knowing if they are, in fact, buying an advantage in terms of windowing. This research question is an especially appealing arena for addressing concurrent verbal protocol implications because interference with the task at hand is a concern with both concurrent verbal protocol (as discussed previously) and with the use of windows. Researchers have suggested that one problem with windows is that the actual manipulation of windows to achieve the task objectives may interfere with the task itself (Shneiderman, 1987).

Windows should, in fact, reduce the load on the user by allowing the user to access multiple files and applications without having to leave the current work space. The use of windows also allows the files to be partially visible so that the user does not have to remember all the items with which he/she is working. This display technique should reduce the short-term memory load of the user (Davies, Bury, and Darnell, 1985). The use of overlapping windows allows the computer display to be analogous to a desk top on which the worker may leave partially overlapping folders on the desk top and open these folders to examine them without having to clean off his/her entire desk. Each window reminds the user, as does a file folder on a desk, which tasks need to be accomplished and what resources are available for completing the tasks. While windows should reduce the load, if the manipulation of windows begins to interfere with the task itself then windows cease to be useful.

Several researchers have suggested that interference problems associated with window manipulation may be eliminated if the system has control of window management. There are, however, problems with allowing the system to manage the windowing. First, the programmer must program the system to manage the windows as the user would want them managed. Although window management strategies have been researched, different window management strategies will not be addressed in the present research; a good source for this research may be found in Hopgood, Duce, Fielding, Robinson, and Williams (1985).

The second problem with allowing the system to control the window management is that the user may not be comfortable turning that responsibility over to the system. Simes and Sirsky (1985) suggest that the locus of control, a dimension of control which describes the extent to which a person feels that he controls his environment (internal locus of control) or that the environment controls him (external locus of control), is an important variable in determining the user's acceptance of a system. Simes and Sirsky suggest that feeling in control of a computer system is particularly important for experienced users. This need for experienced users to have control of the system must be considered in the trade-off between the user's control of the windows and the interference of window manipulation with the actual task. While this trade-off should certainly be considered, the use of overlapping windows, in which the user is given control, is extensive.

One issue to be addressed is whether the large screen reduces window manipulation problems as compared to the smaller screen in an overlapping windows interface. If windowing does distract the user, a larger screen may help the user by providing more area on which to display information and thus less overlap of a finite number of windows. The possibility also exists that disadvantages may be associated with the larger monitor such as the large amount of mouse movement required. The possible advantages and disadvantages associated with a larger screen, however, would likely be highly task dependent. For tasks that require a great deal of window manipulation, particularly those that are much more easily done with windows simultaneously present on the screen, a larger screen would indeed seem to offer an advantage to the user because the user will not be as concerned with how to manipulate the windows in order to accomplish the task. Further, the user may have more information available so that he does not have to rely on his memory for some tasks. For tasks that do not require a great deal of windowing, the larger screen may not offer much advantage in windowing manipulation.

The comparison of small and large monitors is the type of issue that has been addressed by usability research in the past. This question seems to lend itself well to testing a number of usability metrics to help resolve issues concerning the utility of these metrics. More specifically, the advantages and disadvantages of concurrent and retrospective verbal protocol can be addressed in this context. These advantages and disadvantages are summarized in Table 1.

TABLE 1. Possible Advantages and Disadvantages Associated with Concurrent and Retrospective Protocols

CONCURRENT PROTOCOL	
Possible Advantages	Possible Disadvantages
Pinpoints problems as they occur	May interfere with time data
Addresses why the user has difficulty	Traditional statistical analysis may be inappropriate
Gathers information before the user adjusts to the interface	May interfere with cognitive processing which may cause efficiency degradation
	May not generalize to situations in which a user does not think aloud
RETROSPECTIVE PROTOCOL	
Possible Advantages	Possible Disadvantages
Does not interfere with time data	Traditional statistical analysis may be inappropriate
Does not interfere with cognitive processing	Information may be unavailable to subject
Pinpoints problems as they occur	More time consuming than concurrent verbal protocol
Addresses why the user has difficulty	

RATIONALE AND APPROACH

The primary objectives of the present study are to address the implications of using concurrent verbal protocol in an applied human-computer interaction setting and to compare the information which may be discovered when employing two different methods of verbal protocol collection. The study compares the information gathered and possible interference of protocol collection with two verbal protocol procedures, concurrent and retrospective verbal protocol. The research is also intended to discover advantages and disadvantages associated with monitor sizes on tasks of varying window difficulty.

Independent Measures

Verbal protocol. Two levels of verbal protocol were selected, concurrent and retrospective. The concurrent protocol procedure was chosen to be representative of the protocols in use in the software industry. While some probing of subjects may be common, subjects were not interrupted or probed once they began to verbalize as it seemed that this would more likely interfere with the cognitive processing. Interrupting the subjects seemed to more likely lead to the predicted results. Real-time review of videotapes of their own experimental session was chosen as the method for cueing the subjects in retrospective protocol. This cueing method not only provided the subjects with rich cues, but also kept the times allowed for protocol collection reasonably similar between the two protocol conditions.

Monitor size. This study compares a small monitor to a larger monitor. The monitors were chosen to vary only in size. Other monitor parameters were kept as constant as possible. The large monitor was chosen to represent a monitor size often used for desk top publishing.

Task difficulty. The importance of task type in predicting the effects of the concurrent verbal protocol, as well as the effects of a large versus small screen, indicates that task type or difficulty should be varied as an independent variable. Evaluations of the effects of the small versus large screen and concurrent versus retrospective verbal protocol are therefore made with respect to task types.

The dimension along which to vary the tasks is crucial. Task difficulty seems the obvious alternative. Unfortunately, task difficulty in a windowing environment has not previously been objectively defined or quantified. The tasks had to be defined for difficulty by the experimenter. The use of a secondary task to define task difficulty was considered for defining the task difficulty but seemed too similar to the methodology employed during concurrent verbal protocol collection. Because using a secondary task is so similar to the collection of concurrent verbal protocol, it seems that defining task difficulty by using a secondary task would insure the predicted results. Subjectively determining window difficulty was also considered but rejected due to problems associated with subjective definition of task difficulty (Mitta, 1988). Finally, an operational definition of task difficulty was chosen. The operational definition takes into account the dimensions of a windowing task that would intuitively make it difficult. The task difficulty was determined by adding together three weighted dimensions: (1) the number of windows that must be opened simultaneously in order for the subject to not have to use short-term memory to transfer information from one window to another, (2) the total number of windows required to complete the task, and (3) the total number of operations performed in the windows. These weightings were intuitively derived. Because the task difficulty is defined in terms of windowing, the first dimension, simultaneous windows, is weighted most heavily (7), the number of windows next (2), and the operations inside the windows least (1). The selected tasks range in difficulty from 7 to 57, with 7 being least difficult and 57 being most difficult. The tasks divide into three arbitrary levels of difficulty, Low (1-20), Medium (21-40), and High (41-60), with four tasks at each level. The task difficulty levels-- Low, Medium, and High-- were block randomized throughout the session. All subjects, however, saw the tasks in the same order.

In order to generate meaningful tasks in a windowing environment, the tasks were imbedded in a cooperative work scenario. Some example tasks are included in Appendix A. The subjects were told that they are coauthoring a document with two other authors. They were told that in order to complete the document, they must perform given tasks which involve their coauthors' drafts, coauthors' comments, and other pertinent documents and facts, as well as their own drafts. This task situation was chosen due to its applicability to a real work situation, as well as the fact that it provided a good environment for creating both simple and more complex tasks that seemed realistic.

Dependent Measures

Task completion time. Task completion times were employed because they have been the most heavily used and successful metrics for windowing research. Time is an especially appealing measure since Deffner (1987) found that subjects who were concurrently verbalizing employed less efficient strategies; time can be viewed as a measure of efficiency. Unfortunately, while task completion time has been the most heavily used metric for measuring the usability of windowing systems, time measures may be inappropriate for use during concurrent verbal protocol collection, as are timed benchmark tasks, since most researchers feel that verbalization time interferes with task completion time. Some researchers, however, have found no differences due to verbalization. Time measures were collected to discover if the verbalization increases the task completion time.

Windowing errors. Another measure that was considered important for measuring possible interference of verbal protocol, as well as possible advantages offered by a larger screen, is the number of windowing errors made by the subjects. The subjects' data were analyzed for windowing errors, which included opening an incorrect window, covering a necessary window with another window and having to retrace in order to return to that window, closing a window which was not to be closed and having to reopen that window, and any other errors which became apparent during data analysis. The errors were analyzed in terms of total number of errors, as well as the kinds of errors. Both the efficiency and error data were collected by reviewing video tapes made during the experimental session. These tapes were reviewed without sound so that the reviewer was not aware whether the subject was in the concurrent or retrospective verbal protocol condition.

Steps to completion. The number of window operations used to complete each task have also been used by researchers to discover the usability in a windowing environment. This measure seems to be more appropriate for comparing concurrent and retrospective verbal protocol collection since the time measure may not be appropriate. Several approaches using windowing operations as a dependent measure were considered. In light of Deffner's (1987) findings regarding efficiency, these measures were concerned with measuring efficiency. Discovering the most efficient strategy for doing a windowing task using a state-to-state measure of strategy was considered. After defining the most efficient strategy for doing the task, each subject's pathway (or strategy) through the task would be compared to the most efficient strategy by comparing each state through which the subject traveled to each state of the most efficient strategy. This measure was rejected, however, because data collected during pretesting revealed that expert users had many efficient strategies by which to complete the tasks. Another more global way of defining strategy efficiency was finally adopted. This method defines the efficiency of each subject's strategy by comparing the number of windowing operations, or steps to completion, required by the subject to complete the task. This method of defining efficiency, then, is not as concerned with the order in which the subject moves through the states but merely the total number of windowing operations required. This dependent measure seemed to be the best for defining the efficiency with which the subjects used the windows.

Subjective measures. Two subjective measures of usability were collected. First, after completing each task, the subjects were asked to rate that task on a scale ranging from Easy (1) to Difficult (7). After all the verbal protocol data were collected, subjects were asked to rate the overall system on a modified Questionnaire for User Interface Satisfaction (QUIS) scale. The developers of QUIS have reported very high internal reliability with QUIS as measured by consistency of the responses to similar questions. The present structure of QUIS, however, may lead to this high internal consistency because related questions are grouped together and the negative response is always on the left for each of the scales. A discussion with a researcher who had administered QUIS verified that subjects were indeed not reading every question, but were instead getting an idea of how they felt about the system and circling those same values. In the present study the structure of QUIS was altered to determine if the high consistency may in part be due to the structure of the questionnaire. The QUIS was modified such that the same questions were included but the order of the questions was altered. Further, the positive and negative sides of the scales were randomly assigned.

Verbal data. Verbal protocol data were recorded on video tape while the subject verbalized. The protocol statements were then placed into one of six categories: Procedure, Procedure + Explanation, Explanation, Reading, Design, and Other. These categories were modified from categories employed by Deffner (1988). The Procedure statements are statements in which subjects simply stated the what they were doing. The Procedure + Explanation is a statement in which the subject said what they was doing and followed with the reason for that procedure. An Explanation statement is a reason for a global strategy without any reference to a specific procedure. A Reading statement is one in which subjects simply read the words presented on the monitor aloud. A Design statement is either a comment on some part of the interface or a suggestion for a possible change in the interface. Finally, the Other category includes statements which do not fall into the first five categories.

METHOD

Subjects

Forty-eight subjects (40 males, 8 females) were solicited by advertisements in the school newspaper at Virginia Tech. They were required to show their proficiency on a windowing benchmark task which was given prior to the experimental tasks (Appendix B). The test was adapted from Gaylin's (1985) windowing benchmark, which includes basic windowing tasks such as opening, closing, and moving, but was expanded to include more complex tasks. Subjects received payment for their participation.

Equipment

The subjects performed the experimental tasks in Microsoft Word (Version 3.02) on a Macintosh II. The two monitors which were compared included the Apple® High-Resolution Monochrome Monitor and the Sigma Designs SilverView Monitor. The Apple® monitor was chosen because it is the standard monitor sold with the Macintosh II computer. The High-Resolution Monochrome Monitor has a 30.48 cm diagonal and has been criticized for not being large enough. The monitor has 30 dots per cm (640 x 480). The SilverView monitor was chosen as the comparison monitor because it has a 53.34 cm diagonal and because it closely resembles the Apple® monitor in resolution. The SilverView monitor has 28 dots per cm (1152 x 780). Both monitors were set at the luminance level of 35 cd/m^2. The luminance modulation was approximately 0.71 for both monitors.

The experimental room was 3.2 m X 3.4 m with the Macintosh II on a typing table against one wall. The table was 76 cm high. The subjects were allowed to adjust their chair to their own preferred height. A Canon VC-30A video camera and Panasonic video cassette recorder were placed over the left shoulder of the subjects for video taping the experimental session. The CRT on which the retrospective subjects were shown their performance was placed on a small table to the left of the subjects. The monitor not in use was covered and next to the monitor the subject was using.

Design

The experimental design is a 2 X 2 X 3 X 4 (Verbal Protocol X Monitor Size X Task Difficulty X Block) mixed factorial design. The design is graphically represented in Figure 1. A between-subjects treatment of verbal protocol was chosen because of the likelihood of unequal transfer of training between the verbal protocol conditions. Deffner (G. Deffner, personal communication, January,1989) recommended that a between-subjects design be used to study concurrent verbal protocol as a method for data collection due to unequal transfer of training. Further, a between-subjects treatment of monitor size was also chosen because subjects using verbal protocol may be less flexible in changes of strategy (Deffner, 1988). Finally, each subject saw all three levels of task difficulty and all 4 four blocks.

Procedure

The subjects read a general description of the experiment and signed an informed consent form. The subjects were required to take the windowing benchmark task in order to prove proficiency on the use of the windowing system in Microsoft Word. If the subjects attained an acceptable score, they did seven warm up tasks and began the actual experimental tasks; otherwise, they were excused (a provision for being excused for not attaining a high enough score on the bench mark task was included in the informed consent to avoid subjects objecting to being excused). The concurrent verbal protocol subjects were told that they were to describe aloud what they are doing and thinking throughout the session but were reminded to do so only before each task and not interrupted during the task. The retrospective verbal protocol subjects were told that they would be video taped while they performed the tasks and that after they had completed the tasks they would watch the video tape and describe aloud what they were doing and thinking during the session. Their verbalizations were restricted to the time during which the tape was played back in order to give them the same opportunity and time to verbalize as the concurrent subjects. They were also reminded to verbalize before each task. The experiment required all of the subjects to work through tasks of varying degrees of difficulty which were presented on-line. The subject had to open the file that contained a task and follow the instructions in that task. After each task was completed, the subjects were required to rate the task on a 7-point bipolar scale which ranged from Simple to Difficult. Finally, after the subjects had completed the tasks and verbalizations, they rated the system on the QUIS (Appendix C) which was modified from the QUIS employed by Chin et al. (1988). The concurrent subjects required approximately 1 1/4 hours to complete the experimental session, while the retrospective subjects required approximately 2 hours. The subjects were debriefed and paid.

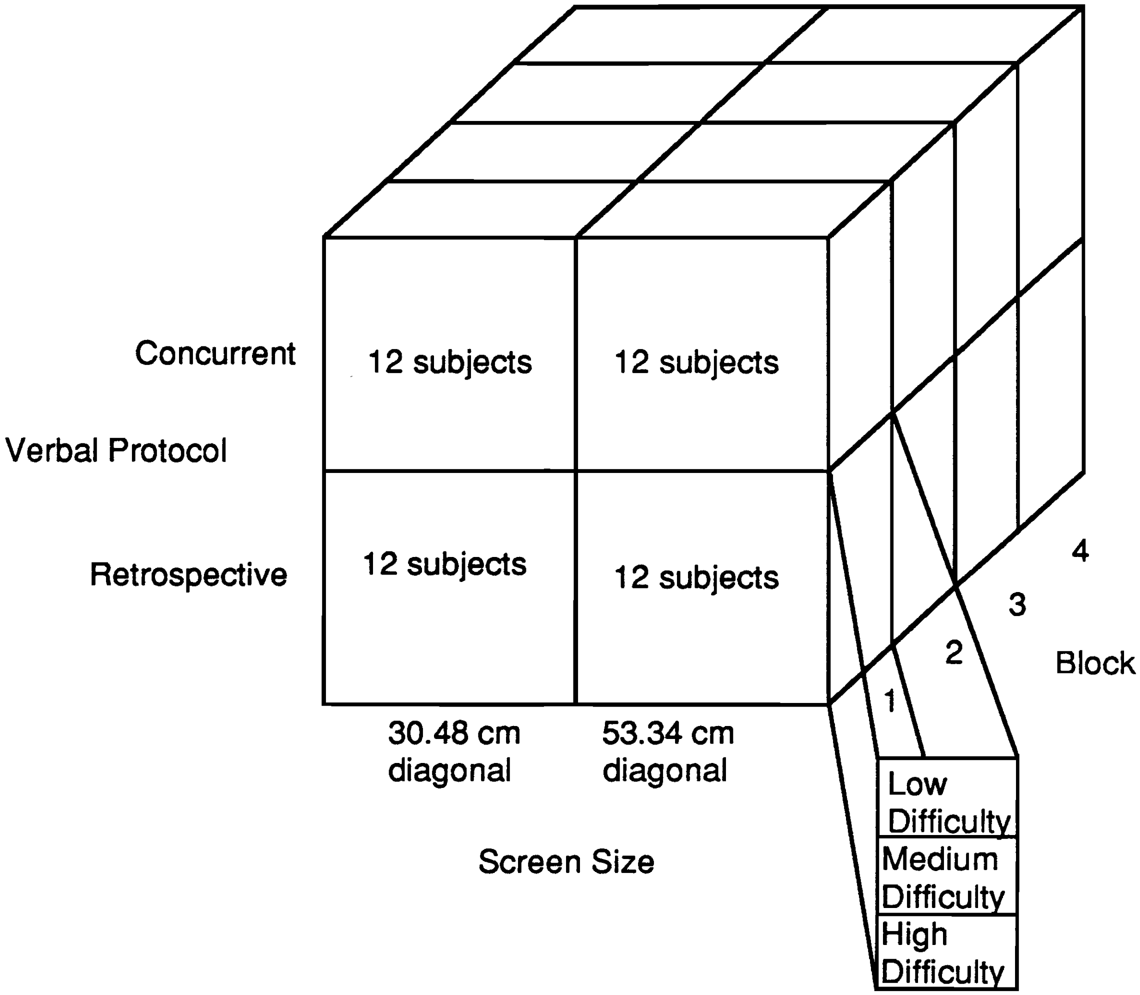


Figure 1. Experimental design.

RESULTS

Task Performance

Task performance was measured by the number of steps the subjects required to complete the tasks and by the task completion times. The errors made by subjects were collected and categorized in terms of the type of the error. There were, however, too few errors made by the subjects to be quantitatively analyzed. A review of the errors did not yield any practical information.

Steps to completion. The steps to completion were analyzed with a 2 X 2 X 3 X 4 (Monitor X Protocol X Difficulty X Block) mixed-factor analysis of variance. The results of this ANOVA are found in Table D1. As the task difficulty increased, more steps were required to complete the tasks (Table D2, Figure 2). The Low Difficulty tasks required fewer steps than did the Medium Difficulty tasks, and the High Difficulty tasks required subjects to use more steps to complete the tasks than did the Medium Difficulty tasks.

While in general, subjects using the Large Monitor needed fewer steps in order to finish the tasks than did subjects using the Small Monitor (Figure 3), this was not true for Low Difficulty tasks (Figure 4). On the Low Difficulty tasks, subjects who were using the Small Monitor did not require significantly more steps than subjects using the Large Monitor (Table D3). However, once the task difficulty increased to the Medium level, subjects using the Large Monitor did not need as many steps to complete the tasks as subjects using the Small Monitor.

The subjects in general decrease the number of steps needed to complete

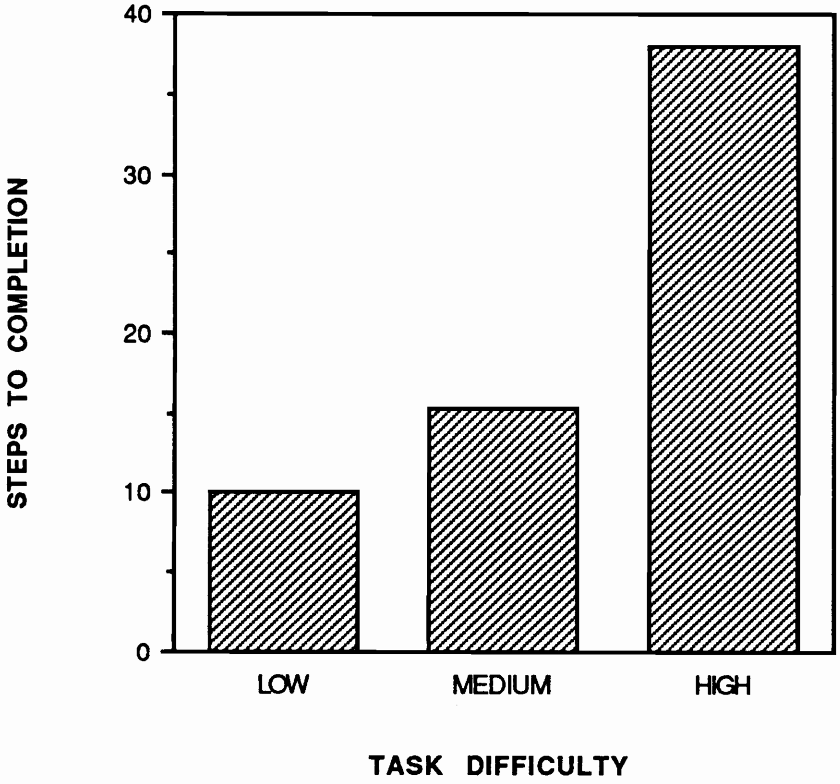


Figure 2. Effect of Difficulty on steps to completion.

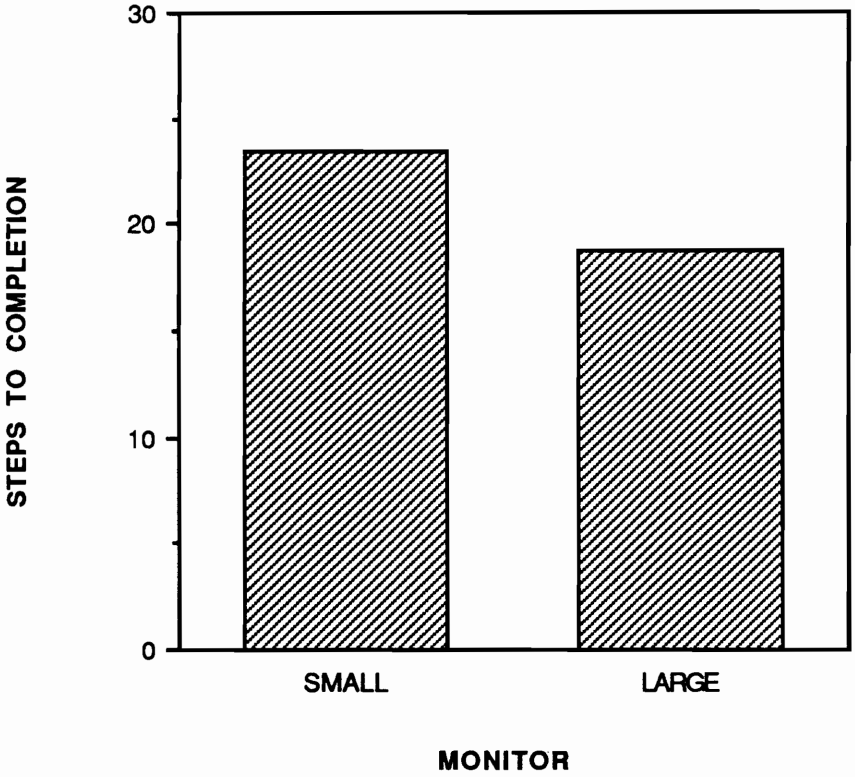


Figure 3. Effect of Monitor on steps to completion.

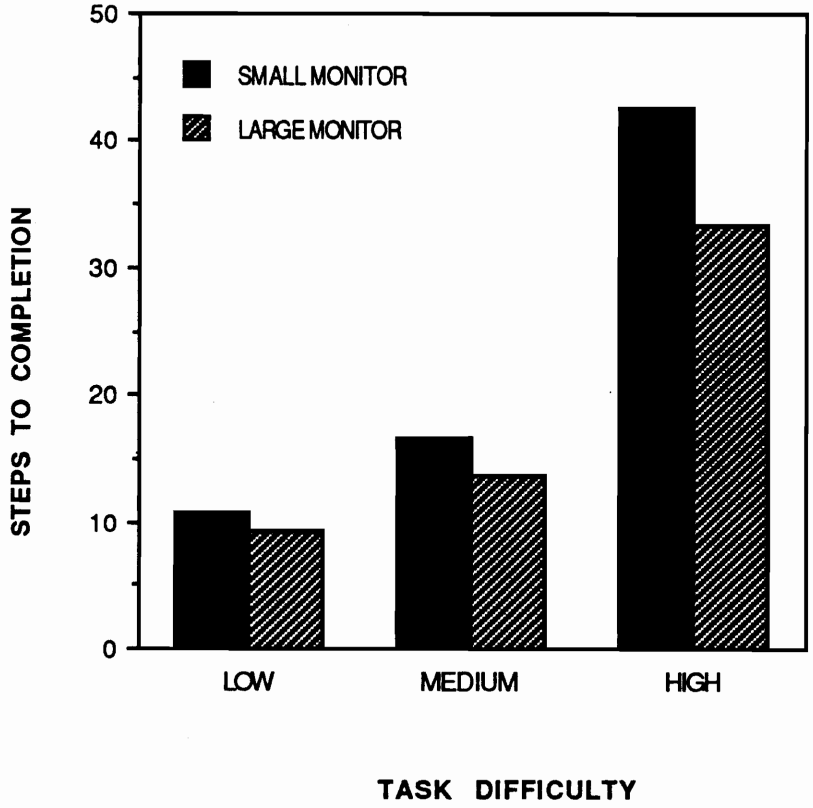


Figure 4. Difficulty X Monitor interaction for steps to completion.

the tasks across time (Table D4, Figure 5). While it is true that the steps do decrease over time, if these data are examined more closely, it can be seen that the steps to completion decrease dramatically over time for the high difficulty tasks but do not decrease over time for the Low and Medium Difficulty tasks (Tables D5 and D6, Figure 6).

Finally, if the data are examined at an even more detailed level, it can be seen that there is no difference in Monitors across Blocks when tasks are Low Difficulty (Table D7, Figure E1). However, for Medium Difficulty tasks and High Difficulty tasks, there is a significant M X B interaction (Table D7). For Medium Difficulty tasks, there is no real difference between monitors during Blocks 1 and 4, but the Large Monitor requires fewer steps during Blocks 2 and 3 (Table D8). For the High Difficulty tasks, on the other hand, the superiority of the Large Monitor exists during all four blocks (Table D8), and is greater during Block 1 as indicated by the significant M X B interaction (Table D7).

Task completion time. The task completion times were analyzed with a 2 X 2 X 3 X 4 (Monitor X Protocol X Difficulty X Block) mixed-factor analysis of variance. A summary table of this ANOVA is found in Table D9.

In light of the steps to completion data, it is not surprising that the High Difficulty tasks took subjects longer to complete than did Medium Difficulty tasks and that Medium Difficulty tasks took longer than did Low Difficulty tasks (Table D10, Figure 7).

When considered without regard for the Task Difficulty, the task completion time trends over blocks are not clear. The third block of tasks required subjects

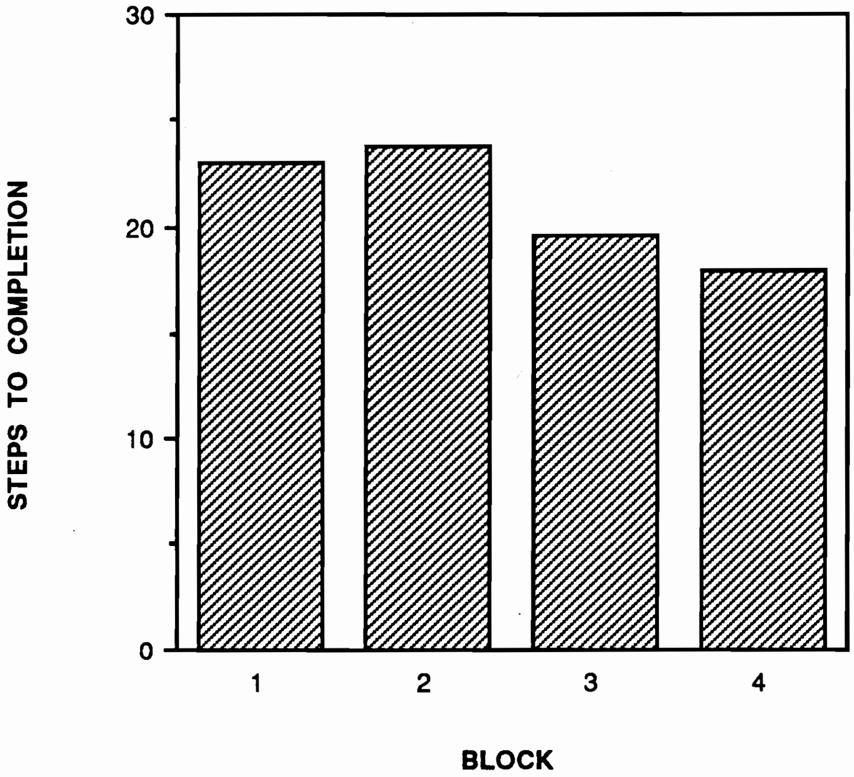


Figure 5. Main effect of Block for steps to completion.

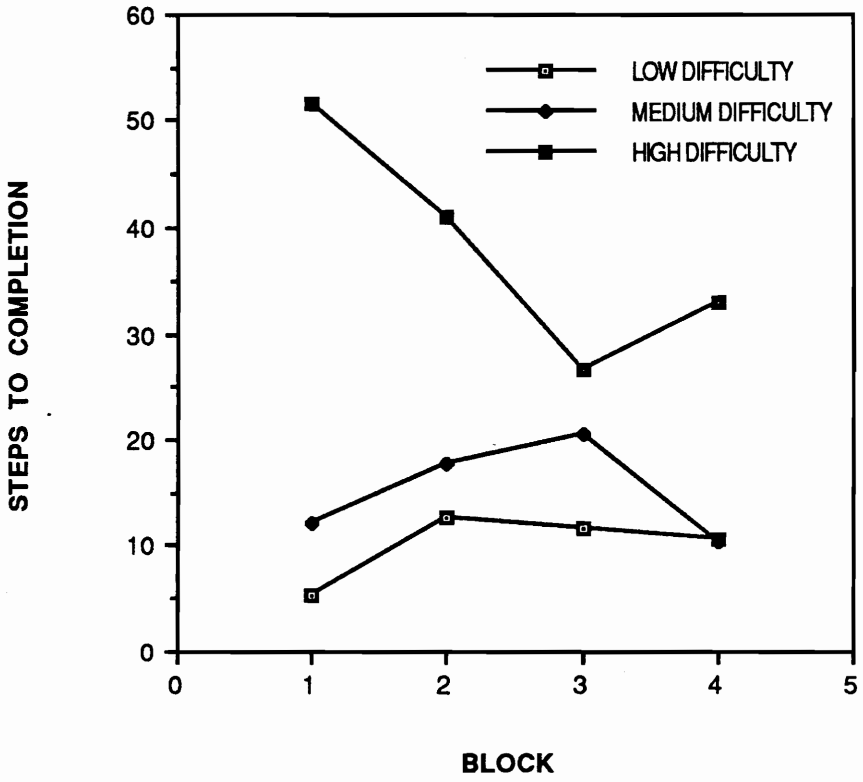


Figure 6. Difficulty X Block interaction for steps to completion

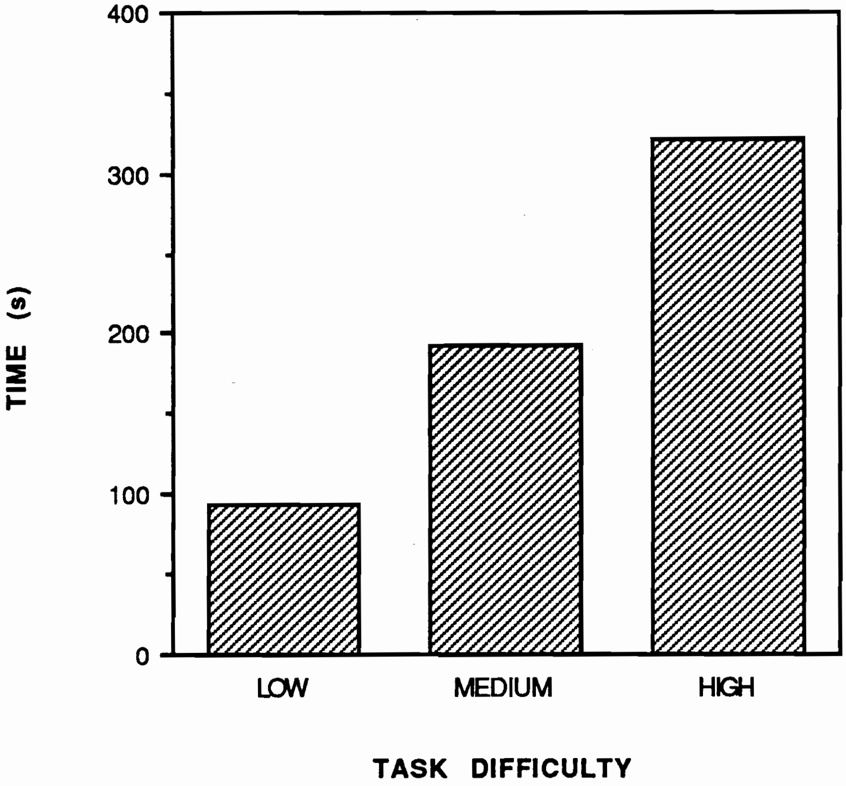


Figure 7. Effect of Difficulty on task completion time.

less time than did other blocks of tasks (Table D11, Figure 8). When the task completion time trends over blocks were examined along with the task difficulty, trends became apparent (Tables D12 and D13, Figure 9). The time required for subjects to complete the High Difficulty tasks decreased over time; again, large decreases were evident in earlier blocks of tasks. On the other hand, the subjects required more time to complete the Low and Medium Difficulty tasks over time.

The relationship between the level of task difficulty and trends across time are similar for task completion time and steps to completion. In fact, the correlation between the task completion times and steps to completion across cells of the original analysis is quite high ($r = 0.95$, $p < 0.0001$).

Subjective Data

Two forms of subjective data were collected. First, after completion of each task, subjects rated the difficulty of that task. Second, at the end of the experimental session, subjects completed the QUIS (Questionnaire for User Satisfaction) in order to give a global evaluation of the experimental session.

Difficulty ratings. The Difficulty Ratings were analyzed with a 2 X 2 X 3 X 4 (Monitor X Protocol X Difficulty X Block) mixed factor analysis of variance. The results of this ANOVA are found in Table D14.

In general, subjects did not employ the full range of the task difficulty ratings. The low end of the scale was well represented, but the highest average rating for any given task was only 4.5. Subjects rated the Low Difficulty tasks as less difficult than the Medium Difficulty tasks, and the High as more difficult than

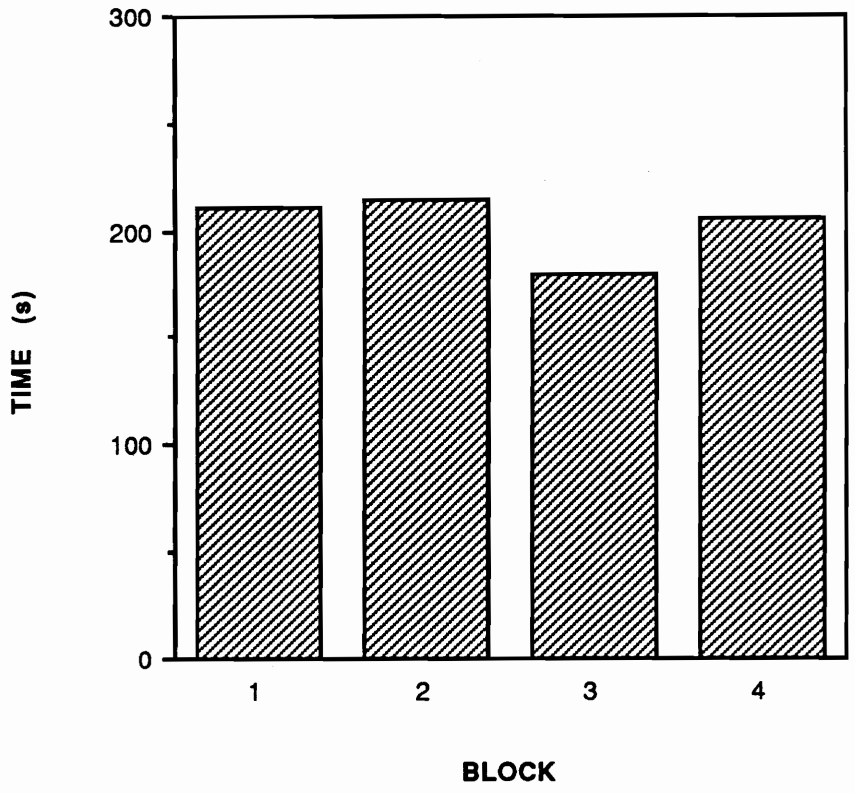


Figure 8. Effects of Block on task completion time.

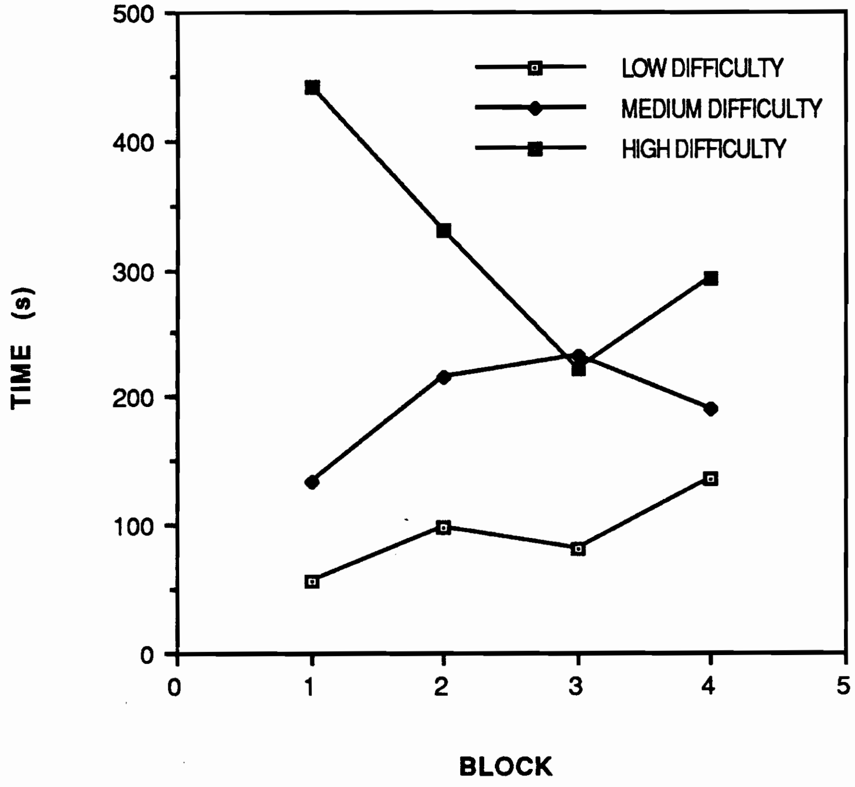


Figure 9. Difficulty X Block interaction for task completion time.

the Medium Difficulty tasks (Table D15, Figure 10). Obviously this lends further validity to the operational definition of the task difficulties.

The subjects who used the Small Monitor rated the tasks as more difficult than subjects who used the Large Monitor (Figure 11). This higher difficulty rating was, however, related to the task difficulty. Subjects using the Small Monitor rated the Medium and High Difficulty tasks as more difficult than subjects using the Large Monitor (Table D16, Figure 12). There was no difference in the difficulty ratings of the Low Difficulty tasks between the Small and Large Monitors.

Examining the subjects' task difficulty ratings over blocks reveals that subjects rated the earliest block of tasks as the easiest and the second block as most difficult (Table D17, Figure 13). If the data are more closely examined, no obvious trends become apparent in the interaction between Blocks and and Difficulty (Figure 14).

Finally, it can be seen that there is no interaction between Monitor and Block for the Low Difficulty tasks (Table D18, Figure E2). There is, however, an M X B interaction for the Medium Difficulty tasks (Figure E2), with the Small Monitor rated as more difficult only for Block 2 and 3 (Table D19). Finally, there is no M X B interaction for High Difficulty tasks; the Small Monitor is rated more difficult across all blocks (Table D20). The effect of block and the interactions with which block is involved are largely unexplained.

QUIS. The modified QUIS (Questionnaire for User Interface Satisfaction), the more global measure of subjective response to interface differences, did not show any differences for Monitor (Table D21), for Protocol (Table D22), or for the combinations of Monitor and Protocol (Table D23). This indicates that

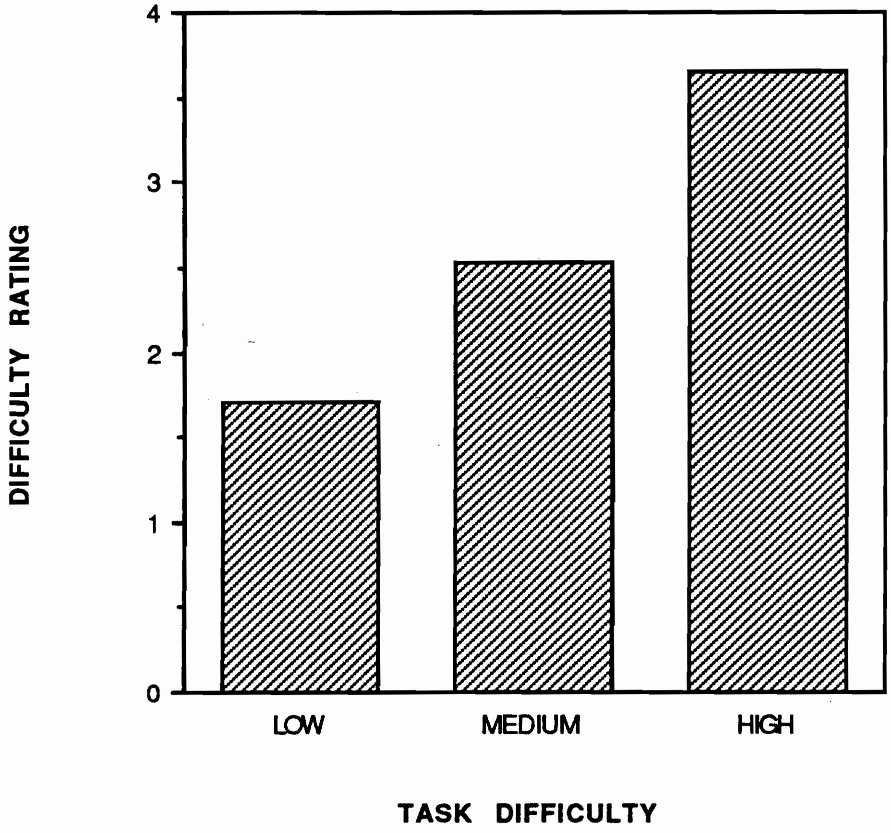


Figure 10. Effect of Difficulty for difficulty ratings.

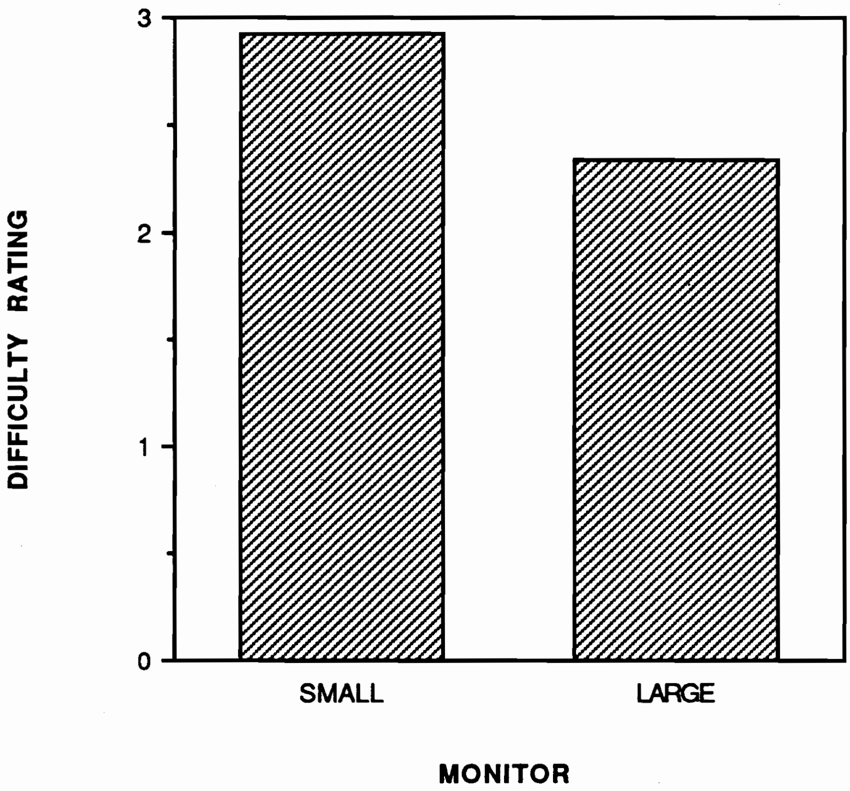


Figure 11. Effect of monitor size on task difficulty ratings.

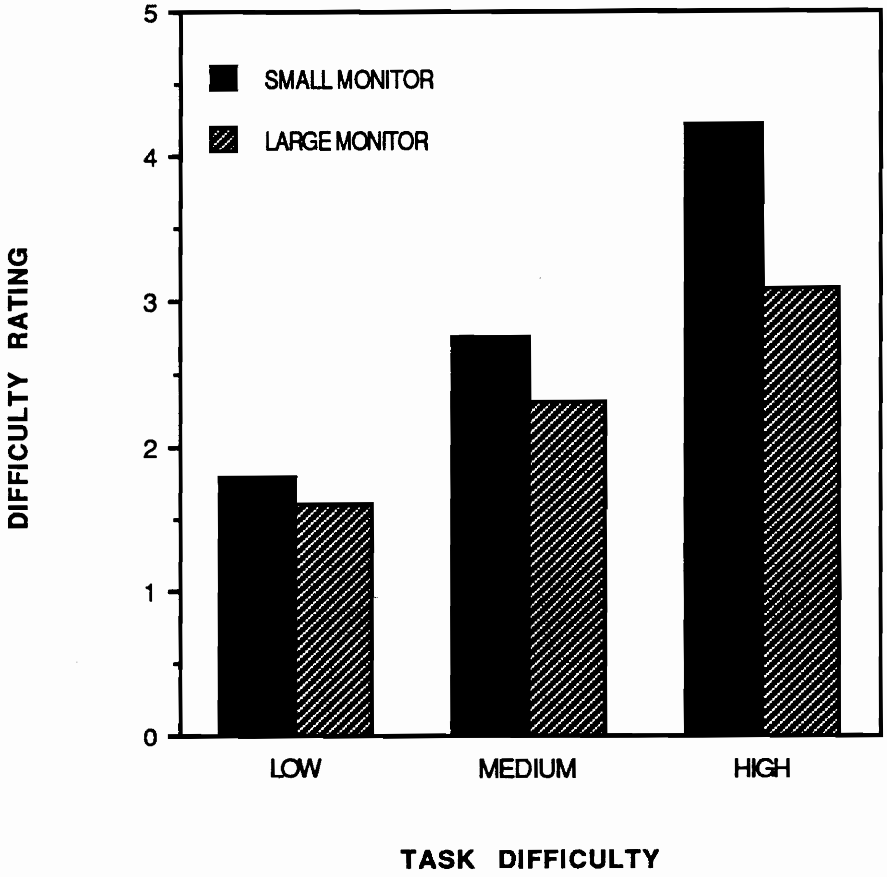


Figure 12. Monitor X Difficulty interaction for task difficulty ratings.

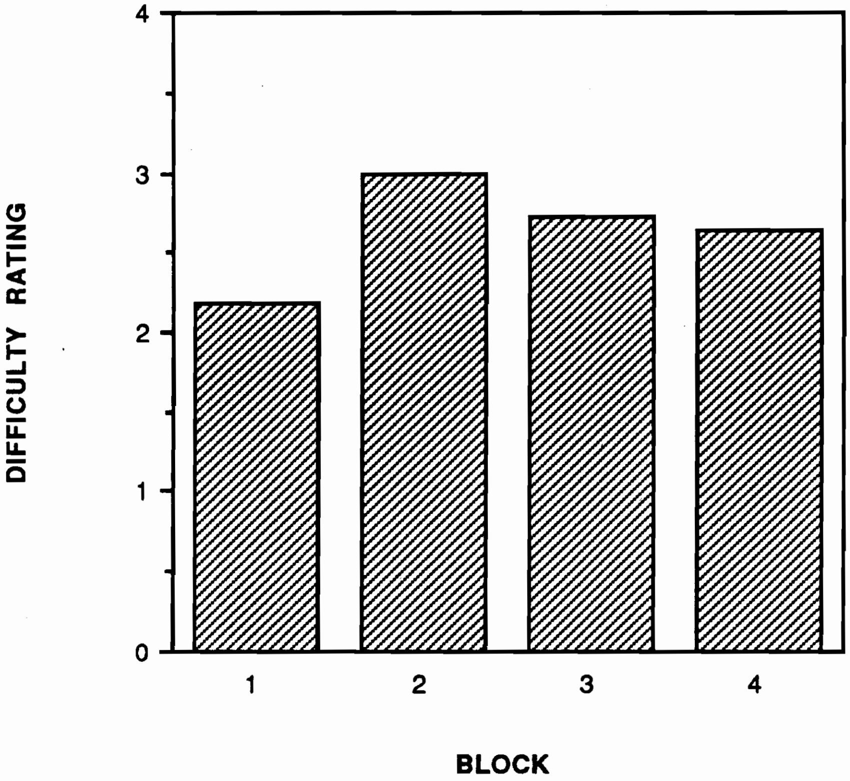


Figure 13. Effect of Block on task difficulty ratings.

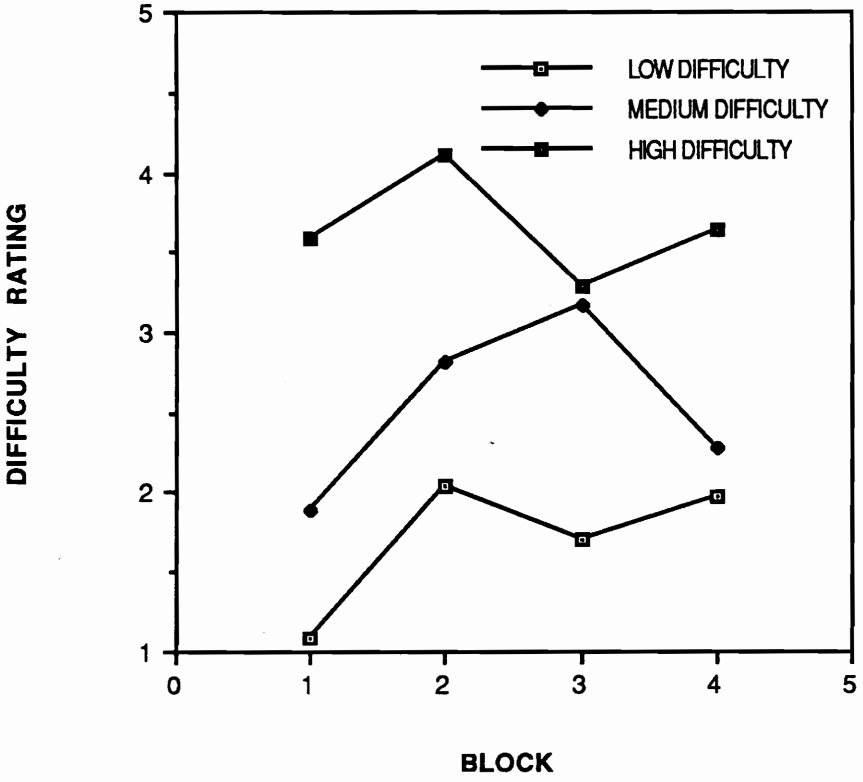


Figure 14. Difficulty X Block interaction for task difficulty ratings.

differences between the two monitors, the protocols or the combinations of protocols were not detected by the 27 QUIS items. Although QUIS was modified in order to eliminate the structural component contributing to the high internal reliability, measures of internal reliability were quite high (Cronbach alphas for each of the monitor-protocol combinations were .83, .86, .84, and .78). This indicates that the internal reliability of QUIS remains high regardless of a structural change.

Verbal Protocol Data

Two types of Verbal Protocol data were collected and analyzed. First, the number of statements made by the subjects were analyzed without regard for the contents of those statements. Next, each of the statements was placed into one of six categories: Procedure, Procedure + Explanation, Explanation, Reading, Design, and Other. Examples of each of these are shown in Table 2.

Number of Verbalizations. The numbers of verbalizations were analyzed with a 2 X 2 X 3 X 4 (Monitor X Protocol X Difficulty X Block) mixed factor analysis of variance. The results of this ANOVA are found in Table D24.

The verbalization data indicated that the concurrent protocol subjects gave many more verbalizations than did the retrospective protocol subjects (Figure 15). Further, in general, the subjects gave more verbalizations during High Difficulty tasks than during the Medium Difficulty tasks, and more during Medium than during the Low Difficulty tasks (Table D25, Figure 16). These differences were also significant for the concurrent protocol subjects alone (Tables D26 and D27, Figure 19). However, the retrospective subjects do not

TABLE 2

Examples of each Category of Verbal Protocol Statements

Procedure

I'm saving the window.

Procedure + Explanation

I'm shrinking the window so that I can put all the mail on the screen.

Explanation

I realized that I could have copied instead of typing.

Reading

When you have completed the task type "DONE" in the space provided.

Design

I would like to be able copy the time directly from the alarm clock.

Other

Darn.

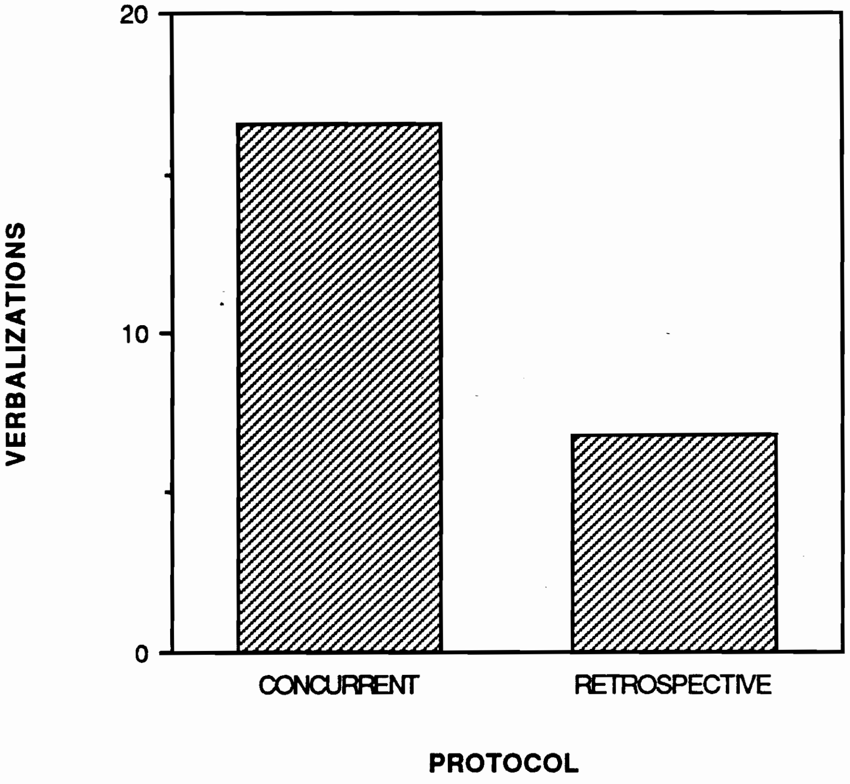


Figure 15. Effect of Protocol on number of verbalizations.

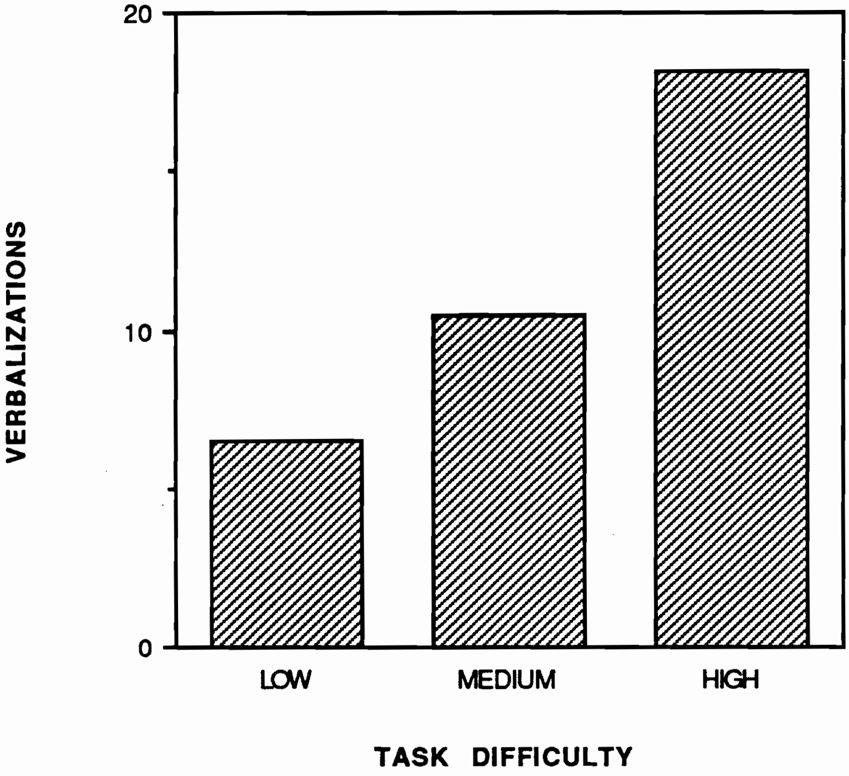


Figure 16. Effect of Difficulty on number of verbalizations.

have significantly more verbalizations for the Medium Difficulty tasks than for the Low Difficulty tasks (Tables D26 and D27, Figure 17).

Subjects seem to be decreasing the number of verbalizations over time as shown in Figure 18 (Table D28). The relationship of the task difficulty to verbalizations over time must be explored further to understand the data. The subjects generally decrease the number of verbalizations over time for the High Difficulty tasks (Tables D29 and D30, Figure 19), whereas the verbalizations generally increase over time for the Low and Medium Difficulty tasks. The interpretation for Medium Difficulty is less clear.

Finally, if Protocol is also considered in this relationship, as shown in Figure E3, it can be seen that the decrease of verbalizations over time for the High Difficulty tasks is greater for the concurrent protocol subjects than for the retrospective subjects (Tables D31 and D32). The trends across blocks are similar for Concurrent and Retrospective subjects for the Low and Medium Difficulty tasks; the Retrospective subjects give fewer verbalizations overall, but follow the same pattern of increasing and decreasing over Blocks as the Concurrent subjects.

The relationship among Monitor, Protocol, and Block is illustrated in Figure E4. While further breakdown of the data indicated that Concurrent Protocol subjects using the Large Monitor produced more verbalizations than those using the Small in the third block of tasks, this difference did not exist for retrospective subjects (Tables D33 and D34). This result is not considered particularly meaningful.

While the total verbalizations made by subjects are of interest, the kinds of information conveyed by subjects during the experiment is more crucial to

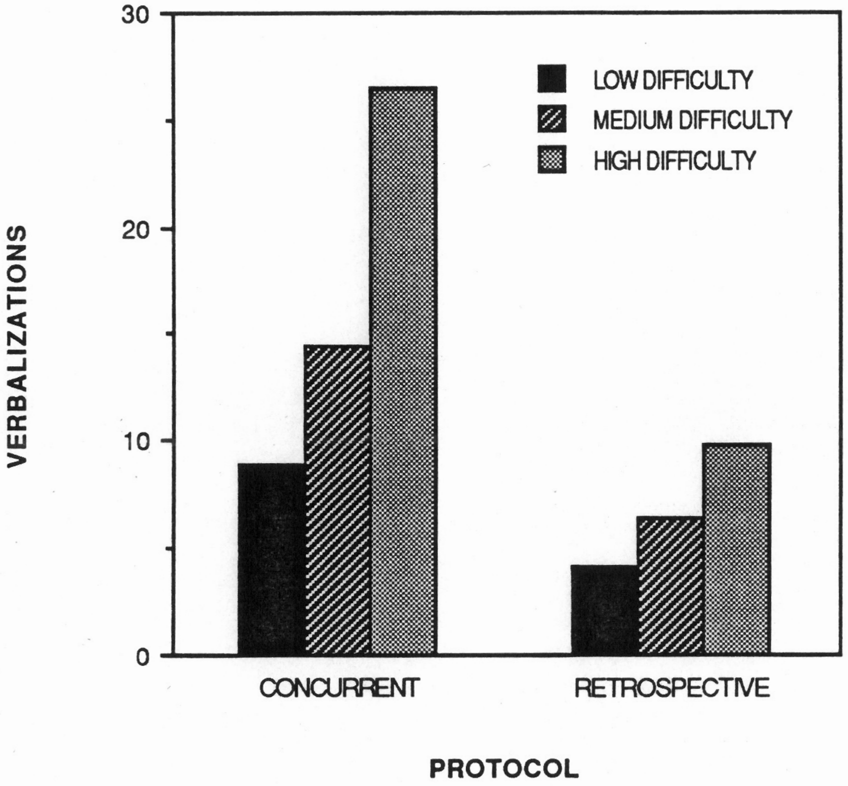


Figure 17. Protocol X Difficulty interaction for number of verbalizations.

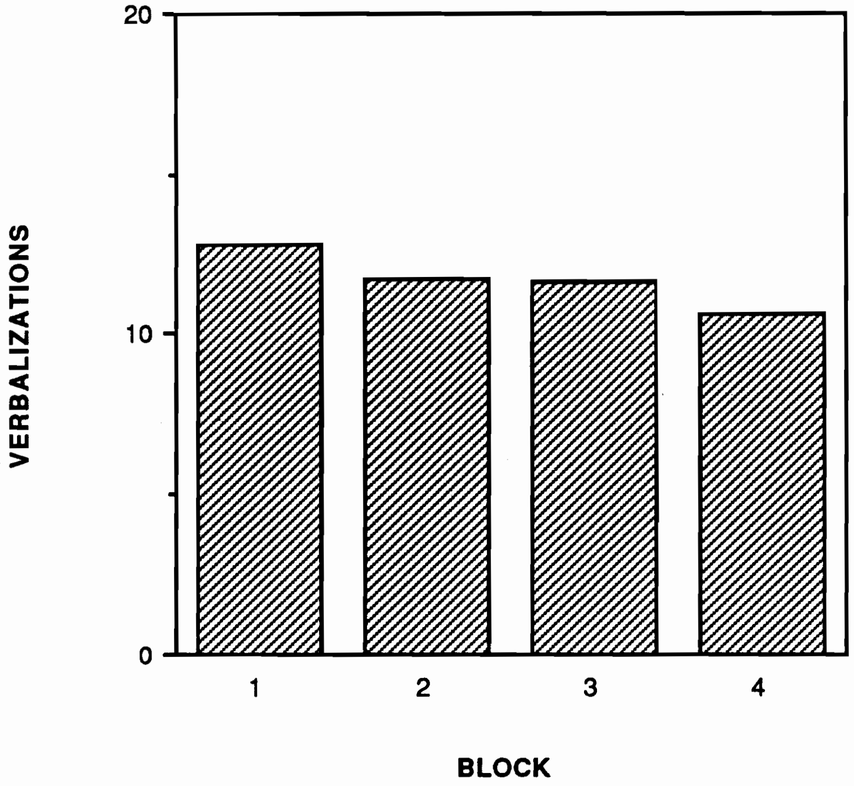


Figure 18. Effect of Block on number of verbalizations.

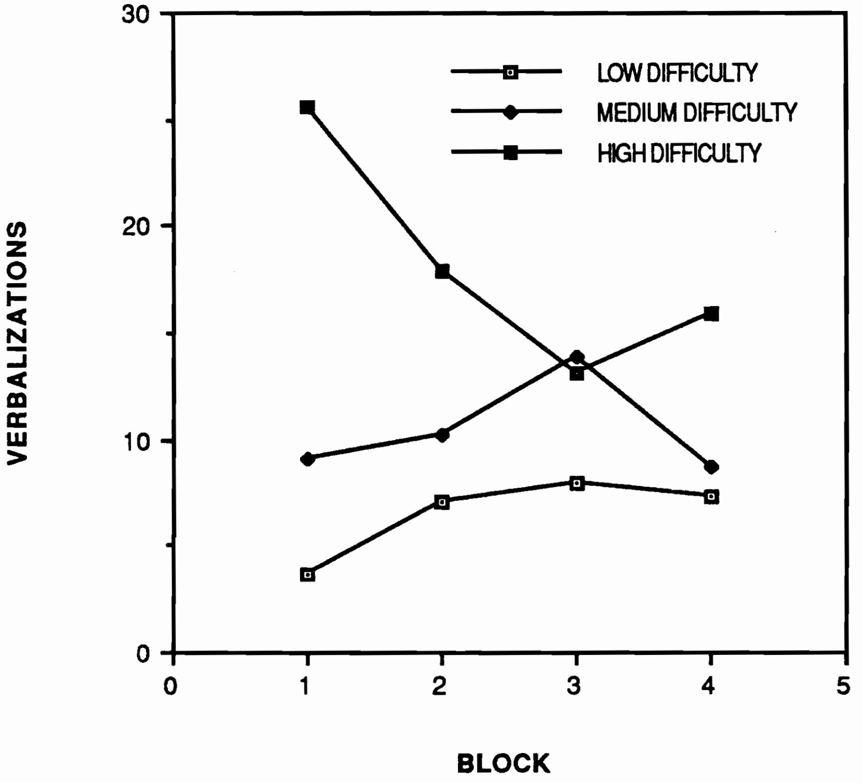


Figure 19. Difficulty X Block interaction for number of verbalizations.

understanding the differences between Retrospective and Concurrent verbal protocol techniques. This information is discussed below.

Verbal protocol categories. Each of the verbal protocol statements was categorized into one of the six categories. These frequencies were then recorded in a Monitor X Protocol X Category matrix and analyzed with a 2 X 2 X 6 Sutcliffe chi-square which allows frequency data to be analyzed for interactions as well as main effects (Sutcliffe, 1957). The summary table of the Sutcliffe chi-square, Table D35, indicates that all effects are significant except the M X P interaction.

The subjects' overall pattern of verbalizations are shown in Figure 20. The kinds of information provided by the retrospective and concurrent subjects are quite different. In general, the verbalizations given by concurrent subjects were low level types of verbalizations. The concurrent subjects gave more Procedure and Reading statements than did the retrospective subjects (Table D36, Figures 21 and 22). The retrospective subjects gave more Procedure + Explanation, Explanation, and Design statements than did the concurrent subjects (Table D36, Figures 23 through 25).

The monitor on which the subjects performed the tasks had some influence over the types of statements made (Table D37). Subjects using the small monitor made more Procedure statements than subjects using the large monitor (Figure 26). Conversely, subjects using the large monitor made more design statements than subjects using the small monitor (Figure 27). Several possible explanations for this influence of monitor will be discussed later.

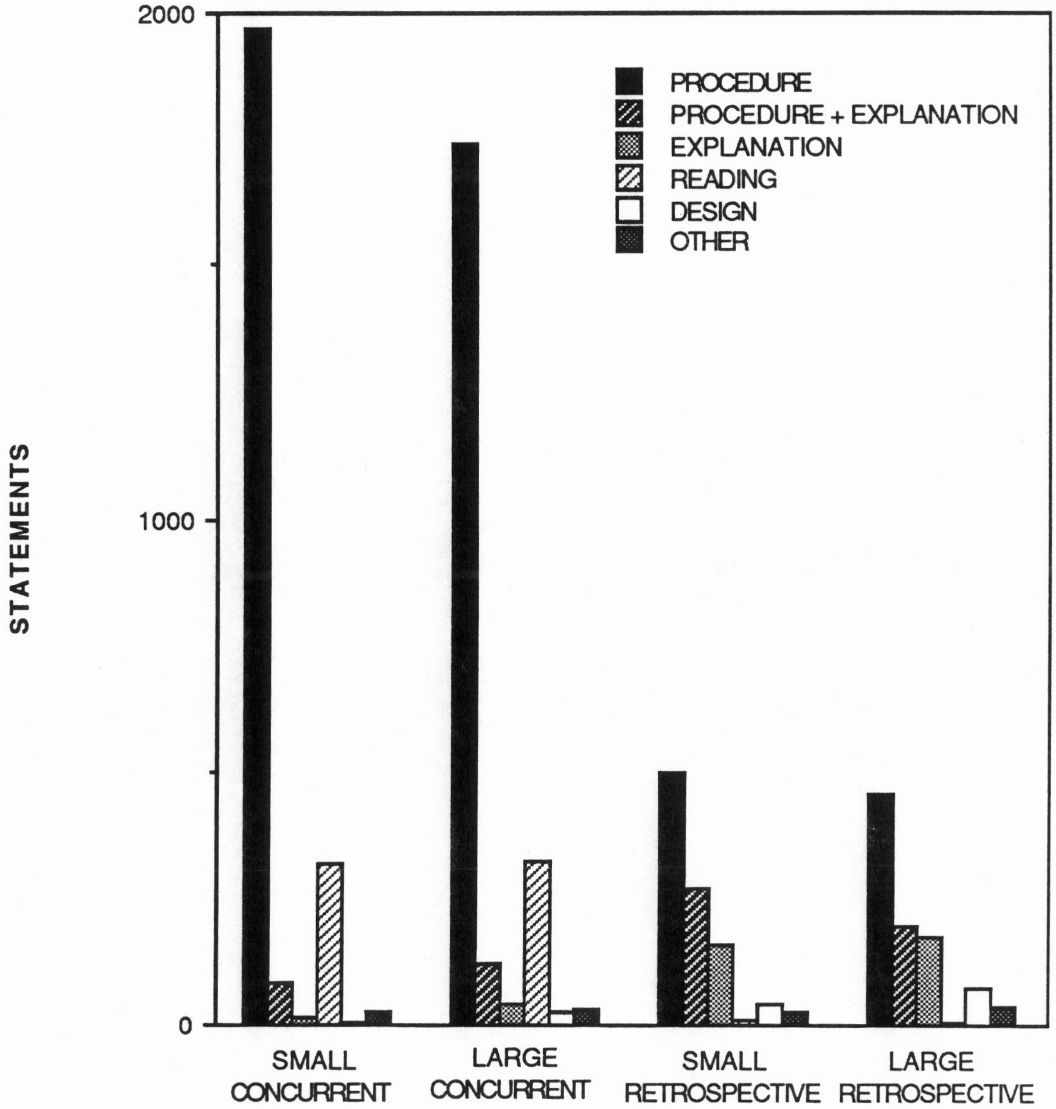


Figure 20. Monitor X Protocol X Category interaction for total number of statements.

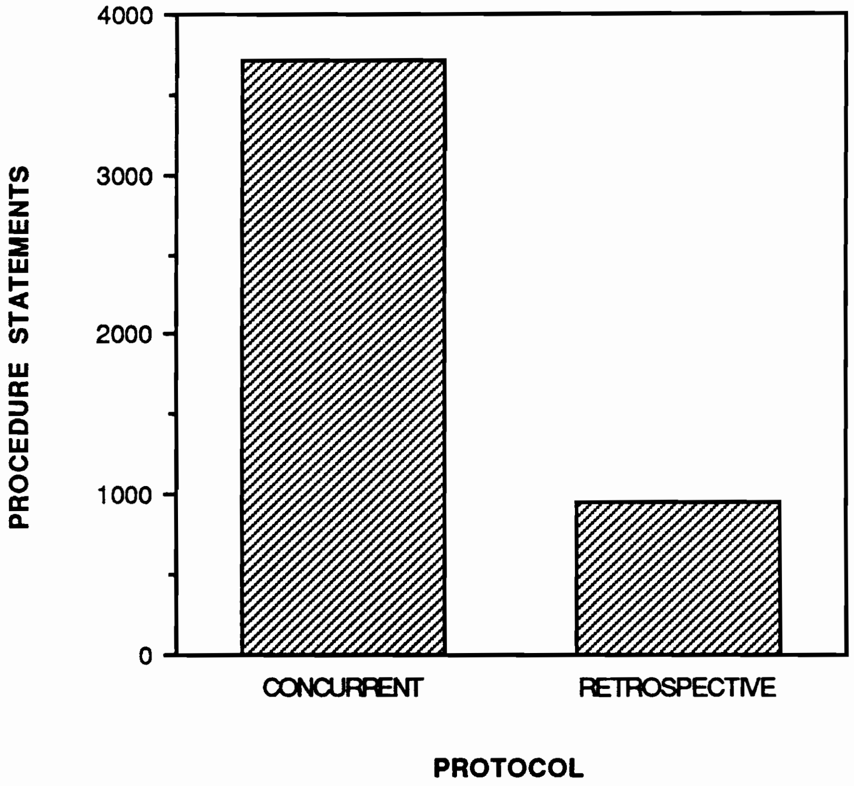


Figure 21. Effect of Protocol on number of Procedure statements.

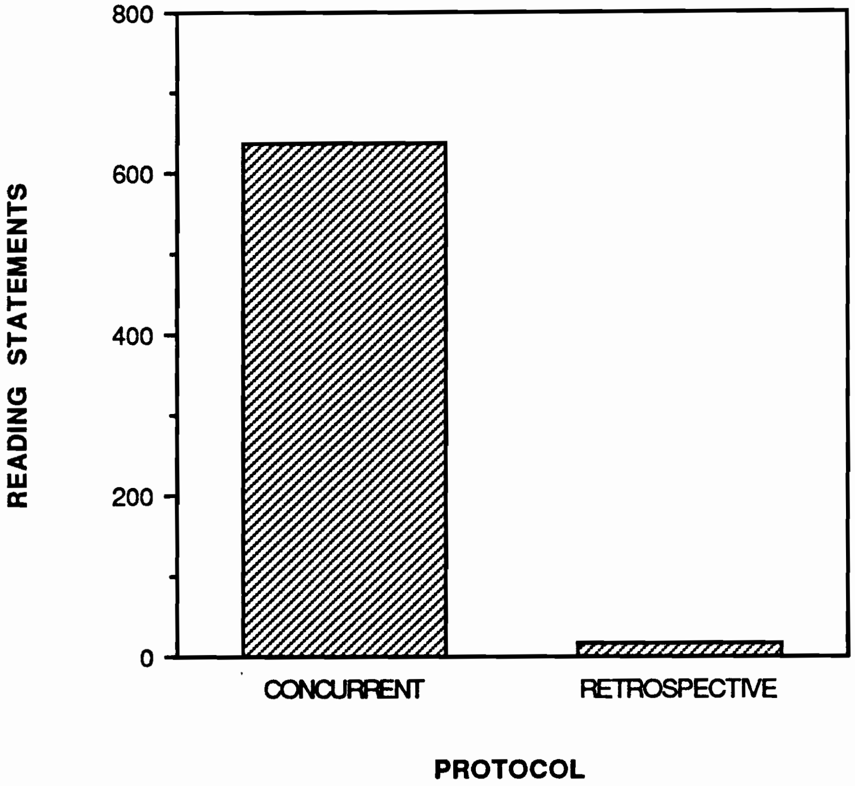


Figure 22. Effect of Protocol on number of Reading statements.

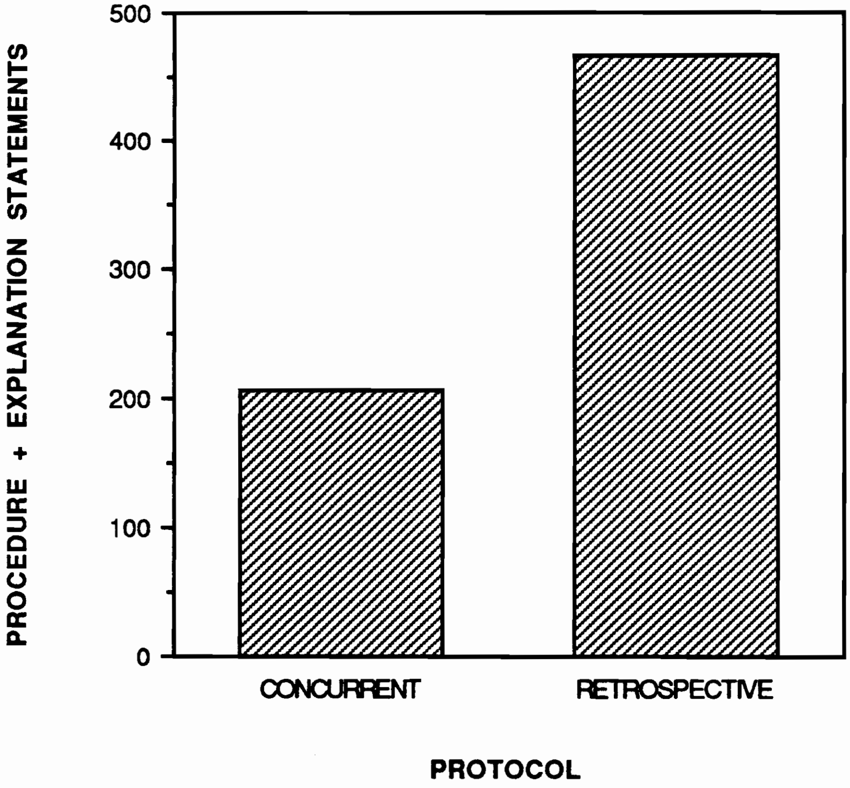


Figure 23. Effect of Protocol on number of Procedure + Explanation statements.

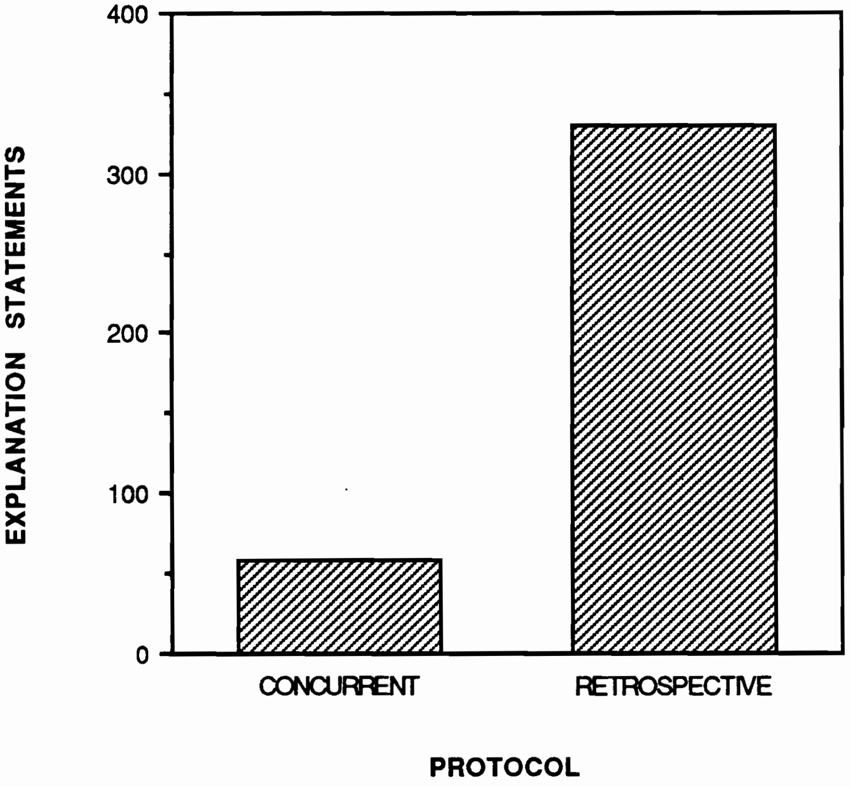


Figure 24. Effect of Protocol on number of Explanation statements.

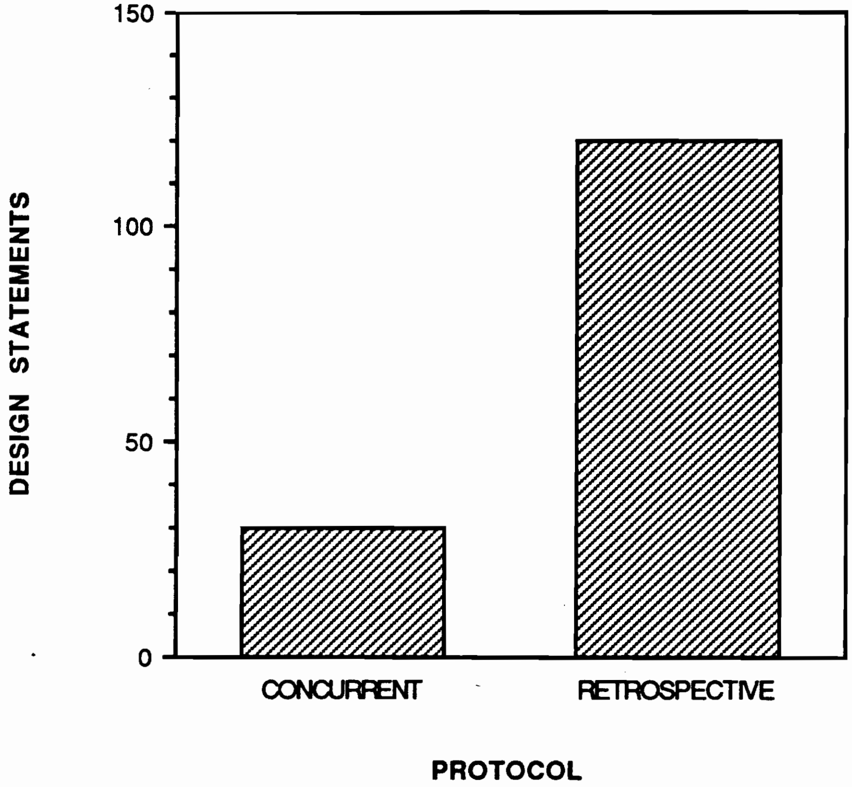


Figure 25. Effect of Protocol on number of Design statements.

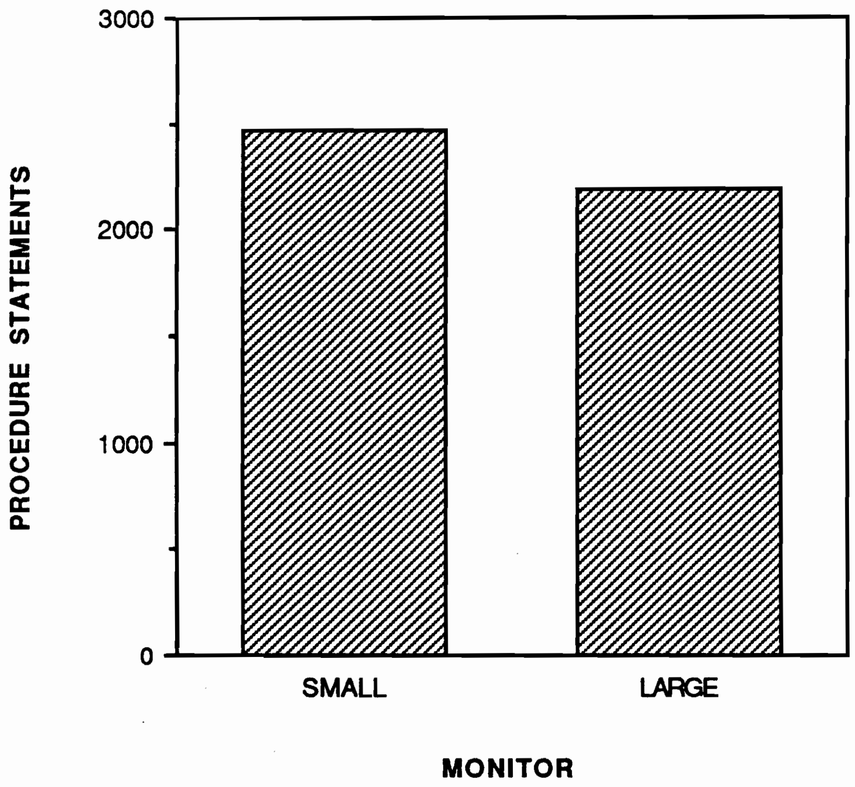


Figure 26. Effect of Monitor on number of Procedure statements.

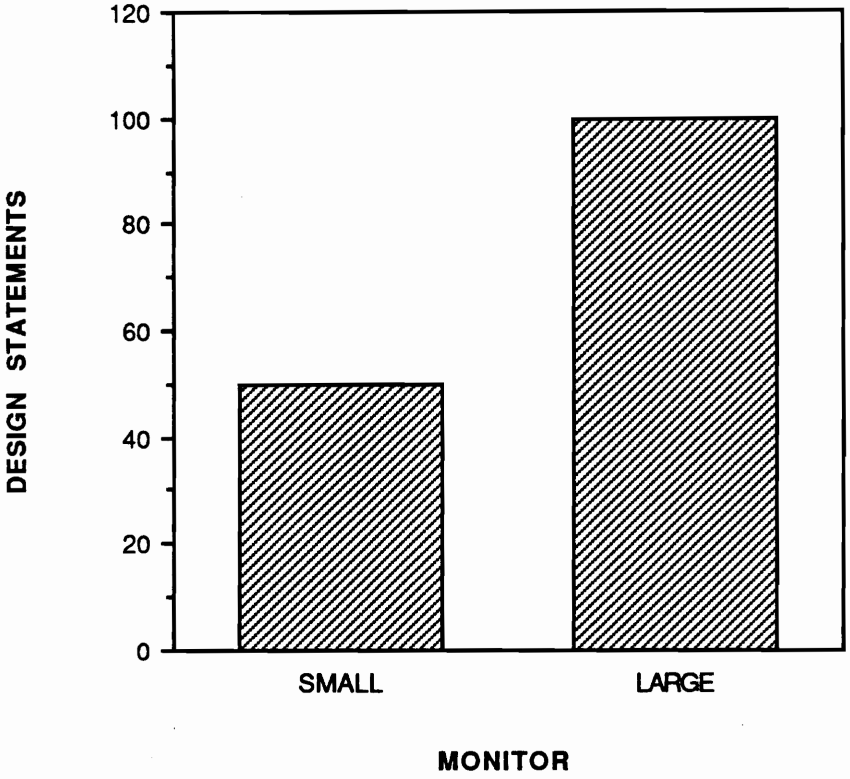


Figure 27. Effect of Monitor on number of Design statements

One rather peculiar relationship existed between the monitor the subjects were using and the protocol method which they were following. Subjects using concurrent verbalization tended to give more PE statements for the large monitor than the small, but those performing retrospective protocol gave more PE statements for the small monitor (Table D38, Figure 28).

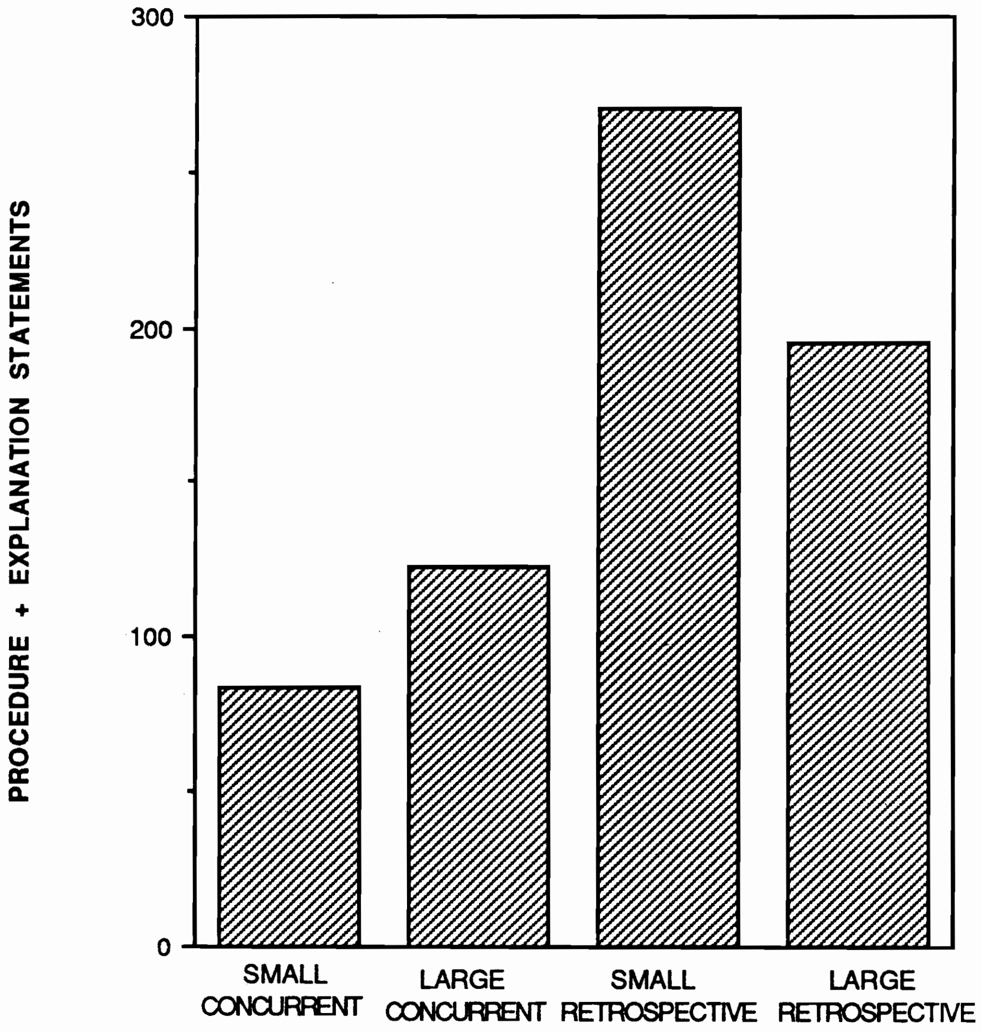


Figure 28. Monitor X Protocol interaction for number of Procedure + Explanation statements.

DISCUSSION

Performance Data

As previously discussed, the use of performance data has been very successful as a metric for usability. Certainly the performance data in this experiment proved to be sensitive to changes brought about by the separate factors such as Task Difficulty, Protocol, and Monitor. The steps to completion and task completion times served to measure effects of the Monitor size and Task Difficulty on the subjects' performance. The errors, however, did not seem to be a useful measure since there were too few to analyze or interpret.

The small number of errors resulted in the inability to draw conclusions about the usability of the different monitor size these errors. Further, due to the few number of errors, there could be no conclusions drawn about the interference of protocol methods on the subjects' ability to complete the tasks in an error-free manner. These results by no means suggest that errors are a poor measure of usability in all situations. They do suggest, however, that in certain situations, errors may not be good measures of usability. A probable explanation for the low error rate is that the subjects' screening was so stringent that those subjects who passed the pretest were at such a level of proficiency that committing errors was very unlikely. Perhaps errors are not a good measure of usability if the subjects of interest are already expert or highly experienced users.

The Difficulty level of the tasks was reflected in the number of steps the subjects required to complete the tasks. This result is not terribly surprising given that the task difficulty levels were empirically derived from weighted averages of the number of windows that must be opened simultaneously, the total number of windows required to complete the task, and the total number of operations performed in the windows, all of which are closely related to steps required for task completion. The steps to completion data also proved to be sensitive to the monitor that the subjects used. These data showed that for the low difficulty tasks, the extra space provided by the large monitor is not important, but as the task difficulty increases, the large monitor allows the subjects to perform the tasks with fewer steps. Thus, the performance data show that for the low difficulty tasks that do not require more than one window to be present simultaneously, the small monitor is adequate. If the tasks require two or more windows to be present on the monitor, however, the large monitor can increase the efficiency with which subjects complete the tasks as measured by steps to completion.

The fact that some subjects decrease over time the number of steps required to complete the task indicates that the subjects may be learning windowing management strategies. This decrease in number of steps is apparent only for the high difficulty tasks; that is, these strategies are not necessary for the low and medium difficulty tasks which do not require as many simultaneously opened windows. Further, the decrease over time in steps to completion is greater for those subjects using the small monitor. Perhaps this larger decrease is due to the relative importance of windowing strategies for the two monitors. Again, it seems that the strategies may be more important for the small monitor because of the smaller work area. The large monitor has more flexibility with regard to the work area so that poor strategies are not as costly in terms of steps to completion as on the small monitor.

Regardless of the development of window management strategies, subjects using the small monitor always required more steps than those using the large monitor for the high difficulty tasks. This result indicates that while subjects decrease their steps over time, they cannot completely overcome the disadvantages of the small monitor on the high difficulty tasks, even with the learning of coping strategies.

The lack of effect of protocol on steps to completion and task completion time is disappointing. As with all research, however, simply not finding a significant difference is not conclusive evidence that no difference exists. As discussed earlier, some researchers (Lewis, 1982) have suggested that task completion times may be increased due to the collection of concurrent protocol. Further, these researchers believe that time is therefore an inappropriate measure of usability when used simultaneously with concurrent protocol. The present results indicate that any increased task completion time due to collection of concurrent protocol is not a severe problem. Perhaps the measure of task completion time may be used simultaneously with concurrent protocol for accurate measures of learnability and throughput. The use of task completion time with concurrent verbal protocol increases the utility of concurrent verbal protocol substantially.

While the lack of performance degradation for concurrent protocol subjects may be disappointing, it may not be completely surprising. According to Ericsson and Simon (1984), the lack of performance differences would certainly be expected if subjects limited their verbalizations to those easily accessed from short-term memory. In other words, if the subjects only volunteer low level verbalizations and are not forced to give verbalizations that require further cognitive processing, then no performance differences should occur. The verbal data indicate that subjects who were required to give concurrent verbalizations did, in fact, give very low level processing. These data did not support Deffner's finding of efficiency degradation as a result of concurrent verbal protocol. One possible explanation for the different results may be in the nature of the two tasks. Deffner's task was a cognitive problem solving task with minimal procedures, while the tasks in the present study had a large procedural component. Perhaps when subjects were asked to verbalize in Deffner's study, they were forced to verbalize higher order information because there were no procedure type statements to be made, thus resulting in performance degradation. Subjects in this study, however, could give procedure statements to fulfill the verbalization requirement which did not cause performance degradation.

Thus, the types of tasks being studied should be considered in deciding a protocol method. Concurrent protocol may be best for tasks that are mainly procedural but not with tasks that are heavy in cognitive problem solving. Retrospective protocol may be best for tasks that have both procedure and strategy development. For tasks that are completely cognitive problem solving, retrospective is appropriate but may not yield all the thought processes that the subject has used.

One of the major issues in this research was whether concurrent verbal protocol interferes with the task performance and whether the performance during concurrent verbal protocol might be generalized to tasks in which users do not think aloud. No performance differences between subjects using concurrent and retrospective protocols were found. This result indicates that the performance during verbal protocol collection is generalizable to situations in which the user is not required to verbalize. As discussed earlier, however, it seems that subjects using concurrent protocol gave very low level information. Perhaps if the subjects had somehow been forced to give more detailed and explanatory information, then performance would have been degraded. This approach might be interesting in further research. The present study, however, did not show any interference from concurrent protocol.

Another interesting explanation for the lack of performance differences between the concurrent and retrospective subjects is that the retrospective subjects' knowledge that they would be verbalizing after the task affected their performance. Perhaps the retrospective subjects were rehearsing what they planned to say during retrospective verbalization while they were doing the tasks. If the subjects were rehearsing, then their performance might not be different than the concurrent subjects who were actually verbalizing. While this seems unlikely, given the differences in the content of the verbalization, future research might include a retrospective condition in which the subjects would be unaware of the subsequent retrospective data collection.

Subjective Data

QUIS was not sensitive enough to measure the differences in usability between the small and large monitors. While QUIS has been shown to discriminate differences between user's favorite and least favorite interfaces (Chin et al., 1988), the lack of sensitivity to the monitor differences seems to indicate that QUIS is not a good measure of usability for comparisons of similar interfaces. While the measures of internal reliability were lower than those previously cited in research for the original QUIS, the internal reliability is still high. The high internal reliability indicates that QUIS is a good instrument as a global measure of usability. More research is needed to demonstrate its validity and sensitivity, however.

As discussed earlier, the subjects did not employ the entire range of the difficulty ratings. Previous research has indicated that subjects tend to use the lower end of task difficulty rating scales (Mitta, 1988). The present subjects did, however, use enough of the rating scale to discriminate differences in perceived difficulty. The difficulty ratings did seem to be a sensitive measure of the subject's perceptions of usability.

The subjects' ability to discriminate the levels of difficulty with the difficulty ratings suggests that the original operational definitions are valid, as discussed earlier. Subjects consistently rated the high difficulty tasks as more difficult than the medium, and the medium as more difficult than the low difficulty tasks. This consistency indicates the very strong effect of the level of task difficulty on the difficulty rating, which also lends confidence that the subjects were giving thought to their difficulty ratings.

While the data indicate that subjects were making valid difficulty ratings, the subjects may have had some difficulty with anchoring the ratings. They did not use the full range of ratings. This result suggests that if researchers intend to use rating scales as a measure of usability, they should allow the users to experience the entire range of ratings in order to help them anchor their ratings.

While the subjects who used the small monitor rated the tasks as more difficult than subjects using the large monitor, the difference was not present for the low difficulty tasks. The monitor size became a factor in the difficulty rating of tasks only as the difficulty of the tasks increased. This interaction suggests that tasks which require very little simultaneous windowing are not made easier by the use of a large monitor. However, when two or three simultaneous windows are necessary, users begin to perceive the advantages offered by a large monitor. When four simultaneous windows are required, the small monitor makes the tasks much more difficult for subjects than the large monitor. The fact that the monitor only became a factor as the task difficulty increased has implications for the choice of monitor size for designers. While the issue of monitor size was by no means fully researched because only two monitor sizes were evaluated, the results clearly indicate that the types of tasks and number of windows required for these tasks should be incorporated in determining the size of monitor to purchase. While these monitors may be fairly representative of many of the monitors in use, there are certainly smaller monitors, the Macintosh 23 cm diagonal, for example. It could be that the 23 cm monitor would be a disadvantage even on the low difficulty tasks. Further, it could be that the 53.34 cm monitor tested would not have an advantage over a 43 cm diagonal monitor for high difficulty tasks. The results indicate that monitor size is an issue when two or more windows are present.

Verbal Data

As discussed earlier, two types of verbal data were collected. The numbers of verbalizations were merely the amount of verbalization without regard for content. Perhaps the most obvious and maybe surprising information in the number of verbalizations data is that concurrent protocol subjects make more statements than do retrospective subjects. The majority of these verbalizations are procedural, however, and seem to accompany steps of the tasks. It seems that many of the trends of the verbalization, especially those for the concurrent subjects, are driven by the number of steps to completion.

As reported earlier, the subjects decreased their verbalization over time for the high difficulty tasks while increasing them for the low and medium difficulty tasks. First, it could be that since the number of verbalizations is related to steps to completion ($r = 0.97$, $p < 0.0001$), that as the subjects used more efficient strategies and fewer steps for the high difficulty tasks that there were fewer verbalizations. Further, as they began to learn more efficient strategies for doing the high difficulty tasks, they generalized these strategies to the low and medium tasks. Strategies which are more efficient for tasks with many windows often prove to be less efficient for tasks with fewer windows so that subjects actually use more steps, and perhaps verbalize more.

The idea that verbalizations are driven by Procedure statements is further supported by the fact that the trends that were found with steps to completion are represented in the data of the concurrent subjects more strongly than those of the retrospective subjects. For example, concurrent-subject data revealed that there were more verbalizations for high difficulty tasks than for medium, and more verbalizations for medium than for low which reflects the trends in the number of steps to completion. The retrospective subjects, on the other hand, had more verbalizations in the high difficulty tasks than the low and medium, but no differences existed between the low and medium difficulty tasks and the steps to completion trends were still present. It seems that the verbalizations for concurrent subjects more closely mirrored the trends in steps to completion, and that the concurrent had more Procedure statements. This relationship suggests that the number of verbalizations may follow the steps to completion rather than vice versa.

While the results indicate that more statements are made by subjects using concurrent protocol, it is the information in these statements that may be of more interest to the researcher. The information gathered from the two protocol techniques is obviously quite different. The results indicate that concurrent subjects are giving relatively low level verbalizations in the form of Procedure and Reading Statements, while the retrospective subjects gave more complex statements (Procedure + Explanation, Explanation , and Design). These differences could be due to the amount of effort the subject can put into the verbal portion of the task. Subjects in the concurrent protocol condition seem to be attending to the experimental tasks and give little thought to the comments they are giving. The retrospective subjects, on the other hand, can give their full attention to the verbalization and in doing so give richer information.

The performance decrement for concurrent protocol subjects was not found. Perhaps an explanation for this lack of performance decrement is that the verbalization was, in essence, a secondary non-interfering task. When the subjects were attending to the experimental tasks, they gave less attention to the verbalization process. This approach resulted in a performance decrement in the verbalization such that subjects gave only low level verbalization. Therefore, there was no performance decrement on the tasks. This result is also in keeping with the Ericsson and Simon idea that verbalizations that do not require further processing beyond short-term memory will not interfere with task performance.

The fact that the monitor had an effect on the kinds of information that subjects gave indicates that subjects may be sensitive to factors outside the actual protocol procedure which may influence their verbalizations. The greater number of Design statements made by subjects using the large monitor suggests that the unusual monitor drew the subjects' attention to design issues. For example, many subjects using the large monitor commented that the large screen made the tasks easier. Subjects in the small monitor condition were not as aware of screen size as a design issue since most were using the same size monitor to which they were accustomed. Further, once the subjects were aware of design issues in general, they were possibly more likely to comment on these issues. The greater number of Procedure statements given by subjects may, however, simply be due to the extra steps incurred on the small monitor. Perhaps P statements often accompanied a step towards the task completion, and because there were more steps to completion for the small monitor, there were more procedure statements in this condition. The other categories of statements may have been more independent of the number of steps required by the monitor, and thus did not increase for the small monitor.

The information that was gained through the retrospective protocol indicates that the subjects seem to have the ability to recall the reasoning as well as the procedures of the tasks done during an experimental session. The concern of some researchers that these thoughts may not be available retrospectively does not seem to be a serious issue. The subjects often recalled the tasks, the procedure they had employed, and the mistakes that they had made as soon as they were shown the task, without having to go all the way through the task. In general, while more time consuming, there do not seem to be any major methodological problems associated with the collection of retrospective protocols.

Verbal Protocol for Usability Testing

Many of the disadvantages of both these methods seem to not be serious problems for concurrent and retrospective protocols. In referring to the original list of disadvantages for the two methods of protocol in Table 1, several of these disadvantages are not significant factors in this research. No time differences were detected between the retrospective and concurrent subjects, which suggests that time data may be appropriate with concurrent protocols. There was no evidence of efficiency degradation for concurrent protocol subjects. It seems that the information of interest is still available retrospectively, which makes retrospective protocol more attractive. The extra time needed to carry out research is still a factor for retrospective protocol. It seems the major issue for deciding between the two methods of protocol may be in the question which the researcher intends to answer. If procedural, low-level information is required, and nothing more, than concurrent protocol seems to be the most efficient method. If the researcher desires higher level explanations, however, then retrospective protocol should be considered because both procedural and higher-level explanations may be achieved with retrospective protocol.

Another issue which may affect the use of verbal protocol in usability testing is the time constraint under which the experimenter must work. Retrospective protocol, by its very nature, requires more time than does concurrent protocol. The subjects must complete the tasks and then go through them a second time in order to verbalize. The experimenter must decide if the information retrieved from a retrospective protocol is worth the extra time invested. Perhaps the subjects do not need to be shown the entire task procedure but only a skeleton of the procedure in which the cueing is minimal. For example, the retrospective subject could verbalize his general strategy, make suggestions for improving the design, and then be allowed to fast forward through procedural videotape that serves no cueing purpose. This approach may be quite successful because retrospective subject needed little cueing to remind them of how they had proceeded through the tasks and why they had chosen certain strategies. One problem with the use of minimal cues may be that subjects tend to underestimate the time they spent going down the wrong path. For this approach to be successful, it seems that subjects should not have the ability to fast forward during a critical incident. Future research might indicate how successful this approach might be as well as how to determine when subjects should not have the ability to fast forward.

Both concurrent and retrospective protocols produce qualitative rather than quantitative data. The information learned from these protocols cannot directly be used for quantitative verification of Schackel's usability goals. The current research indicates that collection of these protocols does not interfere with traditional measures of these goals such as time and error measurements. Information to help reach these goals may be obtained through the protocol procedures, however. For example, concurrent protocol may reveal that the procedures that the subjects employ are not the most efficient methods for task completion. This information could be used to change the interface such that the inefficient strategies are avoided, thereby increasing throughput. The kinds of information learned from concurrent verbal protocol might be more useful towards the end of the design cycle when working prototypes exist. Retrospective protocol, on the other hand, may reveal more basic information about the desires and expectations of the interface. This information would be more useful early in the design phase.

CONCLUSIONS

The implications of this research are in two major directions. First, the results of the data address the issue of whether to employ a large or small monitor. Secondly, the data address the methodological issue of data collection with concurrent versus retrospective protocol.

Monitor Size

The data did not reveal any evidence that the large monitor is ever a disadvantage. The large monitor did reduce the steps to completion and difficulty ratings of tasks that required four or more simultaneous windows. The results were not as clear cut for tasks that required two or three simultaneous windows. Obviously if a user is considering whether to use a large screen, the tasks that the end user will be doing should be considered. The results of this study indicate that with tasks requiring two or three simultaneous windows, the user may want to use a large monitor. When the tasks require four or more windows, a large monitor should be seriously considered.

Verbal Protocol

One of the major issues of this research was whether concurrent verbal protocol interferes with the processing required to complete a task and whether information learned through concurrent verbal protocol can be generalized to situations in which users do not think aloud. No performance differences between subjects using concurrent and retrospective protocols were found, It appears that this information can be generalized. Further, no task completion time differences were found. Verbal protocol appears to be a good method for collecting usability data if the researcher is only interested in very procedural information.

The concern that retrospective subjects may not recall their procedures or reasoning for given task strategies did not seem to be a problem. The subjects often recalled the tasks, the procedure they had employed, and the mistakes that they had made as soon as they were shown the task, without having to go all the way through the task. If a researcher is interested in richer information such as explanation and design issues, retrospective verbal protocol is the method of choice for verbal protocol collection.

The results of this research do not indicate that one protocol is always good and the other bad, but instead, indicate that the conditions under which the data is collected such as time constraints, as well as the questions that need to be addressed, should determine the protocol to be employed.

REFERENCES

Bailey, W. A., Knox, S. T. and Lynch, E. F. (1988). Effects of interface design upon user productivity. In *CHI '88 Proceedings.* New York: ACM (pp. 207-212).

Barnard, P., Wilson, M. and MacLean, A. (1986). The elicitation of system knowledge by picture probes. In *CHI '86 Proceedings.* New York: ACM (pp. 235-240).

Bennett, J. C. (1984). Managing to meet usability requirements: Establishing and meeting software development goals. In J. Bennett, P. Case, J. Sandelin, and M. Smith (Ed.) *Visual display terminals.* Englewood,NJ: Prentice Hall (pp.161-184).

Borenstein, N.S. (1985). The evaluation of text editors: In *CHI'85 Proceedings.* New York: ACM (pp. 99-105).

Brewley W. L., Roberts, T. L., Schroit, D. and Verplank, W. L. (1983). Human factors testing in the design of Xerox's 8010 "Star" office workstation. In *CHI'83 Proceedings.* New York: ACM. (pp. 72-77).

Butler, K. A. (1985). Connecting theory and practice: A case study of achieving usability goals. In *CHI '85 Proceedings.* New York: ACM (pp. 85-88).

Card, S. K., Moran,T. P., and Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. In *Communications of the ACM.* July, (pp. 396-410).

Chin, J. P., Diehl, V. A., and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human computer interface. In *CHI '88 Proceedings.* New York: ACM (pp. 13-28).

Chin, J. P., Norman, K. L., and Shneiderman, B. (1987). *Subjective user evaluation of CF Pascal programming tools.* (CAR-TR-304; CS-TR-1980), University of Maryland, College Park, Maryland.

Coleman, W. D., Williges, R. C., and Wixon, D. R. (1985). Collecting detailed user evaluations of software interfaces. In *Proceedings of the Human Factors Society.* Santa Monica, CA: Human Factors Society (pp.240-244).

Cordes, R. E. (1984). Application of magnitude estimation for evaluating software ease-of-use. In G. Salvendy (Ed.) *Human-computer interaction.* Amsterdam: Elsevier Science Publishers B.V., (pp.199-203).

Davies, S. E., Bury, K. F., and Darnell, M. J. (1985). An experimental comparison of a windowed versus non-windowed operating system environment. In *Proceeding of the Human Factors Society.* Santa Monica, CA: Human Factors Society (pp. 250-54).

Deffner, G. (1988). Concurrent thinking aloud: An on-line tool for studying representations used in text understanding. *TEXT, 8* (4), 351-367.

Deffner, G. (1987). Eye movement pattern as indicator of solution strategy in solving n-term series tasks. Presented at ECEM Gottingen.

del Galdo, E. M., Williges, R. C., Williges, B. H., and Wixon, D. R. (1986). An evaluation of critical incidents for software documentation design. In *Proceedings of the Human Factors Society.* Santa Monica, CA: Human Factors Society (pp.19-23).

Ericsson, K.A. and Simon, H.A. (1980). Verbal reports as data. *Psychological Review, 87,* 215-251.

Ericsson, K.A. and Simon, H.A. (1984). *Protocol Analysis.* Cambridge: MIT Press.

Gaylin, K. B. (1985). *Creating an empirically-based windowing benchmark task.* (Digital Equipment Corporation Technical Report DEC-TR-370) Human Engineering Research.

Good, M. D., Spine, T. M., Whiteside, J., and George, P. (1986). User-derived impact analysis as a tool for usability engineering. In *CHI '86 Proceedings.* New York: ACM (pp. 241-246).

Good, M. D., Whiteside, J. A., Wixon, D. R., and Jones, S. J. (1984). Building a user-derived interface. In *Communications of the ACM,* October. (pp.1032-1043).

Gould, J. D. and Lewis, C. (1983). Designing for usability--key principles and what designers think. In *CHI '83 Proceedings.* New York: ACM (pp. 50-53).

Hopgood, F. R. A., Duce, D. A., Fielding, E. V. C., Robinson, K., and Williams, A. S. (1985). *Methodology of window management.* Berlin: Springer-Verlag.

Lewis, C. (1982). *Using the "thinking aloud" method in cognitive interface design.* (IBM Research Report RC 9265 (#40713)) IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y.

Lund, M. L. (1985). Evaluating the user interface: The Candid Camera approach. In *CHI '85 Proceedings*. New York : ACM (pp.107-113).

Mitta, D. A. (1988). *Expert system usability: Modeling and analysis of human-advisor interaction*. Unpublished doctoral dissertation. Virginia Polytechnic Institute and State University, Blacksburg, VA.

Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

O'Malley, C., Draper, S. and Riley, M. (1984). Constructive interaction: A method for studying user-computer user interaction. In B. Shackel (Ed.) *Interact '84*. London: North Holland (pp.1-5).

Roberts, T. L. and Moran, T. P. (1983). The evaluation of text editors: Methodology and empirical results. In *Communications of the ACM*, April. New York: ACM (pp. 265-282).

Root, R. W. and Draper, S. (1983). Questionnaires as a software evaluation tool. In *CHI '83 Proceedings*. New York: ACM (pp. 83-87).

Shackel, B. (1984). The concept of usability. In J. Bennett, P. Case, J. Sandelin, and M. Smith (Ed.) *Visual display terminals.* (pp. 45-87) Englewood, NJ: Prentice Hall.

Shneiderman, B. (1987). *Designing the user interface: Strategies for effective human-computer interaction.* Reading, MA: Addison-Wesley.

Simes, D. K. and Sirsky, P. A. (1985). Human factors: An exploration of the psychology of human computer dialogues. In H.R. Hartson (Ed.) *Advances in human computer interaction*, Volume 1. Norwood, NJ: Ablex Publishing Corporation. (pp. 49-104).

Sutcliffe, J. P. (1957). A general method of analysis of frequency data for multiple classification designs. *Psychological Bulletin, 34*, 134-137.

Whiteside, J., Jones, S., Levy, P. S., and Wixon, D. (1985). User performance with command, menu, and iconic interfaces. In *CHI '85 Proceedings.* New York: ACM (pp.185-191).

APPENDIX A Example Tasks

Task 1

Your first task is to log your mail. In order to do this you must locate each item of mail. Presently, you have two items, **Boss Mail 1** and **Boss Mail 2** . Please open each item and place the time (including the time in seconds) at which you are opening it in the appropriate spaces provided in the mail item. You must retrieve the time from the **Control Panel** found under the apple. You should not read the mail but only place the opening time in it. You will be reading the mail later.

When you have completed the task, type "DONE" in the space provided below, close the all the opened windows but stay in the MS-Word application and then go on to **Task 2**.

.

__________.

Task 2

Before you begin to work on your document, you remember that you must calculate the time that you spent on each activity for the last week at work. In order to do this, open your file called **Time** and calculate the "percentage of contracted" for each activity using the system's calculator. The method with which to calculate the "percentage contracted " is found below the table. After you have calculated the number, enter it in the "percentage of contracted" column.

When you have completed the task, type "DONE" in the space provided below, close the all the opened windows but stay in the MS-Word application and then go on to **Task 3**.

.

________.

Task 3

You must set up an appointment with your two other coauthors and your boss to discuss some deadlines for your document. Fortunately, all four of you keep your calendars on-line. Please check all four of your calendars (**Pam's Calendar, Bob's Calendar**, **Boss's Calendar**, and **My Calendar**) to discover when all three of you are available to meet on October 11. When you do find a time when you are all free, enter "meeting with coauthors" on all of the calendars.

When you have completed the task, type "DONE" in the space provided below, close the all the opened windows but stay in the MS-Word application and then go on to **Task 4**.

.

________.

Task 4

One section of your document should contain a section which helps the new employees find housing. This section provides information for both renting and owning homes. Pam was responsible for researching the apartment locaters (list is called **Apartment Locaters**), You were responsible for apartments (a list called **Apartments**), and Bob was responsible for collecting the names of Realtors (list is called **Realtors**). While the authors were responsible for only one of these areas, if they happened to discover information related to another area, then it was included along with the author's own information. Three lists resulted which include the author's own area of responsibility as well as some extra information. Please make a master list of the housing information that includes all of the information obtained by all the authors but has no overlap. The master list should be included after its explanation which is located in the **Housing** file.

When you have completed the task, type "DONE" in the space provided below, close the all the opened windows but stay in the MS-Word application and then go on to **Task 5**.

____

APPENDIX B Window Benchmark

List of Tasks

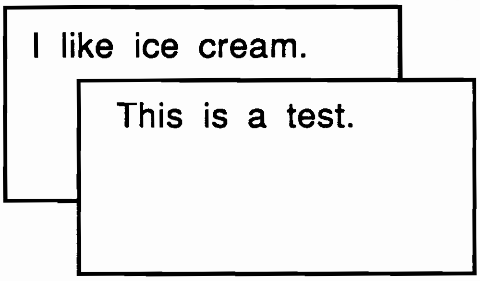
1. Create a MS-Word window in the center of the screen which is about 1/4 the size of the screen.

2. Create a second MS-Word window (also about 1/4 the size of the screen) in the lower left corner of the screen.

3. Type THIS IS A TEST in the center window.

4. Type I LIKE ICE CREAM in the lower left window.

5. Move the window in which you typed I LIKE ICE CREAM such that it is overlapped by the first window as displayed in the diagram below.



6. Make the window in which you typed I LIKE ICE CREAM overlap the window in which you typed THIS IS A TEST.

7. Create a third MS-Word window to the left of the two windows that you created.

8. Use Alarm Clock to find the appropriate time; type the time into the new window.

9. Save, name, and close all the windows except the instructions window (name the windows " **A**", "**B**" , and "**C**".

10. Open **Window 1**, **Window 2**, and **Window 3**.

11 Find the window which has additional graphical information in it and close the other two.

12. Create a new MS-Word window and copy the extra graphical information into the new window.

13. Save and close the new window as **Graphic**; also close the other window.

14. Open the system's calculator and calculate the the answers to the problems in the file called **Problems**. Place the calculated answers in the appropriate places in the **Problems** document.

15. Close <u>all</u> the remaining windows and exit MS-WORD.

APPENDIX C QUIS

QUIS

Computer keeps you informed about what it is doing
never always
1 2 3 4 5 6 7 8 9

System reliability
unreliable reliable
1 2 3 4 5 6 7 8 9

Computer terminology is related to the task you are doing
always never
1 2 3 4 5 6 7 8 9

Exploring new errors by trial and error is
easy difficult
1 2 3 4 5 6 7 8 9

Overall the system was
easy difficult
1 2 3 4 5 6 7 8 9

Over all the system was
stimulating dull
1 2 3 4 5 6 7 8 9

System speed
too slow fast enough
1 2 3 4 5 6 7 8 9

Organization of information of screen
confusing very clear
1 2 3 4 5 6 7 8 9

Use of terms throughout system
consistent inconsistent
1 2 3 4 5 6 7 8 9

Help messages on screen
helpful unhelpful
1 2 3 4 5 6 7 8 9

Tasks can be performed in a straight-forward manner
always never
1 2 3 4 5 6 7 8 9

Position of messages on screen
inconsistent consistent
1 2 3 4 5 6 7 8 9

Error messages
unhelpful helpful
1 2 3 4 5 6 7 8 9

Correcting your mistakes
difficult easy
1 2 3 4 5 6 7 8 9

Overall the system was
terrible wonderful
1 2 3 4 5 6 7 8 9

Characters on computer screen
easy to read hard to read
1 2 3 4 5 6 7 8 9

Overall the system had
inadequate power adequate power
1 2 3 4 5 6 7 8 9

Remembering the name and use of commands
easy difficult
1 2 3 4 5 6 7 8 9

Supplemental reference materials
confusing clear
1 2 3 4 5 6 7 8 9

Highlighting on the screen simplifies the task
very much not at all
1 2 3 4 5 6 7 8 9

Messages on screen which prompt user for input
confusing clear
1 2 3 4 5 6 7 8 9

Sequence of screens
very clear confusing
1 2 3 4 5 6 7 8 9

Overall system was
satisfying frustrating
1 2 3 4 5 6 7 8 9

Overall system was
rigid flexible
1 2 3 4 5 6 7 8 9

System tended to be
quiet noisy
1 2 3 4 5 6 7 8 9

Experienced and inexperienced users' needs are taken into consideration
never always
1 2 3 4 5 6 7 8 9

Learning to operate the system
easy difficult
1 2 3 4 5 6 7 8 9

APPENDIX D ANOVA and Newman-Keuls Tables

TABLE D1

Summary Table of Analysis of Variance on Steps to Completion.

Source of Variance	df	MS	F	p
Monitor (M)	1	3015.84	32.71	< 0.0001
Protocol (P)	1	73.67	0.80	0.3763
M x P	1	220.03	2.39	0.1296
Subjects (S/M,P))	44	92.21		
Difficulty (D)	2	42653.08	1326.24	< 0.0001
M x D	2	840.01	26.12	< 0.0001
P x D	2	45.23	1.41	0.2505
M x P x D	2	65.54	2.04	0.1362
D x S/M,P	88	32.16		
Block (B)	3	1110.75	50.05	< 0.0001
M x B	3	24.41	1.10	0.3517
P x B	3	33.18	1.49	0.2189
M x P x B	3	25.41	1.14	0.3335
B x S/M,P	132	22.19		
D x B	6	3048.48	116.72	< 0.0001
M x D x B	6	200.32	7.67	< 0.0001
P x D x B	6	25.21	0.97	0.4493
M x P x D x B	6	21.76	0.83	0.5453
D x B x S/M,P	264	26.12		
Total	575			

TABLE D2

Newman-Keuls Test for Effect of Difficulty for Steps to Completion

Difficulty	Steps	
Low Difficulty	10.04	A
Medium Difficulty	15.22	B
High Difficulty	38.05	C

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D3

Simple-Effect F-tests for Monitor at Each Level of Difficulty, Steps to Completion

Source of Variance	df	MS	F	p
M for Low Difficulty	1	92.13	2.86	0.0943
M for Medium Difficulty	1	441.05	13.71	0.0004
M for High Difficulty	1	4162.69	129.44	< 0.0001
D x S/M,P	88	32.16		

TABLE D4

Newman-Keuls Test for Effect of Block, Steps to Completion

Block	Steps	
Block 1	23.07	A
Block 2	23.79	A
Block 3	19.56	B
Block 4	17.99	C

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D5

Simple-Effect F-tests for Block at Each Level of Difficulty, Steps to Completion

Source of Variance	df	MS	F	p
B for Low Difficulty	3	512.02	19.60	< 0.0001
B for Medium Difficulty	3	1043.80	39.96	< 0.0001
B for High Difficulty	3	5651.91	216.38	< 0.0001
D x B x S/M,P	264	26.12		

TABLE D6

Newman-Keuls Test for Block at Each Difficulty Level, Steps to Completion

Low Difficulty	Steps	
Block 1	5.27	A
Block 2	12.54	B
Block 3	11.65	B
Block 4	10.69	B
Medium Difficulty	Steps	
Block 1	12.23	A
Block 2	17.79	B
Block 3	20.42	C
Block 4	10.46	A
High Difficulty	Steps	
Block 1	51.71	A
Block 2	41.04	B
Block 3	26.62	C
Block 4	32.83	D

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D7

Simple-Effect F-tests for Monitor X Block Interaction at Each Level of Difficulty, Steps to Completion

Source of Variance	df	MS	F	p
M x B for Low Difficulty	3	24.75	.95	0.4169
M x B for Medium Difficulty	3	109.55	4.19	0.0064
M x B for High Difficulty	3	290.74	11.13	< 0.0001
D x B x S/P,M	264	26.12		

TABLE D8

Newman-Keuls Test for Monitors and Blocks for Medium and High Difficulty Levels, Steps to Completion

Medium Difficulty			High Difficulty		
Block 1	Steps		Block 1	Steps	
Small Monitor	12.19	A	Small Monitor	60.00	A
Large Monitor	12.17	A	Large Monitor	43.42	B
Block 2			Block 2		
Small Monitor	20.79	A	Small Monitor	44.00	A
Large Monitor	14.79	B	Large Monitor	38.08	B
Block 3			Block 3		
Small Monitor	23.04	A	Small Monitor	30.62	A
Large Monitor	17.79	B	Large Monitor	22.62	B
Block 4			Block 4		
Small Monitor	10.83	A	Small Monitor	36.20	A
Large Monitor	10.08	A	Large Monitor	29.46	B

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D9

Summary Table of Analysis of Variance on Task Completion Times

Source of Variance	df	MS	F	p
Monitor (M)	1	403.34	0.02	0.8768
Protocol (P)	1	31417.56	1.89	0.1758
M x P	1	3844.00	0.23	0.6327
Subjects (S/M,P)	44	16595.08		
Difficulty (D)	2	2511049.25	664.29	< 0.0001
M x D	2	1782.95	0.47	0.6255
P x D	2	1838.45	0.49	0.6165
M x P x D	2	1245.98	0.33	0.7201
D x S/M,P	88	3780.02		
Block (B)	3	37895.72	13.23	< 0.0001
M x B	3	396.05	0.14	0.9370
P x B	3	442.60	0.15	0.9266
M x P x B	3	6035.51	2.11	0.1024
B x S/M,P	132	2865.03		
D x B	6	254096.69	107.71	< 0.0001
M x D x B	6	4089.81	1.73	0.1134
P x D x B	6	1096.52	0.46	0.8341
M x P x D x B	6	3276.58	1.39	0.2192
D x B x S/M,P	264	2359.18		
Total	575			

TABLE D10

Newman-Keuls Results for Difficulty, Task Completion Time

Difficulty Level	Time(s)	
Low Difficulty	93.71	A
Medium Difficulty	192.97	B
High Difficulty	321.80	C

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D11

Newman-Keuls Test of Blocks, Task Completion Time

Block	Task Completion Time	
Block 1	211.31	A
Block 2	214.62	A
Block 3	179.03	B
Block 4	206.34	A

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D12

Simple-Effect F-tests for Block at Each Level of Difficulty, Task Completion Times

Source of Variance	df	MS	F	p
B for Low Difficulty	3	52662.96	22.32	< 0.0001
B for Medium Difficulty	3	87620.04	37.14	< 0.0001
B for High Difficulty	3	405805.06	172.01	< 0.0001
D x B x S/M,P	264	2359.18		

TABLE D13

Newman-Keuls Test for Block at Each Difficulty Level, Task Completion Times

Low Difficulty	Time(s)	
Block 1	57.40	A
Block 2	98.40	B
Block 3	82.60	B
Block 4	136.46	C
Medium Difficulty	Time(s)	
Block 1	134.15	A
Block 2	215.08	B
Block 3	231.98	B
Block 4	190.67	C
High Difficulty	Time(s)	
Block 1	442.37	A
Block 2	330.40	B
Block 3	222.52	C
Block 4	291.87	D

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D14

Summary Table Analysis of Variance on Task Difficulty Ratings

Source of Variance	df	MS	F	p
Monitor (M)	1	50.17	7.55	0.0087
Protocol (P)	1	2.01	0.30	0.5855
M x P	1	6.25	0.94	0.3376
Subjects (S/M,P)	44	6.65		
Difficulty (D)	2	184.42	182.52	< 0.0001
M x D	2	11.04	10.93	< 0.0001
P x D	2	0.07	0.07	0.9288
M x P x D	2	1.00	0.99	0.3739
D x S/M,P	88	1.01		
Block (B)	3	15.89	16.32	< 0.0001
M x B	3	0.96	0.98	0.4036
P x B	3	.44	0.45	0.7183
M x P x B	3	1.11	1.11	0.3485
B x S/M,P	132	0.97		
D x B	6	7.00	11.98	< 0.0001
M x D x B	6	2.01	3.44	0.0027
P x D x B	6	1.12	1.92	0.0776
M x P x D x B	6	1.03	1.77	0.1052
D x B x S/M,P	264	0.58		
Total	575			

TABLE D15

Newman-Keuls Test Results for Difficulty, Task Difficulty Ratings

	Difficulty Rating	
Low Difficulty	1.70	A
Medium Difficulty	2.54	B
High Difficulty	3.66	C

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D16

Simple-Effect F-tests for Monitor at Each Level of Difficulty, Task Difficulty Ratings

Source of Variance	df	MS	F	p
M for Low Difficulty	1	1.88	1.88	0.1749
M for Medium Difficulty	1	9.63	9.63	0.0026
M for High Difficulty	1	60.75	60.75	< 0.0001
D x S/M,P	88	1.01		

TABLE D17

Newman-Keuls Results for Block, Task Difficulty Ratings

Block	Difficulty Rating	
Block 1	2.19	A
Block 2	2.99	B
Block 3	2.72	C
Block 4	2.63	C

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D18

Simple-Effect F-tests Monitor X Block Interaction at Each Level of Difficulty, Task Difficulty Ratings

Source of Variance	df	MS	F	p
M x B for Low Difficulty	3	.96	1.65	0.1782
M x B for Medium Difficulty	3	3.28	5.65	0.0009
M x B for High Difficulty	3	.74	1.27	0.2851
D x B x S/M,P	264	.58		

TABLE D19

Newman-Keuls Test Results for Monitors and Blocks at the Medium Difficulty Level, Task Difficulty Ratings

Medium Difficulty Level					
Block 1	Rating		Block 2	Rating	
Small Monitor	1.96	A	Small Monitor	3.37	A
Large Monitor	1.83	A	Large Monitor	2.25	B
Block 3	Rating		Block 4	Rating	
Small Monitor	3.46	A	Small Monitor	2.25	A
Large Monitor	2.87	B	Large Monitor	2.29	A

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D20

Simple-Effect F-tests Monitor for Each Block at High Difficulty, Task Difficulty Ratings

High Difficulty				
Source of Variance	df	MS	F	p
M for Block 1	1	21.33	36.77	< 0.0001
M for Block 2	1	7.52	12.96	< 0.0003
M for Block 3	1	14.08	24.27	< 0.0001
M for Block 4	1	20.02	34.52	< 0.0001
D x B x S/M,P	264	.58		

TABLE D21

Summary of MANOVA Rating Scales for 27 QUIS Items for Monitor

H = ANOVA SS&CP Matrix for:		Monitor
E = ANOVA SS&CP Matrix for:		Sub(Monitor X Protocol)
P = Rank of (H+E)	=	27
Q = Hypothesis DF	=	1
NE = DF of E	=	44
S = MIN (P,Q)	=	1
M = .5(ABS(P-Q)-1)	=	12.5
N = .5(NE-P)	=	8.5
Wilks' Criterion	L = DET(E)/DET(H+E) =	.3877

F approximation: F (27,18) = 1.05 p = 0.4644

TABLE D22

Summary of MANOVA Rating Scales for 27 QUIS Items for Protocol

H = ANOVA SS&CP Matrix for:		Protocol
E = ANOVA SS&CP Matrix for:		Sub(Monitor X Protocol)
P = Rank of (H+E)	=	27
Q = Hypothesis DF	=	1
NE = DF of E	=	44
S = MIN (P,Q)	=	1
M = .5(ABS(P-Q)-1)	=	12.5
N = .5(NE-P)	=	8.5
Wilks' Criterion	L = DET(E)/DET(H+E) =	.4565
F approximation:	F (27,18) = 0.79	p = 0.7135

TABLE D23

Summary of MANOVA Rating Scales for 27 QUIS Items for Protocol x Monitor

H = ANOVA SS&CP Matrix for:		Protocol X Monitor
E = ANOVA SS&CP Matrix for:		Sub(Monitor X Protocol)
P = Rank of (H+E)	=	27
Q = Hypothesis DF	=	1
NE = DF of E	=	44
S = MIN (P,Q)	=	1
M = .5(ABS(P-Q)-1)	=	12.5
N = .5(NE-P)	=	8.5
Wilks' Criterion	L = DET(E)/DET(H+E) =	.3069

F approximation: F (27,18) = 1.51 p = 0.1853

TABLE D24

Summary of Analysis of Variance for Number of Verbalizations

Source of Variance	df	MS	F	p
Monitor (M)	1	9.00	0.02	0.8973
Protocol (P)	1	13786.67	25.79	< 0.0001
M x P	1	1.17	0.00	0.9628
Subjects (S/M,P)	44	534.47		
Difficulty (D)	2	6684.60	74.39	< 0.0001
M x D	2	29.78	0.33	0.7188
P x D	2	1838.33	20.46	< 0.0001
M x P x D	2	25.75	0.29	0.7515
D x S/M,P	88	89.86		
Block (B)	3	112.21	3.78	0.0121
M x B	3	31.23	1.05	0.3715
P x B	3	29.39	0.99	0.3991
M x P x B	3	99.79	3.37	0.0207
B x S/M,P	132	29.65		
D x B	6	851.38	26.41	< 0.0001
M x D x B	6	12.56	0.39	0.8854
P x D x B	6	210.42	6.53	< 0.0001
M x P x D x B	6	25.21	0.78	0.8848
D x B x S/M,P	264	32.24		
Total	575			

TABLE D25

Newman-Keuls Results for Difficulty, Number of Verbalizations

	Verbalizations	
Low Difficulty	6.52	A
Medium Difficulty	10.43	B
High Difficulty	18.11	C

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D26

Simple-Effect F-tests for Difficulty at Each Protocol Level, Number of Verbalizations

Source of Variance	df	MS	F	p
D for Concurrent	2	22595.19	251.45	< 0.0001
D for Retrospective	2	20103.13	223.72	< 0.0001
D x S/M,P	88	89.86		

TABLE D27

Newman-Keuls Results for Levels of Difficulty for Each Protocol Level, Number of Verbalizations

Concurrent	Number of Verbalizations		Retrospective	Number of Verbalizations	
Low Difficulty	8.85	A	Low Difficulty	4.12	A
Medium Difficulty	14.44	B	Medium Difficulty	6.43	A
High Difficulty	26.45	C	High Difficulty	9.78	B

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D28

Newman-Keuls Results for Effect of Blocks, Number of Verbalizations

Block	Verbalizations	
Block 1	12.76	A
Block 2	11.74	A B
Block 3	11.64	A B
Block 4	10.60	B

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D29

Simple-Effect F-tests for Blocks at Each Level of Difficulty, Number of Verbalizations

Source of Variance	df	MS	F	p
B for Low Difficulty	3	177.13	5.97	0.0007
B for Medium Difficulty	3	266.28	8.98	< 0.0001
B for High Difficulty	3	1371.56	46.26	< 0.0001
B x S/M,P	132	29.65		

TABLE D30

Newman-Keuls Results for Blocks at Each Level of Difficulty, Number of Verbalizations

Low Difficulty	Verbalizations	
Block 1	3.69	A
Block 2	7.08	B
Block 3	7.96	B
Block 4	7.33	B
Medium Difficulty	Verbalizations	
Block 1	9.02	A
Block 2	10.25	A
Block 3	13.81	B
Block 4	8.65	A
High Difficulty	Verbalizations	
Block 1	25.58	A
Block 2	17.90	B
Block 3	13.15	C
Block 4	15.83	B

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D31

Simple-Effect F-tests for P x B for Each Level of Difficulty, Number of Verbalizations

Source of Variance	df	MS	F	p
P x B for Low Difficulty	3	68.95	2.14	0.0955
P x B for Medium Difficulty	3	53.82	1.67	0.1738
P x B for High Difficulty	3	327.46	10.16	< 0.0001
D x B x S/M,P	264	32.24		

TABLE D32

Newman-Keuls Results for Protocols and Blocks for High Difficulty, Number of Verbalizations

			High Difficulty			
Concurrent				Retrospective		
	Block 1	38.83	A	Block 1	13.33	A
	Block 2	34.96	B	Block 2	10.83	A B
	Block 3	20.08	C	Block 3	6.21	C
	Block 4	22.92	B C	Block 4	8.75	B C

Note: Means with the same letters are not significantly different, p > 0.05.

TABLE D33

Simple-Effect F-tests for P x M for Each Block, Number of Verbalizations

Source of Variance	df	MS	F	p
P x M for Block 1	1	64.00	2.16	0.1440
P x M for Block 2	1	47.84	1.62	0.2053
P x M for Block 3	1	173.36	5.85	0.0169
P x M for Block 4	1	15.34	.52	0.4721
B x S/M,P	132	29.65		

TABLE D34

Newman-Keuls Results for Protocols and Monitors, Block 3, Number of Verbalizations

		Block 3			
Concurrent Verbalizations	Number of Verbalizations		Retrospective	Number of	
Small Monitor	15.22	A	Small Monitor	6.79	A
Large Monitor	18.50	B	Large Monitor	5.86	A

Note: Means with the same letters are not significantly different, $p > 0.05$.

TABLE D35

Summary Sutcliffe Chi-Square Table for 2 X 2 X 6 (Monitor X Protocol X Category)

Variable	Chi-Square	df	p
Monitor (M)	6.14	1	< .02
Protocol (P)	1139.90	1	< .001
Category (C)	13955.24	5	< .001
M x P	0.82	1	> .05
M x C	64.91	5	< .001
P x C	6213.57	5	< .001
M x P x C	37.43	5	< .001
Total	21418.01	23	

TABLE D36

Summary Sutcliffe Chi-Square Table for Protocol for each level of Category

Variable	Chi-Square	df	p
P for Procedure	1630.08	1	< 0.001
P for Procedure & Explanation	99.68	1	< 0.001
P for Explanation	188.80	1	< 0.001
P for Reading	590.56	1	< 0.001
P for Design	54.00	1	< 0.001
P for Other	0.32	1	> 0.05

TABLE D37

Summary Sutcliffe Chi-Square Table for Monitor for each Category

Variable	Chi-Square	df	p
M at Procedure	15.74	1	< .01
M at Procedure Explanation	2.04	1	> .05
M at Explanation	3.92	1	> .05
M at Reading	0.01	1	> .05
M at Design	16.66	1	< .01
M at Other	1.68	1	> .05

TABLE D38

Summary Sutcliffe Chi-Square Table for Monitor X Protocol for each Category

Variable	Chi-Square	df	p
M x P for Procedure	7.67	1	> .05
M x P for Procedure & Explanation	19.64	1	< .01
M x P for Explanation	0.10	1	> .05
M x P for Reading	0.08	1	> .05
M x P for Design	0.68	1	> .05
M x P for Other	0.13	1	> .05

APPENDIX E Figures for Higher Order Interactions

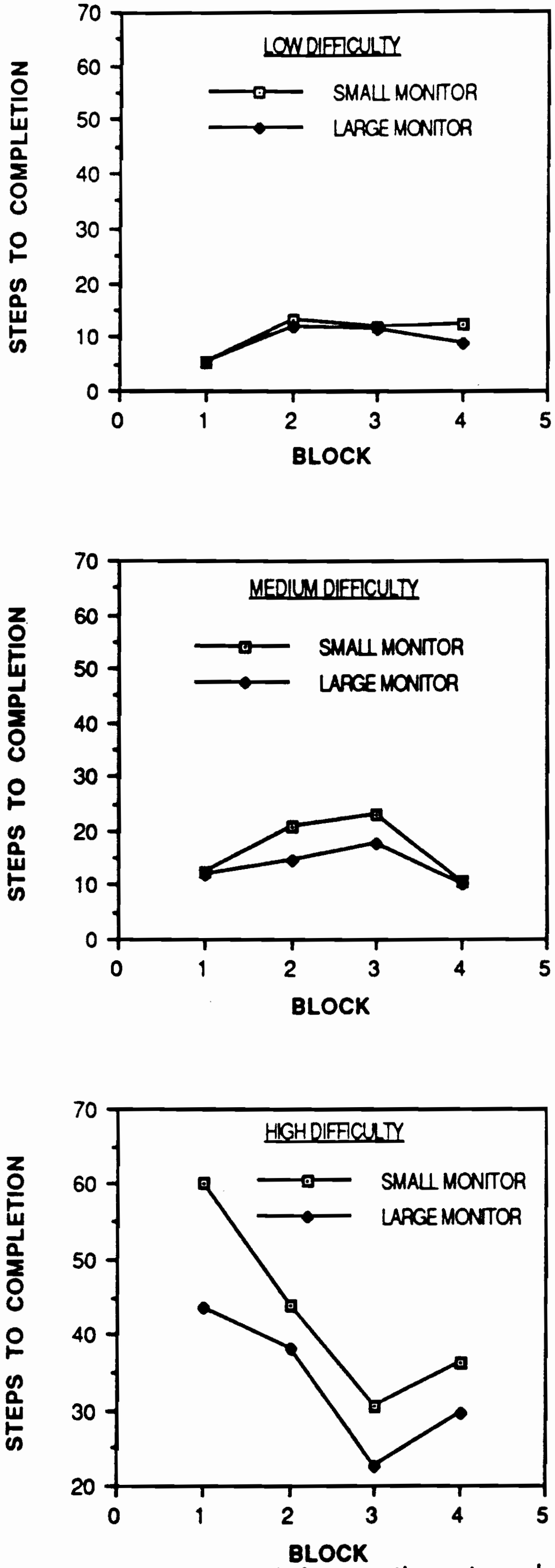


Figure E1. Monitor X Block Interaction at each level of Difficulty for steps to completion.

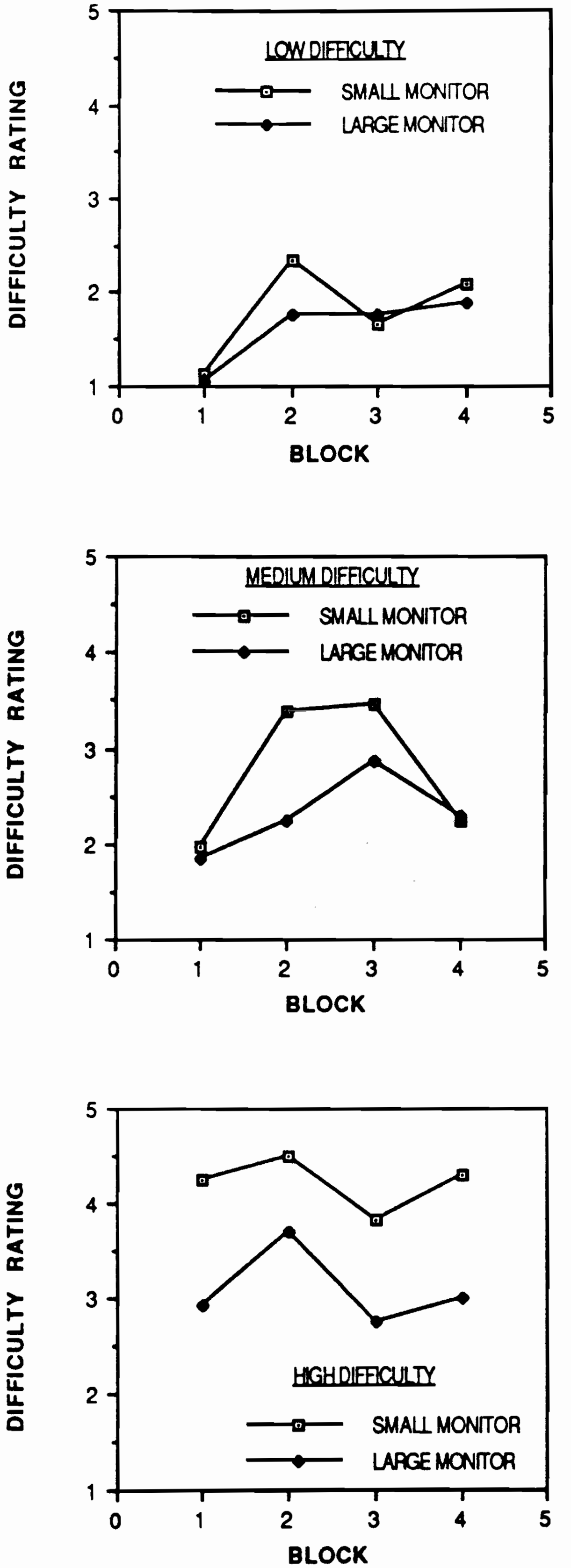


Figure E2. Monitor X Block interaction at each Difficulty Level for task difficulty ratings.

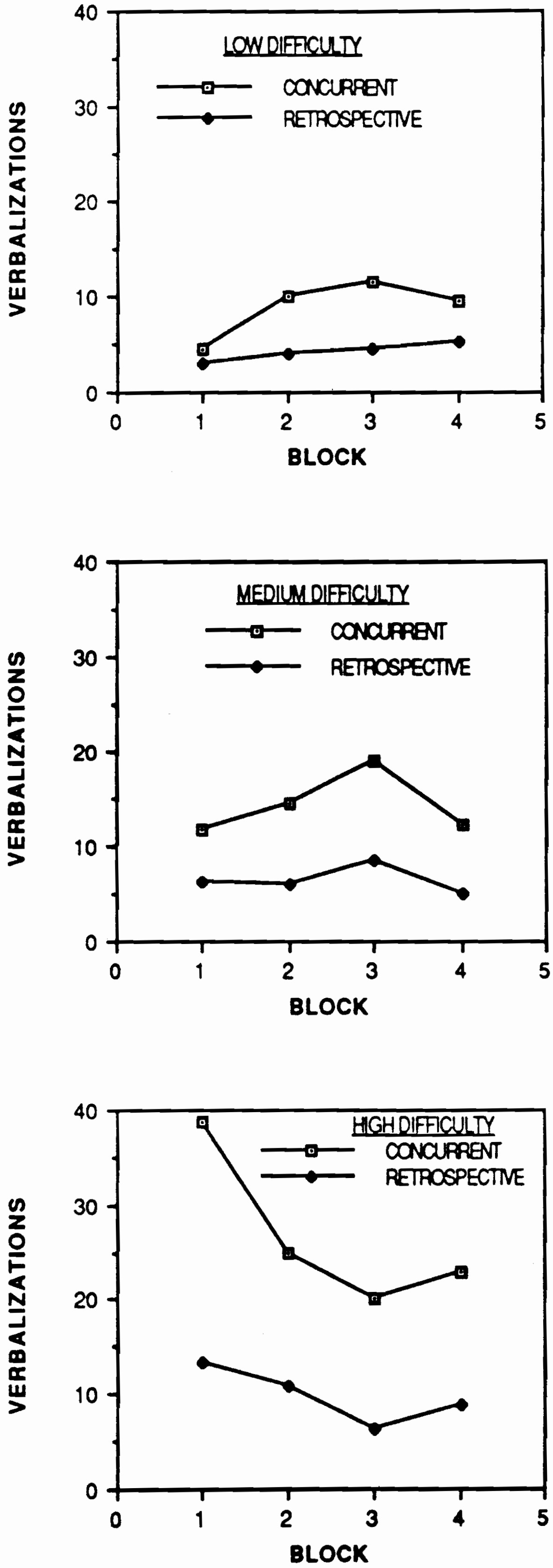


Figure E3. Protocol X Block interaction for each level of Difficulty.

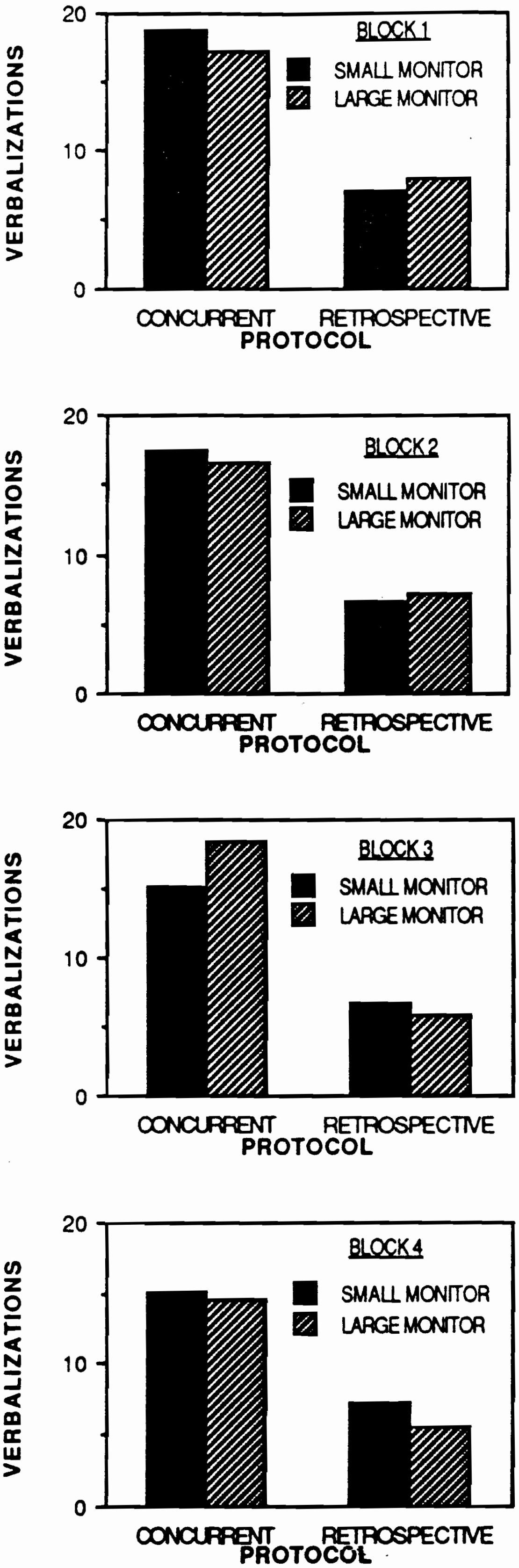


Figure E4. Monitor X Protocol interaction for Block.

Vita

Victoria Ann Bowers

January, 1990

Vicki Bowers was born on November 21, 1960 in Saint Louis, Missouri. She attended Wake Forest University from 1979 to 1983, from which she received her B.A. in psychology. In 1986 she received her M.S. in psychology from State University of New York, College at Cortland. She entered the Displays and Controls Laboratory at Virginia Tech in 1986 under the direction of Dr. Harry L. Snyder. Her dissertation concerns the use of verbal protocol for comparing window usability. She is a member of Alpha Pi Mu, the Human Factors Society, and the Association of Computing Machinery. She received her Ph.D. in Industrial Engineering and Operations Research, Human Factors Engineering Option, in January, 1990.

Victoria Ann Bowers