EXAMINING THE RELATIONSHIP BETWEEN PERFORMANCE MEASURES AND USER EVALUATIONS IN A TRANSFER OF TRAINING PARADIGM

by

William D. Coleman

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Industrial Engineering and Operations Research

APPROVED:

Robert C. Williges, Chairman

Dennis R. Wixon

Beverly H. Williges

August 1985
Blacksburg, Va.

ACKNOWLEDGEMENTS

Support for this research was provide by a MCC/CFM funded Digital Equipment Corporation Human Factors Internship Program. The program was administered by the Digital Equipment Corporation Software Human Engineering Group under the technical direction of Dr. John Whiteside.

Several individuals deserve gratitude for assistance rendered during this research. Dr. Robert Williges provided insights into the utility of my research and gave generous guidance which motivated me when thesis completion seemed an unattainable goal.
and Dr. Dennis Wixon provided timely advise during data collection, analysis, and interpretation. Their experience and expertise proved invaluable.

I would like to thank for both technical and emotional support. Finally I would like to thank my fiancee , whose devotion made even difficult times go by quickly.

TABLE OF CONTENTS

Page

Appendix Page

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Evaluation is a critical component of software interface design. Evaluative measures can be collected throughout the design cycle. Williges, Williges and Elkerton (in press) suggest the application of both formative and summative evaluation in the design of software interfaces. Formative evaluation is iterative, with successive designs assessed and revised based on a series of design subgoals. Summative evaluation focuses on testing the finished product.

Frequently, only user performance (objective) measures are considered during the evaluation process. However, of equal importance to interface evaluation are user attitude (subjective) measures. In the evaluation of software interfaces, researchers have several choices for both objective and subjective measures. Error rate, number of tasks completed, and task completion time are typical of objective measures collected. Researchers may compute a single objective measure, such as a user performance score (Whiteside, Jones, Levy, and Wixon, 1985), from a combination of other measures. Subjective evaluations usually consist of verbal reports, scale rating, or preference rankings. When both objective and subjective measures are collected, disagreement between the sets of measures is often found. Williges et al.

(in press) suggest that the lack of agreement between measures may be caused by a difference between satisfaction and usability or some insensitivity of the subjective measures.

The usefulness of subjective evaluations is determined, in part, by the specificity of the information gained. To identify problematic interface components user evaluations must generate detailed information. Formal instruments for collecting evaluations are typically global in nature, requiring the user to evaluate the entire interface on a single subjective measure, such as satisfaction or preference. This type of evaluation fails to give the designer information necessary to isolate aspects of the interface which should be targeted for improvement (Ives, Olson, and Baroudi, 1983).

Research Problem

The aim of this research was two-fold. First, a preliminary research study developed a methodology for the systematic collection of detailed subjective evaluations of software interfaces. Second, the main thesis topic was to explore the relationship between these detailed subjective evaluations and traditional objective measures in the context of a text editing benchmark task. Given these two research intentions, background literature concerning the evaluation of software interfaces using benchmark tasks and the development of subjective evaluation matrices for software interfaces is presented.

Benchmark Tasks to Evaluate Software Interfaces

Several researchers have developed and applied procedures, referred to as "benchmark tasks," for the purpose of objectively evaluating software interfaces. Benchmarking was originally applied to evaluate the usability of software products by Roberts (1979). The major goal behind benchmarking procedures is to provide standardized techniques which allow for comparisons to be made across a wide variety of products of a single class (e.g. editors).

In the development of text editor benchmark procedures, Roberts (1979) focused on characteristics common to all editors. Three criteria guided the evolution of these benchmark techniques; 1) Objectivity; the method should not be biased against any editor, 2) Thoroughness; the method should consider several dimensions of editor usage, and 3) Ease of use; the application of the benchmark procedures should be easy and require no special skills. The four dimensions of editor usage assessed were; 1) expert performance time, 2) expert errors, 3) novice learning time, and 4) functionality. In order to evaluate each editor on these four dimensions, Roberts created a taxonomy of 212 editing tasks. Expert performance time was computed as the total time taken by experts to accomplish 53 editing tasks. These tasks were selected from a set of core tasks considered executable on all editors. To assess novice learning time, 23 core tasks were used. Finally, functionality was measured by determining how many of the 212 tasks in the task taxonomy could be performed with the editor. Employing these benchmark procedures, Roberts (1979) and Roberts and Moran (1983) compiled a comparative data base of objective measures for nine text editors on the four dimensions discussed.

Expanding on techniques used by Roberts (1979), Whitside, Jones, Levy, and Wixon (1985), developed three sets of computer benchmark tasks. These task were; 1) file manipulation; adapted from Magers (1983), 2) electronic mail, and 3) general office procedures. Employing these standardized techniques, Whiteside _et al_., (1985), were able to compare successfully the usability of three vastly different software interface styles; command, menu, and iconic.

Several researchers have emphasized the need for benchmark tasks which more closely represent the actual daily use of interfaces (Whiteside, Archer, Wixon, and Good, 1982; Williges _et al_., in press). Gaylin (1985), successfully developed an empirically based benchmark task for the evaluation of windowing interfaces. In the process of building this benchmark, Gaylin (1985) measured command usage frequencies both objectively and subjectively. Command usage frequency was measured objectively by observing the daily activities of windowing system users. Subjective estimates of command usage were obtained from these same users by questionnaire. Analysis of the objectively and subjectively determined frequencies revealed that individual users were quite poor at judging command usage frequencies (r = 0.27). To determine if the empirically developed benchmark represented actual command usage frequencies, Gaylin (1985), observed two experienced users as they completed the benchmark. The frequency with which they used windowing commands during the benchmark was highly correlated with the frequency with which the original users used the commands in their daily work (r = 0.95).

Subjective Evaluations of Software Interfaces

Several researchers (Dzida, Herda, and Itzfeldt, 1978; Ives, *et al*. 1983; Nickerson, 1981) report that subjective evaluations of an interface reflect how well that interface fulfills user needs. Dzida, *et al*. (1978) further assert that user evaluations are critical to the development of user-oriented design guidelines. However, the literature on systematic collection of user subjective evaluations of software interfaces is sparse.

The application of subjective evaluations to the assessment of ease of use has been suggested by several researchers (Bennett, 1984; Shackel, 1984) In an effort to quantify subjective ease of use, Cordes (1984), applied techniques from the area of magnitude estimation. By using a standardized modulus to calibrate ease of use evaluations, Cordes was able to compare ratings across two different interfaces. These evaluations led to the successful redesign of hard to use interface components.

Rosson (1984) examined the effects of three user characteristics on ease of use ratings given a computer text editor. These user characteristics were; 1) experience with the editor evaluated, 2) experience with other text editors, and 3) type of work performed. Analysis of the ease of use ratings revealed that more sophisticated users (e.g. researchers and programmers) found the editor more difficult to use than less sophisticated users (e.g. secretaries). Further analysis of user comments indicated that this effect was due primarily to the increased application of the more complex editor characteristics (e.g. macros) by researchers and programmers.

In addition, Rosson (1984) solicited three types of written comments. These comments were entitled; 1) likes, 2) dislikes and 3) suggested improvements. The analysis of the comments revealed that the number of suggested improvements was found to be positively correlated with the diversity of user experience with other editors. The researcher cites two factors which may have caused this effect. First, the more diverse a user's experience the more comprehensive the baseline to which they can compare any individual editor. Second, more diversity may make the user a better judge of the feasibility of a suggested improvement.

Root and Draper (1983) examined several issues related to the effectiveness of questionnaires as a software evaluation tool. Their questionnaire focused on obtaining information from users pertaining to various commands included in a specific text editor. However, since actual command names were an integral part of their questionnaire, their methodology is restricted to within-editor comparisons. Nevertheless, the questionnaire identified several commands with which users had difficulty.

Root _et al_. (1983) further investigated three techniques for questionnaire administration. These techniques differed in terms of the recency of user experiences with the interface. The recency variable was manipulated between three discrete levels, "hot", "cold", and "ultra-cold." The "hot" condition required subjects to complete an online version of a standard editing task prior to making interface evaluations. In the "cold" condition, subjects completed the standard task on paper before their evaluations. Finally, in the "ultra-cold" treatment, subjects simply evaluated the editor with no recent experience. The results of these comparisons revealed that sensitivity differences, in terms of the rating variances, existed between the groups. However, the relative rankings of the various commands did not differ among administration techniques.

In another attempt to assess the effect of recency of experience on subjective evaluations, Rushinek, Rushinek, and Stutz (1984), employed an online evaluation program. The program was automatically invoked at the end of a student's CAI lesson, and both scale ratings and free form comments were collected. These researchers claim that their online evaluations were superior to manual paper/pencil evaluations. In addition, Rushinek, et al. (1984) assert that the evaluations resulted in redesign of the CAI lessons which, in turn, improved system effectiveness. However, it is unclear how these researchers came to their conclusions since none of the necessary data appear in their report.

User satisfaction with management information systems (MIS's) has been studied quite comprehensively by

Ives, Olson, and Baroudi (1983). The major premise upon which these individuals based their work was that global estimates of user satisfaction were ineffective as an evaluation tool. These researchers made this statement based on the fact that global evaluations failed to generate information identifying the specific source of user satisfaction or dissatisfaction. To obtain detailed evaluations, Ives *et al*. (1983) asked users to evaluate their management information system in terms of 39 different factors on a set of semantic differential scales. These factors included items addressing documentation and support such as "training provided users" and "vendor support" as well as items concerning system output such as "accuracy" or "precision of output". These researchers were able to describe effectively overall user satisfaction with the information systems they evaluated. The major appeal of Ives *et al*. (1983) approach lies in the ability to build an evaluative data base which spans a diverse group of information systems.

Development of a Subjective Evaluation Metric

Preliminary research was conducted to develop a methodology which could be used to build instruments for collecting detailed subjective evaluations. Basically, the methodology determined which aspects of a text editor users would evaluate and then defined how these evaluations would be made. First, a taxonomy of editing functions was created for users to evaluate. Second, procedures from the field of attitude measurement were used to build a set of 7-point bipolar adjective scales on which users could evaluate the editing functions. Third, the scales and editing functions were used to evaluate an existing editor.

Function taxonomy. The editing function taxonomy describes a text editor in terms of the common editing functions performed by users. For example, the DELETE function consists of anything implemented on a specific editor which could be employed by a user for the purpose of deleting text from a file. Several researchers have developed formal taxonomies to facilitate both the analysis of and the communication about human computer interfaces (Cohill, 1984; Lenorovitz, Phillips, Andrey, and Kloster, 1984).

Lenorovitz *et al*. (1984) assembled a generic taxonomy of logically associated terms to describe the user-system interface (USI) for task analytic purposes. The taxonomy, referred to as USI Action Taxonomy, is a combination of four sub-taxonomies. These sub-taxonomies are the; 1) Computer-Internal Taxonomy, 2) Computer-Output Taxonomy, 3) Human-Internal Taxonomy, and 4) Human-Input Taxonomy. The Computer-Internal Taxonomy consists of the automated aspects of the USI which are invisible to the user, while the Computer-Output Taxonomy describes the methods by which the computer attempts to communicate with the user. Of major interest are the human related sub-taxonomies. The Human-Internal Taxonomy defines actions internal to the user which are transparent to the computer. These include such actions as perception, information processing, and decision making. As the researchers point out, this sub-taxonomy expands on the earlier work of Berliner, Angell and Shearer (1964). Finally, the Human-Input Taxonomy describes the generic methods by which users accomplish their goals. Perhaps the largest contribution of Lenorovitz et al. lies in the fact that the researchers developed definitions of their terms. As they point out, the addition of these definitions should facilitate communication between designers and end users about the USI.

Cohill (1984) built a taxonomy to represent the functional aspects of interfaces dealing with computer data bases. The main purpose behind the development of this taxonomy was to list and define the functional needs of a wide range of data base interfaces. Cohill proposed that this taxonomy be used to develop a single, comprehensive data base interface. Cohill further identified several other uses of this taxonomy, perhaps the most important being the ability to compare across interfaces of a single class (e.g. editors, operating systems). These comparisons could be made on each individual function in terms of user performance measures, user evaluations, and ease of implementation.

In contrast to the taxonomies of Cohill (1984) and Lenorovitz *et al*. (1984), the taxonomy described in Coleman *et al*. (1984), emphasized the use of empirical procedures. In the initial compilation of the taxonomy, research dealing with the description of text editor environments was surveyed (Meyrowitz and Van Dam, 1982; Roberts and Moran, 1983). Suggestions for the function taxonomy were also collected from text editor users and designers. The responses received were incorporated into the initial list, which was assessed for comprehensiveness by members of the Human Engineering Research Group at Digital Equipment Corporation. As displayed in Table 1, the final list contained 16 editing functions. If one compares the taxonomy developed in the preliminary work with the taxonomies of Lenorovitz *et al*. (1984) and Cohill (1984), similarities become evident. The taxonomy which resulted from this pilot work appears as a mixture of the other two taxonomies, in that it describes the functional aspects (i.e. Cohill) of a text editor in relationship to tasks commonly performed by users (i.e Lenorovitz).

TABLE 1

Final List of 16 Editing Functions Chosen for Collecting Subjective Evaluations

FUNCTION	DEFINITION
TRAVEL	used to change the position of the cursor in a file; includes fine, coarse, forward, and backward movement
SEARCH	used to locate a specific target such as a string of characters
VIEW	used to examine visually what a file contains
DELETE	used to remove portions of a file
INSERT	used to put new information, text, into a file; does not include transferring text into the file from an external file
COPY	used to duplicate text at another location within the file; includes only duplications within a file
MOVE	used to relocate text within a file
REPLACE	used to substitute one piece of text for another; combines delete and insert into a single function
CUSTOMIZE	used to modify the interface environment; includes creating special commands
REQUEST	used to get help from the system in performing any task
RECOVER	used to recover from any user mistakes; includes ability to cancel or undo an operation and to analyze and interpret errors

TABLE 1

Continued

FUNCTION	DEFINITION
INITIATE	used to start an interaction with the interface
TERMINATE	used to end an interaction with an interface
WRITE	used to copy information between files; includes moving information from file being edited to an external file
INCLUDE	used to copy information from an external file into the file being edited
FORMAT	used to format information within a file while in that file

Bipolar adjective scales. The use of bipolar adjective scales for the measurement of subjective reactions was first proposed by Osgood, Tannenbaum, and Succi (1957). These researchers employed bipolar adjective scales as a means of semantically differentiating objects. Application of bipolar adjective scales to a multitude of different concepts has consistently revealed that the variance in the ratings on these scales can be accounted for by three dimensions, EVALUATION, POTENCY, and ACTIVITY (EPA). Each dimension consists of bipolar adjective scales whose ratings co-vary. For example the scales GOOD-BAD and SAFE-DANGEROUS are included in the EVALUATION dimension; the scales POWERFUL-POWERLESS and HEAVY-LIGHT fall on the POTENCY dimension; and the scales ACTIVE-PASSIVE and FAST-SLOW fall on the ACTIVITY dimension.

These bipolar adjective scales have been applied to the assessment of general affective reactions to computers by Zoltan and Chapanis, (1982). Lucas (1977) and Kerber (1983) have used bipolar adjective scales to compare affective reactions to specific applications of computers, e.g. *bookkeeping*, decision making, medical *interviewing*. More specific applications have included the evaluation of Management Information Systems (Ives *et al*., 1983), keyboard designs, (Burke, Muto, and Gutman, 1984) and operating systems, (Whiteside, Wixon, and

Jones, in press).

The development of the bipolar adjective scales used in this research, as with the function list, emphasized the use of empirical procedures. An initial list of 64 adjectives was assembled by examining articles dealing with user evaluations (Kerber, 1983; and Ives et al., 1983). Computer users were then queried for recommendations. Users generated 22 new adjectives; thus the comprehensive list contained 86. The adjective list was then refined based on user-perceived similarity, as determined by sorting, and user-perceived importance, as indicated by rating scale. The result was a list of 17 adjectives. Finally, these adjectives were paired with antonyms (Bolander, Varner, Pine 1981) to form bipolar adjective rating scales. The list of 17 adjectives, their importance ratings, and their antonyms are displayed in Table 2.

As shown in Table 2, the four adjectives rated as most important for describing a satisfactory text editor interface were DEPENDABLE, USEFUL, FAST, and CONSISTENT. This result indicates that users may prefer more concrete adjectives for evaluating interfaces. In addition, these adjectives seem to describe measurable and adjustable interface parameters, which is consistent with the assertion that user evaluations of an interface reflect how well the interface fulfills user needs (Dzida et al.,

TABLE 2

Seventeen Adjectives, Their Mean Importance to the Description of a Satisfactory Interface (as rated by text editor users, maximum value = 7), and Antonyms Selected for Bipolar Scales

ADJECTIVE	IMPORTANCE	ANTONYM
✓ DEPENDABLE	6.8889	UNDEPENDABLE
✓ USEFUL	6.7407	USELESS
✓ FAST	6.6296	SLOW
✓ CONSISTENT	6.5926	INCONSISTENT
✓ COMPLETE	6.0000	INCOMPLETE
MAINTAINABLE	5.9259	UNMAINTAINABLE
ADAPTIVE	5.8889	UNADAPTIVE
✓ FRIENDLY	5.8519	UNFRIENDLY
INTERPRETABLE	5.7778	UNINTERPRETABLE
✓ SIMPLE	5.4444	COMPLICATED
INTELLIGENT	5.3704	UNINTELLIGENT
CONCISE	5.1111	REDUNDANT
UNCLUTTERED	5.1111	CLUTTERED
COOPERATIVE	4.9259	UNCOOPERATIVE
SAFE	4.9259	UNSAFE
✓ NATURAL	4.5741	UNNATURAL
✓ PLEASING	4.2593	IRRITATING

1978).

It was felt that the use of user-perceived similarity and importance ratings would increase the face validity of the instrument for users. These procedures are quite different from the selection technique used in the creation of a semantic differential. Semantic differentials are traditionally created by selecting several scales from each of the EPA dimensions.

Preliminary evaluations of the subjective metric. To study the instrument developed, 27 users were asked to complete three questionnaires on an existing editor. First, overall satisfaction with the editor was rated on a 7-point rating scale, 1 represented extremely unsatisfied, 4 neutral (no opinion), and 7 extremely satisfied. Second, users evaluated the editor in terms of the 16 editing functions on two 7-point rating scales anchored by the bipolar adjective pairs BAD/GOOD and UNIMPORTANT/IMPORTANT. Finally, users evaluated the editor in terms of the 16 editing functions on the 17 bipolar adjective rating scales developed previously. Specifically, these data were used to; 1) examine the relationship between global and detailed evaluations of an interface, 2) assess the usefulness of the instrument developed, and 3) examine the appropriateness of the bipolar rating scales for the evaluation of the 16 editing functions.

Analysis of user responses revealed no significant correlation existed between the average goodness ratings (detailed ratings) assigned the 16 editing functions and overall user satisfaction (global ratings) with the editor ($p > 0.05$). However, during the post experiment interviews, the majority of users felt that the functions were related to how satisfied they were with the editor. Three possible explanations are offered for the lack of statistical relationship between detailed and global evaluations. First, actually seeing the function name may have caused users to remember their experience differently, perhaps better than when asked to evaluate the editor as a whole. This could happen if the function provided users with a retrieval cue for past experiences with the editor. Second, it is also possible that the function list was not exhaustive enough for general users. Finally, a third possibility is that rather restricted variance of the overall satisfaction scale may have attenuated the between-measure correlation. In general, the results indicate a difference exists between global and detailed evaluations. However, identifying the reason for this difference will require further research.

One indication of whether an editing function is problematic may be its mean rank relative to other editing functions. To explore this, a mean rank across all 17 bipolar scales was determined for each editing function. As illustrated in Table 3, the mean ranks reveal that the functions for the evaluated editor seem to fall into two groups, functions 1-7 and functions 8-14. This result may suggest that effort should be devoted to the improvement of this second group of functions.

Further analysis of the bipolar adjective scale ratings revealed two interesting results. First, the relative rankings of the 16 editing functions based on the goodness scale ratings were significantly correlated with the relative rankings based on a mean of the 17 bipolar adjective scale ratings (Table 3). This relationship also was supported by the fact that all individual bipolar adjective scales correlated significantly ($p < 0.05$) with the function goodness ratings, as shown in Table 4. These two findings seem to support the procedure by which the scales were selected. Second, the 17 scales were used for extreme ratings with the same relative frequency on all functions. Assuming extreme (non-neutral) ratings indicate scale appropriateness for rating a function, this result suggests that all functions could be evaluated with the same set of scales.

In summary, the development and application of the

TABLE 3

Mean Rank of 16 Editing Functions Across 17 Bipolar Adjective Scales

FUNCTION	MEAN RANK
MOVE	3.059
DELETE	3.529
TRAVEL	3.706
SEARCH	3.882
INITIATE	3.882
TERMINATE	4.412
INSERT	5.118
RECOVER	7.000
INCLUDE	7.235
FORMAT	8.294
WRITE	8.765
VIEW	9.647
CUSTOMIZE	9.824
REQUEST	10.330
COPY	10.350
REPLACE	10.820

TABLE 4

Rank Order of 17 Bipolar Adjective Scales in Terms of Highest Correlation to GOOD/BAD Scale (n=336)

SCALE	r
PLEASING................IRRITATING	0.455
FRIENDLY................UNFRIENDLY	0.400
COMPLETE................INCOMPLETE	0.383
COOPERATIVE..........UNCOOPERATIVE	0.376
DEPENDABLE............UNDEPENDABLE	0.375
SIMPLE.................COMPLICATED	0.353
CONSISTENT............INCONSISTENT	0.352
NATURAL..................UNNATURAL	0.351
INTELLIGENT..........UNINTELLIGENT	0.329
INTERPRETABLE......UNINTERPRETABLE	0.326
FAST..........................SLOW	0.292
ADAPTIVE................UNADAPTIVE	0.255
USEFUL.....................USELESS	0.239
CONCISE..................REDUNDANT	0.211
UNCLUTTERED..............CLUTTERED	0.201
SAFE........................UNSAFE	0.194
MAINTAINABLE........UNMAINTAINABLE	0.153

methodology justify four conclusions.

1) User suggestions can be successfully incorporated into the development of an instrument designed for collecting their evaluations of software interfaces.

2) Users prefer concrete adjective scales for evaluating interfaces (e.g. DEPENDABLE, USEFUL, FAST).

3) An instrument which describes an interface in terms of the tasks users perform encourages detailed evaluations and seems to identify aspects of the interface which users view as problematic.

4) The same set of 17 bipolar adjective rating scales is appropriate for evaluating all 16 editing functions.

Purpose of Thesis Research

The major intention of the thesis was to refine and evaluate a methodology for the collection of detailed subjective evaluations. A secondary aim was to examine the relationship between objective (performance) measures and subjective measures. These goals were addressed in a transfer of training paradigm with inexperienced computer users being taught to use two text editors. A transfer of training paradigm was used to allow subjects to use both editors evaluated. Several research hypotheses were evaluated;

1) The relationship between subjective and objective measures would vary across bipolar scales.

2) Learning time would be shorter on the second editor learned.

3) Transferring between editors would result in negative transfer, on both objective and subjective measures, for both editors.

4) The effects of transferring between editors would be larger on the subjective measures than on the objective measures.

5) Subjective re-evaluations of a users original editor, after exposure to both editors, would be significantly different from prior evaluations.

6) Detailed subjective evaluations would be more sensitive to differences between editors than global evaluations.

METHOD

Subjects

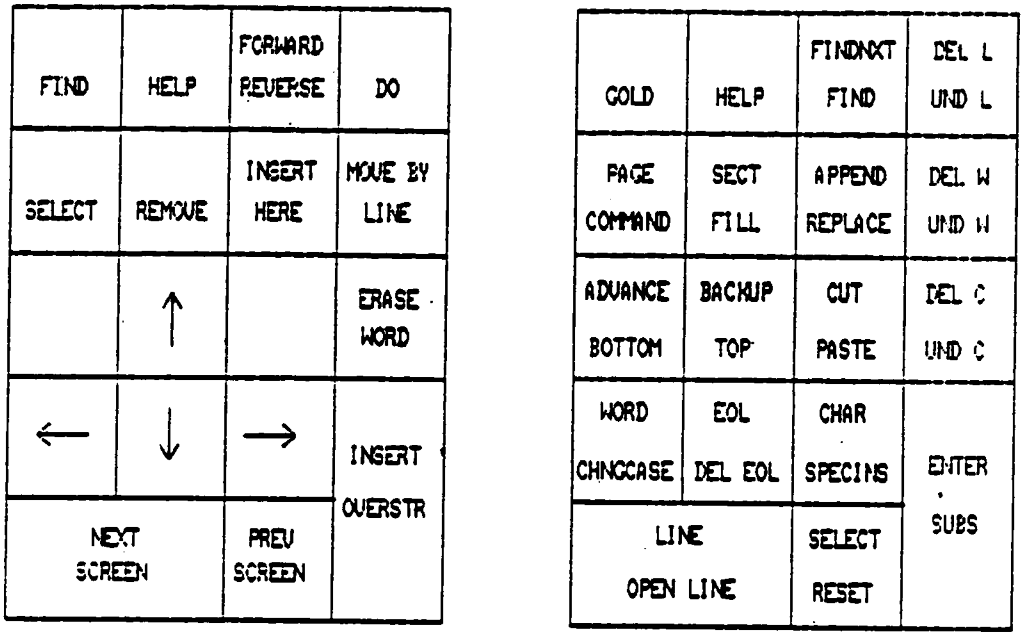
Sixteen students at Virginia Polytechnic and State University, received payment for their participation in this experiment. Selection of subjects was restricted by the requirement that a subject have no previous experience with interactive computers. This restriction was implemented to control for previous, as well as intervening, usage of similar computer software products. The experimenter randomly divided these subjects into two groups. A subject's group determined the direction of transition between two editors. Prior to participation subjects were required to sign an informed consent form (Appendix A).

Editors

This experiment evaluated two full screen text editors. The selection of these editors was motivated by two factors. First, both editors were implemented on the same system, which assured roughly equivalent system response times and allowed for the use of a single CRT terminal type. Second, the editors were diverse enough in their implementation to allow for meaningful comparisons. To access the editing commands of both editors subjects used a combination of the keypad and keyboard keys on a computer terminal. Prior to each session the experimenter provided subjects with the appropriate keypad template indicating the location of the keypad editing commands. These keypad templates, as shown in Figure 1, give some indication of the relative complexity of the two editors. The design philosophy behind EDITOR A emphasized simplicity, with only the most frequently used commands included on the keypad. Less frequently used commands were accessed through the use of a command line. EDITOR B was based on an alternate philosophy, with the majority of commands implemented on the keypad. Each keypad key contained two commands differentiated by a mode key. In addition to the keypad template each subject was given a summary sheet containing a short description of the various basic editing tasks. The summary sheets for the two editors are contained in Appendix B.

Experimental Design

The design required that each subject complete four experimental sessions, one on each of four consecutive days (see Table 5). The procedures followed on days 1 and 3 were identical. The basic procedures employed on days 2 and 4 were also identical, with the exception of two additional requirements on day 4.



EDITOR A

EDITOR B

FIGURE 1

Keypad Templates for the Two Editors Evaluated

TABLE 5

General Experimental Procedures

DAY 1

EDITOR	ACTIVITIES
1st	Complete Roberts' editor training procedures
1st	Complete objective and subjective evaluations

DAY 2

EDITOR	ACTIVITIES
1st	Complete practice session
1st	Complete objective and subjective evaluations

DAY 3

EDITOR	ACTIVITIES
2nd	Complete Roberts' editor training procedures
2nd	Complete objective and subjective evaluations

DAY 4

EDITOR	ACTIVITIES
2nd	Complete practice session
2nd	Complete objective and subjective evaluations
1st	Complete subjective evaluations
1st&2nd	Complete comparative subjective evaluations

Training

Subjects received training on the two editors in the order determined by their assigned experimental group. To accomplish this training, the experimenter followed the general guidelines developed by Roberts (1979). These guidelines outline procedures and tasks employed to train computer novices on the use of any text editor. Roberts created these standardized training procedures for the purpose of evaluating an editor in terms of the novice learning time for a set of benchmark editing tasks.

The experimenter taught each subject, according to Roberts guidelines, how to accomplish a standard set of 31 core editing tasks. Completion of the core tasks dictated that a subject acquire the minimum skills necessary to work with a given computer text editor. The experimenter trained subjects employing a fixed sequence of five cycles for task presentation. Each cycle required that the experimenter train and quiz the subject on a small subset of the core editing tasks. Prior to each quiz a subject was allowed to practice. When the subject felt they had practiced a sufficient amount they notified the experimenter. The experimenter then administered a short quiz consisting of tasks which may or may not have been taught previously. Table 6 contains the core editing tasks, the cycle in which they were taught, and the cycle in which they were tested. As evident from Table 6, subjects were actually tested on only 23 of the 31 core editing tasks in accordance with Roberts (1979) guidelines.

During testing, the experimenter deviated from Roberts' procedures by bringing any uncompleted tasks to a subject's attention. This occurred at least once per subject and was done in order to ascertain whether the subject was incapable of completing or simply overlooked the task. While this reduced the variance in the dependent measure of tasks completed it also allowed the experimenter to determine and eliminate any difficulties each subject experienced.

The data collected consisted of two dependent measures, total time and tasks completed. Total time, which included training, practice, and quiz time, was determined by stopwatch and recorded along with tasks completed on Roberts' standard coding sheet. Mean learning time per task was computed by dividing total time by the total number of tasks completed. The training methodology is described here only in brief, Roberts (1979) should be consulted for a more complete description.

TABLE 6

Training and Testing Order for Editing Tasks According to Roberts (1979) Specifications

TASK	CYCLE(S) TRAINED	CYCLE(S) TESTED
Get document	1, 2, 3, 4, 5	1, 2, 3, 4, 5
Save document	1, 2, 3, 4, 5	1, 2, 3, 4, 5
Insert line	1, 2	1, 2, 5
Insert paragraph	1, 2	1, 2, 3
Insert character	3	3
Insert sentence	4	
Insert word	3	1, 4
Delete paragraph	2	2, 4
Delete character	3	3, 5
Delete word	3	2
Delete sentence	4	1
Delete line	2	
Replace line	2	2
Replace character	3	1, 2
Replace word	3	3, 5
Replace sentence	4	3, 4
Split word	3	3, 5
Split sentence	4	4
Split paragraph	4	4, 5
Split line	4	
Merge word	3	2, 3, 5
Merge sentence	4	
Merge paragraph	4	1, 2, 3, 4
Merge line	4	
Move sentence	4	5
Move line	4	2
Move paragraph	4	3, 4
Copy sentence	4	
Copy line	4	
Copy paragraph	4	
Find string	5	4, 5

Practice

All subjects completed one practice session with each of the two editors. The practice session with an editor occurred on the day following training. Practice required subjects to complete seven sets of editing tasks, each set located in a separate computer file. The type and number of editing tasks in each computer file, as illustrated in Table 7, were chosen to be representative of the tasks which subjects had been taught previously during training. However, this practice was an extension beyond Roberts (1979) methodology since Roberts procedures only evaluate initial learning time for novices. The experimenter created two identical sets of editing tasks, on files of varying content, in order to present subjects with different files for each of the two practice sessions. At the start of a practice session the experimenter provided a subject with a paper copy of each electronic file with all the necessary editing changes, marked and highlighted.

The nature of the practice sessions required that subjects complete all editing tasks. The experimenter brought any omitted tasks to the subject's attention, and assisted in their completion if necessary. Since

TABLE 7

Type and Number of Editing Tasks in Seven Computer Files Used for Practice Sessions

TASK		COMPUTER FILE						
		1	2	3	4	5	6	7
Get	document	1	1	1	1	1	1	1
Save	document	1	1	1	1	1	1	1
Insert	character	1	2	1	1	2	1	2
Insert	word	2	1	2	1	3	1	3
Insert	line					1	1	1
Insert	paragraph			1	1	1		
Delete	character		1		2	1	1	
Delete	word	2				2		
Delete	sentence	2		1	1	2	1	2
Delete	paragraph					1		1
Replace	character	1	1		1		2	
Replace	word	1	1	2	1	1	2	1
Replace	sentence		1		1		1	
Split	word		2	1	1	1	2	1
Split	sentence	1	1	2	1	1	1	1
Merge	word			1	1		3	1
Merge	paragraph			1	1			
Move	sentence	1	1	1	1	1	1	1
Move	line	1	1		1			
Move	paragraph		1				1	
Find	string						3	4

subjects finished all tasks, the only dependent measure collected was total time to complete the seven sets of editing tasks.

Evaluations of Editors

Every subject participated in both an objective and a subjective editor evaluation as part of each experimental session. This resulted in two of each type of evaluation per editor for every subject. The objective evaluation, also referred to as a performance test, required a subject to attempt several sets of editing tasks, each set requiring the use of a different editing function. The subjective evaluation, dictated that a participant complete a set of questionnaires. The objective and subjective evaluations were designed to allow for detailed, systematic comparisons to be made both between editors and between objective and subjective measures. Again, these evaluations were an extension beyond the procedures specified by Roberts (1979) for editor evaluations.

Objective evaluations. The objective evaluations specified that the subject attempt tasks which required the use of nine editing functions. These nine functions, shown in Table 8, are a subset of an editing function taxonomy empirically developed in a previous study, (Coleman, Wixon, and Williges, 1984). The adaptation of

TABLE 8

Nine Editing Functions Used to Partition Both Objective and Subjective Evaluations of Editors

FUNCTION	DEFINITION
TRAVEL	used to change the position of the cursor in the file
SEARCH	used to find a specified target such as a string of characters
DELETE	used to delete text from the file
INSERT	used to insert new text into a file
MOVE	used to move a section of text to another location within a file
REPLACE	used to replace one piece of text with another
WRITE	used to save a file
INCLUDE	used to get a file into the editor
FORMAT	used to adjust text within the file

this taxonomy to the present research is discussed in Appendix C. The experimenter randomized the order in which subjects encountered the nine editing functions. In general, a subject attempted all the tasks requiring a particular editing function prior to attempting tasks requiring the next function. However, there were four exceptions. The performance of each set of tasks required traveling from one target to another as well as including and saving an electronic file. Therefore, subjects performed the TRAVEL, INCLUDE and WRITE functions with each task set. In addition, for the sake of realism, the experimenter instructed subjects to reformat each electronic file when finished with a set of tasks. This reformatting required the use of the FORMAT function. Therefore, while the FORMAT and TRAVEL functions were evaluated separately from all other functions, they were also imbedded in each function task set. The WRITE and INCLUDE task sets consisted of seven subtasks, each performed on a separate computer file. Table 9 contains a list of all editing function task sets used during objective editor evaluations.

While participating in the objective evaluations, subjects were instructed to avoid asking the experimenter for assistance. Therefore, if subjects were unable to complete any given task they simply went to the next task. The experimenter videotaped the subject's

TABLE 9

Specific Editing Tasks Performed by Users During Objective Evaluations of Editors by Editing Function

EDITING FUNCTION	EXAMPLE TASKS
TRAVEL	1. go to 4th paragraph (down 32 lines) 2. go to end of file (down 72 lines) 3. go to end of 3rd paragraph (up 67 lines) 4. go to start of 1st paragraph (up 27 lines)
SEARCH	1. forward find, word 2. forward find, string 3. backward find, string (reverse direction) 4. forward find, word (reverse direction) 5. backward find, word (reverse direction)
DELETE	1. delete word 2. delete paragraph 3. delete character (merge word) 4. delete line of text 5. delete sentence 6. delete blank line (merge paragraph) 7. delete character
INSERT	1. insert paragraph 2. insert character 3. insert word 4. insert blank line (split paragraph) 5. insert sentence 6. insert character (split word) 7. insert line of text
MOVE	1. move paragraph 2. move sentence 3. move sentence 4. move line 5. move paragraph
REPLACE	1. replace paragraph 2. replace word 3. replace sentence 4. replace character 5. replace line

TABLE 9

Continued

EDITING FUNCTION	EXAMPLE TASKS
FORMAT	1. format paragraph
	2. format paragraph
	3. format paragraph
	4. format paragraph
WRITE	1. write file out of editor
	2. write file out of editor
	3. write file out of editor
	4. write file out of editor
	5. write file out of editor
	6. write file out of editor
	7. write file out of editor
INCLUDE	1. get file into editor
	2. get file into editor
	3. get file into editor
	4. get file into editor
	5. get file into editor
	6. get file into editor
	7. get file into editor

performance to allow for delayed coding and analysis. The videotaped image recorded was identical to what appeared on the subjects video terminal screen throughout the entire objective evaluation.

Subjective evaluations. The subjective evaluations were collected both concurrent with and following the objective evaluations. In the concurrent subjective evaluations, participants rated how they felt about the editor in terms of performing each of the nine editing functions. Each function was evaluated immediately after a participant completed the entire set of editing tasks representing that function. In general, it was hoped that these immediate evaluations would reflect the subjects' attitudes more accurately.

Subjects rated each function on 12 bipolar adjective scales (see Table 10). A description of the selection of these scales from the results of Coleman et al. (1984) is given in Appendix C. The direction of the adjective anchors was reversed on a subset of these scales in order minimize response bias. A participant rated an editing function on each seven point scale by circling the appropriate number which described how they felt about that editing function. A single scale, as shown in Figure 2, was anchored at opposite ends by the bipolar adjectives and by adverbs describing various gradients of the end point adjectives.

TABLE 10

Bipolar Adjective Scales Used to Collect User Subjective Evaluations of Editors

	ADJECTIVE SCALES
SCALE	NEGATIVE..............POSITIVE
1	USELESS..................USEFUL
2	UNDEPENDABLE.........DEPENDABLE
3	INCONSISTENT.........CONSISTENT
4	UNINTERPRETABLE...INTERPRETABLE
5	COMPLEX..................SIMPLE
6	UNSAFE.....................SAFE
7	SLOW.......................FAST
8	UNNATURAL...............NATURAL
9	INCOMPLETE.............COMPLETE
10	DISGUSTING.............PLEASING
11	UNCOOPERATIVE.......COOPERATIVE
12	UNSATISFACTORY.....SATISFACTORY

```
useless                                                                         useful
     1           2          3          4          5          6          7
:----------:----------:----------:----------:----------:----------:----------:
extremely    quite     slightly   neutral   slightly    quite     extremely
```

FIGURE 2

Bipolar Adjective Scale

Two additional sets of subjective evaluations were collected following the completion of an editor's objective evaluation. First, each subject ranked the nine editing functions from one to nine, with one represented the function that they liked to perform the most and nine the one they liked to perform the least. These rankings were included to allow for a comparison between absolute ratings of each function on the twelve bipolar scales and its relative ranking as compared to the other functions. Second, each subject evaluated the entire editor in terms of the 12 bipolar rating scales used to evaluate the editing functions during the concurrent subjective evaluations. These evaluations would be used to compare detailed and global evaluations. A complete set of questionnaires used in this study are presented in Appendix D.

RESULTS

The results section is divided into four segments. The first segment presents results from the analysis of both learning and practice data. Second, an overview of the procedures used in the compilation and computation of various objective performance measures of test data is presented. Third, the results of the test data are described in terms of both subjective and objective measures. Finally, the relationships among the various subjective measures are presented.

Training and Practice Performance

Analyses were conducted to discover any performance differences which existed between the two editors during learning and practice. Since the results of these analyses were not of primary interest to the comparisons between objective and subjective measures, they are discussed only briefly.

Learning data. A learning rate measure was calculated for each subject by procedures specified in Roberts (1979). Completion time was divided by tasks completed resulting in a mean learning time per task in minutes. An analysis of variance (ANOVA) on this learning rate measure revealed a significant main effect of EDITOR, $F(1,14) = 14.65$, $p < 0.0018$, favoring EDITOR A (EDITOR A = 2.77 minutes per task; EDITOR B = 3.21

minutes per task). However, as shown in Table 11, this effect is restricted primarily to the second editor encountered (Tukey, alpha=0.05). This was supported by separate comparisons of the first and second editor. These comparisons revealed that 75 percent of the variance of the EDITOR main effect was due to the second editor encountered. Further examination of Table 11 indicates that both subject groups improved significantly in their performance from first to second editor (Tukey, alpha=0.05). A complete ANOVA summary table for the learning data analysis is presented in Appendix F.

Practice data. The dependent measure collected during practice was simply the total time taken by subjects to complete seven sets of editing tasks. The results of an ANOVA on total time to completion indicated a significant main effect of EDITOR $F(1,14) = 15.19$, $p < 0.0016$, again in the direction of EDITOR A (EDITOR A = 55.75 minutes; EDITOR B = 63.313 minutes). As shown in Table 12, the effect of editor is limited primarily to the second editor encountered (Tukey, alpha=0.05). In addition, only subjects transitioning from EDITOR B to EDITOR A, experienced positive transfer effects (Tukey, alpha=0.05). No statistically significant difference in the performance of the subjects transitioning from EDITOR A to EDITOR B was measured. A complete ANOVA summary table describing this analysis is given in Appendix F.

TABLE 11

Differences Between Means Applicable to the EDITOR by ORDER Interaction for Learning Rate Measure in Minutes per Task

EDITOR DIFFERENCES, EDITOR A - EDITOR B

		MEANS	
EDITOR	DIFFERENCE	EDITOR A	EDITOR B
FIRST	0.321	3.487	3.808
SECOND	0.561*	2.054	2.615

TRANSITION DIFFERENCES, FIRST - SECOND EDITOR

ORDER	DIFFERENCE
A --> B (3.487) (2.615)	0.872*
B --> A (3.808) (2.054)	1.754*

* significant, alpha=0.05

TABLE 12

Differences Between Means Applicable to the EDITOR by ORDER Interaction for Total Practice Time in Minutes

EDITOR DIFFERENCES, EDITOR A - EDITOR B

		MEANS	
EDITOR	DIFFERENCE	EDITOR A	EDITOR B
FIRST	5.875	62.000	67.875
SECOND	9.250*	49.500	58.750

TRANSITION DIFFERENCES, FIRST - SECOND EDITOR

ORDER	DIFFERENCE
A --> B (62.000) (58.750)	3.250
B --> A (67.875) (49.500)	18.375*

* significant, alpha=0.05

Data Compilation and Reduction

The following subsections describe compilation and reduction procedures applied to the data collected during the daily objective evaluations of the two editors.

Videotape analysis. To analyze a subject's performance as recorded on videotape, the experimenter collected four dependent measures; total time on task, tasks completed, total errors committed, and total error correction time. The reliability of the videotape coding process was determined by recoding two videotapes. One tape was selected from each of two separate subjects, who participated at different times in the experiment. The original tape coding occurred immediately following the subjects' participation, while recoding occurred at the conclusion of the experiment. A reliability coefficient, simply the correlation between first and second coding, was computed on the four dependent measures. The reliability coefficients for the four dependent measures of total time on task, tasks completed, total errors committed, and total error correction time were 0.99, 1.00, 0.94, and 0.99, respectively. These high reliability estimates suggest that the tape coding process was quite consistent.

$$S = 1/T * PC$$

S = Performance Rate (User Performance Score)
T = Total Time on Task
P = Percent of Task Completed
C = Expert Completion Time

FIGURE 3

Performance Rate Measure Developed by Whiteside, Jones, and Levy (1984)

Determination of a performance rate. A single measure incorporating total time and tasks completed was computed to represent a subject's performance with each editing function task set. This objective measure, defined in Figure 3, was developed by Whiteside et al. (1984). Percent of tasks completed, time on task, and a reference time constant are combined to express performance as a function of the rate of task completion per unit time. The subject's time on task was substituted directly into the equation for the letter "T". The percent of tasks completed or "P", was derived empirically by determining the proportion of total time on task spent on each task component by an expert. These proportions were then totalled for each subject dependent on the task components they completed. The time constant "C" was defined by the fastest completion time for an expert for each task set. A more detailed description of the determination of the values for "P" and "C" is given in Appendix E.

Categorization of error data. Error data were classified by two methods. First, errors were categorized by the editing function task set in which they occurred. This parallels the classification of the dependent measures of time on task and tasks completed utilized in the computation of a subject's performance rates. Second, errors were classified by type, each type representing one of the nine particular editing functions. Errors classified by this scheme were not confined to any particular task set, i.e. an error committed while using the TRAVEL function during the MOVE function task set would be analyzed with the TRAVEL function data. Each error type was subjectively determined by the experimenter. The analysis of the error measures revealed that no differences existed between the two methods of error classification, in terms of the location or significance of any effects. Therefore, the first method was chosen for further comparisons involved in this research.

Comparisons Between Objective and Subjective Evaluations

The major results of the experiment, in terms of comparisons between objective and subjective evaluations, follow in two sections. The first section, analysis of overall effects, examines the relationship between objective and subjective measures for the entire experiment. The ensuing section, analysis of transfer effects, focuses on the consequences of transitioning between the two editors on both objective and subjective measures.

Analysis of overall effects. The overall analysis examined a set of 15 dependent measures, 3 objective and 12 subjective. The 3 objective measures were, performance rate, errors committed, and error correction time. The subjective measures were simply the rating scales shown previously in Table 10. Protection against the occurrence of Type I errors during analysis was maintained separately for the objective and subjective families of measures at an alpha level of 0.05. The inflation of alpha error was controlled across the three objective measures by requiring a significance level of 0.0167 for null hypothesis rejection on each measure. To safeguard the familywise alpha level of 0.05 for the subjective measures a single ANOVA was performed with rating scale considered a factor having 12 levels. To determine the locus of any significant ($p < 0.05$) interactions involving scale, twelve separate ANOVAs were conducted, each with a different rating scale as the dependent measure. The p-values for all significant effects indicated by the analysis of both subjective and objective measures are presented in Table 13. The means applicable to each of the illustrated p-values are discussed in the following text and tables as appropriate. The complete set of ANOVA summary tables for the analyses of overall effects are contained in Appendix F.

As illustrated in Table 13, both simple main effects and main effects described in terms of interactions existed in the data. As would be expected, all three

TABLE 13

P-values for Significant Differences Indicated by Objective and Subjective Measures

OBJECTIVE MEASURES	DAY	EDITOR	EDITING FUNCTION	EDxEF
RATE MEASURE	0.0001	0.0029	0.0001	0.0001
ERRORS COMMITTED	0.0001	NS	0.0001	0.0001
E. CORRECTION TIME	0.0001	NS	0.0001	0.0001

SUBJECTIVE MEASURES	DAY	EDITOR	EDITING FUNCTION	EDxEF
OVERALL ANALYSIS	NS	0.0341	0.0001	0.0145
INDIVIDUAL ANALYSES				
INTERPRETABLE		NS	NS	NS
USEFUL		NS	0.0001	NS
DEPENDABLE		NS	0.0009	NS
CONSISTENT		NS	0.0177	NS
SAFE		NS	0.0005	NS
NATURAL		NS	NS	0.0007
PLEASING		NS	0.0001	0.0001
SIMPLE		0.0196	0.0001	0.0042
FAST		0.0042	0.0001	0.0241
COMPLETE		0.0487	0.0025	0.0367
COOPERATIVE		0.0302	0.0176	0.0034
SATISFACTORY		0.0285	0.0076	0.0016

NS:Nonsignificant

objective measures indicated an effect of practice, in a main effect of DAY ($p < 0.0001$). Furthermore, these effects were equal regardless of editor or editing function as demonstrated by the lack of any significant interactions involving DAY. The magnitudes of the mean differences are shown in Table 14, with performance rate increasing from DAY 1 to DAY 2 and errors committed and error correction time decreasing. No practice effect was evident on the subjective measures, as illustrated by the absence of a significant DAY by SCALE interaction ($p > 0.05$).

Four of the subjective scales, USEFUL-USELESS, DEPENDABLE-UNDEPENDABLE, CONSISTENT-INCONSISTENT, and SAFE-UNSAFE, indicated only a main effect of EDITING FUNCTION. Paired comparisons (Tukey, alpha=0.05) were conducted on these scales. The analysis of the USEFUL scale revealed that regardless of editor, the SEARCH function was rated as significantly less useful than all other functions. The analysis of the DEPENDABLE scale indicated several differences. Both the FORMAT and WRITE functions were rated as significantly more dependable than the TRAVEL function. In addition, the FORMAT function was rated as significantly more dependable than the INSERT function. The paired comparison analysis of the CONSISTENT scale ratings indicated that for both editors, the TRAVEL function was considered to be less

TABLE 14

Means for Day Main Effect, Objective Measures

	RATE MEASURE	ERRORS COMMITTED	ERROR CORRECTION TIME
DAY 1	34.615	2.8	16.257
DAY 2	43.538	2.2	9.571

consistent than the INCLUDE function. No significant differences ($p > 0.05$) were revealed between functions on the SAFE scale despite the fact that a main effect of editing function was found.

Closer examination of Table 13 reveals that several additional dependent measures manifested main effects of EDITOR and/or EDITING FUNCTION. However, these measures also revealed a significant EDITOR by EDITING FUNCTION interaction, indicating that these effects were restricted to specific editor and editing function combinations. All three objective measures, performance rate, errors committed, and error correction time revealed a significant EDITOR by EDITING FUNCTION interaction. This same EDITOR by EDITING FUNCTION interaction was shown by seven subjective rating scales, NATURAL-UNNATURAL, PLEASING-DISGUSTING, SIMPLE-DIFFICULT, FAST-SLOW, COMPLETE-INCOMPLETE, COOPERATIVE-UNCOOPERATIVE, and SATISFACTORY-UNSATISFACTORY. Paired comparison tests (Tukey, alpha=0.05) were conducted to determine the locus of the editor by editing function interaction on these ten dependent measures (see Table 15). Cells containing means, upper mean EDITOR A, lower mean EDITOR B, signify significant effects.

Inspection of the significant effects in Table 15 reveals that total agreement does not exist between objective and subjective measures. This was demonstrated

TABLE 15

Significant Mean Differences Indicated by Objective or Subjective Measures in Terms of the EDITOR by EDITING FUNCTION Interaction

	TRAVEL	SEARCH	DELETE	INSERT	MOVE	REPLACE	WRITE	INCLUDE	FORMAT
OBJECTIVE <u>MEASURES</u>									
RATE MEASURE		36.6 28.2		48.2 38.4		48.2 40.8		43.4 25.8	37.9 26.4
ERRORS COMMITTED	1.5 0.4	1.1 2.8			4.4 2.8				
CORRECTION TIME		11.1 26.6			26.6 15.4				
SUBJECTIVE <u>MEASURES</u>									
NATURAL								5.3 4.3	
PLEASING		6.0 5.3						5.6 4.3	
SIMPLE								6.6 5.5	6.8 6.1
FAST	5.5 4.3							6.3 5.5	
COMPLETE	5.9 5.3								
COOPERATIVE	6.0 5.4	6.3 5.7						6.4 5.7	
SATISFACTORY		6.3 5.5						6.3 5.4	

top value EDITOR A
bottom value EDITOR B

clearly on the TRAVEL function, where the objective measure of errors committed indicated an effect in favor of EDITOR B, while three subjective measures manifested an effect in favor of EDITOR A. Furthermore, the objective and subjective measures dispute the existence of a difference between editors on the INSERT, REPLACE and MOVE functions. On the INSERT and REPLACE functions, an objective difference was revealed by the performance rate measure, while no congruent difference was shown on the subjective dimension. Similarly, no subjective difference was evident on the MOVE function, while the error measures indicated that subjects committed fewer errors and spent less time correcting errors with the MOVE function as implemented on EDITOR B.

Agreement between objective and subjective measures was demonstrated on the SEARCH, INCLUDE, and FORMAT functions. On these functions, performance rate and at least one rating scale indicated a difference between editors in favor of EDITOR A. Further agreement was found between the objective and subjective measures in that neither the rate measure nor any adjective rating scale indicated that a difference existed between editors on the DELETE and WRITE functions.

In order to explore the relationship between objective and subjective measures further, correlations between the performance rate measure and the subjective measures, which indicated a significant EDITOR by EDITING FUNCTION interaction, were computed. It is immediately apparent from the results of this analysis, as shown in Table 16, that the relationship between objective and subjective measures is complex. However, the correlations are all fairly low, in the range of $r = 0.25$ to $r = 0.46$.

Analysis of transfer effects. These analyses investigated the effects of transferring between two editors on the same 15 dependent measures, 3 objective and 12 subjective, studied in the overall analysis. The transfer analyses included only the data from days two and three. These days represented a subject's second day with their first editor and first day with their second editor. Protection against Type I error was maintained separately for the families of subjective and objective measures by utilizing the same procedures employed in the overall analysis. The complete ANOVA summary tables for all transfer analyses are contained in Appendix F.

To explore the objective measures for evidence of transfer effects, interactions involving the factor ORDER were examined. As shown in Table 17, three interactions involving ORDER were significant ($p < 0.0167$). The objective measures of errors committed and error correction time indicated a significant EDITOR by ORDER interaction, while the performance rate revealed both an

TABLE 16

Correlations of Seven Subjective Scales to Performance Rate in Terms of the EDITOR by EDITING FUNCTION Interaction

	TRAVEL	SEARCH	DELETE	INSERT	MOVE	REPLACE	WRITE	INCLUDE	FORMAT
NATURAL			0.25 0.0478	0.29 0.0211				0.29 0.0188	
PLEASING		0.28 0.0258	0.28 0.0276	0.27 0.0323				0.26 0.0388	
SIMPLE			0.39 0.0014	0.46 0.0001	0.43 0.0004	0.30 0.0173		0.33 0.0074	0.39 0.0014
FAST		0.34 0.0058	0.29 0.0172	0.37 0.0029	0.26 0.0348				0.33 0.0082
COMPLETE		0.37 0.0030	0.38 0.0020	0.32 0.0108	0.31 0.0130	0.28 0.0240		0.27 0.0333	
COOPERATIVE			0.39 0.0015	0.32 0.0089	0.38 0.0022	0.38 0.0018		0.28 0.0270	0.26 0.0355
SATISFACTORY		0.36 0.0038	0.37 0.0024	0.39 0.0014	0.29 0.0199	0.43 0.0004			

top value = r
lower value = p-value

TABLE 17

P-values for Transfer Effects for Objective Measures

	RATE MEASURE	ERRORS COMMITTED	ERROR CORRECTION TIME
ED X ORD	7.23 (0.0177)	10.13 (0.0066)	10.63 (0.0057)
EF X ORD	0.57 (0.7968)	1.20 (0.3045)	0.75 (0.6459)
ED X EF X ORD	2.62 (0.0115)	0.64 (0.7420)	1.19 (0.3100)

Main entries are *F*-ratios.
Entries in parentheses represent *p*-values.

TABLE 18

EDITOR by ORDER Transfer Effects, Difference Scores*, Errors and Error Correction Time

DIRECTION OF TRANSFER: EDITOR A TO EDITOR B

ERROR CORRECTION TIME (in seconds)
-5.5277 (9.9305 - 15.4583)

DIRECTION OF TRANSFER: EDITOR B TO EDITOR A

NUMBER OF ERRORS COMMITTED	ERROR CORRECTION TIME (in seconds)
-0.7639 (2.1250 - 2.8889)	-5.4028 (9.7500 - 15.1528)

*sign indicates type of transfer

EDITOR by EDITING FUNCTION by ORDER interaction. Examination of the EDITOR by ORDER interaction displayed in Table 18 indicated that both treatment groups encountered significant negative transfer on at least one error measure. The two groups experienced increases in error correction time, but only subjects transitioning from EDITOR B to EDITOR A committed significantly more errors. Again, the performance rate measure indicated a significant EDITOR by EDITING FUNCTION by ORDER interaction. As illustrated in Table 19, subjects transitioning from EDITOR A to EDITOR B, experienced negative transfer on three editing functions, SEARCH, INCLUDE, and FORMAT. Subjects transitioning from EDITOR B to EDITOR A experienced positive transfer on one editing function, INCLUDE.

There were no transfer effects in the subjective data as evidenced in the lack of any significant interactions involving ORDER by SCALE (see Table 20).

The effect of transitioning between editors was explored further using post-experimental evaluations. These evaluations required subjects to re-evaluate the first editor they encountered after they had used both editors. As evident in Table 21, only interactions involving DAY by SCALE and EDITOR by DAY by SCALE were significant. Differences were located on four scales, USEFUL-USELESS, SIMPLE-DIFFICULT, FAST-SLOW, and

TABLE 19

EDITOR by EDITING FUNCTION by ORDER Transfer Effects, Difference Scores*, Rate Measure

DIRECTION OF TRANSFER: EDITOR A TO EDITOR B

EDITING FUNCTION	RATE MEASURE
SEARCH	-14.875
INCLUDE	-16.875
FORMAT	-24.250

DIRECTION OF TRANSFER: EDITOR B TO EDITOR A

EDITING FUNCTION	RATE MEASURE
INCLUDE	+16.625

*sign indicates type of transfer

TABLE 20

Significance Levels of Interactions Involving ORDER by SCALE

	F-RATIO	P-VALUE
ORDER X SCALE	1.43	0.1647
ED X ORD X SCALE	1.49	0.1408
EF X ORD X SCALE	0.87	0.7965
ED X EF X ORD X SCALE	0.90	0.2239

TABLE 21

Significance Levels of Interactions Involving DAY by SCALE, Day 2 versus Post-experimental Evaluations

	F-RATIO	P-VALUE
DAY X SCALE	3.18	0.0007
ED X DAY X SCALE	2.80	0.0024
EF X DAY X SCALE	0.86	0.8229
ED X EF X DAY X SCALE	1.21	0.0942

SATISFACTORY-UNSATISFACTORY (see Table 22). EDITOR A was rated as more USEFUL, while EDITOR B was rated as less SIMPLE, less FAST, and less SATISFACTORY.

Comparisons Among Subjective Measures

The description of the interrelationships of various subjective measures is presented in three subsections. First, comparisons of both detailed and global subjective measures are presented. Next, the relationship of the SATISFACTORY-UNSATISFACTORY scale with all other scales is described. Finally, the association between subjective scales and preference ratings is defined.

Comparisons of detailed and global subjective measures. Subjects were requested to make both detailed and global evaluations of the editors studied during this research. The detailed evaluations, as already discussed, were collected on twelve scales across nine editing functions. To make global evaluations subjects simply rated the entire editor on the same twelve scales. As evident in the results of these analyses displayed in Table 23, only three of the interactions involving the factor scale were directly comparable between the detailed and global evaluations. Examination of the results reveals that both the detailed and global subjective measures indicated no significant interactions of DAY by SCALE or EDITOR by DAY by SCALE. However, of

TABLE 22

Day 2 Versus Post-Experimental Evaluations
Difference Scores*

DIRECTION OF TRANSFER: EDITOR A TO EDITOR B

USEFUL	SIMPLE	FAST	SATISFACTORY
+0.26 (6.72-6.46)	---	---	---

DIRECTION OF TRANSFER: EDITOR B TO EDITOR A

USEFUL	SIMPLE	FAST	SATISFACTORY
----	-0.40 (6.18-5.76)	-0.31 (5.97-5.47)	-0.36 (6.07-5.71)

*Sign indicates type of transfer
Entries in parentheses represent means.

more interest is the fact that the detailed and global evaluations disagree on the existence of an EDITOR by SCALE interaction. The detailed evaluations, while indicating differences between the two editors, also revealed that the differences existed at specific editor by editing function combinations.

Relationship of all other scales to satisfaction scale. A stepwise regression procedure was utilized to determine any existing relationship between user ratings on 11 bipolar-adjective scales and evaluations on a single scale of satisfaction. This analysis was conducted both jointly and separately for the two editors. As shown in Table 24, the results across the three analyses are the same with the exception of the USEFUL-USELESS scale which drops out of the separate analyses for the two editors. Five scales describe 62 percent of the variance in the satisfaction ratings, DEPENDABLE-UNDEPENDABLE, FAST-SLOW, COMPLETE-INCOMPLETE, PLEASING-DISGUSTING and COOPERATIVE-UNCOOPERATIVE.

A principal components factor analysis was conducted to determine any underlying dimensions in the subjective data. Using the Kaiser Criterion (Thorndike, 1978), factor eigenvalue greater than 1.0, two factors were extracted (see Table 25). These two component factors accounted for 54 percent of the variance. The two

TABLE 23

Significance of Effects for both Detailed and Global Subjective Measures

	DETAILED EVALUATIONS	GLOBAL EVALUATIONS
ED X SCALE	1.97 (0.0341)	1.05 (0.4019)
EF X SCALE	3.79 (0.0001)	____
ED X EF X SCALE	1.37 (0.0145)	____

Main entries are *F*-ratios.
Entries in parentheses represent *p*-values.

TABLE 24

Relationship of 11 Bipolar Scales to SATISFACTION Scale

		REGRESSION p-values		
SCALE	r	OVERALL	EDITOR A	EDITOR B
USEFUL	0.30	0.0266	--	--
DEPENDABLE	0.51	0.0001	0.0001	0.0348
CONSISTENT	0.36	--	--	--
INTERPRETABLE	0.44	--	--	--
SIMPLE	0.45	--	--	--
SAFE	0.34	--	--	--
FAST	0.57	0.0001	0.0028	0.0001
NATURAL	0.49	--	--	--
COMPLETE	0.63	0.0001	0.0001	0.0093
PLEASING	0.63	0.0001	0.0001	0.0001
COOPERATIVE	0.67	0.0001	0.0001	0.0024
	R-SQUARED	0.6239	0.6202	0.6217

TABLE 25

Principal Components Analysis of All Subjective Scales Ratings

SCALE	FACTOR 1	FACTOR 2	COMMUNALITY
USEFUL	0.3459	0.7649	0.7047
DEPENDABLE	0.6858	0.0415	0.4721
CONSISTENT	0.5776	-0.1875	0.3687
INTERPRETABLE	0.6471	-0.2505	0.4815
SIMPLE	0.6847	-0.2942	0.5554
SAFE	0.5792	-0.3373	0.4493
FAST	0.6324	0.1654	0.4272
NATURAL	0.6859	-0.0881	0.4782
COMPLETE	0.7803	0.0826	0.6157
PLEASING	0.7207	0.1720	0.5490
COOPERATIVE	0.8543	-0.0473	0.7168
SATISFACTORY	0.8090	0.2245	0.7049
Eigenvalue	5.5143	1.0092	
Variance Explained	45.95 %	8.41 %	

factors were identified as SATISFACTION and USEFULNESS. Examination of the factor loadings in Table 25 reveals that all scales except USEFUL-USELESS have loadings above 0.5 on FACTOR 1 "SATISFACTION". The scale USEFUL-USELESS has a very high loading (0.76) on FACTOR 2 "USEFULNESS".

Relationship of bipolar adjective scales and preference ratings, within-editor. To determine the relationship between adjective scale ratings and user preferences for the nine editing functions, a stepwise regression procedure was used. This analysis was performed on each editor individually and then collapsed across the two editors. The results of these analyses are displayed in Table 26. Examination of Table 26 reveals that regardless of editor, the same three scales describe a significant amount of the variance in user preferences. These scales, which account for 20 percent of the total variance in preference rankings are; SIMPLE-DIFFICULT, FAST-SLOW, and PLEASING-DISGUSTING.

Relationship of bipolar adjective scales and preference ratings, between-editor. No direct statistical comparison was possible between user preferences for the two editors and adjective scale ratings. The results of user preferences for the two editors, by editing function, are shown in the upper portion of Table 27. The lower portion of Table 27 contains the scales which indicated a significant

TABLE 26

Relationship of 12 Bipolar Scales to User Preferences Rankings

	REGRESSION		
	p-values		
SCALE	OVERALL	EDITOR A	EDITOR B
USEFUL	--	--	--
DEPENDABLE	--	--	--
CONSISTENT	--	--	--
INTERPRETABLE	--	--	--
SIMPLE	0.0001	0.0069	0.0001
SAFE	--	--	--
FAST	0.0019	0.0391	0.0069
NATURAL	--	--	--
COMPLETE	--	--	--
PLEASING	0.0001	0.0001	0.0001
COOPERATIVE	--	--	--
SATISFACTORY	--	--	--
R-SQUARED	0.1910	0.1766	0.2039

TABLE 27

User Preferences and Location of Significant Differences Between Editors on Adjective Scale Ratings, by Editing Function

	TRAVEL	SEARCH	DELETE	INSERT	MOVE	REPLACE	WRITE	INCLUDE	FORMAT
EDITOR A	75	100	31	25	81	18	56	87	100
EDITOR B	12	--	12	--	6	6	--	6	--
NO PREFERENCE	12	--	56	75	12	75	43	6	--
NATURAL								A	
PLEASING		A						A	
SIMPLE								A	A
FAST	A							A	
COMPLETE	A								
COOPERATIVE	A	A						A	
SATISFACTORY		A						A	

difference between editors, the letter "A" signifies that EDITOR A was rated higher. Inspection of Table 27 reveals no consistent relationship between that magnitude of user preference and the number of scales indicating a difference between editors. An example is evident if the TRAVEL and INCLUDE functions are compared. EDITOR A's FORMAT function was preferred by all users and one scale indicated a difference, while four scales manifested differences on the TRAVEL function with only 75 percent preference. Furthermore, 81 percent of the users preferred EDITOR A's MOVE function, yet no scale indicated a similar effect.

DISCUSSION

Overall Relationship Between Objective and Subjective Measures

Both objective and subjective measures indicated that differences exist between the two editors evaluated. However, the two types of dependent measures were not in total agreement about the location or magnitude of these differences. This lack of concurrence was immediately evident in terms of the differences between editors, indicated by these measures. An extreme case of divergence was shown on the TRAVEL function, where significantly fewer errors were committed on EDITOR B, yet three subjective measures indicated that EDITOR A was superior. Less extreme examples were visible on the INSERT and MOVE functions where significant differences indicated by an objective measure were not manifested by any subjective measures.

One possible explanation for the failure of the objective and subjective measures to reveal similar differences is unequal sensitivity. To explore this possibility, correlations were computed between the performance rate measure and the subjective measures. These correlations would reveal any underlying relationship which existed between objective and subjective measures. A strong correlation between measures, despite a disagreement on the existence of a significant difference, could indicate that the subjective measures were insensitive. Although over half of these correlations were significant, none of the coefficients were above 0.46. The low correlations suggested that discrepancies between the two types of measures could not be totally explained by insensitivity of the subjective measures. This result implied that the measures, at least in some instances, seemed to be scaling qualitatively different effects. Evidence in support of this hypothesis was found on the TRAVEL and WRITE functions, where no significant correlations between measures were found. Additionally, on three occasions when both the objective and subjective measures indicated a difference between editors the measures were uncorrelated.

Finally, it was hypothesized that the relationship between the objective and subjective measures would vary across the different bipolar scales. Examination of the correlations between the performance rate measure and the individual adjective scales revealed low correlations in the range of 0.25 to 0.46. This result makes it difficult to suggest that the relationship between subjective and objective measures differed across adjective scales.

Transfer effects.

Several predictions were made about the effects of transferring between editors. It was expected that novice users would learn their second editor faster than they learned their first. All users did learn their second editor faster than their first, regardless of the order in which they encountered the editors.

Both the objective and subjective evaluations of the editors, collected immediately after learning, were expected to show negative transfer. These negative transfer effects were to be more severe on the subjective measures. Close examination of the data collected revealed an effect on the objective measures exclusively. Furthermore, only subjects transferring from EDITOR A to EDITOR B experienced negative transfer. The subjects transitioning from EDITOR B to EDITOR A encountered positive transfer.

Finally, a third hypothesis concerned users' post-experiment evaluations of their original editor. It was predicted that user evaluations of their original editor would change after they had used a second editor. Indeed this was the case. The group transferring from EDITOR A to EDITOR B evaluated their original editor as more useful while subjects transitioning from EDITOR B to EDITOR A evaluated their original editor as more complicated, slower, and less satisfactory.

Interrelationships of Subjective Measures

Comparison of detailed and global subjective evaluations. One of the major hypotheses of this research was that detailed evaluations would allow the designer to identify the specific aspects of an interface which are viewed by the user as problematic, whereas global evaluations would fail to generate this information. A comparison between evaluations revealed that global measures were insensitive to differences between the editors suggested by detailed evaluations. The detailed evaluations further indicated that the editors differed on specific editing functions. Therefore, it seems obvious that detailed evaluations are more useful to interface testing than global evaluations.

Describing user satisfaction. Determining user satisfaction is of utmost importance to the evaluation of a software interface. An attempt to describe user satisfaction in terms of the other bipolar adjective scales, accounted for a substantial portion of the variance in the satisfaction ratings. This analysis indicated that the scales DEPENDABLE, FAST, COMPLETE, PLEASING, and COOPERATIVE described 62 percent of the variance. In an earlier study (Coleman *et al*. 1984) experienced computer users ranked the adjectives DEPENDABLE, FAST and COMPLETE as extremely important to assessing their satisfaction with text editors. These users also rated USEFUL as quite important. However, in the present study USEFUL appeared on a separate dimension from satisfaction. Additionally, the adjective scales PLEASING and COOPERATIVE were rated as least important by experienced users, but described a significant portion of the variance in the present study. These results suggest a difference between novice and experienced users in terms of which system qualities are important to their satisfaction. However, the nature of the results from the two studies, one a regression analysis solution, the other a list of importance ratings, make the comparisons somewhat suspect. The issue needs to be resolved in future research.

Dimensions in subjective evaluations. The application of factor analytic techniques distinguished two dimensions in the subjective evaluations, "SATISFACTION" and "USEFULNESS." The separation of these dimensions indicates that subjective evaluations of satisfaction and usefulness are fairly independent. This result also suggests that usefulness of an interface does not guarantee user satisfaction.

Describing user preferences. Only three adjective scales described significant portions of the variance in user preference rankings within-editor, SIMPLE-DIFFICULT, FAST-SLOW, and PLEASING-DISGUSTING. Additionally, with all three of these scales in the regression equation only 20 percent of the variance in the preference rankings could be described. The between-editor scale ratings and preference rankings were also compared, but revealed no consistent relationship between measures. These results suggest that preference rankings and adjective scale ratings measure subjective evaluations differently. Preference rankings are a relative judgements. However, the rating of each editing functions on the adjective scales are absolute judgements. Future research should explore the relationship between these measures further. One possible method would be to require subjects to indicate the magnitude of their preferences by positioning all items evaluated in relative location on each adjective scale.

Refining the subjective rating scales. Results of four analyses can be used to refine the set of subjective scales used to collect user evaluations of software interfaces: 1) the multiple regression analysis, which described satisfaction; 2) the overall ANOVA, which identified differences between the two editors; 3) the ANOVA of subjective measures on post-experiment evaluations, which indicated changes in subjective evaluations of subjects original editor; and 4) the principal components analysis, which identified two dimensions in the subjective data. If the results of the two ANOVAs and the regression analysis are compared, it is evident that six scales account for significant amounts of variance in a least two of these analyses, SIMPLE, FAST, COMPLETE, COOPERATIVE, PLEASING, and SATISFACTORY. The scales DEPENDABLE and NATURAL contribute to the variance in at least one of the analyses. The scale USEFUL contributed significant variance in both the ANOVA of transfer effects on evaluations of subjects' initial editor and the Principal Components. The results of these analyses revealed that three scales failed to contribute to the variance in any analysis SAFE-UNSAFE, INTERPRETABLE-UNINTERPRETABLE, and CONSISTENT-INCONSISTENT. In general, these analyses suggest that the scales SAFE, INTERPRETABLE, and CONSISTENT could be dropped from the instrument.

Design and Evaluation of Benchmark Tasks

The purpose of developing and applying benchmark tasks is to allow for meaningful comparisons between systems similar in application. In addition, benchmark procedures should enable the compilation of a database on similar systems evaluated by different researchers. To meet this second goal, painstaking effort must be devoted to minimizing variance between experimenters.

The values of Roberts learning measure obtained in the present study for EDITOR A and EDITOR B were 3.5 and

3.8 minutes per task, respectively. These values are significantly smaller than the learning rate of 4.9 minutes per task for EDITOR A reported by Good (1984). Additionally, these values are smaller than anything described in Roberts and Moran (1983). Closer examination of the data in Good (1984) and the present experiment revealed that differences existed on both dependent measures, tasks completed, and task completion time. As mentioned in the METHOD section, the experimenter deviated slightly from Roberts' testing procedures. Uncompleted tasks were brought to a subject's attention to determine whether the subject was incapable of completing or simply overlooked the task. This deviation should only cause an inflation of the number of tasks completed. However, examination of the data showed that time to completion was also shorter in this study. Examination of Roberts' methodology suggests three major factors which may have contributed to the differences between studies. These areas are; 1) task coding, 2) teaching abilities of individual experimenters, and 3) selection of subjects.

Task coding. To code tasks accomplished, as specified by Roberts (1979), the experimenter assigns a "1" to completed tasks, a "0" to uncompleted tasks, and "1/2" for partially completed tasks. However, there are several instances in Roberts' tasks where the boundaries which define each task are unclear. For example, task A requires a subject to delete a blank line in order to join two paragraphs. The subject successfully deletes the line and moves on to the next task. Experimenter 1 codes this as a 1, the subject completed the whole task. Experimenter 2 codes this as 1/2, because the subject joined the paragraphs, but did not reformat the block of text formed. The boundaries of the task are not only different for the experimenters, they are unclear to the subject. In addition, this type of unclear boundary may cause differences in recording of task completion time. Suppose the subject asks the experimenter, after joining the two paragraphs, if the task is completed. Experimenter 1 would say "yes", while Experimenter 2 would say "no" requiring the subject to reformat the paragraph.

Teaching abilities of individual. Variations between experimenters, in terms of teaching skills, may also influence results. As an experimenter uses Roberts (1979) benchmark, it is quite likely that they will improve in their ability to train novice computer users. Good (1984) did not indicate that the experimenter used in his evaluation of EDITOR A trained more than the four subjects required by Roberts (1979) guidelines. However, the experimenter involved in the present study taught 4 pretest, and 16 experimental subjects on each editor evaluated. Although Roberts (1979) taught a large number of subjects also, there is no indication that she taught a large number of subjects on any one editor as was done in this research. To assess the existence of an experimenter learning effect the pilot data were examined. Unfortunately only the pretest results for EDITOR B, 5.23 minutes per task, were available. This figure is much closer to the value of 4.9 minutes per task obtained by Good (1984), as well as the general results reported in Roberts and Moran (1983). These results suggest that the large difference between the data in the present experiment and that of Good (1984) may be due to different levels of teaching ability of the two experimenters.

Selection of subjects. Roberts and Moran (1983) imply that any research which generates data with less variance than their research must have involved a restricted sample of subjects. Roberts (1979) used secretaries as subjects yet gives no indication how it was determined that an unrestricted sample was selected. The present research used a relatively homogeneous sample of college students, which may have contributed to the between experiment differences.

Suggested improvements. Based on the discussed sources variation, several suggestions follow which may lead to improvement in the standardization of benchmark procedures. First, a standardized coding method must be developed which includes clear specifications of task boundaries. Second, tasks should be taught and tested in a logical order. Tasks which have a low probability of being completed based on the subject's previous experience (i.e. search by content) should not be tested until the subject has been exposed to the necessary skills. Third, subject's task proficiency should be examined in a final comprehensive quiz, administered after all training is completed. Fourth, the experimenter's training technique should be relatively stable prior to editor evaluation. Lastly, a testable criterion for the selection of subjects should be developed. This must include a standardized form for collecting demographic data which allows the experimenter to assess objectively a subject's relationship to these criterion.

CONCLUSIONS

Although some agreement existed among objective and subjective measures, differences were found both in terms of magnitude and location of effects. These disagreements seem to be due to insensitivity on the part of the subjective measures as well as real differences in what the types of measures quantify. The present data does not suggest that the association between objective and subjective evaluations varies across the different bipolar adjective scales. Given that the measures are not completely redundant, it seems clear that interface testing should include both objective and subjective evaluations.

Research effort needs to be devoted to the development of empirically derived benchmark tasks. These benchmarks should specify systematic administration procedures which facilitate the collection of both objective and subjective measures. Furthermore, before a task is considered a "benchmark", an extensive evaluation process is necessary. This process should test extensively the benchmark's reliability and ease of use across different interfaces and evaluators.

Global subjective evaluations were insensitive to differences between the two editors indicated by detailed evaluations. The detailed evaluations revealed that both subjective and objective between editor differences were due to specific editing functions. In conclusion, the methodology developed as a result of this research seems to be a viable one for systematically collecting detailed user evaluations.

REFERENCES

Bennet, J. (1984). Managing to meet usability requirements: Establishing and meeting software development goals. In J. Bennet, D. Case., J. Sandelin, and M. Smith (Eds.) Visual display terminals: usability issues and health concerns (pp. 161-184). Englewood Cliffs, New Jersey:Prentice Hill.

Berliner, D. C., Angell, D., and Shearer, J. W. (1964). Behaviors, measures and instruments for performance evaluation in simulated environments, In Proceedings of the Symposium and Workshop on the Quantification of Human Performance. University of New Mexico.

Bolander, D. O., Varner, D. D., and Pine, E. (1981). Instant synonyms and antonyms. Little Falls, New Jersey:Career Publishing.

Burke, T. M., Muto, W. H., and Gutman, J. C. (1984). Effects of keyboard height on typist performance and preferences. In Proceedings of the Human Factors Society 28th Annual Meeting, (pp. 272-276). Santa Monica, CA.:Human Factors Society.

Cohill, L. F., (1984). A taxonomy of user-computer interface functions. In G. Salvendy (ED.) Human-Computer Interaction. B.V. Amsterdam:Elsevier Science Publishers.

Coleman, W. D., Wixon, D. R., and Williges, R. C. (1984). Collecting detailed user evaluations of software interfaces. (Tech. Report DEC TR290). Maynard, Massachuesetts:Digital Equipment Corporation.

Cordes, R. E. (1981). Use of magnitude estimation for evaluating product ease of use. Paper presented at Ergonomics and Health Aspects in Modern Offices, Turin, Italy.

Dzida, W., Herda, S., and Itzfeldt, W. D. (1979). User-perceived quality of interactive systems. IEEE Transactions on Software Engineering, 4, 270-276.

Elkerton, J., and Williges, R. C. (1984). Information retrieval strategies in a file-search environment. Human Factors, 26, 171-184.

Gaylin, K. (1985). Creating an empirically-based windowing benchmark task. (Tech. Report DEC TR370). Maynard, Massachuesetts: Digital Equipment Corporation.

Good, M. (1985). The use of logging data in the design of a new text editor. In Proceedings CHI'85 Human Factors in Computing Systems (pp. 93-98). New York:Association for Computing Machinery.

Ives, B., Olson, M. H., and Baroudi, J. J. (1983). The measurement of user information satisfaction, Communications of the ACM, 26, 785-793.

Kerber, K. (1983). Attitudes towards specific uses of the computer; quantitative, decision-making and record-keeping applications. Behavior and Information Technology, 2, 197-209.

Lenorovitz, D. R., Phillips, A. D., Andrey, R. S., and Kloster, G. V. (1984). A taxonomic approach to characterizing human-computer interfaces. In G. Salvendy (ED.) Human-Computer Interaction. B.V., Amsterdam:Elsevier Science Publishers.

Lucas, R. W. (1977). A study of patient's attitudes to computer interrogation. International Journal of Man Machine Studies, 9, 69-86.

Magers, C. (1983). An experimental evaluation of on-line help for non-programers. In Proceedings CHI'83 Human Factors in Computing Systems (pp. 277-281). New York:Association for Computing Machinery.

Meister, D. (1985). Behavior Analysis and Measurement Methods. New York:Wiley.

Meyrowitz, N., and A. Van Dam, (1982). Interactive Editing Systems: Part 1, Computing Surveys, 14, 321-345.

Nickerson, R. S., (1981) Why interactive computer systems are sometimes not used by people who might benefit from them. International Journal of Man-Machine Studies, 15, 469-493.

Osgood, C. E., Tannenbaum, P. H., and Succi, G. J. (1957). The measurement of meaning. Urbana:University of Illinois Press.

Roberts, T. L. (1979). Evaluation of computer text editors (Report SSL-79-9). Palo Alto, California:Xerox.

Roberts, T. and Moran, T. P. (1983). The evaluations of text editors: Methodology and Empirical Results, Communications of the ACM, 26, 265-283.

Root, R. W., and Draper, S. (1983). Questionnaires as a software evaluation tool. In Proceedings CHI'83 Human Factors in Computing Systems (pp. 83-87). New York:The Association for Computing Machinery.

Rosson, B. M. (1984). Effects of experience on learning, using, and evaluating a text editor. Human Factors, 26, 463-476.

Rushinek, A., Rushinek, S. F., and Stutz, J. (1984). A methodology for interactive evaluation of user reactions to software: an empirical analysis of system performance, interaction and routine. International Journal of Man-Machine Studies, 20, 169-188.

Thorndike, R. M. (1978). Correctional techniques for research. New York:Garner press, Inc.

Shackel, B. (1984). The concept of usability. In J. Bennet, D. Case., J. Sandelin, and M. Smith (Eds.) Visual display terminals: usability issues and health concerns (pp. 45-88). Englewood Cliffs, New Jersey:Prentice Hill.

Whiteside, J., Archer, N., Wixon, D., and Good, M. (1982). How people really use text editors. SIGOA Newsletter, 3, 29-40.

Whiteside, J., Jones, S., Levy, P., and Wixon D. (1985). User performance with command, menu, and iconic interfaces. In Proceedings CHI'85: Human Factors in Computer Systems (pp. 185-191). New York:Association for Computing Machinery.

Williges, R. C., Williges, B. H., and Elkerton, J. (in press). Software Interface Design. In G. Salvendy (Ed.) *Handbook of Human Factors / Ergonomics*. New York:Wiley.

Zoltan, E., and Chapanis, A. (1982). What do professional persons think about computers? *Behaviour and Information Technology*, *1*, 55-68.

Appendix A

Participants Informed Consent

PARTICIPANT'S STATEMENT OF INFORMED CONSENT

You are asked to participate in a study of how people use computer text editors. The purpose of the study is to give us information we need to make computers easier to use.

We are not evaluating you, rather we are studying how easy computers are to use. All information you give us and all data that we collect concerning your task will be held in strict confidence. We will use the information for statistical and summary purposes only, and will make certain that your name is not associated with your records. To the best of our knowledge, there are no physical or psychological risks associated with the procedures in our study.

As a participant in this study, you have certain rights. These rights will now be explained to you, and you will be asked for your signature, indicating that you consent to participation in this research.

1. You have the right to stop the experiment in which you are participating at any time if you feel that it is not agreeable to you. Should you terminate the experiment, you will receive pay only for the proportion of time you participated.

2. You have the right to see your data and to withdraw them from the experiment if you feel that you should.

3. You have the right to be informed of the results of the overall experiment. If you wish to receive a summary of the results, please indicate your address (three months hence) with your signature below. A summary will be sent to you. If you should then like further information, please contact the Human Factors Laboratory and a full report will be made available to you.

4. You have the right to call either Dr. Robert Williges, the principle investigator, at 961-6270 or Mr. Charles Waring, Institutional Review Board Chairman, at 961-5283, with your concerns about any aspect of the experiment if you feel uncomfortable talking with the experimenter.

The faculty and graduate students involved greatly appreciate your help as a participant. If you have any question about the experiment or your rights as a participant, please do not hesitate to ask. We will do our best to answer them, subject only to the constraint that we do not want to pre-bias the experimental results.

Your signature below indicates that you have read your rights as a participant as stated above and that you consent to participation. If you include your printed name and address below, a summary of the experimental results will be sent to you.

____
Signature

____
____
____
Print address above
(3 months hence) if
you would like to be
informed of results.

Appendix B

Editing Task Summary Sheets

Editor Summary Sheet

Commands to modify text

insert into document: position cursor: type text

delete text: position cursor at beginning: SELECT
position cursor at end: REMOVE
or ERASE WORD

move text: position cursor at beginning: SELECT
position cursor at end: REMOVE
position cursor at destination: INSERT HERE

copy text: position cursor at beginning: SELECT
position cursor at end: REMOVE, INSERT HERE
position cursor at destination: INSERT HERE

Commands to get and save a document

get document: DO, get file name.ext, RETURN

save document: DO, write file name.ext, RETURN

Finding and displaying text

position cursor:
or NEXT SCREEN
or PREV SCREEN

search for a string:

forward FIND, type text, RETURN

Formatting text

position cursor on paragrapgh to be formatted: DO, fill, RETURN

Error correction

delete during typein: DELETE
cancel SELECT: SELECT
cancel FIND: delete all text after prompt, RETURN
cancel DO: delete all text after prompt, RETURN

Editor Summary Sheet

Commands to modify text

insert into document:	position cursor: type text
delete text:	position cursor at beginning: SELECT position cursor at end: CUT
or	DEL W
move text:	position cursor at beginning: SELECT position cursor at end: CUT position cursor at destination: GOLD, PASTE
copy text:	position cursor at beginning: SELECT position cursor at end: CUT, GOLD, PASTE position cursor at destination: GOLD, PASTE

Commands to get and save a document

get document:	GOLD, COMMAND, include name.ext =name, ENTER
save document:	GOLD, COMMAND, write name.ext, ENTER

Finding and displaying text

position cursor:	
or	ADVANCE, SECT
or	BACKUP, SECT
search for a string:	
forward	ADVANCE, GOLD, FIND, type text, ENTER
backwards	BACKUP, GOLD, FIND, type text, ENTER

Formatting text

	position cursor at beginning: SELECT position cursor at the end: GOLD,FILL

Error correction

delete during typein:	DELETE
cancel SELECT:	GOLD, RESET
cancel FIND:	delete all text after prompt, ENTER
cancel COMMAND:	delete all text after prompt, ENTER
delete on command line	DELETE, (DO NOT USE ARROW KEYS)

Appendix C

Modifications of Questionnaire

The primary questionnaire used to collect subjective evaluations experiment consisted of a modified version of an instrument resulting from preliminary work reported in Coleman, Wixon and Williges(1984). The modified instrument was designed to represent a text editor in terms of the core editing tasks performed by novice users. The core editing tasks required that novices use only nine of the 16 editing functions from the comprehensive list. The functions from the instrument which were not used were COPY, CUSTOMIZE, VIEW, REQUEST, RECOVER, INITIATE, and TERMINATE.

If one compares the nine functions selected for the core editing instrument with the importance rankings of the 16 editing functions from the pilot study, displayed in Table Cl, similarities become evident. Eight out of the 10 most important editing functions from the point of view of experienced users are included in the core editing instrument, the exceptions being TERMINATE and RECOVER. TERMINATE was not included because during core learning and practice the subjects never activated or deactivated the editor. In addition, the RECOVER function was not used since the users were not required to recover from any editor failures. The FORMAT function, although it was rated as the least important function by experienced users, was included in the core evaluation instrument. Most of the experienced users of the editor evaluated in preliminary work used a format process external to the editor. However, the novice users were required to reformat the computer files they edited while they were in the editor.

The list of semantic differentials was also shortened. Instead of the original 17 scales only 11 scales were used. The usage frequency of the extreme scale regions (1,2 or 6,7) was computed for each semantic differential. The 11 most frequently used scales were chosen. The usage frequency and rank order of all 17 scales can be seen in Table C2. A large drop in usage frequency occurs after the 11th scale (150 down to 109). In addition, the scale SATISFACTORY-UNSATISFACTORY was added for the purpose of determining which bipolar scales described user satisfaction.

TABLE C1

The Importance of 16 Text Editor Functions, as Rated by Text Editor Users, (n=27), (out of a maximum of 7 which equals extremely important)

RANK	FUNCTION	MEAN
1.0	MOVE	6.9630
2.5	TRAVEL	6.8519
2.5	SEARCH	6.8519
4.5	DELETE	6.7778
4.5	TERMINATE	6.7778
6.0	INCLUDE	6.6666
8.0	RECOVER	6.4444
8.0	WRITE	6.4444
8.0	INSERT	6.4444
10.0	REPLACE	6.3704
11.0	CUSTOMIZE	6.3333
12.0	INITIATE	6.1481
13.0	VIEW	5.8571
14.0	COPY	5.7407
15.0	REQUEST	5.6296
16.0	FORMAT	5.0374

TABLE C2

Ranked Usage Frequency of 17 Bipolar Adjective Scales in Terms of Non-neutral ratings.

RANK	SCALE
1.0	USEFUL..................USELESS
2.0	DEPENDABLE.........UNDEPENDABLE
3.0	CONSISTENT.........INCONSISTENT
4.0	INTERPRETABLE...UNINTERPRETABLE
5.0	SIMPLE..............COMPLICATED
6.0	COMPLETE.............INCOMPLETE
7.0	SAFE.....................UNSAFE
8.0	FAST.......................SLOW
9.0	NATURAL...............UNNATURAL
10.0	COOPERATIVE.......UNCOOPERATIVE
11.0	PLEASING.............IRRITATING
12.0	ADAPTIVE.............UNADAPTIVE
13.0	FRIENDLY.............UNFRIENDLY
14.0	UNCLUTTERED...........CLUTTERED
15.0	CONCISE...............REDUNDANT
16.0	INTELLIGENT.......UNINTELLIGENT
17.0	MAINTAINABLE.....UNMAINTAINABLE

Appendix D

Questionnaires

Evaluate the text editor specified by the experimenter.

Evaluate the editor in terms how well you were able to accomplish 9 editing functions with it. The functions are;

TRAVEL...
This function is used to change the position of the cursor in the file.

SEARCH...
This function is used to find a specified target such as a string of characters.

DELETE...
This function is used to delete text from the file.

INSERT...
This function is used to insert new text into a file.

MOVE...
This function is used to move a section of text to another location within a file.

REPLACE...
This function is used to replace one piece of text with another.

WRITE...
This function is used to save a file.

INCLUDE...
This function is used to get a file into the editor.

FORMAT...
This function is used to adjust text within the file.

The editing functions are printed, one to a page.

Evaluate the each editing function on the 12 scales below it. These scales are;

```
useless                                                             useful
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely

undependable                                                     dependable
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely

consistent                                                     inconsistent
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely

uninterpretable                                               interpretable
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely

simple                                                          complicated
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely

unsafe                                                                 safe
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely

fast                                                                   slow
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely

unnatural                                                           natural
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely

complete                                                         incomplete
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely

disgusting                                                         pleasing
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely

cooperative                                                       obstinate
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely
satisfactory                                                 unsatisfactory
   1         2          3          4          5          6          7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely   quite     slightly   neutral    slightly    quite     extremely
```

Make your evaluations by circling the scale number which appropriately describes how you feel about the editor in terms of accomplishing that editing function.

Below are examples of incorrect and correct markings of the type of scales you will be asked to use in your evaluations of a text editor. Please indicate your evaluations by carefully circling the appropriate number on the scale.

CORRECT MARKING OF A SCALE

good						bad
1	2	3	4	[✓ 5]	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

INCORRECT MARKING OF A SCALE

good						bad
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

TRAVEL This function is used to change the position of the cursor in the file.

useless useful
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

undependable dependable
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

consistent inconsistent
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

uninterpretable interpretable
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

simple complicated
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

unsafe safe
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

fast slow
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

unnatural natural
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

complete incomplete
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

disgusting pleasing
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

cooperative obstinate
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

satisfactory unsatisfactory
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

SEARCH This function is used to find a specified target such as a string of characters.

useless						useful
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

undependable						dependable
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

consistent						inconsistent
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

uninterpretable						interpretable
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

simple						complicated
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

unsafe						safe
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

fast						slow
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

unnatural						natural
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

complete						incomplete
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

disgusting						pleasing
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

cooperative						obstinate
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

satisfactory						unsatisfactory
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

DELETE This function is used to delete text from the file.

useless useful
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

undependable dependable
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

consistent inconsistent
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

uninterpretable interpretable
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

simple complicated
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

unsafe safe
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

fast slow
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

unnatural natural
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

complete incomplete
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

disgusting pleasing
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

cooperative obstinate
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

satisfactory unsatisfactory
1 2 3 4 5 6 7
: ---------: ---------: ---------: ---------: ---------: ---------: ---------:
extremely quite slightly neutral slightly quite extremely

INSERT This function is used to insert new text into a file.

useless						useful
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

undependable						dependable
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

consistent						inconsistent
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

uninterpretable						interpretable
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

simple						complicated
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

unsafe						safe
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

fast						slow
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

unnatural						natural
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

complete						incomplete
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

disgusting						pleasing
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

cooperative						obstinate
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

satisfactory						unsatisfactory
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

MOVE This function is used to move a section of text to another location within a file.

useless						useful
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

undependable						dependable
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

consistent						inconsistent
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

uninterpretable						interpretable
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

simple						complicated
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

unsafe						safe
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

fast						slow
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

unnatural						natural
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

complete						incomplete
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

disgusting						pleasing
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

cooperative						obstinate
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

satisfactory						unsatisfactory
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

REPLACE This function is used to replace one piece of text with another.

useless						useful
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

undependable						dependable
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

consistent						inconsistent
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

uninterpretable						interpretable
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

simple						complicated
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

unsafe						safe
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

fast						slow
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

unnatural						natural
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

complete						incomplete
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

disgusting						pleasing
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

cooperative						obstinate
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

satisfactory						unsatisfactory
1	2	3	4	5	6	7
extremely	quite	slightly	neutral	slightly	quite	extremely

WRITE This function is used to save a file.

```
useless                                                           useful
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

undependable                                                  dependable
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

consistent                                                  inconsistent
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

uninterpretable                                            interpretable
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

simple                                                       complicated
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

unsafe                                                              safe
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

fast                                                                slow
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

unnatural                                                        natural
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

complete                                                      incomplete
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

disgusting                                                      pleasing
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

cooperative                                                    obstinate
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely

satisfactory                                              unsatisfactory
    1          2          3          4          5          6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely    quite    slightly   neutral   slightly    quite   extremely
```

INCLUDE This function is used to get a file into the editor.

```
useless                                                          useful
    1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

undependable                                                 dependable
   1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

consistent                                                 inconsistent
   1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

uninterpretable                                           interpretable
   1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

simple                                                     complicated
   1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

unsafe                                                             safe
   1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

fast                                                               slow
   1         2         3         4         5         6       - 7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

unnatural                                                       natural
   1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

complete                                                     incomplete
   1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

disgusting                                                     pleasing
   1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

cooperative                                                   obstinate
   1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely

satisfactory                                             unsatisfactory
   1         2         3         4         5         6         7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite   slightly  neutral  slightly   quite    extremely
```

FORMAT This function is used to adjust text within the file.

```
useless                                                          useful
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

undependable                                                 dependable
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

consistent                                                 inconsistent
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

uninterpretable                                           interpretable
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

simple                                                      complicated
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

unsafe                                                             safe
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

fast                                                               slow
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

unnatural                                                       natural
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

complete                                                     incomplete
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

disgusting                                                     pleasing
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

cooperative                                                   obstinate
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely

satisfactory                                             unsatisfactory
    1          2          3          4          5          6          7
:---------:---------:---------:---------:---------:---------:---------:
extremely   quite    slightly  neutral   slightly   quite    extremely
```

Please evaluate the ENTIRE EDITOR on the scales below. To indicate you evaluation, circle the number on each scale you feel best describes the editor.

useless   useful
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

undependable   dependable
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

consistent   inconsistent
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

uninterpretable   interpretable
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

simple   complicated
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

unsafe   safe
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

fast   slow
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

unnatural   natural
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

complete   incomplete
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

disgusting   pleasing
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

cooperative   obstinate
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

satisfactory   unsatisfactory
1 2 3 4 5 6 7
:---------:---------:---------:---------:---------:---------:---------:
extremely quite slightly neutral slightly quite extremely

Presented below are the 9 editing functions you have learned to perform. Please RANK each of the editing functions with regard to how well you liked to perform it with the editor you are now working on. Rank the editing functions from 1 to 9, with 1 representing the editing function you liked to perform the best and 9 representing the function you liked to perform least.

Editing Task	Rank
TRAVEL	____
SEARCH	____
DELETE	____
INSERT	____
MOVE	____
REPLACE	____
WRITE	____
INCLUDE	____
FORMAT	____

Presented below are the 9 editing functions you have laerned to perform with two text editors. For each editing function, please indicate the editor you would prefer to use by placing a check in the column of the preferred editor.

To peform the following editing function I prefer to use

	First Editor Learned	Second Editor Learned	No Preference
TRAVEL	____	____	____
SEARCH	____	____	____
DELETE	____	____	____
INSERT	____	____	____
MOVE	____	____	____
REPLACE	____	____	____
WRITE	____	____	____
INCLUDE	____	____	____
FORMAT	____	____	____

Appendix E

Rate Measure Computations

To determine the performance rate measure developed by Whitside, Jones, and Levy (1985), the researcher has several options for assigning values to the letters "P" and "C". As shown in Figure E1 "P" represents percent of task completed and "C" represents an expert's fastest task completion time.

To calculate "P" the experimenter can assign the components of the task equal or proportional weights. The researcher determines proportional weights either by subjectively estimating the contribution of each task component to the whole task or by empirically deriving a task's component's weight through the observation of an expert's performance. The weights used in this experiment were empirically derived for each editor by observation of an expert's performance. The time constant "C" can also be ascertained by one of two methods. The first method selects a single value of "C" as determined by the fastest possible expert time on a given class of interface. The second method selects a separate value for "C" for each editor based on an expert's time on that editor. The calculation of the rate measure for this experiment was determined by a single value of "C" representing the fastest expert time given the two editors examined.

Appendix F

ANOVA Summary Tables

Core Editing Task Learning

Dependent Measure: Mean Learning Time per Task

Source		df	SS	F-ratio	p-value
Between subjects					
ORDER	(ORD)	1	0.116	0.29	0.5961
SUBJECTS/ORD	(SUBJ/ORD)	14	5.500		
Within subjects					
EDITOR	(ED)	1	1.558	14.65	0.0018
ED X ORD		1	13.798	129.74	0.0001
ED X SUBJ/ORD		14	1.489		
TOTAL		31	22.460		

Practice Data

Dependent Measure: Total Time to Complete All Practice Tasks

Source		df	SS	F-ratio	p-value
Between subjects					
ORDER	(ORD)	1	22.781	0.32	0.5787
SUBJECTS/ORD	(SUBJ/ORD)	14	986.688		
Within subjects					
EDITOR	(ED)	1	457.531	15.19	0.0016
ED X ORD		1	935.281	31.05	0.0001
ED X SUBJ/ORD		14	421.688		
TOTAL		31	2823.969		

Overall Analysis

Dependent Measure: Rate Measure

Source	df	SS	F-ratio	p-value
Between Subjects				
SUBJECTS (SUBJ)	15	9000.250		
Within Subjects				
EDITOR (ED)	1	6724.000	12.61	0.0029
ED X SUBJ	15	7999.833		
EDITING FUNCTION(EF)	8	32419.326	40.74	0.0001
EF X SUBJ	120	11936.563		
DAY (D)	1	11466.840	283.99	0.0001
D X SUBJ	15	605.660		
ED X EF	8	4013.563	7.08	0.0001
ED X EF X SUBJ	120	8501.104		
ED X D	1	150.063	2.91	0.1088
ED X D X SUBJ	15	744.438		
EF X D	8	352.347	0.73	0.6636
EF X D X SUBJ	120	7226.652		
ED X EF X D	8	202.250	0.39	0.9221
ED X EF X D X SUBJ	120	7705.750		
TOTAL	575	109078.638		

Overall Error Analysis

Dependent Measure: Number of Errors Committed

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
SUBJECTS (SUBJ)	15	241.929		
<u>Within Subjects</u>				
EDITOR (ED)	1	0.043	0.02	0.8985
ED X SUBJ	15	38.651		
EDITING FUNCTION (EF)	8	1597.399	48.83	0.0001
EF X SUBJ	120	490.712		
DAY (D)	1	49.585	19.07	0.0001
D X SUBJ	15	38.998		
ED X EF	8	123.691	6.29	0.0001
ED X EF X SUBJ	120	294.865		
ED X D	1	0.002	0.00	0.9836
ED X D X SUBJ	15	59.471		
EF X D	8	27.274	1.19	0.3099
EF X D X SUBJ	120	343.392		
ED X EF X D	8	25.358	1.05	0.4032
ED X EF X D X SUBJ	120	362.420		
TOTAL	575	3693.790		

Overall Error Analysis

Dependent Measure: Error Correction Time

Source		df	SS	F-ratio	p-value
Between Subjects					
SUBJECTS	(SUBJ)	15	8769.248		
Within Subjects					
EDITOR	(ED)	1	42.793	0.11	0.7459
ED X SUBJ		15	5891.679		
EDITING FUNCTION	(EF)	8	25951.024	9.93	0.0001
EF X SUBJ		120	39216.142		
DAY	(D)	1	6540.766	42.25	0.0001
D X SUBJ		15	2322.040		
ED X EF		8	8590.628	3.02	0.0040
ED X EF X SUBJ		120	42715.649		
ED X D		1	47.266	0.11	0.7420
ED X D X SUBJ		15	6304.762		
EF X D		8	3828.156	1.48	0.1731
EF X D X SUBJ		120	38915.788		
ED X EF X D		8	3055.969	1.17	0.3227
ED X EF X D X SUBJ		120	39171.753		
TOTAL		575	231363.665		

Subjective Evaluations

Dependent Measure: Bipolar Adjective Scale Ratings

Source	df	SS	F-ratio	p-value
Between Subjects				
SUBJECTS (SUBJ)	15	2418.579		
Within Subjects				
EDITOR (ED)	1	72.931	11.75	0.0037
ED X SUBJ	15	93.138		
EDITING FUNCTION(EF)	8	178.086	4.14	0.0002
EF X SUBJ	120	644.895		
DAY (D)	1	35.593	12.37	0.0031
D X SUBJ	15	43.144		
SCALE (SC)	11	752.664	11.77	0.0001
SC X SUBJ	165	958.831		
ED X EF	8	81.428	4.01	0.0003
ED X EF X SUBJ	120	304.294		
ED X D	1	0.422	0.13	0.7246
ED X D X SUBJ	15	49.083		
ED X SC	11	33.624	1.97	0.0341
ED X SC X SUBJ	165	255.751		
EF X D	8	24.225	1.90	0.0667
EF X D X SUBJ	120	191.664		
EF X SCALE	88	285.653	3.79	0.0001
EF X SC X SUBJ	1320	1129.976		
D X SC	11	7.088	0.85	0.5867
D X SC X SUBJ	165	124.398		
ED X EF X D	8	15.302	1.37	0.2154
ED X EF X D X SUBJ	120	167.235		
ED X EF X SC	88	79.422	1.37	0.0145
ED X EF X SC X SUBJ	1320	866.911		
ED X D X SC	11	9.821	0.89	0.5482
ED X D X SC X SUBJ	165	164.785		
EF X D X SC	88	52.376	1.10	0.2521
EF X D X SC X SUBJ	1320	714.013		
ED X EF X D X SC	88	45.049	1.02	0.4326
ED X EF X D X SC X SUBJ	1320	662.803		
TOTAL	6911	10463.185		

Overall Subjective Analysis

Dependent Measure: USELESS/USEFUL

Source		df	SS	F-ratio	p-value
<u>Between Subjects</u>					
SUBJECTS	(SUBJ)	15	50.498		
<u>Within Subjects</u>					
EDITOR	(ED)	1	1.460	2.57	0.1296
ED X SUBJ		15	8.512		
EDITING FUNCTION	(EF)	8	45.035	4.78	0.0001
EF X SUBJ		120	141.299		
DAY	(DAY)	1	0.002	0.01	0.9304
D X SUBJ		15	3.304		
ED X EF		8	2.118	0.74	0.6519
ED X EF X SUBJ		120	42.660		
ED X D		1	0.002	0.00	0.9575
ED X D X SUBJ		15	8.859		
EF X D		8	3.889	1.03	0.4164
EF X D X SUBJ		120	56.556		
ED X EF X D		8	4.264	1.26	0.2690
ED X EF X D X SUBJ		120	50.625		
TOTAL		575	419.082		

Overall Subjective Analysis

Dependent Measure: UNDEPENDABLE/DEPENDABLE

Source		df	SS	F-ratio	p-value
<u>Between Subjects</u>					
SUBJECTS	(SUBJ)	15	167.900		
<u>Within Subjects</u>					
EDITOR	(ED)	1	1.266	1.89	0.1893
ED X SUBJ		15	10.040		
EDITING FUNCTION	(EF)	8	26.108	3.59	0.0009
EF X SUBJ		120	109.059		
DAY	(DAY)	1	2.377	3.63	0.0761
D X SUBJ		15	98.178		
ED X EF		8	5.281	1.35	0.2255
ED X EF X SUBJ		120	58.663		
ED X D		1	0.085	0.03	0.8581
ED X D X SUBJ		15	38.554		
EF X D		8	7.233	1.81	0.0809
EF X D X SUBJ		120	59.823		
ED X EF X D		8	4.462	0.87	0.5460
ED X EF X D X SUBJ		120	77.149		
TOTAL		575	577.873		

Overall Subjective Analysis

Dependent Measure: INCONSISTENT/CONSISTENT

Source		df	SS	F-ratio	p-value
Between Subjects					
SUBJECTS	(SUBJ)	15	300.734		
Within Subjects					
EDITOR	(ED)	1	4.516	2.06	0.1722
ED X SUBJ		15	32.957		
EDITING FUNCTION	(EF)	8	24.910	2.44	0.0177
EF X SUBJ		120	153.313		
DAY	(DAY)	1	6.891	7.80	0.0136
D X SUBJ		15	13.248		
ED X EF		8	2.500	0.15	0.9960
ED X EF X SUBJ		120	243.278		
ED X D		1	1.891	1.30	0.2720
ED X D X SUBJ		15	21.804		
EF X D		8	9.813	1.17	0.3200
EF X D X SUBJ		120	125.299		
ED X EF X D		8	5.125	0.62	0.7615
ED X EF X D X SUBJ		120	124.431		
TOTAL		575	1070.707		

Overall Subjective Analysis

Dependent Measure: UNINTERPRETABLE/INTERPRETABLE

Source		df	SS	F-ratio	p-value
Between Subjects					
SUBJECTS	(SUBJ)	15	304.512		
Within Subjects					
EDITOR	(ED)	1	4.877	4.34	0.0549
ED X SUBJ		15	16.873		
EDITING FUNCTION	(EF)	8	11.680	1.07	0.3903
EF X SUBJ		120	164.097		
DAY	(DAY)	1	3.210	2.95	0.1064
D X SUBJ		15	16.318		
ED X EF		8	8.389	0.96	0.4737
ED X EF X SUBJ		120	131.611		
ED X D		1	0.085	0.09	0.7704
ED X D X SUBJ		15	14.443		
EF X D		8	4.868	0.66	0.7293
EF X D X SUBJ		120	111.354		
ED X EF X D		8	6.618	0.89	0.5276
ED X EF X D X SUBJ		120	111.604		
TOTAL		575	910.540		

Overall Subjective Analysis

Dependent Measure: DIFFICULT/SIMPLE

Source		df	SS	F-ratio	p-value
Between Subjects					
SUBJECTS	(SUBJ)	15	272.429		
Within Subjects					
EDITOR	(ED)	1	18.418	6.82	0.0196
ED X SUBJ		15	40.498		
EDITING FUNCTION	(EF)	8	65.941	6.88	0.0001
EF X SUBJ		120	143.839		
DAY	(DAY)	1	9.252	16.66	0.0010
D X SUBJ		15	8.332		
ED X EF		8	22.253	3.01	0.0042
ED X EF X SUBJ		120	111.080		
ED X D		1	0.210	0.19	0.6682
ED X D X SUBJ		15	16.484		
EF X D		8	6.608	1.80	0.0835
EF X D X SUBJ		120	155.059		
ED X EF X D		8	7.337	1.92	0.0625
ED X EF X D X SUBJ		120	57.219		
TOTAL		575	834.957		

Overall Subjective Analysis

Dependent Measure: UNSAFE/SAFE

Source		df	SS	F-ratio	p-value
Between Subjects					
SUBJECTS	(SUBJ)	15	234.104		
Within Subjects					
EDITOR	(ED)	1	3.063	2.02	0.1755
ED X SUBJ		15	22.715		
EDITING FUNCTION	(EF)	8	30.170	3.85	0.0005
EF X SUBJ		120	117.552		
DAY	(DAY)	1	3.361	2.82	0.1136
D X SUBJ		15	17.861		
ED X EF		8	9.281	1.80	0.0839
ED X EF X SUBJ		120	77.441		
ED X D		1	2.250	2.42	0.1410
ED X D X SUBJ		15	13.972		
EF X D		8	3.858	0.74	0.6579
EF X D X SUBJ		120	78.420		
ED X EF X D		8	4.344	1.33	0.2344
ED X EF X D X SUBJ		120	48.934		
TOTAL		575	667.326		

Overall Subjective Analysis

Dependent Measure: SLOW/FAST

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
SUBJECTS (SUBJ)	15	233.639		
<u>Within Subjects</u>				
EDITOR (ED)	1	25.840	11.38	0.0042
ED X SUBJ	15	34.049		
EDITING FUNCTION (EF)	8	102.639	6.38	0.0001
EF X SUBJ	120	241.361		
DAY (DAY)	1	3.063	3.35	0.0872
D X SUBJ	15	13.715		
ED X EF	8	21.535	2.31	0.0241
ED X EF X SUBJ	120	139.576		
ED X D	1	0.111	0.11	0.7426
ED X D X SUBJ	15	14.889		
EF X D	8	17.688	2.59	0.0121
EF X D X SUBJ	120	102.535		
ED X EF X D	8	9.389	1.37	0.2154
ED X EF X D X SUBJ	120	102.611		
TOTAL	575	1062.639		

Overall Subjective Analysis

Dependent Measure: UNNATURAL/NATURAL

Source		df	SS	F-ratio	p-value
<u>Between Subjects</u>					
SUBJECTS	(SUBJ)	15	570.776		
<u>Within Subjects</u>					
EDITOR	(ED)	1	0.293	0.11	0.7500
ED X SUBJ		15	41.790		
EDITING FUNCTION	(EF)	8	18.514	1.93	0.0611
EF X SUBJ		120	143.708		
DAY	(DAY)	1	3.516	1.80	0.2000
D X SUBJ		15	29.345		
ED X EF		8	20.472	3.69	0.0007
ED X EF X SUBJ		120	83.194		
ED X D		1	3.516	2.18	0.1609
ED X D X SUBJ		15	24.234		
EF X D		8	4.875	0.77	0.6301
EF X D X SUBJ		120	95.014		
ED X EF X D		8	2.375	0.46	0.8826
ED X EF X D X SUBJ		120	77.625		
TOTAL		575	1119.248		

Overall Subjective Analysis

Dependent Measure: INCOMPLETE/COMPLETE

Source		df	SS	F-ratio	p-value
<u>Between Subjects</u>					
SUBJECTS	(SUBJ)	15	335.443		
<u>Within Subjects</u>					
EDITOR	(ED)	1	11.391	4.60	0.0487
ED X SUBJ		15	37.137		
EDITING FUNCTION	(EF)	8	24.024	3.20	0.0025
EF X SUBJ		120	112.698		
DAY	(DAY)	1	6.460	4.16	0.0594
D X SUBJ		15	23.290		
ED X EF		8	10.094	2.14	0.0367
ED X EF X SUBJ		120	70.628		
ED X D		1	0.016	0.01	0.9124
ED X D X SUBJ		15	18.734		
EF X D		8	2.087	0.45	0.8879
EF X D X SUBJ		120	64.413		
ED X EF X D		8	2.406	0.88	0.5369
ED X EF X D X SUBJ		120	41.094		
TOTAL		575	764.915		

Overall Subjective Analysis

Dependent Measure: DISGUSTING/PLEASING

Source		df	SS	F-ratio	p-value
Between Subjects					
SUBJECTS	(SUBJ)	15	328.167		
Within Subjects					
EDITOR	(ED)	1	13.444	4.25	0.0571
ED X SUBJ		15	47.500		
EDITING FUNCTION	(EF)	8	62.243	6.32	0.0001
EF X SUBJ		120	147.646		
DAY	(DAY)	1	1.778	3.40	0.0849
D X SUBJ		15	7.833		
ED X EF		8	27.868	6.09	0.0001
ED X EF X SUBJ		120	68.688		
ED X D		1	1.778	2.34	0.1468
ED X D X SUBJ		15	11.389		
EF X D		8	5.097	1.87	0.0702
EF X D X SUBJ		120	40.792		
ED X EF X D		8	7.222	2.77	0.0076
ED X EF X D X SUBJ		120	39.111		
TOTAL		575	810.556		

Overall Subjective Analysis

Dependent Measure: UNCOOPERATIVE/COOPERATIVE

Source		df	SS	F-ratio	p-value
Between Subjects					
SUBJECTS	(SUBJ)	15	359.583		
Within Subjects					
EDITOR	(ED)	1	10.028	5.73	0.0302
ED X SUBJ		15	26.250		
EDITING FUNCTION	(EF)	8	21.878	2.44	0.0176
EF X SUBJ		120	134.510		
DAY	(DAY)	1	2.007	2.09	0.1685
D X SUBJ		15	14.382		
ED X EF		8	14.441	3.08	0.0034
ED X EF X SUBJ		120	70.281		
ED X D		1	0.007	0.01	0.9362
ED X D X SUBJ		15	15.715		
EF X D		8	1.962	0.45	0.8896
EF X D X SUBJ		120	65.649		
ED X EF X D		8	4.150	1.44	0.1856
ED X EF X D X SUBJ		120	43.128		
TOTAL		575	783.972		

Overall Subjective Analysis

Dependent Measure: UNSATISFACTORY/SATISFACTORY

Source		df	SS	F-ratio	p-value
<u>Between Subjects</u>					
SUBJECTS	(SUBJ)	15	219.568		
<u>Within Subjects</u>					
EDITOR	(ED)	1	11.960	5.87	0.0285
ED X SUBJ		15	30.568		
EDITING FUNCTION	(EF)	8	30.597	2.77	0.0076
EF X SUBJ		120	165.792		
DAY	(DAY)	1	0.766	1.14	0.3030
D X SUBJ		15	10.095		
ED X EF		8	16.613	3.36	0.0016
ED X EF X SUBJ		120	74.104		
ED X D		1	0.293	0.30	0.5934
ED X D X SUBJ		15	14.790		
EF X D		8	8.625	2.83	0.0066
EF X D X SUBJ		120	45.764		
ED X EF X D		8	2.660	0.71	0.6857
ED X EF X D X SUBJ		120	56.507		
TOTAL		575	688.707		

Transfer of Training

Dependent Measure: Rate Measure

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
ORDER (ORD)	1	62.347	0.19	0.6714
SUBJECTS (SUBJ/ORD)	14	4648.486		
<u>Within Subjects</u>				
EDITOR (ED)	1	4433.681	117.69	0.0001
ED X ORD	1	272.222	7.23	0.0177
ED X SUBJ/ORD	14	527.431		
EDITING FUNCTION(EF)	8	17643.090	31.01	0.0001
EF X ORD	8	326.965	0.57	0.7968
EF X SUBJ/ORD	112	7965.389		
ED X EF	8	2486.757	5.91	0.0001
ED X EF X ORD	8	1102.715	2.62	0.0115
ED X EF X SUBJ/ORD	112	5894.194		
TOTAL	287	45363.378		

Transfer of Training

Dependent Measure: Number of Errors Committed

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
ORDER (ORD)	1	0.000	0.00	1.0000
SUBJECTS (SUBJ/ORD)	14	97.986		
<u>Within Subjects</u>				
EDITOR (ED)	1	0.681	0.22	0.6496
ED X ORD	1	32.000	10.13	0.0066
ED X SUBJ/ORD	14	44.208		
EDITING FUNCTION (EF)	8	765.924	31.81	0.0001
EF X ORD	8	28.938	1.20	0.3045
EF X SUBJ/ORD	112	337.139		
ED X EF	8	63.757	2.73	0.0086
ED X EF X ORD	8	14.938	0.64	0.7420
ED X EF X SUBJ/ORD	112	326.417		
TOTAL	287	1711.986		

Transfer of Training, Error Data

Dependent Measure: Error Correction Time

Source	df	SS	F-ratio	p-value
Between Subjects				
ORDER (ORD)	1	4.253	0.01	0.9185
SUBJECTS (SUBJ/ORD)	14	5489.493		
Within Subjects				
EDITOR (ED)	1	0.281	0.00	0.9708
ED X ORD	1	2150.587	10.63	0.0057
ED X SUBJ/ORD	14	2832.521		
EDITING FUNCTION (EF)	8	18489.875	7.62	0.0001
EF X ORD	8	1823.528	0.75	0.6459
EF X SUBJ/ORD	112	33965.819		
ED X EF	8	5823.000	2.01	0.0515
ED X EF X ORD	8	3454.819	1.19	0.3100
ED X EF X SUBJ/ORD	112	40562.292		
TOTAL	287	114596.469		

Transfer of Training

Dependent Measure: Bipolar Adjective Scale Ratings

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
ORDER (ORD)	1	20.014	0.19	0.6668
SUBJECTS (SUBJ/ORD)	14	1448.518		
<u>Within Subjects</u>				
EDITOR (ED)	1	65.836	18.13	0.0008
ED X ORD	1	17.940	4.94	0.0432
ED X SUBJ/ORD	14	50.849		
EDITING FUNCTION(EF)	8	97.405	3.32	0.0019
EF X ORD	8	34.541	1.18	0.3200
EF X SUBJ/ORD	112	411.258		
SCALE (SC)	11	381.052	9.54	0.0001
SC X ORD	11	57.066	1.43	0.1647
SC X SUBJ/ORD	154	559.295		
ED X EF	8	42.031	2.59	0.0125
ED X EF X ORD	8	19.958	1.23	0.2892
ED X EF X SUBJ/ORD	112	227.510		
ED X SC	11	28.911	3.48	0.0003
ED X SC X ORD	11	12.362	1.49	0.1408
ED X SC X SUBJ/ORD	154	116.380		
EF X SC	88	150.581	2.57	0.0001
EF X SC X ORD	88	50.973	0.87	0.7965
EF X SC X SUBJ/ORD	1232	820.242		
ED X EF X SC	88	62.066	1.06	0.3381
ED X EF X SC X ORD	88	52.833	0.90	0.7282
ED X EF X SC X SUBJ/ORD	1232	662.823		
TOTAL	3455	5390.444		

Transfer of Training, Subjective Data DAY2-DAY4b

Dependent Measure: Bipolar Adjective Scale Ratings

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
EDITOR (ED)	1	42.003	0.57	0.4610
SUBJECTS (SUBJ/ED)	14	1023.502		
<u>Within Subjects</u>				
DAY	1	0.396	0.12	0.7325
DAY X ED	1	8.069	2.48	0.1379
DAY X SUB(ED)	14	45.613		
EDITING FUNCTION(EF)	8	57.803	1.82	0.0813
EF X ED	8	83.557	2.62	0.0113
EF X SUBJ/ED	112	445.733		
SCALE (SC)	11	339.517	9.36	0.0001
SC X ED	11	25.388	0.70	0.7384
SC X SUBJ/ED	154	507.922		
DAY X EF	8	29.601	2.11	0.0407
DAY X EF X ED	8	8.386	0.60	0.7786
DAY X EF X SUBJ/ED	112	196.642		
DAY X SC	11	23.496	3.18	0.0007
DAY X SC X ED	11	20.725	2.80	0.0024
DAY X SC X SUBJ/ED	154	103.533		
EF X SC	88	179.600	3.15	0.0001
EF X SC X ED	88	59.331	1.04	0.3793
EF X SC X SUBJ/ED	1232	797.531		
DAY X EF X SC	88	33.788	0.86	0.8229
DAY X EF X SC X ED	88	47.850	1.21	0.0942
DAY X EF X SC X SUBJ/ED	1232	552.339		
TOTAL	3455	4632.389		

Day 2 Compared to Post-Experiment

Dependent Measure: USELESS/USEFUL

Source	df	SS	F-ratio	p-value
Between Subjects				
EDITOR (ED)	1	0.014	0.01	0.9936
SUBJECTS/ED (SUBJ/ED)	14	26.986		
Within Subjects				
DAY (D)	1	1.125	6.96	0.0195
D X ED	1	1.389	8.59	0.0110
D X SUBJ/ED	14	2.264		
EDITING FUNCTION (EF)	8	39.188	5.53	0.0001
EF X ED	8	8.549	1.21	0.3021
EF X SUBJ/ED	112	99.264		
D X EF	8	0.438	0.15	0.9962
D X EF X ED	8	0.549	0.19	0.9917
D X EF X SUB/ED	112	40.236		
TOTAL	287	220.000		

Day 2 Compared to Post-Experiment

Dependent Measure: UNDEPENDABLE/DEPENDABLE

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
EDITOR (ED)	1	0.781	0.12	0.7333
SUBJECTS/ED (SUBJ/ED)	14	90.549		
<u>Within Subjects</u>				
DAY (D)	1	2.531	4.21	0.0593
D X ED	1	0.003	0.01	0.9405
D X SUBJ/ED	14	8.410		
EDITING FUNCTION (EF)	8	13.250	1.69	0.1075
EF X ED	8	13.125	1.68	0.1115
EF X SUBJ/ED	112	109.514		
D X EF	8	2.750	0.65	0.7380
D X EF X ED	8	4.153	0.97	0.4596
D X EF X SUB/ED	112	59.653		
TOTAL	287	304.719		

Day 2 Compared to Post-Experiment

Dependent Measure: INCONSISTENT/CONSISTENT

Source	df	SS	F-ratio	p-value
Between Subjects				
EDITOR (ED)	1	0.018	0.00	0.9698
SUBJECTS/ED (SUBJ/ED)	14	131.153		
Within Subjects				
DAY (D)	1	0.056	0.07	0.7935
D X ED	1	0.014	0.01	0.9405
D X SUBJ/ED	14	10.931		
EDITING FUNCTION (EF)	8	23.278	2.28	0.0266
EF X ED	8	9.986	0.98	0.4564
EF X SUBJ/ED	112	142.847		
D X EF	8	3.444	0.45	0.8853
D X EF X ED	8	13.486	1.78	0.0883
D X EF X SUB/ED	112	106.069		
TOTAL	287	441.278		

Day 2 Compared to Post-Experiment

Dependent Measure: UNINTERPRETABLE/INTERPRETABLE

Source	df	SS	F-ratio	p-value
Between Subjects				
EDITOR (ED)	1	0.087	0.01	0.9304
SUBJECTS/ED (SUBJ/ED)	14	153.771		
Within Subjects				
DAY (D)	1	0.781	1.00	0.3343
D X ED	1	0.014	0.00	0.9478
D X SUBJ/ED	14	10.931		
EDITING FUNCTION (EF)	8	14.632	1.81	0.0820
EF X ED	8	8.882	1.10	0.3686
EF X SUBJ/ED	112	113.041		
D X EF	8	3.813	0.56	0.8100
D X EF X ED	8	5.340	0.78	0.6195
D X EF X SUB/ED	112	95.625		
TOTAL	287	406.913		

Day 2 Compared to Post-Experiment

Dependent Measure: DIFFICULT/SIMPLE

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
EDITOR (ED)	1	1.681	0.23	0.6417
SUBJECTS/ED (SUBJ/ED)	14	104.042		
<u>Within Subjects</u>				
DAY (D)	1	0.889	0.82	0.3810
D X ED	1	6.125	5.64	0.0324
D X SUBJ/ED	14	15.208		
EDITING FUNCTION (EF)	8	28.438	3.78	0.0006
EF X ED	8	18.132	2.41	0.0193
EF X SUBJ/ED	112	105.208		
D X EF	8	15.299	2.68	0.0100
D X EF X ED	8	4.438	0.78	0.6245
D X EF X SUB/ED	112	80.042		
TOTAL	287	379.500		

Day 2 Compared to Post-Experiment

Dependent Measure: UNSAFE/UNSAFE

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
EDITOR (ED)	1	1.681	0.17	0.6890
SUBJECTS/ED (SUBJ/ED)	14	140.875		
<u>Within Subjects</u>				
DAY (D)	1	0.056	0.05	0.8344
D X ED	1	4.014	3.28	0.0918
D X SUBJ/ED	14	17.153		
EDITING FUNCTION (EF)	8	13.611	2.39	0.0206
EF X ED	8	8.069	1.41	0.1981
EF X SUBJ/ED	112	79.875		
D X EF	8	7.444	2.34	0.0232
D X EF X ED	8	5.736	1.80	0.0842
D X EF X SUB/ED	112	44.597		
TOTAL	287	323.111		

Day 2 Compared to Post-Experiment

Dependent Measure: SLOW/FAST

Source	df	SS	F-ratio	p-value
Between Subjects				
EDITOR (ED)	1	6.420	1.21	0.2890
SUBJECTS/ED (SUBJ/ED)	14	73.993		
Within Subjects				
DAY (D)	1	3.337	4.48	0.0528
D X ED	1	5.837	7.83	0.0142
D X SUBJ/ED	14	10.438		
EDITING FUNCTION (EF)	8	33.340	2.93	0.0053
EF X ED	8	8.433	0.74	0.6568
EF X SUBJ/ED	112	159.569		
D X EF	8	5.382	1.05	0.4046
D X EF X ED	8	5.632	1.10	0.3707
D X EF X SUB/ED	112	71.875		
TOTAL	287	384.247		

Day 2 Compared to Post-Experiment

Dependent Measure: UNNATURAL/NATURAL

Source	df	SS	F-ratio	p-value
Between Subjects				
EDITOR (ED)	1	21.125	1.11	0.3092
SUBJECTS/ED (SUBJ/ED)	14	265.597		
Within Subjects				
DAY (D)	1	7.347	8.54	0.0111
D X ED	1	0.500	0.58	0.4585
D X SUBJ/ED	14	12.042		
EDITING FUNCTION (EF)	8	23.375	2.81	0.0072
EF X ED	8	25.750	3.09	0.0035
EF X SUBJ/ED	112	116.653		
D X EF	8	3.778	0.72	0.6713
D X EF X ED	8	2.125	0.41	0.9149
D X EF X SUB/ED	112	73.208		
TOTAL	287	551.500		

Day 2 Compared to Post-Experiment

Dependent Measure: INCOMPLETE/COMPLETE

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
EDITOR (ED)	1	7.031	0.79	0.3905
SUBJECTS/ED (SUBJ/ED)	14	125.354		
<u>Within Subjects</u>				
DAY (D)	1	0.032	0.05	0.8324
D X ED	1	0.281	0.42	0.5282
D X SUBJ/ED	14	9.410		
EDITING FUNCTION (EF)	8	5.500	0.95	0.4768
EF X ED	8	4.000	2.68	0.0098
EF X SUBJ/ED	112	80.833		
D X EF	8	8.625	0.69	0.6972
D X EF X ED	8	3.125	0.97	0.4619
D X EF X SUB/ED	112	45.028		
TOTAL	287	289.219		

Day 2 Compared to Post-Experiment

Dependent Measure: DISGUSTING/PLEASING

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
EDITOR (ED)	1	8.681	0.66	0.4285
SUBJECTS/ED (SUBJ/ED)	14	182.750		
<u>Within Subjects</u>				
DAY (D)	1	5.556	3.12	0.0992
D X ED	1	4.500	2.53	0.1343
D X SUBJ/ED	14	24.944		
EDITING FUNCTION (EF)	8	16.486	3.21	0.0025
EF X ED	8	19.194	3.74	0.0007
EF X SUBJ/ED	112	71.875		
D X EF	8	2.069	0.70	0.6915
D X EF X ED	8	4.500	1.52	0.1579
D X EF X SUB/ED	112	41.431		
TOTAL	287	381.986		

Day 2 Compared to Post-Experiment

Dependent Measure: UNCOOPERATIVE/COOPERATIVE

Source	df	SS	F-ratio	p-value
<u>Between Subjects</u>				
EDITOR (ED)	1	9.753	0.96	0.3444
SUBJECTS/ED (SUBJ/ED)	14	142.604		
<u>Within Subjects</u>				
DAY (D)	1	1.837	1.67	0.2177
D X ED	1	0.003	0.00	0.9560
D X SUBJ/ED	14	15.438		
EDITING FUNCTION (EF)	8	7.549	1.66	0.1172
EF X ED	8	4.840	1.06	0.3952
EF X SUBJ/ED	112	63.833		
D X EF	8	5.132	1.67	0.1132
D X EF X ED	8	4.090	1.33	0.2351
D X EF X SUB/ED	112	43.000		
TOTAL	287	298.080		

Day 2 Compared to Post-Experiment

Dependent Measure: UNSATISFACTORY/SATISFACTORY

Source	df	SS	F-ratio	p-value
Between Subjects				
EDITOR (ED)	1	10.125	1.15	0.2391
SUBJECTS/ED (SUBJ/ED)	14	93.750		
Within Subjects				
DAY (D)	1	0.347	0.41	0.5343
D X ED	1	6.125	7.16	0.0181
D X SUBJ/ED	14	11.972		
EDITING FUNCTION (EF)	8	18.757	2.61	0.0119
EF X ED	8	13.938	1.94	0.0613
EF X SUBJ/ED	112	100.750		
D X EF	8	5.215	1.51	0.1607
D X EF X ED	8	3.063	0.89	0.5290
D X EF X SUB/ED	112	48.278		
TOTAL	287	312.319		

The three page vita has been removed from the scanned document. Page 1 of 3

The three page vita has been removed from the scanned document. Page 2 of 3

The three page vita has been removed from the scanned document. Page 3 of 3

EXAMINING THE RELATIONSHIP BETWEEN PERFORMANCE MEASURES AND USER EVALUATIONS IN A TRANSFER OF TRAINING PARADIGM

by

William D. Coleman

ABSTRACT

User evaluations which generate detailed information can identify problematic aspects of software interfaces. In a preliminary study (Coleman, Wixon, and Williges, 1984), a methodology was developed for the systematic collection of detailed subjective evaluations of software interfaces. This methodology created a taxonomy of editing functions for users to evaluate and a set of bipolar scales on which they could make their evaluations. The present research investigated the utility of this methodology, while comparing two text editors within the context of a benchmark editing task. In addition, the detailed subjective measures collected were compared with more traditional objective measures.

The results of this research revealed that global subjective evaluations were insensitive to differences between two editors indicated by detailed evaluations. Examination of the detailed subjective evaluations indicated that the differences between editors could be

attributed to specific editing functions. The objective measures also indicated very specific differences between the two evaluated editors. Examination of the relationship between the objective and subjective measures indicated that the measures differed on both the magnitude and location of effects. Closer inspection of the data revealed that insensitivity on the part of the subjective measures could not account for all disagreement between measures. On several occasions the objective and subjective measures seemed to measure qualitatively different effects. Given that the measures were not completely redundant it was concluded that both objective and subjective measures should be collected during interface evaluation.