Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface

John P. Chin, Virginia A. Diehl and Kent L. Norman

Human-Computer Interaction Laboratory
Department of Psychology
University of Maryland
College Park, MD 20742

ABSTRACT

This study is a part of a research effort to develop the Questionnaire for User Interface Satisfaction (QUIS). Participants, 150 PC user group members, rated familiar software products. Two pairs of software categories were compared: 1) software that was liked and disliked, and 2) a standard command line system (CLS) and a menu driven application (MDA). The reliability of the questionnaire was high, Cronbach's alpha=.94. The overall reaction ratings yielded significantly higher ratings for liked software and MDA over disliked software and a CLS, respectively. Frequent and sophisticated PC users rated MDA more satisfying, powerful and flexible than CLS. Future applications of the QUIS on computers are discussed.

KEYWORDS: User Satisfaction, User Interface Questionnaire, Design Tool

INTRODUCTION

There are many possible ways to evaluate the human-computer interface. There are five different types of dependent measures for evaluating interfaces [10]. For many tasks, speed and accuracy are two related performance measures which affect a person's attitude toward the system. The time it takes to learn a system and the retention of acquired knowledge over time also affect the utility of a system. User acceptance of a system (i.e., subjective satisfaction) is also a critical measure of a system's success. Although a system may be evaluated favorably on every performance measure, the system may not be used very much because of the user's dissatisfaction with the system and its interface.

A large number of questionnaires have been developed to assess the user's subjective satisfaction of the system and related issues. However, few have focused exclusively on user evaluations of the interface. This paper concerns the development of a measurement tool, called the Questionnaire for User Interface Satisfaction (QUIS). QUIS measures the user's subjective rating of the human-computer interface. A brief literature review will be presented, followed by a description of QUIS's development in a previous study. The present study, involving the administration of the current QUIS (5.0) to a large user group, will then be discussed.

Review of the Literature

In the past, several questionnaires have been developed to assess users' perceptions of systems. Recently, literature reviews found weaknesses in many of the subjective evaluation measurement tools [3, 5]. Problems ranged from a lack of validation [4] to low reliabilities [6]. Problems with respondents marking the same response for many of the questions inflated reliability values [5]. One study suffered from a small sample size and a nonrepresentative population [1]. Thus, the range of problems has been diverse.

Past studies have examined the types of questions that would be appropriate for questionnaires. Checklist questionnaires were not sufficient in evaluating systems since they did not indicate what new features were needed [8]. Open-ended questions were suggested as a possible supplement for checklists. Users preferred concrete adjectives for evaluations [2]. In addition, specific evaluation questions appeared to be more accurate than global satisfaction questions.

In general, the research regarding questionnaires for evaluating computer systems has steadily improved by increasing sample size and the number of

different systems evaluated. In one study, 4,597 respondents evaluated 179 different systems ranging from micros to large mainframes [9]. Many have demonstrated more concern for reliability and validity issues. However, few studies have had a sustained development of a questionnaire, and although many of the surveys consider several issues associated with general subjective satisfaction of the system, few if any directly focus on the interface. This research effort attempts to address these issues.

Review of the Development Process

The original questionnaire consisted total of 90 questions [10]. Five questions were overall reaction ratings of the system. The remaining 85 items were organized into 20 different groups, which had a main component question followed by related subcomponent questions. The questionnaire's short version had only the 20 main questions listed along with the five overall questions. Each of the questions had rating scales ascending from 1 on the left to 10 on the right and anchored at both endpoints with adjectives (e.g., inconsistent/consistent). These adjectives were always positioned so that the scale went from negative on the left to positive on the right. In addition, each item had "not applicable" as a choice. Instructions also encouraged raters to include any written comments. No empirical work was done to assess its reliability or validity.

The original questionnaire was modified and expanded to three sections in the Questionnaire for User Interface Satisfaction (QUIS 3.0). In Section I, three questions concerned the type of system being rated and the amount of time spent on the system. In Section II, four questions dealt with the user's past computer experience. Section III was a modified version of the original questionnaire with the rating scales changed from 1 through 10 to 1 through 9.

QUIS's generalizability could be established by having different user populations evaluate different systems. The development of QUIS included respondents who were: 1) students, 2) computer professionals, 3) computer hobbyists, and 4) novice users. Moreover, it was important to administer the QUIS under different experimental conditions: 1) strictly controlled experiments with a small number of subjects exposed to a system for a very short period of time, 2) less rigidly controlled manipulations with a medium number of participants who used a system for a limited time, and 3) a field study having no control with volunteers who have used a system extensively. The characteristics of versions 3.0 and 4.0 were based on student evaluations of a system in a moderately controlled situation. The present sampling domain included computer professionals and hobbyists who had extensive and uncontrolled use of the evaluated systems.

Since a questionnaire's reliability is related to the number of items and scaling steps, the larger the number of items and scaling steps, the higher the reliability of the questionnaire [7]. Thus, QUIS began with a large number of questions. In addition, 10 point scales were used since more than 10 steps would contribute little to reliability [7]. However, QUIS had to be shortened to improve the percentage of completed questionnaires. Thus, successive versions of the questionnaire had less items while maintaining a high reliability.

The QUIS (3.0) and a subsequent revised version (4.0) was administered to an introductory computer science class learning to program in CF PASCAL [3]. Participants, 155 males and 58 females, were assigned to either the interactive batch run IBM mainframe or an interactive syntax-directed editor programming environment on an IBM PC. During class time, they evaluated the environment they had used during the first 6 weeks of the course with QUIS (3.0). A multiple regression of the sub-component questions of QUIS (3.0) with each main component question was used to eliminate questions with low beta weights, thereby reducing the number of ratings from 103 to 70 in QUIS (4.0) while retaining the same basic organization. Next, they evaluated the other environment with QUIS (4.0) after switching to the other environment for six weeks.

The participants' exam and project grades were used as a reference point for establishing validity since an effective interface might be associated with better performance. Higher satisfaction ratings and performance were expected for the interactive syntax-directed editor programming environment. However, subjective ratings did not correspond with the students' performance in the class. Problems in the syntax-directed editor's interface had led to higher satisfaction ratings for the mainframe. Although the performance measures (exam and project grades) failed to help establish validity, QUIS diagnosed interface problems in the syntax-editing programming environment.

QUIS's reliability was high. Cronbach's alpha, an estimation of reliability based on the average intercorrelation among items, indicated that QUIS (3.0) had an overall reliability of .94, with interitem alpha values varying by .002. QUIS (4.0) had an overall reliability of .89, with the

values of alpha ranging between .89 and .90. A small drop in reliability in QUIS (4.0) was still respectable considering that 33 items had been eliminated. The small variability of the alpha of each item indicates high internal consistency.

The Present Study

Although QUIS (4.0) was reliable, the sample of students evaluating a programming environment's interface limited the results' generalizability to the academic community. This study examined the reliability of QUIS (5.0) with other user populations (professionals & hobbyists) and products (commercially distributed software).

Four groups rated the following: 1) a liked product; 2) a disliked product; 3) a command line system (CLS); and 4), a Menu Driven Application (MDA). This study examines the reliability and external validity of QUIS, by comparing the ratings from the liked vs. disliked groups and from a mandatory CLS and a voluntarily chosen MDA.

Questionnaire for User Interface Satisfaction (5.0)

• OVERALL REACTIONS TO THE SOFTWARE

terrible wonderful
0 1 2 3 4 5 6 7 8 9

difficult easy
0 1 2 3 4 5 6 7 8 9

frustrating satisfying
0 1 2 3 4 5 6 7 8 9

inadequate power adequate power
0 1 2 3 4 5 6 7 8 9

dull stimulating
0 1 2 3 4 5 6 7 8 9

rigid flexible
0 1 2 3 4 5 6 7 8 9

• SCREEN

Characters on the computer screen
hard to read easy to read
0 1 2 3 4 5 6 7 8 9

Highlighting on the screen simplifies task
not at all very much
0 1 2 3 4 5 6 7 8 9

Organization of information on screen
confusing very clear
0 1 2 3 4 5 6 7 8 9

Sequence of screens
confusing very clear
0 1 2 3 4 5 6 7 8 9

• TERMINOLOGY AND SYSTEM INFORMATION

Use of terms throughout system
inconsistent consistent
0 1 2 3 4 5 6 7 8 9

Computer terminology is related to the task you are doing
never always
0 1 2 3 4 5 6 7 8 9

Position of messages on screen
inconsistent consistent
0 1 2 3 4 5 6 7 8 9

Messages on screen which prompt user for input
confusing clear
0 1 2 3 4 5 6 7 8 9

Computer keeps you informed about what it is doing
never always
0 1 2 3 4 5 6 7 8 9

Error messages
unhelpful helpful
0 1 2 3 4 5 6 7 8 9

• LEARNING

Learning to operate the system
difficult easy
0 1 2 3 4 5 6 7 8 9

Exploring new features by trial and error
difficult easy
0 1 2 3 4 5 6 7 8 9

Remembering names and use of commands
difficult easy
0 1 2 3 4 5 6 7 8 9

Tasks can be performed in a straight-forward manner
never always
0 1 2 3 4 5 6 7 8 9

Help messages on the screen
unhelpful helpful
0 1 2 3 4 5 6 7 8 9

Supplemental reference materials
confusing clear
0 1 2 3 4 5 6 7 8 9

• SYSTEM CAPABILITIES

System speed
too slow fast enough
0 1 2 3 4 5 6 7 8 9

System reliability
unreliable reliable
0 1 2 3 4 5 6 7 8 9

System tends to be
noisy quiet
0 1 2 3 4 5 6 7 8 9

Correcting your mistakes
difficult easy
0 1 2 3 4 5 6 7 8 9

Experienced and inexperienced users' needs are taken into consideration
never always
0 1 2 3 4 5 6 7 8 9

METHOD

Subjects

The participants, 127 males, 14 females, and 9 not reporting their gender, were members/affiliates of a local PC User's Group, ranging from ages 14 to 78. The level of computer experience varied widely; 11% had used only PC-DOS systems and 32% had used over six other types. More than 75% of the respondents had used a word processor, file manager, spreadsheet, modem, and a hard disk drive. Among these respondents, 27 rated the command line system, MS-DOS™, and 25 evaluated the menu driven system, WordPerfect™. In addition, 35 respondents rated a software product they liked and 18 evaluated one they disliked. A total of 46 different products were evaluated.

Materials

Participants completed a short version of QUIS (5.0) consisting of 27 rating scales with a 10 point scale from 0 to 9. Number two pencils were used to mark optical scanning sheets containing 10 alternatives for each question. The background information section of QUIS 4.0 was altered to suit the software and hardware configurations being evaluated. A principle component factor analysis of the data from versions 3.0 and 4.0 lead to a reorganization of the main component questions. QUIS (3.0) had 7 factors while QUIS 4.0 had 6 factors. Each group of items was given a heading based on an aspect of the user interface being described. When an item did not clearly fall within a factor, intuition determined the placement of an item under a particular heading. An item concerning the system noisiness was added for a total of 21 main component items.

Procedure

Distribution of about 500 questionnaires during the group's monthly meeting occurred as attendees entered the auditorium. Four different instructions accompanied the questionnaire which asked raters to evaluate: 1) a product they liked, 2) a product they disliked, 3) MS-DOS™, and 4) WordStar™, WordPerfect™, Lotus™, DBase™ or any comparable software product. Next, a prepared statement was read to the participants, who read and followed the instructions on the questionnaire's cover page. After the meeting, about 30% of the questionnaires were completed and returned.

RESULTS

Reliability

The overall reliability of version 5.0 using Cronbach's alpha was .94. Interitem alpha values varied by only .006. The mean ratings varied between 4.72 and 7.02, while standard deviations ranged from 1.67 to 2.25.

Factor Analysis

A principle components factor analysis was performed on the 21 main component questions to determine if the factor analysis of versions 3.0 and 4.0 corresponded with the data from version 5.0 (See Table 1). The items under the Learning and System Capabilities headings matched, with the exception of "experienced and inexperienced users' needs are taken into consideration" which factored with the Learning items. The items under Terminology and System Information were grouped together with the exceptions of "computer keeps you informed of what it is doing" and "error messages." The items under the Screen heading did not match the original organization. The four latent factors may be named: 1) Learning, 2) Terminology and Information flow, 3) System Output, and 4) System Characteristics, respectively. Both "error messages" and "highlighting" do not fit any of the four factors very well.

Table 1

Sorted Rotated Factor Loadings of QUIS 5.0

Header	Question	F1	F2	F3	F4
Learn	Learning the system	.84	.00	.00	.00
Learn	Remembering names...	.78	.28	.00	.00
Learn	Exploring ...by trial & error	.75	.00	.28	.00
Sys	Experienced & inexperienced users' needs ...consideration	.66	.31	.00	.28
Learn	Tasks are straight-forward	.64	.38	.31	.00
Learn	Reference materials	.61	.27	.00	.27
Learn	Help messages	.58	.46	.00	.00
Terms	Use of sys terms	.00	.79	.00	.27
Terms	Position of messages	.31	.79	.25	.00
Screen	Organization of screen	.26	.77	.29	.00
Screen	Sequence of screens	.40	.72	.00	.00
Terms	Terminology is task related	.00	.68	.00	.27
Terms	Prompts for user input	.44	.61	.35	.00
Terms	Computer informs you...	.00	.35	.74	.00
Screen	Characters on screen	.00	.29	.69	.00
Sys	System speed	.00	.00	.65	.50
Sys	System tends ...noisy/quiet	.00	.00	.00	.79
Sys	System reliability	.00	.00	.44	.69
Sys	Correcting your mistakes	.49	.39	.00	.58
Terms	Error messages	.43	.40	.48	.00
Screen	Highlighting on the screen simplifies task	.48	.42	.00	.00

Note: Loadings=0 if less than 0.25. N=96.

Liked vs. Disliked

Liked and disliked ratings were compared on the six overall reactions and 21 main component questions

(See Table 2). All of the means from the liked system evaluations were higher than those from the disliked systems. Liked ratings were significantly ($p<.001$) higher than disliked in the overall reactions: 1) "terrible / wonderful," 2) "frustrating / satisfying," 3) "dull / stimulating," and 4) "rigid / flexible." Although the main component questions were not significantly different at $p<.001$, the differences were in the same direction on "learning to operate the system," "exploring new features by trial and error," tasks can be done in a straaight-forward manner," "system speed," "system reliability," "error correction," and "experienced and inexperienced users" at $p<.05$.

Table 2

Like vs. Dislike Groups Mean Ratings

	Like	Dislike	
Overall Reactions to the System			
terrible/wonderful	7.21	4.44	***
frustrating/satisfying	7.12	3.29	***
dull/stimulating	6.68	3.75	***
inadequate power/adequate power	7.00	5.06	**
rigid/flexible	6.28	3.52	***
Learning			
Learning to operate the system	5.67	3.53	**
Exploring ...by trial and error	5.62	3.76	**
Tasks can be...straight-forward	5.94	4.65	*
System Capabilities			
System speed	6.09	4.29	*
System reliability	7.45	6.35	*
Correcting your mistakes	6.64	4.71	**
Experienced & inexperienced users' needs...consideration	5.63	3.88	**

note: * for $p<.05$, ** for $p<.01$, *** for $p<.001$

Command Line Systems & Menu Driven Applications

CLS and MDA ratings were compared in an item analysis. T-tests performed on the overall reaction and the main component questions revealed many differences (See Table 3). In general, all the MDA mean ratings were higher than CLS. All of the overall reaction items were significant at the .001 level, with the exception of "easy/difficult" and "inadequate power/adequate power." Eight of the 21 main component items were significantly different ($p<.001$): 1) "information organization," 2) "screen sequence," 3) "position of messages, 4) "status of computer," 5) "error messages," 6) "help," 7) "error correction," and 8) "experienced and inexperienced users."

Table 3

Mean Ratings of Command Line System & Menu Driven Applications

	CLS	MDA	
Overall Reactions to the System			
terrible/wonderful	5.33	7.36	***
frustrating/satisfying	5.07	6.84	***
dull/stimulating	4.65	5.83	*
inadequate power/adequate power	4.96	7.75	***
rigid/flexible	4.33	6.88	***
Screen			
Characters on the computer screen	6.08	7.62	*
Highlighting ...simplifies task	5.00	6.72	*
Organization of... screen	4.36	7.40	***
Sequence of screens	5.18	7.20	***
Terminology and System Information			
Use of terms throughout system	6.42	7.54	*
Computer terminology is related to the task you are doing	5.46	6.63	*
Position of messages on screen	6.00	8.04	***
Messages prompting ...input	4.77	6.44	**
Computer keeps you informed...	4.19	6.71	***
Error messages	3.54	5.80	***
Learning			
Learning to operate the system	3.56	5.08	**
Exploring ...by trial and error	4.35	5.56	*
Tasks can be performed in a straight-forward manner	4.74	6.16	**
Help messages on the screen	3.74	6.16	***
Supplemental reference materials	4.30	5.84	**
System Capabilities			
System speed	5.31	6.84	**
Correcting your mistakes	5.24	7.04	***
Experienced & inexperienced users' needs are taken into consideration	3.80	6.00	***

note: * for $p<.05$, ** for $p<.01$, *** for $p<.001$

DISCUSSION

Summary of the Results

Successive versions of QUIS maintained a high reliability as the number of items decreased. Low variability of the reliability values indicate high internal consistency. A factor analysis revealed that both the learning and terminology sections corresponded well with the latent factors. System capability questions appeared to break down into two different factors: one concerning the system output and the other focusing on system characteristics. However, two questions concerning error messages and highlighting on the screen did not seem to fit any category. The item analyses show that the QUIS has good discriminability in the overall reaction ratings for the following: 1) like vs. dislike and 2) command

line system (CLS) vs. menu driven application (MDA). Both like and MDA groups had consistently higher ratings compared to dislike and CLS groups, respectively.

Although there were strong differences found between like vs. dislike and CLS vs. MDA in the overall reactions, more significant differences were found between CLS vs. MDA in the specific main component questions in comparison to the like vs. dislike group. Evaluating a large number of different software products may account for the lack of significant differences in the like vs. dislike groups, since each product differs in its strengths and weaknesses. Aggregation of the evaluations across different products may have cancelled the rating differences between like and dislike groups.

MDA was rated higher than CLS for many reasons. Shneiderman (1987) lists five advantages of MDA: 1) shortening of learning time, 2) reduction of keystrokes, 3) structuring of decision-making, 4) permitting the use of dialog management, and 5) support for error handling. MDA's higher ratings in error messages, help, and feedback is evidence of MDA's superiority in dialog management and error handling. Moreover, MDA's higher ratings in organization of screen information and the sequence of screens indicate that good design of menus can avoid the pitfalls of MDA: 1) slowing down frequent users and 2) getting lost while navigating through the menus. Surprisingly, although CLSs are known for their flexibility and appeal to "power" users [10], the overall ratings of MDA suggests that these frequent and sophisticated users rated a MDA more satisfying, powerful and flexible than a CLS.

Future Studies

Although this study established external validity, no attempt to establish any construct or predictive validity was done. There are two reasons for the difficulty in establishing validity: 1) a lack of theoretical constructs about interfaces to test with QUIS, and 2) a lack of established questionnaires for cross-validating purposes. Future validation studies of the questionnaire include the use of a standard interface to calibrate QUIS ratings. Calibration can be accomplished by comparing successive ratings with corresponding degradations of an interface standard. Comparisons between respondents along successive ratings in calibration will also allow assessment of interrater reliability of the questionnaire.

All previous questionnaires have been paper and pencil tasks. A computerized questionnaire would facilitate customizing of questions for particular systems and data collection by eliminating encoding errors. Presently a computerized IBM™ PC version of QUIS has been implemented and distributed.

ACKNOWLEDGEMENTS

We thank Yuri Gawdiak and Steven Versteeg for their help collecting data. Funding was provided by NSF, AT&T and the University of Maryland Computer Science Center.

REFERENCES

1. Bailey, J. E., & Pearson, S. W. Development of a tool for measuring and analyzing computer user satisfaction. *Management Science*, 29, 5, (May 1983), 530-545.

2. Coleman, W. D., Williges, R. C., & Wixon, D. R. Collecting detailed user evaluations of software interfaces. *Proceedings of the Human Factors Society - 29th Annual Meeting - 1985*, 240-244.

3. Chin, J. P., Norman, K. L., & Shneiderman, B. *Subjective user evaluation of CF PASCAL programming tools.* Technical Report (CAR-TR-304), Human-Computer Interaction Laboratory, University of Maryland, College Park, MD 20742, 1987.

4. Gallagher, C. A. Perceptions of the value of a management information system. *Academy of Mangement Journal*, 17, 1, (1974), 46-55.

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

6. Larcker, D. F. & Lessig, V. P. Perceived usefulness of information: A psychometric examination. *Decision Science*, 11, 1, (1980), 121-134.

7. Nunnally, J. C. *Psychometric Theory*, McGraw-Hill Book Company, New York, 1978.

8. Root, R. W., & Draper, S. Questionnaires as a software evaluation tool. *CHI'83 Proceedings*, December, 83-87, (1983).

9. Rushinek, A. & Rushinek, S. F. What makes users happy? *Communications of the ACM*, 29, 7, (1986), 594-598.

10. Shneiderman, B. Designing the User Interface: *Strategies for Effective Human-Computer Interaction.* Addison-Wesley Publishing Co., Reading, MA, 1987.