PSYCHOMETRIC EVALUATION OF AN AFTER-SCENARIO QUESTIONNAIRE FOR COMPUTER USABILITY STUDIES: THE ASQ

JAMES R. LEWIS

Abstract: A three-item after-scenario questionnaire was used in three related usability tests in different areas of the United States. The studies had eight scenarios in common. After participants finished a scenario, they completed the After-Scenario Questionnaire (the ASQ). A factor analysis of the responses to the ASQ items revealed that an eight-factor solution explained 94 percent of the variability of the 24 (eight scenarios by three items per scenario) items. The varimax-rotated factor pattern showed that these eight factors were clearly associated with the eight scenarios. The benefit of this research to system designers is that this three-item questionnaire has acceptable psychometric properties of reliability, sensitivity, and concurrent validity, and may be used with confidence in other, similar usability studies.

INTRODUCTION

When developing computer systems, it is important to develop a method for measuring user satisfaction with existing systems or prototypes of future systems. The purpose of this report is to describe a psychometric evaluation of a questionnaire which has been used to assess user satisfaction during participation in scenario-based usability studies. This questionnaire is named the After-Scenario Questionnaire (ASQ), since it is administered after each scenario.

Psychometric instruments to evaluate computer-user satisfaction are not new (see Ives, Olson, and Baroudi, 1983 for a review). Lewis (1990) has recently reported favorable psychometric properties of the Post-Study System Usability Questionnaire (PSSUQ). Two other new instruments, the Questionnaire for User Interface Satisfaction (QUIS) (Chin, Diehl, and Norman, 1988) and the Computer User Satisfaction Inventory (CUSI) (Kirakowski and Corbett, 1988) also seem to have good reliability and validity characteristics. None of the above instruments were developed specifically for use during a usability study, although the PSSUQ is intended for use after a usability study. The ASQ was developed to be used immediately following scenario completion in scenario-based usability studies, where a scenario is a collection of related tasks.

The purpose of this report is to discuss the steps by which the ASQ was developed and evaluated. The questionnaire items were treated as constituent items for summative, or Likert scales. The methods for developing such summative scales are explained in detail in Nunnally (1978) and McIver and Carmines (1981). The topics to be covered in this report are item construction, item selection, exploratory factor analysis of the items, estimates of scale reliability, assessment of scale sensitivity, and an estimate of concurrent validity for the ASQ.

ITEM CONSTRUCTION

The items are 7-point graphic scales, anchored at the end

points with the terms "Strongly agree" for 1 and "Strongly disagree" for 7, and a Not Applicable (N/A) point outside the scale, as shown in Figure 1. For a discussion of important properties of rating scales, see Nunnally (1978, pp. 594-602).

FIGURE 1. The After-Scenario Questionnaire (ASQ)

For each of the questions below, circle the answer of your choice.

1. Overall, I am satisfied with the ease of completing the tasks in this scenario.

strongly agree <==========> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

2. Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario.

strongly agree <==========> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

3. Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing the tasks?

strongly agree <==========> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

ITEM SELECTION

The three items were selected on the basis of their content regarding hypothesized constituents of usability. Characteristics such as ease of task completion, time required to complete tasks, and satisfaction with support information (on-line help, system messages, documentation) would be expected to influence a user's perception of system usability.

PSYCHOMETRIC EVALUATION

Scenario-based usability studies were conducted to evaluate the usability characteristics of three office application systems. One important characteristic of usability was user satisfaction, measured using the ASQ after each scenario.

Participants

Forty-eight employees of temporary help agencies participated in the studies, with 15 hired in Hawthorne, New York; 15 hired in Boca Raton, Florida; and 18 hired in Southbury, Connecticut. Each set of participants consisted of one-third clerical/secretarial work experience with no mouse experience (SECNO), one-third business professionals with no mouse experience (BPNO), and one-third business professionals with mouse experience (BPMS). All participants had at least three months experience using some type of computer system. They had no programming training or experience, and had no (or very limited) knowledge of the disk operating systems.

Materials and Apparatus

Three office systems (hereafter referred to as System I, System II and System III) were put together by installing a word processing application, a mail application, a calendar application, and a spreadsheet application on three different platforms that allowed a certain amount of integration among the applications. All three platforms allowed windowing and used a mouse as a pointing device. The systems differed in details of implementation, but were generally similar. The three word processing and spreadsheet applications were quite similar, but the mail and calendar applications differed substantially. Packages were prepared for each participant, consisting of a background questionnaire, 10 scenarios (with each scenario followed by the ASQ), and an overall system rating questionnaire (the PSSUQ). Eight of the ten scenarios were common among the usability studies of the three systems, and are listed in Table 1.

TABLE 1. Scenario Descriptions

Scenario	Component Tasks
Mail (M1A)	Open, reply to, and delete a note.
Mail (M1B)	Open, reply to, and delete a note.
Mail (M2)	Open a note, forward with reply, save and print the note.
Address (A1)	Create, change, and delete address entries.
File Management (F1)	Rename, copy, and delete a file.
Editor (E1)	Create and save a short document.
Editor (E2)	Locate and edit a document, open a note, copy text from the note into the document, save and print the document.
Decision Support (D1)	Create a small spreadsheet, open a document, copy the spreadsheet into the document, save and print the document, save the spreadsheet.

Procedure

Participants began with a brief tour of the lab, a description of the study's purpose and events of the day, and completed

the background questionnaire. Participants using System I began by working on an interactive tutorial shipped with the system, while the other participants were given a brief demonstration about how to move, point and select with a mouse; how to open the icons for each product; and how to maximize and minimize windows. After this system-exploration period (usually about one hour), participants performed the scenarios, completing the ASQ as each scenario was finished. As the participant performed the scenario, an observer logged the participant's activities. If the participant completed the scenario without assistance and without unrecoverable errors, the scenario was recorded as successfully completed. Otherwise, it was recorded as unsuccessfully completed. After all scenarios had been completed (or at the end of the day, if some scenarios still had not been done), participants rated the system using the PSSUQ. It usually took a participant a full day (eight hours) to complete the study.

At the end of the three studies, the responses to the PSSUQ, the ASQ, and the scenario completion data were entered into a data base. From this data base, an exploratory factor analysis, reliability analyses, a sensitivity analysis, and some validity analyses were conducted.

RESULTS

Exploratory Factor Analysis and Reliability Analyses

Factor analysis is a statistical procedure which examines the correlations among variables to test for, or discover clusters of variables (Nunnally, 1978). Since summated (Likert) scales are more reliable than single-item scales (Nunnally, 1978) and it is easier to present and interpret a smaller number of scores, an exploratory factor analysis was conducted on the responses from the ASQ. The intention of this exploration was to discover if there was a statistical basis for combining the three items into a single scale. The SAS (TM) procedure FACTOR was used to perform a principal factor analysis with a varimax rotation (SAS Institute, 1985).

Due to the way the data were collected, either three- or eight-factor solutions were anticipated. The expected three-factor solution would have shown grouping by item, and the expected eight-factor solution would show grouping by scenario. The scree plot did not support a three-factor solution, since third and fourth eigenvalues were almost equal in value. There was a definite break after the eighth eigenvalue. The break after the fifth and sixth eigenvalues were suggestive, but were not as easily interpreted as the eight-factor solution. The seventh and eighth eigenvalues were less than one, but the eigenvalues-less-than-one criterion is only a rough guideline (Cliff, 1987). Therefore, the eight-factor solution seemed to be most appropriate. The solution was varimax rotated, with the rotated factor pattern shown in Table 2. The ITEM column in Table 2 shows the scenario and the ASQ item number. Using a selection criterion of .6 for the factor loadings, the eight factors were clearly associated with the eight common scenarios. The eight factors explained almost all (94 percent) of the total variance. The reliabilities of the scales created by summing the three items into scenario scales were assessed with coefficient alpha (Nunnally, 1978), a measure of internal consistency. All the coefficient alphas exceeded .9, indicating that the scales are acceptably reliable. Coefficient alphas this large are surprising since each scale was based only on three items, and reliability is largely a function of the number of items (Nunnally, 1978).

Table 2. Varimax Rotated Factor Pattern

QUES	FAC1	FAC2	FAC3	FAC4	FAC5	FAC6	FAC7	FAC8
M1A1	-0.06	0.15	-0.00	0.20	**0.80**	0.43	0.22	0.07
M1A2	0.02	0.35	0.05	0.10	**0.73**	0.42	0.05	0.25
M1A3	0.27	0.22	0.16	0.23	**0.76**	0.17	0.16	0.27
M1B1	0.30	0.04	0.07	0.15	0.34	**0.83**	0.11	0.12
M1B2	0.37	0.08	0.02	0.11	0.26	**0.82**	0.05	0.20
M1B3	0.52	-0.01	0.04	0.12	0.39	**0.64**	0.15	0.26
M21	**0.88**	0.12	0.10	0.16	0.12	0.15	0.22	-0.15
M22	**0.89**	0.13	0.04	0.23	0.01	0.24	0.12	0.08
M23	**0.87**	0.02	0.00	0.26	0.01	0.23	0.07	-0.14
A11	-0.04	0.14	**0.88**	0.15	-0.12	-0.01	0.25	0.14
A12	0.01	0.06	**0.86**	0.10	0.10	-0.08	0.13	0.33
A13	0.21	0.02	**0.85**	-0.01	0.21	0.20	0.14	0.13
F11	0.07	**0.91**	0.13	0.09	0.16	-0.03	0.23	0.06
F12	0.14	**0.93**	0.07	0.07	0.07	0.07	0.10	0.15
F13	0.01	**0.87**	0.00	0.18	0.13	0.08	0.07	0.07
E11	0.10	0.24	0.23	0.15	0.11	0.21	**0.87**	-0.00
E12	0.15	0.24	0.18	0.15	0.05	-0.02	**0.90**	0.06
E13	0.38	-0.04	0.22	0.00	0.44	0.09	**0.68**	0.09
E21	0.21	0.26	0.15	**0.80**	0.08	0.28	0.23	0.08
E22	0.28	0.24	0.07	**0.83**	0.07	-0.02	0.09	0.19
E23	0.28	-0.03	0.06	**0.84**	0.25	0.12	0.06	0.19
D11	-0.14	0.19	0.36	0.22	0.07	0.15	0.09	**0.79**
D12	-0.14	0.25	0.15	0.37	0.11	0.12	0.00	**0.82**
D13	0.09	-0.02	0.27	-0.02	0.36	0.20	0.02	**0.76**

Sensitivity Study

Having determined that the three questionnaire items could be reasonably summed into a scale, it was important to determine if the scale was sensitive enough to detect differences as a function of the independent variables of interest. Specifically, could these scores discriminate among the different systems, user groups, or scenarios examined in the three usability studies?

An ANOVA was conducted on the scale scores. Of the 48 participants, 27 completed all the items on the ASQ. Only these data were used in the ANOVA. The main effect of Scenario was highly significant ($F(7,126)=8.92$, $p<.0001$).

The Scenario by System interaction was also significant ($F(14,126)=1.75$, $p=.05$). These results suggest that the ASQ scale score was a reasonably sensitive measure.

Concurrent Validity

For each case in which a scenario was attempted and all ASQ questions were answered, the point-biserial correlation between the summed item scores for the scenario and a 0/1 coding to indicate scenario failure/success was -.40 ($n=48$, $p<.01$). This result shows a tendency for participants who successfully completed a scenario to give lower (more favorable) ratings.

LIMITS TO GENERALIZATION

These findings must be considered as preliminary since the sample size for the factor analysis is smaller than is usually recommended. Nunnally (1978) recommends a minimum of five participants per item, which would be 120 participants for this analysis. On the other hand, the factor structure is very strong and clear.

This instrument was designed to assess the attitude of participants following a scenario completion in a formal usability study. The correlations among items, resultant factors, and validity coefficients may all be influenced by the situation in which the data were collected. While it would be interesting to use the ASQ in a less formal setting, such as a mailed questionnaire or as an instrument in a field study, the results presented in this report cannot be used as justification for such use. If such a study is planned, it is important to examine the instruments discussed in the introduction to determine if a more suitable instrument already exists.

CONCLUSIONS

The psychometric evaluation of this questionnaire shows that the three items can reasonably be condensed into a single scale through summation. This condensation should allow easier interpretation and reporting of results when usability studies use the ASQ. The ASQ seems to be sensitive enough to be a useful measure for usability studies. The ASQ also seems to have reasonable concurrent validity when correlated with scenario completion data. These results should be considered as preliminary since the sample size for the factor analysis was smaller than is generally recommended. On the other hand, the factors and the content of the items summed into scales seem clear. Others who conduct usability studies are encouraged to use this questionnaire unless there is a strong reason to do otherwise.

REFERENCES

Chin, J.P., Diehl, V.A., and Norman, K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In Proceedings of the CHI'88 Conference on Human Factors in Computing Systems (pp. 213-218). New York, NY: Association for Computing Machinery.

Cliff, N. (1987). Analyzing multivariate data. San Diego, CA: Harcourt Brace Jovanovich.

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

Kirakowski, J. and Corbett, M. (1988). The Computer User Satisfaction Inventory (CUSI): Manual and scoring key. Ireland: University College of Cork, Human Factors Research Group.

Lewis, J.R. (1990). A psychometric evaluation of a post-study system usability questionnaire: The PSSUQ (IBM Tech. Report 54.535). Boca Raton, FL: IBM Corp.

McIver, J.P. and Carmines, E.G. (1981). Unidimensional scaling. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-024. Beverly Hills, CA: Sage Publications.

Nunnally, J.C. (1978). Psychometric theory. New York, NY: McGraw-Hill.

SAS Institute. (1985). SAS user's guide: statistics, version 5 edition. Cary, NC: SAS Institute.

ABOUT THE AUTHOR

James R. Lewis

James R. Lewis is a staff human factors engineer for IBM. He earned his Master's Degree in Engineering Psychology at New Mexico State University. His current interests include methods of usability engineering and usability measurement.