# IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use

Technical Report 54.786

James R. Lewis
Human Factors Group
Boca Raton, FL

ABSTRACT

This paper describes recent research in subjective usability measurement at IBM. The focus of the research was the application of psychometric methods to the development and evaluation of questionnaires that measure user satisfaction with system usability. The primary goals of this paper are to (1) discuss the psychometric characteristics of four IBM questionnaires that measure user satisfaction with computer system usability, and (2) provide the questionnaires, with administration and scoring instructions. Usability practitioners can use these questionnaires with confidence to help them measure users' satisfaction with the usability of computer systems.

## Table of Contents

## Introduction

Customers want usable products, and developers strive to produce them. It follows that an important part of modern product engineering, both hardware and software, must be the measurement of usability. Measuring usability is particularly difficult because usability is not a unidimensional product or user characteristic, but emerges as a multidimensional characteristic in the context of users performing tasks with a product in a specific environment (Bevan, Kirakowski, & Maissel, 1991; Shackel, 1984). However, if you are unable to measure usability, how can you judge your product against your competitors', or even your own previous versions of the product? The appropriate measurement methods for assessing usability are not obvious, and are an ongoing concern of human factors engineers involved in the development of computer systems.

### *Subjective and Objective Evaluation*

Most usability evaluations gather both subjective and objective quantitative data in the context of realistic scenarios-of-use, as well as descriptions of the problems representative participants have trying to complete the scenarios. Subjective data are measures of participants' opinions or attitudes concerning their perception of usability. Objective data are measures of participants' performance (such as scenario completion time and successful scenario completion rate).

One area of human factors that has undergone considerable research in subjective and objective evaluation is mental workload (Gopher & Braune, 1984; Wickens, 1984), a multidimensional construct similar to, but more restricted in scope than usability. The mental workload of a task is the extent to which it absorbs an operator's attentional capacity. The construct is most useful in continuous monitoring and control tasks, such as piloting an airplane. Objective measures of mental workload include primary-task performance, secondary-task performance, and physiological measures such as pupil diameter and heart-rate variability (Wickens, 1984). Subjective measures include the Cooper-Harper scale (Cooper & Harper, 1969), Sheridan's dimensional scale (Sheridan, 1980), and the Subjective Workload Assessment Technique (Reid, 1985). For situations in which objective and subjective workload measures agree, the subjective measures are better because they do not disrupt primary task activity, and are easier, quicker and less expensive to obtain. When the objective and subjective measures do not agree, then which measure is "best" depends on the situation. For example, if operator errors can have catastrophic consequences, then objective performance measures should carry more weight in design decisions. If the purpose of the research is to increase operator satisfaction, then designers should give greater consideration to the subjective measures.

Mental workload is an appropriate measure of usability when the task involves continuous demand on a user's attention for monitoring and control. This is rarely the case when assessing the usability of a computer system. Most usability assessment scenarios are sets of tasks in which users solve problems, such as how to create and print a document or how to compose and send a piece of electronic mail. Like tasks suitable for assessment with mental workload measures, researchers can measure usability with

both objective and subjective variables. Objective usability measures include, but are not limited to, scenario completion time, successful scenario completion rate, and time spent recovering from errors (Whiteside, Bennett, & Holtzblatt, 1988). Subjective usability measures are usually responses to Likert-type questionnaire items that assess user attitude concerning attributes such as system ease-of-use and interface likeability (Alty, 1992). Most usability evaluators collect both objective and subjective data. Like mental workload measures, which measure is "best" depends on the purpose of the evaluation. If the development goal of a new system is to increase productivity, then objective measures are of primary importance. If the development goal is user satisfaction, then subjective measures are more important.

### *Research Focus*

The focus of this research was the application of psychometric methods to the development and evaluation of standard questionnaires to assess subjective usability. The goal of psychometrics is to establish the quality of psychological measures (Nunnally, 1978). Is a measure reliable in the sense that it is consistent? Given a reliable measure, is it valid (measures the intended attribute)? Finally, is the measure appropriately sensitive to experimental manipulations? Psychometrics is a well-developed field, but usability researchers have only recently used these methods to develop and evaluate questionnaires to assess usability (Sweeney & Dillon, 1987). In contrast to other recent computer-user satisfaction questionnaires (Chin, Diehl, & Norman, 1988; Kirakowski & Dillon, 1988; LaLomia & Sidowski, 1990) the IBM questionnaires are specifically for use in the context of scenario-based usability testing (Lewis, 1991a; Lewis, 1991b; Lewis, 1991c; Lewis, 1992b; Lewis, Henry, & Mack, 1990), although additional research has indicated that one may be useful as an instrument for field evaluation (Lewis, 1992a). Usability practitioners can use these questionnaires to enhance their current usability methods. (The four IBM questionnaires appear in the appendix.) Before describing the psychometric properties of the IBM questionnaires, I will briefly review the relevant elements of psychometric practice. (For a comprehensive discussion of psychometrics, see Nunnally, 1978.)

***Brief Review of Psychometric Practice***

*Reliability goals*. In psychometrics, a questionnaire's reliability is a quantitative assessment of its consistency. After determining the relationship between items and factors (with factor analysis, discussed below), the items that make up a factor are the component items of a corresponding summative scale. The most common way to estimate the reliability of these types of scales is with coefficient alpha (Nunnally, 1978). Coefficient alpha can range from 0 (no reliability) to 1 (perfect reliability). Measures that can affect an individual's future, such as IQ tests or college entrance exams should have a minimum reliability of .90, and preferably a reliability of .95. For other research or evaluation, measurement reliability should be .70 to .80 (Landauer, 1988; Nunnally, 1978).

*Validity goals*. A questionnaire's validity is the extent to which it measures what it claims to measure. Researchers commonly use the Pearson correlation coefficient to assess criterion-related validity (the relationship between the measure of interest and a different concurrent or predictive measure). This correlation does not have to be large to provide evidence of validity. For personnel selection decisions -- decisions that may have a serious effect on an individual's life -- moderate correlations (with absolute values as small as .30 to .40) are large enough to justify the use of psychometric instruments (such as questionnaires) (Nunnally, 1978). A validity coefficient is concurrent if the measurements occur at the same time, and is predictive if the measure of interest precedes the criterion measurement. An instrument can be reliable without being valid, but cannot be valid unless it is reliable. The statistical reliability of an instrument and the reliability of the criterion measurement set the upper limit of the criterion-related validity between the measures. This is one reason that criterion-related validity coefficients tend to appear low (Nunnally, 1978).

*Number of scale steps*. All other things being equal, the more scale steps the better, but with rapidly diminishing returns. Numerous studies show that the reliability of individual rating scales is a monotonically increasing function of the number of steps (Nunnally, 1978). As the number of scale steps increases from 2 to 20, the increase in reliability is very rapid at first, but tends to level off at about 7. After 11 steps there is little gain in reliability from increasing the number of steps. The question of the number of steps on a rating scale is very important if there is only one scale, but is usually less important when summing scores over a number of scales. Attitude scales tend to be highly reliable because the items tend to correlate rather highly with one another. Reliability, then, usually is not a serious problem in the construction of summated attitude scales (Nunnally, 1978). Many researchers use seven scale steps as the appropriate balance between scale reliability and discriminative demand on the respondent.

*Factor analysis*. Factor analysis is a statistical procedure that examines the correlations among variables to discover clusters of related variables (Nunnally, 1978). Because 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, it is common to conduct a factor analysis to determine if there is a statistical basis for the formation of

summative scales. Generally, a factor analysis requires at least 5 participants per item to ensure stable factor estimates (Nunnally, 1978).

One way to determine the appropriate number of factors in a factor analysis is to graph a scree plot of the factors' eigenvalues (variances) (Cliff, 1987; Nunnally, 1978). The point at which the slope of the line becomes discontinuous (scanning from right to left) indicates an appropriate number of factors for the analysis (Cliff, 1987). Discontinuity analysis is one of the more effective means for determining the number of factors, especially if used in combination with other approaches (Coovert & McNelis, 1988). Another approach is to investigate whether the resulting factor structure is interpretable. Researchers should avoid using the common rule-of-thumb (and common computer package default) to let the number of factors equal the number of eigenvalues that are greater than one (Cliff, 1987)

For these studies, I used principal factors analysis, and varimax-rotated the solutions to achieve easy-to-interpret factor loadings, which express the strength of the relationship between items and factors. For this research, factor loadings that were greater than or equal to .5 indicated a meaningful relationship between the factor and the item (Cliff, 1987).

*Calculating scale scores*. From psychometric theory (Nunnally, 1978), scale reliability is a function of the interrelatedness of scale items, the number of scale steps per item, and the number of items in a scale. If a participant chooses not to answer an item, the effect would be to slightly reduce the reliability of the scale in that instance. In most cases, the remaining items should offer a reasonable estimate of the appropriate scale score. From a practical standpoint, averaging the answered items to obtain the scale score enhances the flexibility of use of the questionnaire, because if an item is not appropriate in a specific context and users choose not to answer it, the questionnaire is still useful. Also, users who do not answer every item can stay in the sample. Finally, averaging items to obtain scale scores does not affect the statistical properties of the scores, and standardizes the range of scale scores. For example, with items based on 7-point scales, all the summative scales would also have scores that range from 1 to 7. This standardization makes scale scores easier to interpret and compare.

## The After-Scenario Questionnaire (ASQ)

In a scenario-based usability study, participants use a product, such as a computer application, to do a series of realistic tasks. The After-Scenario Questionnaire (ASQ) is a three-item questionnaire that IBM usability evaluators have used to assess participant satisfaction after the completion of each scenario. (See the appendix for a copy of the questionnaire.) The items address three important components of user satisfaction with system usability: ease of task completion, time to complete a task, and adequacy of support information (on-line help, messages, and documentation). Because the questionnaire is very short, it takes very little time for participants to complete -- an important practical consideration for usability studies. Usability professionals have used these items (or very similar items) in usability studies at IBM for many years, but a recent

series of studies has provided a database of sufficient size to allow a preliminary psychometric evaluation of the ASQ.

The ASQ items are the constituent items for a summative, or Likert, scale (McIver & Carmines, 1981; Nunnally, 1978). In developing summative scales, it is important to consider item construction, item selection and psychometric evaluation.

### *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 the appendix.

### *Item Selection*

The content of the items reflects components of usability that usability professionals at IBM have generally considered important.

### *Psychometric Evaluation*

*The office-applications studies*. Scenario-based usability studies of three office application systems (Lewis, Henry, & Mack, 1990) provided the data for a psychometric evaluation of the ASQ. 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 at least three months of 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 operating systems.

Popular word-processing applications, mail applications, calendar applications, and spreadsheet applications installed in three different operating environments comprised the three office systems (hereafter referred to as System I, System II and System III). All three environments allowed windowing, used a mouse as a pointing device, and allowed a certain amount of integration among the applications. The systems differed in details of implementation, but were generally similar. The three word-processing and spreadsheet applications were similar, but the mail and calendar applications differed considerably. The studies contained eight scenarios in common, listed in Table 1.

Participants began the study with a brief lab tour, read a description of the study's purpose and the day's agenda, and completed a background questionnaire. Participants using System I completed an interactive tutorial shipped with the system. Tutors provided the other participants with 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 1 hour), participants performed the scenarios, completing the ASQ as they finished each scenario.

While the participant performed the scenario, an observer logged the participant's activities. If the participant completed the scenario without assistance and produced the correct output, then he or she completed the scenario successfully. Either after completing all scenarios or at the end of the workday (with some scenarios never attempted), participants provided an overall system rating with the Post-Study System Usability Questionnaire (PSSUQ) (Lewis, 1992b; Lewis, Henry, & Mack, 1990). Participants usually needed a full work day (8 hours) to complete the study.

At the end of the three studies, the researchers entered the responses to the ASQ, PSSUQ, and the scenario completion data into a database. From this database, it was possible to conduct an exploratory factor analysis, reliability analyses, validity analyses, and a sensitivity analysis.

Table 1. Descriptions of the Eight Office-Applications Scenarios

| Scenario Type | 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. |

*Factor analysis*. Due to the design of this study (eight scenarios and a 3-item questionnaire), either an 8-factor or 3-factor solution would have been reasonable. An 8-factor solution could indicate grouping by scenario, and a 3-factor solution could indicate grouping by item type. Figure 1 shows the scree plot for the eigenvalues.

The scree plot for this analysis did not support a 3-factor solution, but did support an 8-factor solution. The rotated factor pattern is in Table 2. Using a selection criterion of .5 for the factor loadings (indicated with bold type), a clear relationship existed between the factors and the scenarios. The eight factors accounted for almost all (94%) of the variance in the data.

*Reliability*. For the eight summative scales derived from the eight factors, all the coefficient alphas exceeded .90. Coefficient alphas this large were surprising because each scale contained only three items, and reliability is largely a function of the number of scale items (Nunnally, 1978).

*Validity*. The correlation between the ASQ scores and scenario failure or success (coded as 0=failure and 1=success) was -.40 ($n$=48, $p$<.01). This result showed that participants who successfully completed a scenario tended to give lower (more favorable) ASQ ratings – evidence of concurrent validity.

*Sensitivity*. Of the 48 participants, 27 completed all of the ASQ items for all of the scenarios. This reduced database was appropriate for an analysis-of-variance (ANOVA) to assess the sensitivity of the ASQ. Specifically, did the ASQ scores discriminate among the different systems, user groups, or scenarios in the three usability studies? 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 is a reasonably sensitive measure.

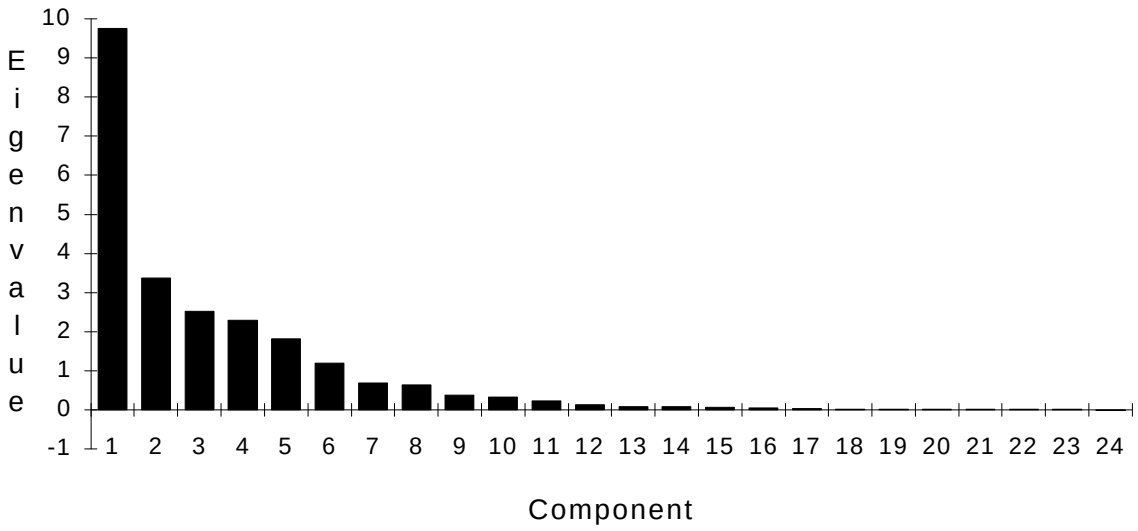


Figure 1. Scree plot for the ASQ principal factors analysis.

### *Discussion*

These findings have limited generalizability because the sample size for the factor analysis was relatively small. The usual recommendation would require 120 participants for this analysis (5 participants x 8 scenarios/participant x 3 items/scenario). On the other hand, the resulting factor structure was very clear.

The psychometric evaluation of this questionnaire showed that it is reasonable to condense the three ASQ items into a single scale through summation (or, equivalently, averaging). The available evidence indicates that the ASQ is reliable, valid, and sensitive. This condensation should allow easier interpretation and reporting of results when usability practitioners use the ASQ.

Table 2. Varimax-Rotated Factor Pattern for the Principal Factor Analysis of the ASQ

| ITEM | FAC1 | FAC2 | FAC3 | FAC4 | FAC5 | FAC6 | FAC7 | FAC8 |
|---|---|---|---|---|---|---|---|---|
| M1A-1 | -0.06 | 0.15 | -0.00 | 0.20 | **0.80** | 0.43 | 0.22 | 0.07 |
| M1A-2 | 0.02 | 0.35 | 0.05 | 0.10 | **0.73** | 0.42 | 0.05 | 0.25 |
| M1A-3 | 0.27 | 0.22 | 0.16 | 0.23 | **0.76** | 0.17 | 0.16 | 0.27 |
| M1B-1 | 0.30 | 0.04 | 0.07 | 0.15 | 0.34 | **0.83** | 0.11 | 0.12 |
| M1B-2 | 0.37 | 0.08 | 0.02 | 0.11 | 0.26 | **0.82** | 0.05 | 0.20 |
| M1B-3 | 0.52 | -0.01 | 0.04 | 0.12 | 0.39 | **0.64** | 0.15 | 0.26 |
| M2-1 | **0.88** | 0.12 | 0.10 | 0.16 | 0.12 | 0.15 | 0.22 | -0.15 |
| M2-2 | **0.89** | 0.13 | 0.04 | 0.23 | 0.01 | 0.24 | 0.12 | 0.08 |
| M2-3 | **0.87** | 0.02 | 0.00 | 0.26 | 0.01 | 0.23 | 0.07 | -0.14 |
| A1-1 | -0.04 | 0.14 | **0.88** | 0.15 | -0.12 | -0.01 | 0.25 | 0.14 |
| A1-2 | 0.01 | 0.06 | **0.86** | 0.10 | 0.10 | -0.08 | 0.13 | 0.33 |
| A1-3 | 0.21 | 0.02 | **0.85** | -0.01 | 0.21 | 0.20 | 0.14 | 0.13 |
| F1-1 | 0.07 | **0.91** | 0.13 | 0.09 | 0.16 | -0.03 | 0.23 | 0.06 |
| F1-2 | 0.14 | **0.93** | 0.07 | 0.07 | 0.07 | 0.07 | 0.10 | 0.15 |
| F1-3 | 0.01 | **0.87** | 0.00 | 0.18 | 0.13 | 0.08 | 0.07 | 0.07 |
| E1-1 | 0.10 | 0.24 | 0.23 | 0.15 | 0.11 | 0.21 | **0.87** | 0.00 |
| E1-2 | 0.15 | 0.24 | 0.18 | 0.15 | 0.05 | -0.02 | **0.90** | 0.06 |
| E1-3 | 0.38 | -0.04 | 0.22 | 0.00 | 0.44 | 0.09 | **0.68** | 0.09 |
| E2-1 | 0.21 | 0.26 | 0.15 | **0.80** | 0.08 | 0.28 | 0.23 | 0.08 |
| E2-2 | 0.28 | 0.24 | 0.07 | **0.83** | 0.07 | -0.02 | 0.09 | 0.19 |
| E2-3 | 0.28 | -0.03 | 0.06 | **0.84** | 0.25 | 0.12 | 0.06 | 0.19 |
| D1-1 | -0.14 | 0.19 | 0.36 | 0.22 | 0.07 | 0.15 | 0.09 | **0.79** |
| D1-2 | -0.14 | 0.25 | 0.15 | 0.37 | 0.11 | 0.12 | 0.00 | **0.82** |
| D1-3 | 0.09 | -0.02 | 0.27 | -0.02 | 0.36 | 0.20 | 0.02 | **0.76** |

## The Printer Scenario Questionnaire (PSQ)

The Printer Scenario Questionnaire (PSQ) was an early version of the ASQ. It differed from the ASQ in item format and number of scale steps per item. Unlike the number of scale steps, physical appearance is one of the least important considerations regarding rating scales (Nunnally, 1978). (See the appendix for the PSQ items.) Because it was the participant satisfaction questionnaire for a series of printer studies that took place during 1983 and 1984, a database large enough to permit psychometric evaluation of the PSQ exists. Due to the similarity between the PSQ and ASQ, it is informative to compare their psychometric properties. Also, participants in the printer studies performed most scenarios three times each, with minor variations. Therefore, this database allows psychometric evaluation of the PSQ over three trials.

### *Item Construction*

The items are 5-point scales, anchored at the end points with the terms "Acceptable as is" for 1 and "Needs a lot of improvement" for 5, and an "Unable to evaluate" rating outside the scale, as shown in the appendix.

### *Item Selection*

The content of the items reflects components of usability that usability professionals at IBM have generally considered important.

### *Psychometric Evaluation*

*The printer studies*. Scenario-based usability studies of seven table-top dot-matrix printers (Lewis, 1991a) provided the data for a psychometric evaluation of the PSQ. Seventy employees of temporary help agencies participated in the studies, ten per printer. The studies had four scenarios in common, listed in Table 3.

To start a scenario, a monitor gave the participant a copy of the scenario description. The participant then performed the scenario up to the point of actually running the print job. Monitors received training in problem identification, and were able to judge whether a subsequent print job would succeed or fail. After performing the scenario, the participant completed the PSQ.

Participants used the instructions and operator manuals to complete the scenarios. If a participant was unable to perform a portion of the scenario and asked for assistance, the monitor took the following steps. When the participant requested assistance, the monitor first determined if the participant had tried to locate the required information in the documentation. If the participant had not attempted to locate the information, the monitor asked the participant to try, and did not record an assistance call. If the participant had attempted to locate the information but had not been able to find it, the monitor helped the participant locate the information and recorded the assistance call. If the participant had located the information but still required help, the monitor provided the necessary guidance and recorded the assistance call. The monitor also recorded the number and types of problems that participants experienced. If a participant indicated that a scenario was complete, but had not properly prepared the printer for the job, then

the problem was of high impact. If a participant experienced difficulty with a portion of the scenario, but managed to solve the problem before indicating that the scenario was complete and without asking for assistance, then the problem was of low impact. For a scenario to be successfully completed, a participant needed to complete the scenario with no calls for assistance or high-impact problems.

Table 3. Descriptions of the Four Printer Scenarios

| Scenario Type | Component Tasks |
|---|---|
| Load paper (LP) | Load continuous paper and set the top of form. |
| Self-test (ST) | Run the printer's self-test. |
| End of Forms (EF) | Load paper and continue a print job after reaching the end of forms. |
| Change Ribbon (CR) | Change the printer's ribbon. |

*Factor analyses*. Because the previous analysis of the ASQ showed that items tended to cluster by scenario, the factor analyses of the PSQ were three 4-factor solutions, one for each trial. Table 4 contains the varimax-rotated factor patterns for the PSQ analyses. Using a selection criterion of .5 (in bold type), the factor analyses confirmed the hypothesis that the PSQ items, like the ASQ items, clustered by scenario. The clustering was very strong in the first trial, and generally persisted for the second and third trials.

*Reliability*. The coefficient alphas for these 12 scales (4 scales per trial x 3 trials) ranged from .64 to .93 and averaged .80, indicating that the scales were marginally reliable for the purpose of usability studies.

*Validity*. Collapsing across trials, the correlation between the PSQ scores and scenario failure or success (coded as 0=failure and 1=success) was -.35 ($n$=70, $p$=.001). This result showed that participants who successfully completed a scenario tended to give lower (more favorable) PSQ ratings. The correlation between the PSQ ratings and the number of assists was 0.38 ($n$=70, $p$=.0006), and that with the number of problems was 0.31 ($n$=70, $p$=.004). These results indicate concurrent validity for the PSQ scores.

*Sensitivity*. Of the 70 participants, 53 completed all of the PSQ items for all of the scenarios. This reduced database was appropriate for an ANOVA to assess the sensitivity of the PSQ. The main effects of Scenario ($F$(3,108)=23.12, $p$<.0001) and Trial ($F$(2.72)=45.2, $p$<.0001) were highly significant. The Scenario by Printer interaction was also significant ($F$(18,108)=2.1, $p$=.01). These results suggest that the PSQ scale score is a reasonably sensitive measure.

Table 4. Varimax-Rotated Factor Pattern for the Principal Factor Analyses of the PSQ, by Trial

| TRIAL | ITEM | FAC1 | FAC2 | FAC3 | FAC4 |
|---|---|---|---|---|---|
| 1 | LP-1 | 0.17 | **0.86** | 0.20 | 0.14 |
| | LP-2 | 0.04 | **0.65** | 0.18 | 0.32 |
| | LP-3 | 0.13 | **0.83** | 0.15 | 0.02 |
| | ST-1 | 0.07 | 0.19 | **0.85** | 0.03 |
| | ST-2 | 0.04 | 0.13 | **0.84** | 0.18 |
| | ST-3 | 0.00 | 0.15 | **0.60** | 0.17 |
| | EF-1 | 0.00 | 0.06 | 0.28 | **0.74** |
| | EF-2 | 0.12 | 0.09 | 0.04 | **0.73** |
| | EF-3 | 0.24 | 0.23 | 0.11 | **0.66** |
| | CR-1 | **0.87** | 0.08 | 0.11 | -0.01 |
| | CR-2 | **0.81** | 0.13 | -0.01 | 0.20 |
| | CR-3 | **0.77** | 0.10 | -0.01 | 0.14 |
| 2 | LP-1 | 0.11 | 0.30 | 0.16 | **0.70** |
| | LP-2 | 0.15 | **0.59** | 0.09 | **0.52** |
| | LP-3 | 0.04 | 0.11 | 0.11 | 0.34 |
| | ST-1 | 0.00 | 0.04 | **0.80** | 0.21 |
| | ST-2 | 0.00 | 0.33 | **0.66** | -0.01 |
| | ST-3 | -0.04 | 0.16 | **0.62** | 0.21 |
| | EF-1 | 0.01 | **0.71** | 0.36 | 0.21 |
| | EF-2 | 0.16 | **0.77** | 0.16 | 0.34 |
| | EF-3 | 0.21 | **0.78** | 0.16 | 0.17 |
| | CR-1 | **0.85** | 0.07 | -0.08 | 0.16 |
| | CR-2 | **0.90** | 0.15 | 0.04 | 0.02 |
| | CR-3 | **0.89** | 0.12 | 0.00 | 0.06 |
| 3 | LP-1 | -0.22 | 0.41 | **0.56** | 0.16 |
| | LP-2 | -0.07 | **0.79** | 0.23 | 0.02 |
| | LP-3 | 0.16 | **0.78** | 0.07 | 0.00 |
| | ST-1 | 0.02 | -0.13 | 0.20 | **0.83** |
| | ST-2 | 0.01 | 0.08 | 0.04 | **0.84** |
| | ST-3 | -0.01 | 0.30 | 0.17 | 0.41 |
| | EF-1 | 0.02 | 0.10 | **0.82** | 0.18 |
| | EF-2 | 0.27 | 0.21 | **0.71** | 0.09 |
| | EF-3 | 0.38 | **0.50** | 0.37 | 0.11 |
| | CR-1 | **0.82** | -0.08 | 0.19 | 0.12 |
| | CR-2 | **0.89** | 0.09 | -0.03 | -0.12 |
| | CR-3 | **0.83** | 0.12 | -0.01 | 0.02 |

***Discussion***

The PSQ and ASQ results are comparable. In both studies, the factor analyses showed the pattern of association with scenarios (through three trials with the PSQ). The sample size of the PSQ study (70 participants) exceeded the minimum requirement for its factor analyses (5 participants/item x 3 items/scenario x 4 scenarios = 60 participants). The correlation between the summed PSQ ratings and successful scenario completion was -.35, quite close to that of the ASQ. The ASQ and PSQ analyses of variance had the same patterns as well, with no main effect of product but a significant Product-by-Scenario interaction. These similarities support extended generalization of the ASQ findings, because the same patterns occurred across two studies of different products and user groups.

The only way in which the ASQ and PSQ psychometric evaluations differed was in their reliabilities. Coefficient alpha for the PSQ ranged from .63 to .93, but exceeded .90 for all eight ASQ scales. The scale characteristics that have the largest influence on scale reliability are the number of items in the scale and the number of scale steps per item (Nunnally, 1978). It is likely that the smaller coefficient alphas found for the PSQ were primarily due to the use of 5-point scales rather than 7-point scales. (According to Nunnally, 1978, differences in physical appearance are not usually important.) For this reason, usability practitioners should use the ASQ rather than the PSQ.

In both the ASQ and PSQ ANOVAs, the main effect of primary interest (system for the ASQ, printer for the PSQ) was not significant, but the interaction with scenario was statistically significant. This is consistent with discussions of "ecological" human factors used to help explain inconsistent research findings by pointing out critical task differences (Vicente, 1990). "Instead of asking which is 'best,' one would ask, 'Under what circumstances is Method A [or Product A] better, and under what conditions is B preferable?" (Vicente, 1990, p. 3) Different products face trade-offs in design that affect the usability of the product depending upon the specific task the user is trying to accomplish. If a usability practitioner studies a reasonably broad range of tasks in comparative usability studies, it may be unrealistic to expect a product to excel in every task. Under those conditions, usability practitioners should expect a significant Product-by-Scenario interaction. This also emphasizes the importance of selecting and prioritizing the appropriate scenarios for usability studies.

## The Post-Study System Usability Questionnaire (PSSUQ)

The Post-Study System Usability Questionnaire (PSSUQ) is currently a 19-item instrument for assessing user satisfaction with system usability. (See the appendix for a copy of the questionnaire items.) Participants need more time to complete the PSSUQ than the ASQ (about 10 minutes to complete the PSSUQ), but only complete it once, at the end of a usability study. Completing the PSSUQ allows participants to provide an overall evaluation of the system they used.

After the 48 participants in the office-applications usability study (Lewis, Henry, & Mack, 1990) completed all the scenarios, they rated their system with the PSSUQ. (See the section on the ASQ for more details about the study). This data allowed preliminary psychometric evaluation of the PSSUQ (Lewis, 1992b).

This earlier version of the PSSUQ (Lewis, 1992b) had only 18 items, with the items in a different order than shown in the appendix. Recently, a series of investigations using decision support systems revealed a common set of five system characteristics associated with usability by several different user groups (Doug Antonelli, personal communication, January 5, 1991). The original 18-item PSSUQ addressed four of these five system characteristics. The 19-item version of the PSSUQ contains an additional item to cover the fifth of these five system characteristics.

### *Item Construction*

The items are 7-point graphic scales, anchored at the end points with the terms "Strongly agree" for 1, "Strongly disagree" for 7, and a "Not applicable" (N/A) point outside the scale.

### *Item Selection*

A group of usability evaluators selected the items on the basis of their comprehensive content regarding hypothesized constituents of usability. For example, the items assess such system characteristics as ease of use, ease of learning, simplicity, effectiveness, information, and the user interface.

### *Psychometric Evaluation*

*Factor analysis*. The scree plot for an exploratory principal factors analysis of the PSSUQ data indicated that a 3-factor solution was appropriate (see Figure 2), so the overall scale defined by the full set of items contained three subscales. Table 5 shows the varimax-rotated factor pattern, revealing the structure of the subscales. Bold type in Table 5 highlights factor loadings that exceeded .5. Items that loaded highly on two factors were ambiguous regarding the appropriate subscale of which they should be a component, so they did not become a component of any subscale. (See the appendix to examine the content of these items.) One of the most difficult tasks following this type of exploratory factor analysis is naming the factors. After considering a number of alternatives, a group of human factors engineers named the factors (and their corresponding subscales) System Usefulness (SYSUSE), Information Quality (INFOQUAL), and Interface Quality (INTERQUAL). These three factors account for 87% of the variability in the data.

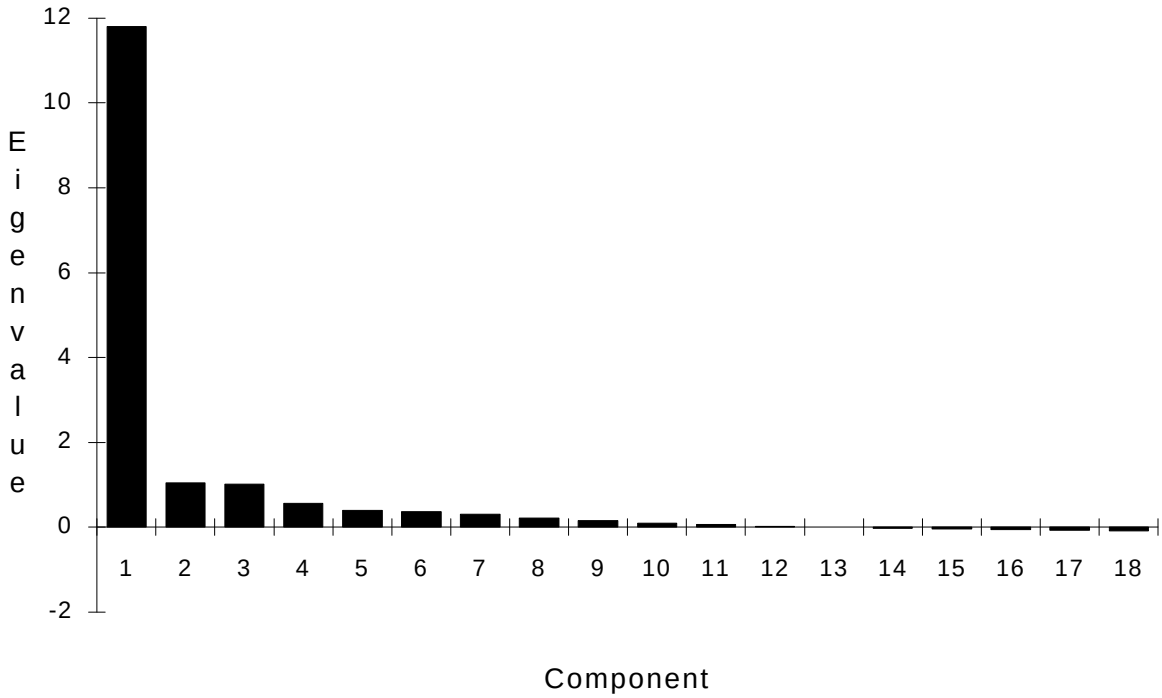


Figure 2. Scree plot for the PSSUQ principal factors analysis.

*Reliability*. Coefficient alpha analyses showed that the reliability of the overall summative scale (OVERALL) was .97, and ranged from .91 to .96 for the three subscales (SYSUSE=.96, INFOQUAL=.91, and INTERQUAL=.91). Therefore, the overall scale and the three subscales have excellent reliability.

*Validity*. Correlation analyses support the validity of the scales. The OVERALL scale correlated highly with the sum of the ASQ ratings that participants gave after completing each scenario ($r(20)=.80$, $p=.0001$). OVERALL also correlated significantly with the percentage of successful scenario completion ($r(29)=-.40$, $p=.026$). The SYSUSE ($r(36)=-.40$, $p=.006$) and INTERQUAL ($r(35)=-.29$, $p=.08$) correlated with the percentage of successful scenario completion.

*Sensitivity*. In the sensitivity ANOVAs, the overall scale and all three subscales indicated significant differences among the user groups (OVERALL: $F(2,29)=4.35$, $p=.02$; SYSUSE: $F(2,36)=6.9$, $p=.003$; INFOQUAL: $F(2,33)=3.68$, $p=.04$; INTERQUAL: $F(2,33)=3.74$, $p=.03$). INFOQUAL showed a significant system effect ($F(2,33)=3.18$, $p=.05$).

Table 5. Varimax-Rotated Factor Pattern for the Principal Factor Analysis of the PSSUQ

| ITEM | SUBSCALE | FAC1 | FAC2 | FAC3 |
|---|---|---|---|---|
| 1 | 1 | **0.77** | 0.26 | 0.43 |
| 2 | 1 | **0.63** | 0.35 | 0.46 |
| 3 | 1 | **0.75** | 0.38 | 0.25 |
| 4 | 1 | **0.81** | 0.45 | 0.07 |
| 5 | 1 | **0.80** | 0.16 | 0.36 |
| 6 | 1 | **0.68** | 0.37 | 0.48 |
| 7 | 1 | **0.69** | 0.46 | 0.40 |
| 8* | N/A | | | |
| 9 | 2 | 0.05 | **0.61** | 0.24 |
| 10 | 2 | 0.36 | **0.71** | 0.23 |
| 11 | 2 | 0.45 | **0.63** | 0.25 |
| 12 | 2 | 0.44 | **0.75** | 0.22 |
| 13 | 2 | 0.43 | **0.70** | 0.32 |
| 14 | 2 | 0.43 | **0.74** | 0.40 |
| 15 | N/A | 0.30 | **0.59** | **0.56** |
| 16 | 3 | 0.30 | 0.36 | **0.75** |
| 17 | 3 | 0.37 | 0.36 | **0.76** |
| 18 | 3 | 0.22 | 0.28 | **0.80** |
| 19 | N/A | **0.58** | 0.21 | **0.64** |

* The first version of the PSSUQ (Lewis, 1992a) did not contain this item.

### *Discussion*

These findings have limited generalizability because the sample size for the factor analysis was relatively small. The usual recommendation would be 90 participants for this questionnaire. However, the factor analysis and reliability analyses suggest that it is reasonable to define three subscales from this set of items. The PSSUQ has reasonable concurrent validity when compared with successful scenario completion rates and the ASQ scores. The overall scale and the subscales are reasonably sensitive. The evidence provided sufficient justification to use the PSSUQ to measure user satisfaction with system usability in usability studies, but also suggested that it would be prudent to collect more data in different circumstances to extend the generalizability of the findings.

## The Computer System Usability Questionnaire (CSUQ)

The PSSUQ research was preliminary for two reasons. First, the sample size for the factor analysis was small, consisting of data from only 48 participants. Second, the PSSUQ data came from a usability study. This setting may have influenced the correlations among the items and, therefore, the resultant factors. The purpose of this research (Lewis, 1992a) was to use a slightly revised version of the PSSUQ, the Computer System Usability Questionnaire (CSUQ) to obtain a database of sufficient size to calculate stable factors from a mailed survey. If the same factors emerged from this research as from the PSSUQ research, the study would demonstrate the potential usefulness of the questionnaire across different user groups and different research settings.

### *Item Selection and Construction*

The CSUQ is identical to the PSSUQ (Lewis, 1991c), except that the wording of the items does not refer to a usability testing situation. For example, Item 3 of the PSSUQ states, "I could effectively complete the tasks and scenarios using this system," but Item 3 of the CSUQ states, "I can effectively complete my work using this system." (See the appendix for the CSUQ items.)

### *Psychometric Evaluation*

*The mail survey using the CSUQ*. The participants were 825 IBM employees who worked at nine IBM development sites: Atlanta, Austin, Bethesda, Boca Raton, Dallas, Raleigh, Rochester, San Jose, and Tucson. I used a random number generator to select the participants' names from the IBM electronic mail directory (CALLUP), and mailed them each a copy of the CSUQ with a cover letter. Responses from the returned questionnaires that arrived within 3 months of mailing made up the database for this study.

*Factor analysis*. Forty-six percent (377) of the participants returned the questionnaire. A principal factor analysis of the returned questionnaires produced the scree plot shown in Figure 3. The scree plot was similar to that found for the PSSUQ, indicating that an appropriate factor analysis should solve for three factors. Table 6 shows the varimax-rotated 3-factor solution. The selection criterion for the factor loadings was 0.5, shown in bold type in the table. The factor analysis showed that Item 8 ("I believe I became productive quickly using this system"), which was not a part of the original PSSUQ, should be part of Factor 1. Item 15 ("The organization of information on the system screens is clear"), which loaded on two factors in the PSSUQ study, loaded on only Factor 2 in the current study. In the PSSUQ study and in the current study, Item 19 ("Overall, I am satisfied with this system") loaded on both Factors 1 and 3, and is not part of any subscale. Otherwise, the factor structure of the CSUQ is very similar to that of the PSSUQ, so the CSUQ and PSSUQ subscales have the same names. The three factors accounted for 98.6% of the variability in the rating data.

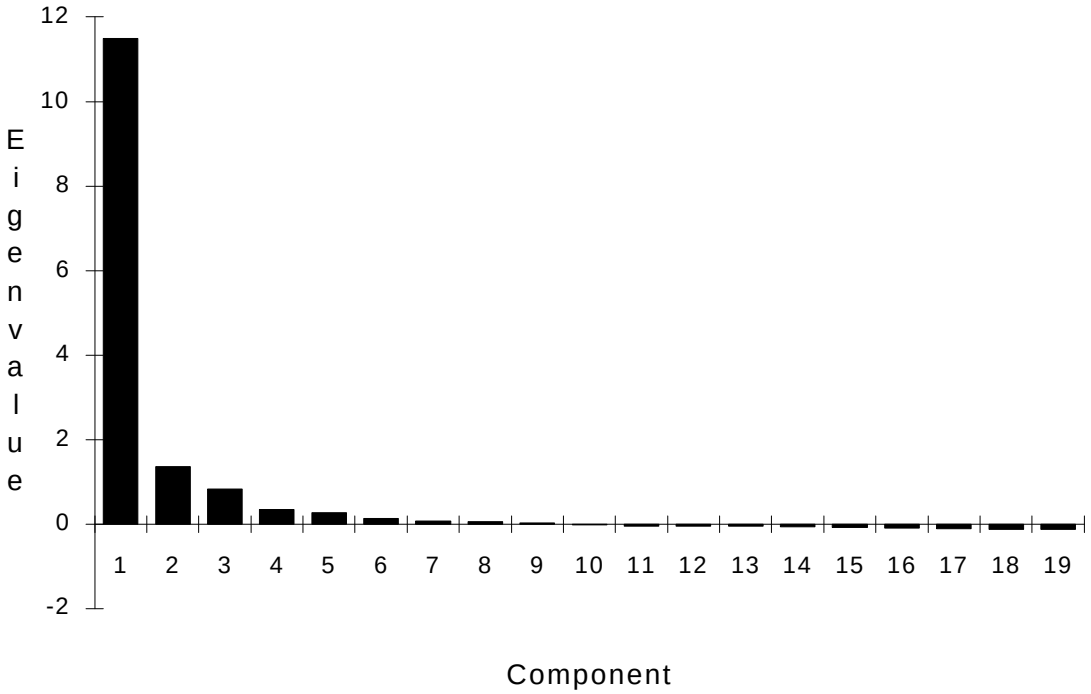


Figure 3. Scree plot for the CSUQ principal factors analysis.

*Reliability*. In all cases, coefficient alpha exceeded 0.89, indicating acceptable scale reliability. The estimates of coefficient alpha for the CSUQ were .93 for SYSUSE, .91 for INFOQUAL, .89 for INTERQUAL, and .95 for the OVERALL scale. The values of coefficient alpha for the CSUQ scales were within 0.03 of those for the PSSUQ scales.

*Validity/Sensitivity*. After establishing scale reliability, the next step in psychometric evaluation is to determine scale validity. However, without a concurrent or predicted measurement, it is impossible to obtain a quantitative measure of validity in the traditional psychometric sense. An indirect way to assess validity is to examine scale sensitivity to variables that should systematically affect the scale. The sensitivity analyses of the PSSUQ (Lewis, 1992b) showed significant effects of user group (business professional with mouse experience, business professional without mouse experience, and secretary/clerk without mouse experience) on the OVERALL, SYSUSE, INFOQUAL, and INTERQUAL scales. The type of computer system the participant used during the study significantly affected the INFOQUAL scale.

Table 6. Varimax-Rotated Factor Pattern for the Principal Factor Analysis of the CSUQ

| ITEM | SUBSCALE | FAC1 | FAC2 | FAC3 |
|---|---|---|---|---|
| 1 | 1 | **0.74** | 0.36 | 0.26 |
| 2 | 1 | **0.69** | 0.41 | 0.16 |
| 3 | 1 | **0.72** | 0.21 | 0.36 |
| 4 | 1 | **0.74** | 0.31 | 0.33 |
| 5 | 1 | **0.77** | 0.30 | 0.32 |
| 6 | 1 | **0.72** | 0.22 | 0.27 |
| 7 | 1 | **0.63** | 0.49 | 0.13 |
| 8 | 1 | **0.66** | 0.39 | 0.26 |
| 9 | 2 | 0.23 | **0.72** | 0.21 |
| 10 | 2 | 0.34 | **0.67** | 0.28 |
| 11 | 2 | 0.23 | **0.81** | 0.20 |
| 12 | 2 | 0.24 | **0.77** | 0.27 |
| 13 | 2 | 0.38 | **0.76** | 0.17 |
| 14 | 2 | 0.40 | **0.73** | 0.18 |
| 15 | 2 | 0.34 | **0.57** | 0.40 |
| 16 | 3 | 0.33 | 0.27 | **0.81** |
| 17 | 3 | 0.38 | 0.26 | **0.81** |
| 18 | 3 | 0.34 | 0.35 | **0.56** |
| 19 | N/A | **0.66** | 0.37 | **0.50** |

A comprehensive listing of the influence of respondent characteristics on the CSUQ scores is outside the scope of this paper. However, the significant findings are similar to those for the PSSUQ. The type of computer that respondents used significantly affected their responses only for the INFOQUAL score ($F(5,311)=2.14$, $p=0.06$). The number of years of experience with their computer system affected respondents' scores for OVERALL ($F(4,294)=3.12$, $p=0.02$), SYSUSE ($F(4,332)=2.05$, $p=0.09$), INFOQUAL ($F(4,311)=2.59$, $p=0.04$) and INTERQUAL ($F(4,322)=2.47$, $p=0.04$). The respondents' range of experience with computer systems (number of different computer systems that they reported having used) affected scores for OVERALL ($F(3,294)=2.77$, $p=0.04$), INFOQUAL ($F(3,311)=2.60$, $p=0.05$) and INTERQUAL ($F(3,322)=2.14$, $p=0.10$). These significant findings provide indirect support to the hypothesis that these scales are valid.

### *Discussion*

The key results from this study are (1) a demonstration of stable factors for the CSUQ (and, by extension, for the PSSUQ) and (2) evidence that the questionnaire works well in non-laboratory settings. The CSUQ scales are comparable to the PSSUQ scales,

both in terms of reliability and validity (indicated by similarity in the sensitivity analyses). These findings substantially enhance the usefulness of the CSUQ and PSSUQ to usability practitioners. Researchers who conduct usability studies (either laboratory or non-laboratory) can use this questionnaire to assess user satisfaction with system usability.

## General Discussion

Although user satisfaction with system usability is only one component of the multifaceted construct of usability (Bevan et al., 1991), it is a very important component in many situations. It is especially important when a primary design goal is user satisfaction. This paper has described the psychometric qualities of four questionnaires that assess user satisfaction with system usability: the ASQ, PSQ, PSSUQ and CSUQ.

The ASQ and PSQ are both after-scenario questionnaires, intended for use in a scenario-based usability testing situation. They contain essentially the same items, but the ASQ uses a 7-point scale and the PSQ uses a 5-point scale. Using data from very different scenario-based usability studies (one a study of software office applications, the other a study of printers), their factor analyses, validity analyses, and sensitivity analyses were virtually identical. Obtaining the same results in different settings with different user groups provides strong evidence that these results are generalizable, and the questionnaires have wide applicability. Because the ASQ has substantially better reliability than the PSQ, usability practitioners should use the ASQ rather than the PSQ as their after-scenario questionnaire.

The PSSUQ and CSUQ are both overall satisfaction questionnaires. The PSSUQ items are appropriate for a usability testing situation, and the CSUQ items are appropriate for a field testing situation. Otherwise, the questionnaires are identical. The psychometric evaluations of the PSSUQ (using data from a usability study) and the CSUQ (using data from a mail survey) were virtually identical. As with the after-scenario questionnaires, this consistency provides strong evidence of generalizability of results and wide applicability of the questionnaires.

Because these questionnaires have acceptable psychometric properties, usability practitioners can use them with confidence as standardized measurements of satisfaction for usability studies and tests (ASQ, PSSUQ) or field research (CSUQ). (Practitioners should note that nothing prevents the addition of items to these questionnaires if a particular situation suggests the need. However, using these questionnaires as the foundation for special-purpose questionnaires ensures that practitioners can score the scales and subscales from the questionnaires, maintaining the advantages of standardized measurement.)

Standardized satisfaction measurements offer many advantages to the usability practitioner (Nunnally, 1978). Specifically, standardized measurements provide:

*Objectivity*. A standardized measurement supports objectivity because it allows usability practitioners to independently verify the measurement statements of other practitioners.

*Quantification*. Standardized measurements allow practitioners to report results in finer detail than they could using only personal judgment. Standardization also permits practitioners to use powerful methods of mathematics and statistics to better understand their results (Nunnally, 1978). Although this position is still controversial among measurement theorists (Lewis, 1989), Nunnally is not alone. For example, Harris (1985) stated:

> That I do not accept [S. S.] Steven's position on the relationship between strength of measurement and "permissible" statistical procedures should be evident from the kinds of data used as examples throughout this Primer: level of agreement with a questionnaire item, as measured on a five-point scale having attached verbal labels . . . . This is not to say, however, that the researcher may simply ignore the level of measurement provided by his or her data. It is indeed crucial for the investigator to take this factor into account in considering the kinds of theoretical statements and generalizations he or she makes on the basis of significance tests. (pp. 326-328)

In other words, the level of measurement (ratio, interval, ordinal) does not limit permissible arithmetic operations or related statistical operations, but does limit the permissible interpretations of the results of these operations. For example, these numbers (ASQ, PSSUQ, CSUQ scores) clearly do not come from a ratio scale (a scale with a known zero point and equal scale intervals). It is most likely that these scales are ordinal. Suppose you compare two products with the PSSUQ, and Product A receives a score of 2.0 versus Product B's score of 4.0. Given a statistically significant comparison, you could say that Product A had more satisfying usability characteristics than Product B (an ordinal claim), but you could not say that Product A was twice as satisfying as Product B (a ratio claim).

*Communication*. It is easier for practitioners to communicate effectively when standardized measures are available. Inadequate efficiency and fidelity of communication in any field is an impediment to progress.

*Economy*. Developing standardized measures requires a substantial amount of work. However, once developed, they are economical. There is rarely any need to re-evaluate standardized measures.

*Scientific generalization*. Scientific generalization is at the heart of scientific work. Standardization is essential for assessing the generalization of results.

In conclusion, these questionnaires should be valuable additions to the repertoire of techniques that usability practitioners apply in the design and evaluation of computer systems.

## Acknowledgments

This paper brings together work that many human factors engineers and other usability professionals (both inside and outside of IBM) reviewed for publication as technical reports and conference papers. These reviews substantially improved the presentation of this research. I do want to acknowledge specifically the contributions of several key individuals. First, this research would never have taken place without the management support of Robert Mack at the T. J. Watson Research Center's User Interface Institute. While I was on assignment at the research center, Dr. Mack also gave me the opportunity to study multivariate analysis (with Jane Monroe) and psychometric methods (with Marvin Sontag) at the Teacher's College/Columbia University Department of Measurement and Statistics. After I returned to Boca Raton, Don Davis, then manager of the Design Center/Human Factors department, gave me the flexibility to continue analyzing the office-applications data collected at the T. J. Watson Research Center. Suzanne Henry made her office-applications ASQ and PSSUQ data available to me, providing a database large enough for a preliminary psychometric evaluation. Richard Granda, Suzanne Henry, Mary LaLomia, Alan Happ, and Sharon Stanners all contributed to the effort to name the PSSUQ factors. I am especially grateful to Dr. LaLomia for her encouragement to pursue this work and for her valuable technical comments.

## References


Alty, J. L. (1992). Can we measure usability? In *Proceedings of the Advanced Information Systems 1992 Conference* (pp. 95-106). London: Learned Information.

Bevan, N., Kirakowski, & Maissel, J. (1991). What is usability? In *Proceedings of the Fourth International Conference on Human Computer Interaction* (pp. 651-655). Stuttgart, Germany: Elsevier.

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

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

Cooper, G. E., & Harper, R. P. (1969). *The use of pilot ratings in the evaluation of aircraft handling qualities* (NASA Ames Technical Report NASA TN-D-5153). Moffett Field, CA: NASA Ames Research Center.

Coovert, M. D., & McNelis, K. (1988). Determining the number of common factors in factor analysis: A review and program. *Educational and Psychological Measurement, 48,* 687-693.

Gopher, D., & Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? *Human Factors, 26,* 519-532.

Harris, R. J. (1985). *A primer of multivariate statistics*. Orlando, FL: Academic Press.

Kirakowski, J., & Dillon, A. (1988). *The computer user satisfaction inventory (CUSI): Manual and scoring key*. Cork, Ireland: Human Factors Research Group, University College of Cork.

LaLomia, M. J., & Sidowski, J. B. (1990). Measurements of computer satisfaction, literacy, and aptitudes: A review. *International Journal of Human-Computer Interaction,* 2, 231-253.

Landauer, T. K. (1988). Research methods in human-computer interaction. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 905-928). New York, NY: Elsevier.

Lewis, J. R. (1989). *The relative reliabilities of mean and median differences as indicators of statistically significant differences for 7-Point scales* (Tech. Report 54.532), Boca Raton, FL: International Business Machines Corporation.

Lewis, J. R. (1991a). An after-scenario questionnaire for usability studies: psychometric evaluation over three trials. *SIGCHI Bulletin, 23*, 79.

Lewis, J. R. (1991b). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *SIGCHI Bulletin, 23*, 78-81.

Lewis, J. R. (1991c). *User satisfaction questionnaires for usability studies: 1991 manual of directions for the ASQ and PSSUQ* (Tech. Report 54.609). Boca Raton, FL: International Business Machines Corporation.

Lewis, J. R. (1992a). *Psychometric evaluation of the computer system usability questionnaire: The CSUQ* (Tech. Report 54.723), Boca Raton, FL: International Business Machines Corporation.

Lewis, J. R. (1992b). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1259-1263). Santa Monica, CA: Human Factors Society.

Lewis, J. R., Henry, S. C., & Mack, R. L. Integrated office software benchmarks: A case study. In *Human-Computer Interaction -- INTERACT '90* (pp. 337-343). Cambridge, England: Elsevier.

McIver, J. P., & 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.

Reid, G. B. (1985). Current status of the development of the subjective workload assessment technique. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp. 220-223). Santa Monica, CA: Human Factors Society.

Shackel, B. (1984). The concept of usability. In J. Bennett, D. Case, J. Sandelin and M. Smith (Eds.) *Visual Display Terminals* (pp. 45-88). Englewood Cliffs, NJ: Prentice-Hall.

Sheridan, T. (1980). Mental workload: What is it? Why bother with it? *Human Factors Society Bulletin, 23*, 1-2.

Sweeney, M., & Dillon A. (1987). Methodologies employed in the psychological evaluation of HCI. In *Proceedings of Human-Computer Interaction -- INTERACT '87* (pp. 367-373).

Vicente, K. J. (1990). A few implications of an ecological approach to human factors. *Human Factors Society Bulletin, 33*, 1-4.

Whiteside, J., Bennett, J., and Holtzblatt, K. (1988). Usability engineering: Our experience and evolution. In M. Helander (Ed.) *Handbook of Human-Computer Interaction* (pp. 791-818). New York, NY: Elsevier.

Wickens, C. D. (1984). *Engineering psychology and human performance*. Columbus, OH: Charles E. Merrill.

# Appendix. The IBM Questionnaires

### *The After-Scenario Questionnaire (ASQ)*

*Administration and Scoring.* Give the questionnaire to a participant after he or she has completed a scenario during a usability evaluation. Average (with the arithmetic mean) the scores from the three items to obtain the ASQ score for a participant's satisfaction with the system for a given scenario. Low scores are better than high scores due to the anchors used in the 7-point scales. If a participant does not answer an item or marks N/A, average the remaining items to obtain the ASQ score.

*Instructions and Items*. The questionnaire's instructions and items are:

For each of the statements below, circle the rating of your choice.

1. Overall, I am satisfied with the ease of completing this task.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

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

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

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

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

***The Printer-Scenario Questionnaire (PSQ)***

*Administration and Scoring*. As indicated in the body of the paper, use the ASQ rather than the PSQ.

*Instructions and Items*. The questionnaire's instructions and items are:

For each of the items below, please circle the response that best describes your experience with the printer for this scenario.

1. Time to Complete Task

   1 = Acceptable as is -- less time than expected
   2 = Acceptable as is -- about right
   3 = Needs slight improvement
   4 = Needs moderate improvement
   5 = Needs a lot of improvement
   - = Unable to evaluate

   Comments:

2. Ease of Performing Tasks

   1 = Acceptable as is -- very easy
   2 = Acceptable as is -- easy
   3 = Needs slight improvement
   4 = Needs moderate improvement
   5 = Needs a lot of improvement
   - = Unable to evaluate

   Comments:

3. Satisfaction with Instructions/Publications

   1 = Acceptable as is -- very satisfied
   2 = Acceptable as is -- satisfied
   3 = Needs slight improvement
   4 = Needs moderate improvement
   5 = Needs a lot of improvement
   - = Unable to evaluate

   Comments:

***The Post-Study System Usability Questionnaire (PSSUQ)***

*Administration and Scoring.* Give the PSSUQ to participants after they have completed all the scenarios in a usability study. You can calculate four scores from the responses to the PSSUQ items: the overall satisfaction score (OVERALL), system usefulness (SYSUSE), information quality (INFOQUAL) and interface quality (INTERQUAL). Because research on an alternative form of the PSSUQ (the Computer System Usability Questionnaire, or CSUQ) confirmed and clarified (and slightly modified) the factor structure of the questionnaire, refer to Appendix Table 1 in the next section of this appendix for the current scoring rules of the PSSUQ.

*Instructions and Items*. The questionnaire's instructions and items are:

This questionnaire, which starts on the following page, gives you an opportunity to tell us your reactions to the system you used. Your responses will help us understand what aspects of the system you are particularly concerned about and the aspects that satisfy you.

To as great a degree as possible, think about all the tasks that you have done with the system while you answer these questions.

Please read each statement and indicate how strongly you agree or disagree with the statement by circling a number on the scale. If a statement does not apply to you, circle N/A.

Please write comments to elaborate on your answers.

After you have completed this questionnaire, I'll go over your answers with you to make sure I understand all of your responses.

Thank you!

1. Overall, I am satisfied with how easy it is to use this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

2. It was simple to use this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

3. I could effectively complete the tasks and scenarios using this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

4. I was able to complete the tasks and scenarios quickly using this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

5. I was able to efficiently complete the tasks and scenarios using this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

6. I felt comfortable using this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

7. It was easy to learn to use this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

8. I believe I could become productive quickly using this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

9. The system gave error messages that clearly told me how to fix problems.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

10. Whenever I made a mistake using the system, I could recover easily and quickly.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

11. The information (such as on-line help, on-screen messages and other documentation) provided with this system was clear.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

12. It was easy to find the information I needed.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

13. The information provided for the system was easy to understand.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

14. The information was effective in helping me complete the tasks and scenarios.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

15. The organization of information on the system screens was clear.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

Note: *The interface includes those items that you use to interact with the system. For example, some components of the interface are the keyboard, the mouse, the screens (including their use of graphics and language).*

16. The interface of this system was pleasant.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

17. I liked using the interface of this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

18. This system has all the functions and capabilities I expect it to have.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

19. Overall, I am satisfied with this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

***The Computer System Usability Questionnaire (CSUQ)***

*Administration and Scoring.* Use the CSUQ rather than the PSSUQ when the usability study is in a non-laboratory setting. Appendix Table 1 contains the rules for calculating the CSUQ and PSSUQ scores.

Appendix Table 1. Rules for Calculating CSUQ/PSSUQ Scores

| Score Name | Average the Responses to: |
|---|---|
| OVERALL | Items 1 through 19 |
| SYSUSE | Items 1 through 8 |
| INFOQUAL | Items 9 through 15 |
| INTERQUAL | Items 16 through 18 |

Average the scores from the appropriate items to obtain the scale and subscale scores. Low scores are better than high scores due to the anchors used in the 7-point scales. If a participant does not answer an item or marks "N/A," then average the remaining item scores.

*Instructions and Items*. The questionnaire's instructions and items are:

This questionnaire (which starts on the following page) gives you an opportunity to express your satisfaction with the usability of your primary computer system. Your responses will help us understand what aspects of the system you are particularly concerned about and the aspects that satisfy you.

To as great a degree as possible, think about all the tasks that you have done with the system while you answer these questions.

Please read each statement and indicate how strongly you agree or disagree with the statement by circling a number on the scale. If a statement does not apply to you, circle N/A.

Whenever it is appropriate, please write comments to explain your answers.

Thank you!

1. Overall, I am satisfied with how easy it is to use this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

2. It is simple to use this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

3. I can effectively complete my work using this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

4. I am able to complete my work quickly using this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

5. I am able to efficiently complete my work using this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

6. I feel comfortable using this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

7. It was easy to learn to use this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

8. I believe I became productive quickly using this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

9. The system gives error messages that clearly tell me how to fix problems.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

10. Whenever I make a mistake using the system, I recover easily and quickly.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

11. The information (such as on-line help, on-screen messages and other documentation) provided with this system is clear.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

12. It is easy to find the information I need.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

13. The information provided with the system is easy to understand.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

14. The information is effective in helping me complete my work.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

15. The organization of information on the system screens is clear.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

Note: *The interface includes those items that you use to interact with the system. For example, some components of the interface are the keyboard, the mouse, the screens (including their use of graphics and language).*

16. The interface of this system is pleasant.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

17. I like using the interface of this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

18. This system has all the functions and capabilities I expect it to have.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**

19. Overall, I am satisfied with this system.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **STRONGLY AGREE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **STRONGLY DISAGREE** |

**COMMENTS:**