International Journal of Human-Computer Interaction
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hihc20

Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies

James R. Lewis

Version of record first published: 22 Jun 2011.

To cite this article: James R. Lewis (2002): Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies, International Journal of Human-Computer Interaction, 14:3-4, 463-488

To link to this article: http://dx.doi.org/10.1080/10447318.2002.9669130

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION, *14*(3&4), 463–488

Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies

James R. Lewis
IBM Corporation

Factor analysis of Post Study System Usability Questionnaire (PSSUQ) data from 5 years of usability studies (with a heavy emphasis on speech dictation systems) indicated a 3-factor structure consistent with that initially described 10 years ago: factors for System Usefulness, Information Quality, and Interface Quality. Estimated reliabilities (ranging from .83–.96) were also consistent with earlier estimates. Analyses of variance indicated that variables such as the study, developer, stage of development, type of product, and type of evaluation significantly affected PSSUQ scores. Other variables, such as gender and completeness of responses to the questionnaire, did not. Norms derived from this data correlated strongly with norms derived from the original PSSUQ data. The similarity of psychometric properties between the original and this PSSUQ data, despite the passage of time and differences in the types of systems studied, provide evidence of significant generalizability for the questionnaire, supporting its use by practitioners for measuring participant satisfaction with the usability of tested systems.

1. INTRODUCTION

1.1. Purpose of This Evaluation

The purpose of this evaluation was to investigate the psychometric characteristics of the Post Study System Usability Questionnaire (PSSUQ) using data from 5 years of lab-based usability evaluation. The research emphasis at the time of development of the PSSUQ was on enterprise-wide and networked office application suites (Lewis, Henry, & Mack, 1990). Over the last 5 years the majority of our use of the instrument has been in the evaluation of speech recognition systems (with a focus on speech dictation). The primary research question was whether the instrument, used for research in an area very different from that for the data used in the previous psychometric evaluations, would exhibit a factor structure, reliability, sensitivity, and norms consistent with the previous research. Replication of the previous findings with this new set of data would provide evidence of significant

Requests for reprints should be sent to James R. Lewis, IBM Corporation, 8051 Congress Ave, Suite 2227, Boca Raton, FL 33487. E-mail: jimlewis@us.ibm.com

generalizability for the questionnaire, supporting its use by practitioners for measuring participant satisfaction with the usability of tested systems.

1.2. History of the PSSUQ

The PSSUQ is a 19-item instrument designed for the purpose of assessing users' perceived satisfaction with their computer systems. It has its origin in an internal IBM project called SUMS (System Usability MetricS), headed by Suzanne Henry in the late 1980s. The mission of SUMS was to document and validate procedures for measuring system usability, including performance, usability problems, and user satisfaction.

At that time, there were a few efforts worldwide to develop instruments for the measurement of user satisfaction with system usability. In particular, the Questionnaire for User Interface Satisfaction (QUIS) at the University of Maryland (Chin, Diehl, & Norman, 1988), the Computer User Satisfaction Inventory (CUSI; Kirakowski & Dillon, 1988), and the Software Usability Measurement Inventory (Kirakowski & Corbett, 1993) at the University College of Cork in Ireland have had a significant influence on usability engineering practices. The System Usability Scale (Brooke, 1996) was also developed during the same time period, but because there has been no peer-reviewed research published on its psychometric properties, it has been less influential. (See LaLomia and Sidowski, 1990, for a review of usability questionnaires published before the PSSUQ; and see Lewis, 1995, for a comparison of the PSSUQ with the QUIS and CUSI.)

At the time we were working on SUMS, however, we did not know about these projects, so we developed our own standardized usability questionnaire. A team of IBM human factors and usability specialists working on SUMS created a pool of items hypothesized to relate to usability, and from those items we selected 18 to use systematically in usability evaluations as an end-of-study questionnaire named the PSSUQ (Lewis, 1991, 1992b).

In a separate unpublished study of customer perception of usability, 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 (quick completion of work, ease of learning, high-quality documentation and online information, and functional adequacy), but did not address rapid acquisition of productivity. We subsequently added an item (Item 8) to the PSSUQ to address this system characteristic and rearranged the order of items to correspond with the questionnaire's factors, producing the current version with 19 items (see the Appendix).

The development of the Computer System Usability Questionnaire (CSUQ) followed the development of the PSSUQ. Its items are identical to those of the PSSUQ except that their wording is appropriate for use in field settings or surveys rather than in a scenario-based usability evaluation, making it, essentially, an alternate form of the PSSUQ. The primary reason for the initial development of the CSUQ was to use it in a mailed questionnaire to obtain sufficient data on the PSSUQ–CSUQ items for a legitimate factor analysis. (At the time, the amount of laboratory data collected with the PSSUQ did not meet the conventional sample size standards for a factor analysis—

data from at least five participants for each item in the questionnaire.) For a discussion of this research and comparison of the PSSUQ and CSUQ items, see Lewis (1995).

1.3. Brief Review of Psychometric Theory and Trade-offs Considered in the Development of the PSSUQ

The primary purpose of this section is to provide a quick review of the basic elements of standard psychometric practice. The section also includes some discussion of trade-offs considered in the development of the PSSUQ (Lewis, 1999b). In this review, the term *scale* typically refers to a composite measurement based on responses to a number of items (a Likert scale). An *item* is a statement for which a participant selects a level of response. A scale *step* is an integer number indicating the participant's level of response to that item. See the Appendix for an example of an item with seven scale steps.

Goals of psychometrics. The goal of psychometrics is to establish the quality of psychological measures (Nunnally, 1978). Is a measure reliable (consistent)? Given a reliable measure, is it valid (measures the intended attribute)? Finally, is the measure appropriately sensitive to experimental manipulations?

Reliability goals. In psychometrics, reliability is quantified consistency, typically estimated using coefficient alpha (Nunnally, 1978). Coefficient alpha can range from 0 (*no reliability*) to 1 (*perfect reliability*). Measures of individual aptitude (such as IQ tests or college entrance exams) should have a minimum reliability of .90 (preferably a reliability of .95). For other research or evaluation, measurement reliability should be at least .70 (Landauer, 1988).

The initial assessments of the PSSUQ and CSUQ scales (Lewis, 1995) produced reliabilities exceeding .85, indicating their suitability for use in research and evaluation.

Validity goals. Validity is the measurement of the extent to which a questionnaire measures what it claims to measure. Researchers commonly use the Pearson correlation coefficient to assess criterion-related validity (the relation between the measure of interest and a different concurrent or predictive measure). Moderate correlations (with absolute values as small as .30–.40) are often large enough to justify the use of psychometric instruments (Nunnally, 1978).

Previous validity assessment of the PSSUQ indicated a significant correlation (r = .80) with other measures of user satisfaction obtained at the completion of each scenario and a significant correlation (r = –.40) with the measure of successful scenario completion (Lewis, 1995).

Sensitivity goals. A questionnaire that is reliable and valid should also be sensitive—capable of detecting appropriate differences. Statistically significant differences in the magnitudes of questionnaire scores for different systems or other usability-related manipulations provide evidence for sensitivity.

Analyses of variance conducted on the data used to assess the original PSSUQ (Lewis, 1995; Lewis et al., 1990) indicated that the PSSUQ was sensitive to user group and system differences. The CSUQ data (Lewis, 1995) indicated significant sensitivity to users' years of experience and breadth of experience with computer systems.

Goals of factor analysis. Factor analysis is a statistical procedure that examines the correlations among variables to discover groups of related variables (Nunnally, 1978). These groups of related variables (typically questionnaire items) become the basis for the development of Likert scales designed to reflect the underlying multidimensional nature of the construct under examination. 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. A weakness of factor analysis is that there are no strong methods for assessing the statistical significance of an estimated principal factor structure (Cliff, 1993), making replication of results with a different sample (as in this study) especially valuable.

Prior work with the PSSUQ and CSUQ indicated that the similarity of their item content resulted in very similar factor structures (Lewis, 1995). That work indicated that the questionnaires tapped into three aspects of a multidimensional construct (presumably usability). 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 PSSUQ–CSUQ scales) System Usefulness (SysUse), Information Quality (InfoQual), and Interface Quality (IntQual).

Number of scale steps. The more scale steps in a questionnaire item the better, but with rapidly diminishing returns (Nunnally, 1978). As the number of scale steps increases from 2 to 20, there is an initial rapid increase in reliability, but it tends to level off at about 7 steps. After 11 steps there is little gain in reliability from increasing the number of steps. The number of steps is important for single-item assessments, but is usually less important when summing scores over a number of items. Attitude scales tend to be highly reliable because the items typically correlate rather highly with one another.

As expected, PSSUQ and CSUQ reliabilities were high (Lewis, 1995), with coefficient alphas exceeding .89 for all scales. A related analysis (Lewis, 1993) showed that the mean difference of 7-point scales correlated more strongly than the mean difference of 5-point scales with the observed significance levels of *t* tests. Because there might be times when practitioners would be interested in item-level comparisons in addition to scale-level comparisons, the current versions of the PSSUQ and CSUQ use 7-point rather than 5-point scales.

Calculating scale scores. From classical 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 should 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 important statistical properties of the scores but does standardize the range of scale scores, making them easier to interpret and compare. For example, with items based on 7-point scales, all the summative scales would have scores that range from 1 to 7. For these reasons, it is the practice in our lab to average the responses given by a participant across the items for each scale.

Based on the factor analyses from Lewis (1995), the rules developed for calculating scale scores for the PSSUQ and CSUQ (see the Appendix for item format and content) were

- Overall: Average the responses to Items 1 through 19.
- SysUse: Average the responses to Items 1 through 8.
- InfoQual: Average the responses to Items 9 through 15.
- IntQual: Average the responses to Items 16 through 18.

Note that this method for calculating scale scores gives equal weight to each item in the scale. Although it is a standard practice to weight items equally, one consequence of this is that the resulting scales will correlate to some extent rather than being statistically independent.

> Although the factors themselves are uncorrelated, this does not mean that estimated factor scores are uncorrelated. … Usually they are only estimated, not obtained directly. … In these cases the estimated factor scores are likely to correlate substantially even if the factors themselves are orthogonal. (Nunnally, 1978, p. 434)

This is not usually a problem as long as (a) the correlations are not too close to 1.0 (avoiding multicollinearity) and (b) the scales are useful in interpreting measurement outcomes.

Control of potential response style or consistency in item alignment. It is a common practice in questionnaire development to vary the tone of items so, typically, one half of the items elicit agreement and the other half elicit disagreement. The purpose of this is to control potential measurement bias due to a respondent's response style. An alternative approach is to align the items consistently.

A potential criticism of the IBM questionnaires is that they do not use the standard control for potential measurement bias due to response style. Our rationale in consistently aligning the items was to make it as easy as possible for participants to complete the questionnaire. With consistent item alignment, the proper way to mark responses on the scales is clearer and requires less interpretive effort on the part of the participant (potentially reducing response errors due to participant confusion). Furthermore, the use of negatively worded items can produce a number of

undesirable effects (Ibrahim, 2001), including "problems with internal consistency, factor structures, and other statistics when negatively worded items are used either alone or together with directly worded stems" (Barnette, 2000, p. 363). The setting in which balancing the tone of the items is likely to be of greatest value is when participants do not have a high degree of motivation for providing reasonable and honest responses (e.g., in many clinical and educational settings).

> Thus, first and foremost, the survey or questionnaire designer must determine if using negatively worded items or other alternatives are needed in the context of the research or evaluation setting. Unless there is some pervasive and unambiguous reason for not doing so, it is probably best that all items be positively or directly worded and not mixed with negatively worded items. (Barnette, 2000, p. 363)

Obtaining reasonable and honest responses is rarely a problem in most usability evaluation settings.

Even if consistent item alignment were to result in some measurement bias due to response style, typical use of the IBM questionnaires is to compare systems or experimental conditions (a relative rather than absolute measurement). In this context of use, any systematic effect of response style (just like the effect of any other individual difference) will cancel out across comparisons.

For example, consider a within-subjects design in which participants provide ratings for two different systems, and the researcher plans to compute a difference score *t* test on the ratings. Assume that each participant has some degree of acquiescence (tendency to agree) that systematically affects his or her responses to the items. Therefore, each participant will produce two scores, each with two components: the true rating (x) and the effect of the response style (b). To get the difference score for the *t* test, the second score is subtracted from the first for each participant, which has the consequence of removing the influence of the individual's response style ($(x_1 + b) - (x_2 + b) = x_1 + b - x_2 - b = x_1 - x_2$).

As a second example, consider a between-subjects design in which a researcher has randomly selected participants from the same population and has randomly assigned them to work with one of two systems. Suppose that one of the measurements is overall system satisfaction obtained with the PSSUQ. As in the previous example, each score from each participant has two components: the true rating (x) and the effect of the response style (b). Averaging across participants, the observed mean for the first system will be Mean(x_1) + Mean(b); and the observed mean for the second will be Mean(x_2) + Mean(b). Note that the expected value of Mean(b) is the same for both groups because the basis for group membership was random assignment from the same population. The value for the numerator of an independent groups *t* test is the difference between the observed means. In this case, that difference would be (Mean(x_1) + Mean(b)) – (Mean(x_2) + Mean(b)), which equals (Mean(x_1) + Mean(b) – (Mean(x_2) – Mean(b)), which equals (Mean(x_1) – Mean(x_2)). In this example as in the first, the computations associated with group comparison have removed the effect of response style. When using the PSSUQ in this way, the presence or absence of an effect of response style on PSSUQ scores is moot.

Nunnally (1978, pp. 658–672) provided a review of the various types of response styles. The major types of styles that have been hypothesized to exist are social desirability (self-desirability), the tendency to guess when in doubt, the tendency to

guess 'true,' the agreement tendency (acquiescence), the extreme response tendency, and the deviant response tendency.

Social desirability is the tendency for some people to rate themselves as excessively good in self-report inventories, so it does not apply to usability questionnaires. Guessing tendencies can influence the scores on performance tests, but do not apply to Likert scales (such as those used in the PSSUQ and other usability questionnaires). Nunnally (1978) discounted the deviant response tendency as actually being a legitimate response style:

> It should be apparent that although the deviant-responding tendency has been discussed frequently as a response style, it is not a response style according to the definition given earlier. It is not an artifact of measurement; rather it comes from a special way of analyzing the valid variance. (p. 671)

The remaining response styles—the agreement tendency and the extreme response tendency—could affect the scores obtained on usability questionnaires. According to Nunnally (1978), however, "The overwhelming weight of the evidence now points to the fact that the agreement tendency is of very little importance either as a measure of personality or as a source of systematic invalidity in measures of personality and sentiments" (p. 669).

The extreme response tendency is the tendency to mark the extremes of rating scales rather than points near the middle of the scale. There is some evidence supporting the existence of this response style (from small correlations found between the degrees of extremeness of ratings made by respondents on different rating tasks such as picture preferences and attitudes toward minorities; Nunnally, 1978). Nunnally provided a method for empirically determining if a set of responses from a questionnaire exhibits evidence for the extreme response tendency. The basis of this method is the computation of the common shared variance between scores based on the full range of each item's scale and scores based on a dichotomous scoring of each item.

An emerging area of research in which extreme response and acquiescence response styles have become issues is the area of cross-cultural research (van de Vijver & Leung, 2001). There is some evidence that responses from members of different cultures exhibit different levels of these response styles, although these differences do not always appear (Grimm & Church, 1999). Matters are even more complicated when the members of the different cultures speak different languages, necessitating translation of items (and consequent uncertainty about the equivalence of items after translation). After collecting data from different cultures, it is possible to test for the effects of differential response sets using structural equations modeling (Cheung & Rensvold, 2000). The problems associated with cross-cultural testing, however, go beyond the issues of response style and include issues such as form invariance (whether the responses from the different cultures lead to the same item-to-factor relation), and the practice of balancing item tone does not guarantee form invariance between cultures.

Use of norms. When a questionnaire has norms, data exist that allow researchers to interpret individual and average scores as greater or smaller than the

expected norm scores (Anastasi, 1976). In some contexts (field studies, standard single-system usability studies), this can be a tremendous advantage. In other contexts (multiple-system comparative usability studies, other types of experiments), it might provide no particular advantage.

Even when norms exist, researchers should be cautious in their use. To apply norms in the usual way requires a substantial correspondence between the conditions under which the normative data were generated and those in the measurement situation. A valid set of norms would require correspondence between the normative and test situations with regard to participant, system, task, and environmental characteristics. Norms are of clear value in many situations, but it is important not to overgeneralize their applicability to usability evaluation.

Rather than using normative data for the purpose of specifying the location of a product on a usability scale (which will generally not be valid for the reasons stated earlier), it is possible to identify patterns in normative data that can be useful in interpreting PSSUQ results obtained in a usability study. For example, in the data used to originally evaluate the PSSUQ and CSUQ, the item that consistently received the poorest ratings was Item 9 ("The system gave error messages that clearly told me how to fix problems"). Another normative pattern was that InfoQual tended to receive poorer ratings than IntQual.

It is important to be careful when interpreting these normative patterns. Although InfoQual scores tend to be poorer than IntQual scores, this is not compelling evidence that the InfoQual of systems is poorer than their IntQual. The underlying distribution of scores might differ simply because the scales contain different items, with the items worded in their specific ways. On the other hand, it is reasonable to interpret patterns that differ markedly from the observed norms. For example, suppose a practitioner has conducted a usability evaluation with a reasonable sample size and finds that the mean InfoQual scores are about equal to the IntQual scores. Depending on the circumstances and the accompanying analysis of usability problems, this could mean that the IntQual is for some reason worse than normal or that the InfoQual is better than normal (e.g., if the developers had made a special effort to provide high-quality documentation).

1.4. Why Apply Classical Test Theory (CTT) Rather than Item Response Theory (IRT)?

For most of the previous century, the basis for psychometric theory was a set of techniques collectively known as CTT. Most of the psychometric training that psychologists (including myself) have received is in the techniques of CTT (Zickar, 1998). For a comprehensive treatment of basic CTT, see Nunnally (1978).

Starting in the last quarter of the 20th century (and accelerating in the last decade) was an alternative approach to psychometrics known as IRT. Although not yet generally accepted in psychology, IRT has had a major impact on educational testing, affecting the development and administration of the Scholastic Aptitude Test, Graduate Record Exam, and Armed Services Vocational Aptitude Battery. Some researchers have speculated that the application of IRT might improve the measurement of usability (Hollemans, 1999).

It is well beyond the scope of this article to explicate all the differences between CTT and IRT (for details, refer to a source such as Embretson & Reise, 2000). One of the key differences is that CTT focuses on scale-level measurement, but IRT focuses on modeling item characteristics. This property of IRT makes it ideal for adaptive computerized testing (Zickar, 1998), which is one of the reasons it has become so popular in large-scale educational testing. On the other hand, obtaining reliable estimates of the parameters of item response models requires data collection from a very large sample of respondents (Embretson & Reise, 2000), which can make IRT unattractive to researchers with limited resources. Furthermore, current IRT modeling procedures do not handle multidimensional measures very well (Embretson & Reise, 2000; Zickar, 1998). In addition to these limitations, the typical conception of the construct of usability is that it is an emergent property that depends on user, system, task, and environmental variables (the same variables that make it so difficult to develop usability norms). There are no existing IRT models that can account for all of these variables, and IRT is better suited for the measurement of latent rather than emergent variables (Embretson & Reise, 2000).

When the development of the PSSUQ began, IRT was virtually unknown in standard psychological psychometrics. IRT has made impressive gains in the last decade, but still does not appear to be adequate to model a construct as intricate as usability. Even if it were adequate, it is not clear that the additional effort involved would be worthwhile. Embretson and Reise (2000) observed that raw (CTT) scores and trait level (IRT) scores based on the same data correlate highly, and "no one has shown that in real data a single psychological finding would be different if IRT scores were used rather than raw scale scores" (p. 324). For these reasons, the analyses in this article will continue to apply CTT techniques to the evaluation of the PSSUQ usability questionnaire.

1.5. Advantages of Using Psychometrically Qualified Instruments

Despite any controversies regarding decisions made in the development of such questionnaires, standardized satisfaction measurements offer many advantages to the usability practitioner. Specifically, standardized measurements provide objectivity, replicability, quantification, economy, communication, and scientific generalization. Standardization also permits practitioners to use powerful methods of mathematics and statistics to better understand their results (Nunnally, 1978).

Although this is also an area of continuing controversy, many researchers hold that the level of measurement of an instrument (ratio, interval, ordinal) does not limit permissible arithmetic operations or related statistical operations, but instead limits the permissible interpretations of the results of these operations (Harris, 1985). Unless the scales meet certain criteria (Cliff, 1996; Embretson & Reise, 2000), measurements using Likert scales developed using CTT are ordinal. Despite this limitation, psychometrically qualified, standardized questionnaires can be valuable additions to a practitioner's repertoire of usability evaluation techniques.

1.6. PSSUQ and CSUQ: Summary of Prior Psychometric Evaluations

As discussed earlier, previous psychometric evaluations (Lewis, 1991, 1992a, 1992b, 1995) indicated that both the PSSUQ and CSUQ produced reliable overall composite scores and had three reliable factors that included the same items.

Investigations into scale validity found that the overall score of the PSSUQ correlated highly with other measures of user satisfaction taken after each scenario. The overall PSSUQ score, SysUse, and IntQual all correlated significantly with the percentage of successful scenario completion. Sensitivity analyses indicated that the PSSUQ and CSUQ scales responded appropriately to manipulations of system and user groups (novices and experts). The normative patterns of relatively poor ratings for Item 9 and InfoQual were consistent for the PSSUQ and the CSUQ.

The similarities between the outcomes for previous psychometric evaluations of the PSSUQ (lab data) and CSUQ (survey data) provided support to the generalizability of the instruments. Replication of the previous evaluations using a completely independent data set (the primary purpose for this investigation) would provide additional support regarding their generalizability of use by usability practitioners.

When applicable, the Results section of this article will include detailed data from the previous evaluations (Lewis, 1995) for comparison with data from this evaluation.

2. METHOD

The data analyzed in this report came from 21 unpublished usability studies conducted in our lab, during which participants completed the PSSUQ (paper-and-pencil administration) at the end of the study. All studies used the same version of the PSSUQ (see the Appendix for details regarding item format and content). Most of the studies (90%) were investigations of speech recognition systems (IBM and non-IBM systems), with an emphasis on speech dictation. The other studies were investigations of a personal communicator (Lewis, 1996) and a pen computing device. The PSSUQ database created from the questionnaires completed for this study had 210 entries from participants of widely varying backgrounds, computer experience, and age. With this database, it was possible to investigate the effect of the following independent variables on the profile of the PSSUQ scales: Study, Developer, Stage of Development, Type of Product, Type of Evaluation, Gender, and Completeness of Response. See section 3.3. Sensitivity for more detailed descriptions of these independent variables and the outcomes of their evaluation.

3. RESULTS AND DISCUSSION

3.1. Factor Analysis

Figure 1 shows the scree plot of the eigenvalues from the analysis. A discontinuity analysis (Coovert & McNelis, 1988) indicated a three-factor solution (note the increase in the difference between the third and fourth eigenvalues relative to the difference between the second and third, which is indicative of a three-factor solution). Table 1 shows the varimax-rotated, three-factor solution for this data as well

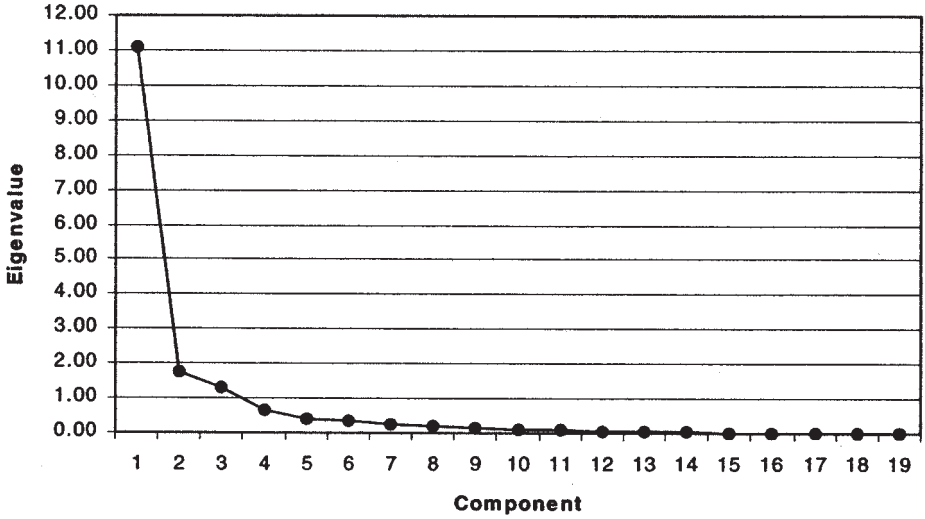

FIGURE 1 Scree plot from factor analysis.

Table 1: Varimax-Rotated, Three-Factor Solutions From Factor Analyses

	PSSUQ (Current)			*PSSUQ (Original)*			*CSUQ (Original)*		
Item	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 1*	*Factor 2*	*Factor 3*
Q1	**0.83**	0.38	0.23	**0.77**	0.26	0.43	**0.74**	0.36	0.26
Q2	**0.62**	0.46	0.20	**0.63**	0.35	0.46	**0.69**	0.41	0.16
Q3	**0.79**	0.35	0.17	**0.75**	0.38	0.25	**0.72**	0.21	0.36
Q4	**0.82**	0.25	0.22	**0.81**	0.45	0.07	**0.74**	0.31	0.33
Q5	**0.82**	0.26	0.32	**0.80**	0.16	0.36	**0.77**	0.30	0.32
Q6	**0.73**	0.40	0.20	**0.68**	0.38	0.48	**0.72**	0.22	0.27
Q7	***0.47***	***0.45***	0.38	**0.69**	0.46	0.40	**0.63**	0.49	0.13
Q8	**0.73**	0.19	0.29	na	na	na	**0.66**	0.39	0.26
Q9	0.32	**0.60**	0.13	0.05	**0.61**	0.24	0.23	**0.72**	0.21
Q10	***0.59***	***0.56***	0.14	0.36	**0.71**	0.24	0.34	**0.67**	0.28
Q11	0.24	**0.89**	0.21	0.45	**0.63**	0.25	0.23	**0.81**	0.20
Q12	0.28	**0.83**	0.15	0.44	**0.75**	0.22	0.24	**0.77**	0.27
Q13	0.32	**0.81**	0.13	0.43	**0.70**	0.32	0.38	**0.76**	0.17
Q14	0.36	**0.79**	0.21	0.43	**0.74**	0.40	0.40	**0.73**	0.18
Q15	0.15	**0.51**	0.47	0.30	***0.59***	***0.56***	0.34	**0.57**	0.40
Q16	0.20	0.19	**0.86**	0.30	0.36	**0.75**	0.33	0.27	**0.81**
Q17	0.36	0.10	**0.86**	0.37	0.36	**0.76**	0.38	0.26	**0.81**
Q18	0.38	0.27	**0.54**	0.22	0.28	**0.80**	0.34	0.35	**0.56**
Q19	**0.76**	0.27	0.37	***0.58***	0.22	***0.64***	***0.66***	0.37	***0.50***

Note. PSSUQ = Post Study System Usability Questionnaire; CSUA = Computer System Usability Questionnaire. Bold type indicates a large factor loading (0.5 or greater, except for Current Q7, with the criterion adjusted to 0.45). Bold italics indicates large factor loadings for an item on more than one factor.

as for the original PSSUQ and CSUQ studies. This three-factor solution explained 72.5% of the variance in the data.

This factor structure was very similar to the structure previously reported for the PSSUQ and CSUQ (Lewis, 1995), with a few minor differences. In this analysis, the 19th item loaded strongly on the first factor (SysUse), whereas in the past it loaded about equally on the first and third factors (SysUse and IntQual). Items 7 and 10 loaded about equally on the first and second factors (SysUse and InfoQual). In the previous evaluations, Item 7 loaded most strongly on SysUse, and Item 10 loaded most strongly on InfoQual. For purposes of the following evaluations, SysUse includes all of its former items (1–8), plus the 19th item. For continuity, I resolved the ambiguities in the factor analysis in favor of the existing PSSUQ scale definitions (Items 9–15 for InfoQual, Items 16–18 for IntQual).

As expected (Nunnally, 1978), an analysis of the correlations among the estimated factor scores showed substantial correlation: SysUse–InfoQual, $r(203) = .72$; SysUse–IntQual, $r(207) = .67$; InfoQual–IntQual, $r(203) = .56$; all ps $< .000002$). In the initial PSSUQ study (Lewis, 1995), the same pairs of estimated factor scores had correlations of .71, .68, and .64, respectively; and in the initial CSUQ study, the correlations were .67, .71, and .61. Therefore, across the studies, the intercorrelations appear to be about .7, .7, and .6, respectively; so the estimated factor scores share about 36% to 50% of their variance. Although these correlations are significantly different from zero, they are not so close to one that they would cause multicollinearity problems during statistical analyses.

3.2. Reliability

Estimates of reliability using coefficient alpha indicated levels of reliability for the overall PSSUQ and its factors that were consistent with previous estimates (shown in Table 2—Values for original PSSUQ and CSUQ are from Lewis, 1995). All the reliabilities exceeded .80, indicating that they have sufficient reliability to be valuable as usability measurements (Anastasi, 1976; Landauer, 1988).

Because it is possible to obtain high reliabilities for a scale by including multiple items that mean the exactly the same (either worded in the same way or in linguistically similar ways), some critics of the PSSUQ have suggested that this could be the basis for its highly reliable scales. Specifically, they have noted the similarity among Items 3 (effective task completion), 4 (quick task completion), and 5 (efficient task completion) in SysUse; and between Items 11 (clear information) and 13

Table 2: Current and Previous Estimates of Reliability for PSSUQ Scales

Study	*Overall*	*SysUse*	*InfoQual*	*IntQual*
PSSUQ (Current)	0.96	0.96	0.92	0.83
PSSUQ (Original)	0.97	0.96	0.91	0.91
CSUQ	0.95	0.93	0.91	0.89

Note. PSSUQ = Post Study System Usability Questionnaire; SysUse = system usefulness; InfoQual - informational quality; IntQual = interface quality; CSUQ = computer system usability questionnaire.

(information easy to understand) in InfoQual (see the Appendix for the complete wording of the items).

To investigate the possibility that the high reliability for the PSSUQ scales was due to these highly similar items, I recalculated the reliabilities for SysUse without Items 3 and 5, InfoQual without Item 13, and the revised Overall scale without these items. Without Items 3 and 5, the reliability of SysUse fell from .96 to .90—still very high. InfoQual declined slightly from .93 to .91. The effect on the overall measurement of removing the three items was to reduce coefficient alpha from .96 to .94—a negligible reduction.

The correlation between the original scores and the revised scores was 0.99 for SysUse, 1.00 for InfoQual, and 1.00 for Overall. The differences between the mean scores for the original and revised versions of these scales (with 99% confidence intervals) were .05 ± .02, –.06 ± .02, and .01 ± .01, respectively. The SysUse mean shifted up slightly (somewhere between .03 and .07—less than one tenth of a scale step), the InfoQual mean shifted down slightly (somewhere between –.04 and –.08—again less than one tenth of a scale step), and the net effect was that the Overall score essentially did not change (somewhere between 0 and .02—less than one fiftieth of a scale step).

3.3. Sensitivity

The mean values of the PSSUQ factors were 2.8 for SysUse, 3.0 for InfoQual, and 2.5 for IntQual, with 2.8 for both (a) the composite score collapsed across all 19 items and (b) the mean of the scale scores. Note that the equality of these two ways of computing the composite scores (averaging across all items or averaging across scale scores) indicates that there is no need to develop weighting schemes to compensate for the difference in the number of items per scale. In all analyses, lower scores indicate better ratings, and alpha was .05.

Analyses of variance conducted to investigate the sensitivity of PSSUQ measures indicated that the following variables significantly affected PSSUQ scores (as indicated by a main effect, an interaction with PSSUQ factors, or both):

- Study (21 levels—the study during which participants completed the PSSUQ).
- Developer (4 levels—the company that developed the product under evaluation).
- Stage of development (2 levels—product under development or available for purchase).
- Type of product (5 levels—discrete dictation, continuous dictation, game, personal communicator, or pen product).
- Type of evaluation (2 levels—dictation study or standard usability evaluation).

The following variables did not significantly affect PSSUQ scores:

- Gender (2 levels—male or female).
- Completeness of responses to questionnaire (2 levels—complete or incomplete).

The details for each of these analyses follows.

Study. Both the main effect, $F(20, 184) = 2.2$, $p = .004$, and the interaction, $F(40, 368) = 3.2$, $p = .000000003$ (see Figure 2) were significant; overall means across the 21 studies (labeled using the letters A through U for reasons of confidentiality) ranged from 1.9 to 4.2.

Note that (a) none of the lines were horizontal and (b) the magnitude of differences among SysUse, InfoQual, and IntQual varied across the studies (in other words, the lines were not parallel)—patterns that indicate scale sensitivity.

Developer. This variable refers to the company that developed the product under study (with companies coded as CIC, CDC, CKC, and CMC for reasons of confidentiality). Both the main effect, $F(3, 201) = 3.4$, $p = .02$, and the interaction, $F(6, 402) = 3.6$, $p = .002$ (see Figure 3) were significant; overall means across developers ranged from 2.5 to 3.3.

The pattern of results were similar to those for Study in that (a) none of the lines were horizontal and (b) the magnitude of differences among SysUse, InfoQual, and IntQual varied across developers—overall patterns indicative of scale sensitivity.

Stage of development. This variable refers to whether the investigated product was in development or available for purchase. Both the main effect, $F(1, 203) = 4.2$, $p = .04$, and the interaction, $F(2, 206) = 3.1$, $p = .05$ (see Figure 4) were significant. Products under development received better ratings than products available for purchase (overall means of 2.6 and 3.0, respectively). The pattern of the interaction (assessed using Bonferroni t tests with $\alpha = .017$) was that mean ratings of IntQual differed by 0.2 (development, 2.4; product, 2.6—$t(207) = 1.1$, $p = .28$), ratings of InfoQual differed by 0.3 (development, 2.9; product, 3.2—$t(207) = 1.7$, $p = .08$), and ratings of SysUse differed by 0.5 (development, 2.6; product, 3.1—$t(207) = 3.1$, $p = .002$).

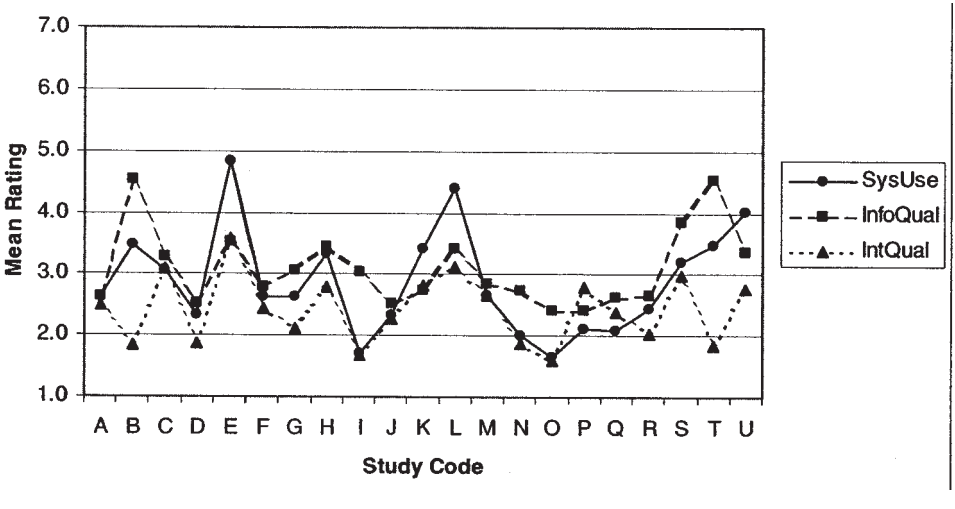

FIGURE 2 Study × Factor interaction.

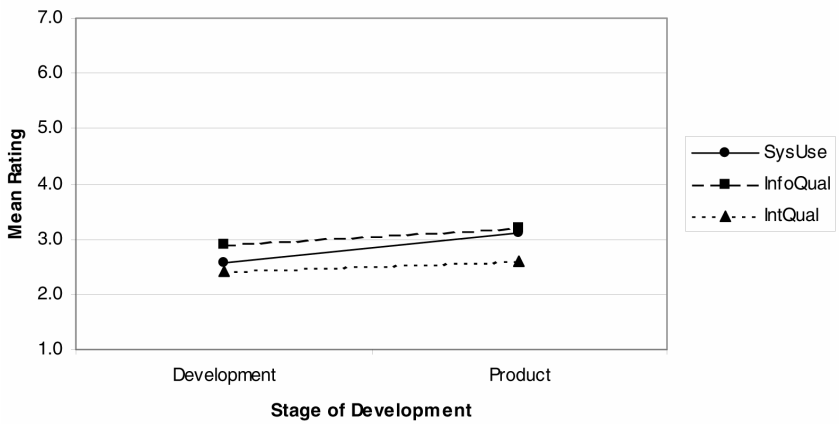

FIGURE 3 Developer × Factor interaction.

This outcome was a somewhat surprising result that might be due to a number of factors. For example, when evaluating a product under development, the range of tasks that the product can perform is more limited than will be the case once the product is complete. This limited functionality affects the number (and possibly the complexity) of tasks that an evaluator can ask participants to perform with the product (which is consistent with the significant difference for SysUse).

Type of product. This variable refers to the type of product under investigation. The types of products included

- Continuous dictation products: Products that allow users to speak continuously when dictating text.
- Discrete dictation products: Products that require users to briefly pause between words when dictating.

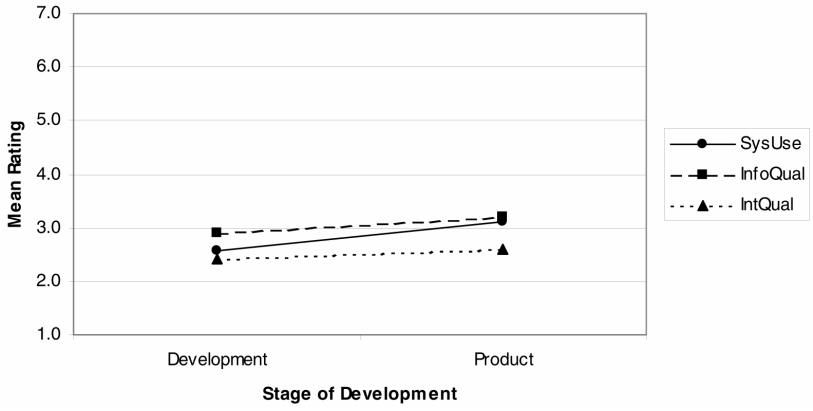

FIGURE 4 Stage of Development × Factor interaction.

- Speech control of computer games: Product that allowed users to control computer games by issuing voice commands.
- Personal communicator: A combination cellular phone and personal digital assistant device.
- Pen computing device: A device for capturing and managing handwritten notes.

The main effect, $F(4, 200) = 1.9$, $p = .11$, was not significant; but the interaction, $F(8, 400) = 2.3$, $p = .02$, was (see Figure 5).

As was the case for the variables of Study and Developer, (a) the lines were not horizontal and (b) differences among the scales were not identical across developers—patterns that provide evidence of sensitivity to the product type.

Type of evaluation. This variable refers to the type of evaluation conducted in the study: *dictation* and *standard.* Dictation refers to the use of a specific protocol for the measurement of dictation speed and accuracy (Lewis, 1997, 1999a). The task in the dictation studies was for a user to dictate from written source text and, in some studies, to also use the system to compose documents. In most dictation studies, participants received training in how to dictate and correct, and rarely consulted any system documentation.

Standard refers to the use of a standard scenario-based usability problem discovery protocol (e.g., see Lewis et al., 1990). In this protocol, the typical procedure was for participants to receive descriptions of tasks to complete with the system under evaluation. In most cases, the tasks were organized within scenarios designed to provide broad functional coverage. In most standard evaluations, participants had access to system documentation and used it as required.

The main effect, $F(1, 203) = .004$, $p = .99$, was not significant; but the interaction, $F(2, 406) = 7.6$, $p = .001$, was (see Figure 6). Post hoc examination of the interaction using Bonferroni t tests (with $\alpha = .008$) indicated that for dictation studies,

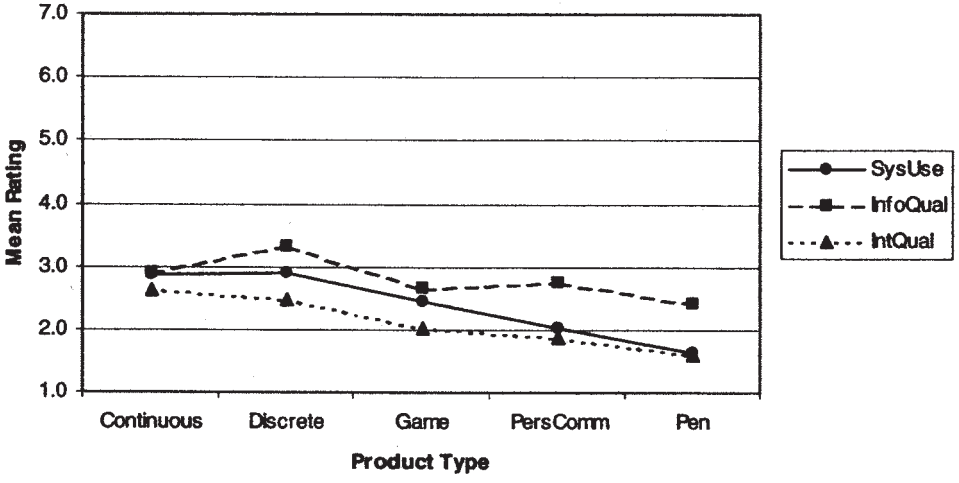

FIGURE 5 Product Type × Factor interaction.

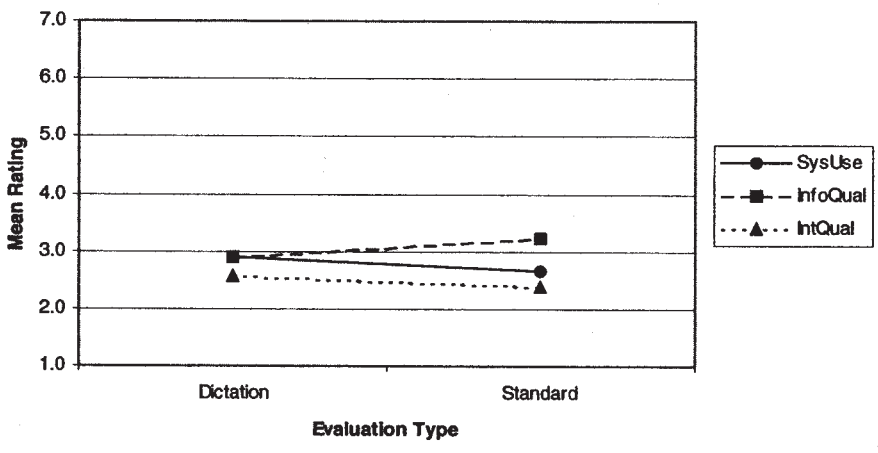

FIGURE 6 Evaluation Type × Factor interaction.

SysUse and InfoQual were not significantly different from one another, $t(123) = .6$, $p = .54$; but both were significantly different from IntQual, $t(127) = 4.2$, $p = .0001$, and $t(123) = 4.8$, $p = .000005$, respectively. For dictation studies, SysUse and IntQual were not significantly different from one another, $t(80) = 1.95$, $p = .05$; but both were significantly different from InfoQual, $t(80) = 5.9$, $p = .0000001$, and $t(80) = 5.28$, $p = .000001$, respectively.

Keeping in mind that these data are not from a designed experiment, it seems reasonable that the difference in the use of system documentation between the evaluation methods (not used in dictation studies, used in standard studies) could account for the difference in the PSSUQ scale patterns. Therefore, these results do not only indicate scale sensitivity by virtue of a significant interaction, but also by virtue of the different behavior of InfoQual as a function of the type of study.

Gender. Neither the main effect, $F(1, 194) = .12$, $p = .74$, nor the interaction, $F(2, 388) = 1.8$, $p = .17$, were significant. The difference between the female and male questionnaire means for each of the PSSUQ scales was only 0.1. Although evidence of gender differences would not affect the usefulness of the PSSUQ, it is of potential interest to practitioners that the instrument does not appear to have an inherent gender bias.

Completeness of responses to questionnaire. Neither the main effect, $F(1, 203) = .26$, $p = .61$, nor the interaction, $F(2, 406) = 1.3$, $p = .28$, were significant (see Figure 7). The difference between the complete and incomplete questionnaire means for each of the PSSUQ scales was only 0.1.

This finding is important because it supports the practice of including partially completed questionnaires when averaging items to compute scale scores (rather than discarding the data from partially completed questionnaires).

Analysis of the distribution of incomplete questionnaires in the analyzed database showed that of 210 total questionnaires, 124 (59%) were complete and 86 (41%)

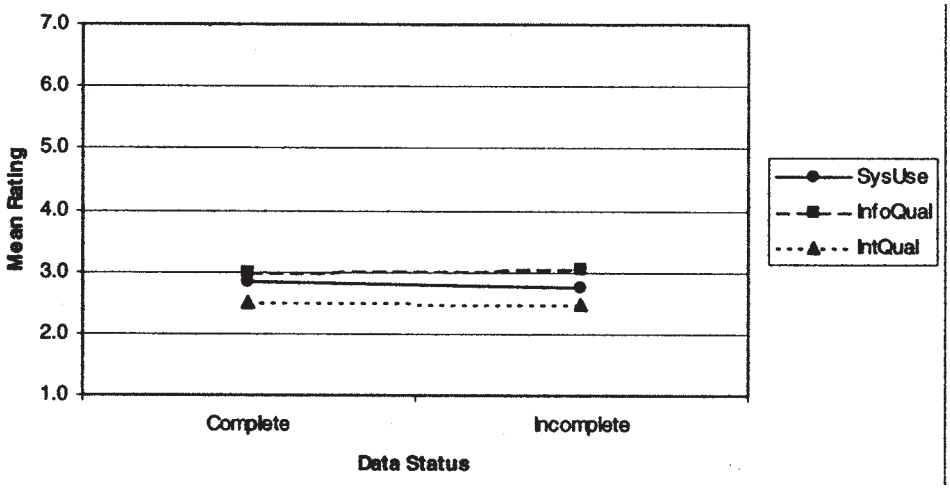

FIGURE 7 Completeness × Factor interaction.

were incomplete. For the incomplete questionnaires, the mean number of items (with 95% confidence interval bounds) for Overall (19 items), SysUse (9 items), InfoQual (7 items), and IntQual (3 items) were, respectively, 15.8 ± 0.5, 8.8 ± 0.1, 4.2 ± 0.5, and 2.9 ± 0.1. Across the incomplete questionnaires, the completion rate for all SysUse and IntQual items exceeded 85% (averaging 95% and 97%, respectively); but the average completion rate for InfoQual items was only 60%. These data indicate that the primary cause of an incomplete questionnaire was the failure to complete one or more InfoQual items. In most cases (78%), these incomplete questionnaires came from dictation studies (which did not typically include documentation) or standard usability studies conducted on prototypes without documentation.

3.4. Norms

Table 3 shows the means and 99% confidence intervals for each item from this PSSUQ data and from the original PSSUQ and CSUQ data sets (Lewis, 1995). Figure 8 illustrates the patterns of the means for these three sets of data.

As discussed previously in Section 1.3. Use of norms, there are probably very few cases in which practitioners can use these norms for the direct assessment of a product under evaluation. The data for this evaluation come from a variety of sources that included different types of products at different stages of development and the performance of different types of tasks. The original PSSUQ data came from a more consistent source, which included the assessment of three different systems using a set of benchmark tasks developed for the study of office systems performed by participants with different levels of computer experience (for details, see Lewis et al., 1990). The original CSUQ data came from a survey conducted over a broad range of users at IBM. The original PSSUQ and CSUQ data are, however, over 10 years old, which casts some doubt on their usefulness as norms for current

Table 3: Means and 99% Confidence Intervals for PSSUQ and CSUQ Norms

	PSSUQ (Current)			*PSSUQ (Original)*			*CSUQ (Original)*		
Item	*Lower Limit*	*Mean*	*Upper Limit*	*Lower Limit*	*Mean*	*Upper Limit*	*Lower Limit*	*Mean*	*Upper Limit*
Q1	2.60	**2.85**	3.09	3.36	**4.00**	4.64	3.12	**3.30**	3.48
Q2	2.45	**2.69**	2.93	3.40	**4.02**	4.64	3.36	**3.54**	3.72
Q3	2.58	**2.85**	3.11	3.07	**3.73**	4.40	2.73	**2.91**	3.09
Q4	2.86	**3.16**	3.45	3.53	**4.15**	4.76	3.09	**3.27**	3.45
Q5	2.79	**3.06**	3.34	3.37	**3.98**	4.59	3.05	**3.23**	3.41
Q6	2.40	**2.66**	2.91	2.75	**3.41**	4.07	2.77	**2.95**	3.13
Q7	2.07	**2.27**	2.48	2.92	**3.57**	4.22	3.61	**3.82**	4.03
Q8	2.54	**2.86**	3.17	na	**na**	na	3.40	**3.61**	3.82
Q9	3.36	**3.70**	4.05	4.38	**4.93**	5.48	4.58	**4.79**	5.00
Q10	2.93	**3.21**	3.49	3.64	**4.18**	4.73	3.82	**4.03**	4.24
Q11	2.65	**2.96**	3.27	3.87	**4.48**	5.09	3.94	**4.15**	4.36
Q12	2.79	**3.09**	3.38	3.42	**4.02**	4.63	4.11	**4.32**	4.53
Q13	2.37	**2.61**	2.86	3.15	**3.79**	4.43	3.95	**4.13**	4.31
Q14	2.46	**2.74**	3.01	2.81	**3.43**	4.04	3.70	**3.88**	4.06
Q15	2.41	**2.66**	2.92	3.02	**3.55**	4.08	3.43	**3.61**	3.79
Q16	2.06	**2.28**	2.49	2.32	**2.91**	3.51	3.01	**3.19**	3.37
Q17	2.18	**2.42**	2.66	2.37	**2.92**	3.47	3.02	**3.20**	3.38
Q18	2.51	**2.79**	3.07	2.44	**3.00**	3.56	3.47	**3.68**	3.89
Q19	2.55	**2.82**	3.09	3.10	**3.69**	4.29	3.13	**3.31**	3.49
SysUse	2.57	**2.80**	3.02	3.26	**3.81**	4.36	3.19	**3.34**	3.49
InfoQual	2.79	**3.02**	3.24	3.58	**4.06**	4.54	3.95	**4.13**	4.31
IntQual	2.28	**2.49**	2.71	2.42	**2.93**	3.43	3.17	**3.35**	3.53
Overall	2.62	**2.82**	3.02	3.30	**3.76**	4.22	3.43	**3.61**	3.79

Note. PSSUQ = Post Study System Usability Questionnaire; CSUQ = Computer System Usability Questionnaire; SysUse = system usefulness; InfoQual = information quality; IntQual = interface quality. Means apear in bold face.

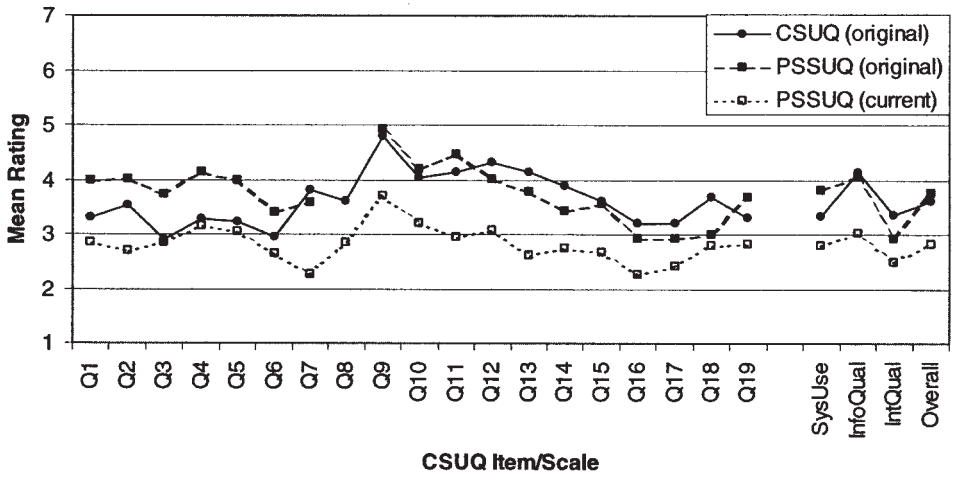

FIGURE 8 Means for three sets of Post Study System Usability Questionnaire–Computer System Usability Questionnaire (PSSUQ–CSUQ) norms.

systems, even if the conditions of evaluation were similar to those for the original PSSUQ and CSUQ.

The consistently better mean ratings in this PSSUQ compared to the original PSSUQ data do not necessarily indicate a wholesale improvement in system usability over the last 10 years (although this is one possible explanation and might contribute to the differences). It is also possible that the differences are due to differences in participant populations, system characteristics, or tasks.

Despite this, there are some interesting and potentially useful patterns in the means from the three sets of data. The item data show substantial correlation, with the greatest correlation between the PSSUQ means (CSUQ–PSSUQ current: $r(19) = .49$, $p = .03$; CSUQ–PSSUQ original: $r(18) = .57$, $p = .01$; PSSUQ current–PSSUQ original: $r(18) = .81$, $p = .0005$). For all three sets of data, the item that received the poorest rating—averaging from .45 to .49 above the next poorest item in the respective set—was Item 9 ("The system gave error messages that clearly told me how to fix problems"). Finally, mean ratings of InfoQual tend to be higher (poorer) then mean ratings of IntQual, with differences for the three data sets ranging from 0.5 to 1.1. There are several ways in which these findings can be of use to practitioners.

The consistently poor ratings for Item 9 indicate

1. This should not surprise practitioners if they find this in their own data.
2. It really is difficult to provide usable error messages throughout a product.
3. It may well be worth the effort to make the effort to focus on providing usable error messages.
4. If practitioners find the mean for this item to be equal to or less than the mean of the other items in InfoQual, they have been successful in addressing the problem.

The consistent pattern of poor ratings for InfoQual relative to IntQual suggest that practitioners who find this in their data should not necessarily conclude that they have poor documentation or a great interface. On the other hand, if this pattern appeared in the first iteration of a usability evaluation and the developers decided to emphasize improvement to the quality of their information, then any significant decline in the difference between InfoQual and IntQual would be evidence of a successful intervention.

3.5. Extreme Response Tendency

The extreme response tendency is the tendency to mark the extremes of rating scales rather than points near the middle of the scale. The procedure for determining if a set of responses from a questionnaire exhibits evidence for the extreme response tendency (Nunnally, 1978) is to

1. Score the responses in two ways—first as the sum of deviations from the center point of the scale using the number of scale steps in the instrument's items, then as the sum of dichotomized scores. Because the PSSUQ uses items with seven

scale steps, the effect of dichotomization is that ratings from 1 to 3 become 0, a rating of 4 becomes 0.5; and ratings from 5 to 7 become 1.

2. Divide the squared correlation between the two sets of scores by the product of their internal reliability coefficients (coefficient alphas) to get an estimate of their shared common variances. If that ratio is considerably lower than 1.0 (Nunnally suggests 0.8 as a criterion), it is reasonable to assume that the extremeness tendency is present to some degree.

The obtained ratios for Overall, SysUse, InfoQual, and IntQual were, respectively, .94, 1.02, 1.18, and .95. Because all ratios were greater than .80, there was no evidence for an extremeness tendency for any of the scales. (I also performed the same procedure with truly dichotomous scores, once scoring the central scale point of 4 as 0 and once scoring it as 1. In both cases, the results were essentially the same as with the procedure discussed earlier—no evidence for an extremeness tendency for any of the scales.)

4. GENERAL DISCUSSION

The primary purpose of this research was to investigate the similarity between the initially published psychometric properties of the PSSUQ (Lewis, 1995) and estimates of the same properties using data from 5 years of lab-based usability evaluation. The key research questions were whether the PSSUQ, used for research in an area very different from that for the previous psychometric evaluations, would exhibit a factor structure, reliability, sensitivity, and norms consistent with the previous research. Replication of the previous findings with this new set of data would provide evidence of significant generalizability for the questionnaire, supporting its use by practitioners for measuring participant satisfaction with the usability of tested systems. Failure to replicate would provide information on appropriate limits of generalization for the psychometric properties of the PSSUQ.

4.1. Results Were Acceptable and Consistent With Previous Research

Although the analyzed data came from studies that differed in both content and protocol from the studies that generated the data for previous analyses, the factor structure, scale reliabilities, and sensitivity analyses were all consistent with prior results (Lewis, 1995), and all reached acceptable levels according to standard psychometric criteria. The reliability of IntQual has been the most variable across studies, possibly because it has the fewest (only three) items. The profiles of item means across the evaluations also showed strong similarity.

The investigation into the effect of removing Items 3 and 5 from SysUse and Item 13 from InfoQual indicated that the high PSSUQ scale reliabilities were not dependent on the inclusion of these items and that the removal of the items had no substantive effect on the scale means. Practitioners can now treat these items as optional when using the PSSUQ, either using the traditional version or a version that has only 16 items (16% shorter). (Personally, I plan to use the shorter version in fu-

ture studies because the optional items do not provide much additional information but do consume study time.)

In summary, the analyses support the continued use by usability practitioners of the PSSUQ and its historical factors as a measure of user perception of and satisfaction with product usability in scenario-based usability evaluations.

4.2. PSSUQ Ratings Were Insensitive to Gender

Although it would have been acceptable to have detected PSSUQ response differences as a function of gender, the insensitivity of the PSSUQ to gender makes it easier to interpret and report PSSUQ scores because there will typically be no reason to present scores broken down by gender. Despite this, practitioners should still plan to include gender as a variable in their analyses for cases in which different genders might react differently to a product.

4.3. PSSUQ Ratings Tended to be Robust Even When Questionnaires Were Incomplete

Based on psychometric theory (CTT), I had hypothesized in earlier articles that the failure to complete all items in the questionnaire should not invalidate the responses or the gathering of the available responses into scale scores by averaging across the available items (Lewis, 1995). The basis for this hypothesis was that, according to CTT, 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 (Nunnally, 1978). If a participant chooses not to answer an item, the effect should be to reduce slightly the reliability of the scale in that instance; and, in most cases, the remaining items should offer a reasonable estimate of the appropriate scale score. The nonsignificant main effect and interaction for the completeness variable supported this hypothesis and, by extension, our current practice of computing mean scale scores from PSSUQs that participants have not fully completed.

This is an important finding because one of the criticisms made by IRT practitioners regarding CTT scales is that "by not including item properties in the model, true score can apply only to a particular set of items or their equivalent" (Embretson & Reise, 2000, p. 53). Taken to its extreme, this would mean that adding or deleting even one item from a scale developed using CTT could render its scores invalid (Hollemans, 1999). The essential equivalence of mean scores for complete and incomplete PSSUQs in this study is consistent with the expectation of equivalence implied by CTT. The questionnaires that result from the decisions of participants to leave some responses blank are apparently equivalent to the standard PSSUQ (at least, when averaged over a number of participants). Practitioners using the PSSUQ need not fear that the failure of participants to complete all items makes the obtained PSSUQ items worthless for the purpose of computing its scale scores.

These data do not provide information concerning how many items a participant can ignore and still produce reliable scale scores. The data do suggest that, in practice, participants typically complete enough items to produce reliable scale scores.

4.4. Absolute Normative Data is of Limited Value, but Normative Patterns Can Be Useful

Practitioners should be cautious when attempting to interpret their own absolute PSSUQ or CSUQ data against the norms presented in this article. On the other hand, relative patterns of data that appear consistently in the norms can be of interpretative value. One example is the consistently poor rating for Item 9 ("The system gave error messages that clearly told me how to fix problems"). Another is the consistent difference between the InfoQual and IntQual scores. Because these patterns appeared consistently in the original and these evaluative studies of the PSSUQ and CSUQ, practitioners should expect to find these patterns in their data. Deviations from these patterns are potentially meaningful, especially if the practitioner has focused on the development of clear error messages and high-quality information and finds that the means for these scores are consistent with the means of the other items and scales and are consistent with observed usability problems.

4.5. Response Styles and PSSUQ Scores

Of the hypothesized response styles (Nunnally, 1978), the ones that might reasonably affect PSSUQ scores are the agreement tendency and the extreme response tendency. The computation of the shared common variance for deviation and dichotomous scores indicated no presence of an extreme response tendency in the current PSSUQ rating data. Nunnally reviewed the evidence for the agreement tendency and concluded that it was of little importance as a source of scale invalidity.

Some recent research (Baumgartner & Steenkamp, 2001; Clarke, 2001; van de Vijver & Leung, 2001) indicates that there could be significant differences among different cultures with regard to the agreement tendency and the extreme response tendency. Practitioners should avoid using the PSSUQ for the purpose of comparing different cultural groups unless there is evidence that the groups do not differ in these tendencies or the practitioners are prepared to test after the fact for model equivalence (Cheung & Rensvold, 2000). Note that this is not a limitation that applies only to the PSSUQ, but is a limitation of any similar questionnaire if used to compare groups from different cultures.

The strongly nonsignificant outcomes for PSSUQ sensitivity to gender do suggest that, at least in the culture of the United States, there is no difference between men and women with regard to these potential tendencies when completing the PSSUQ. There also appears to be no such difference between participants who complete all of the PSSUQ items at the end of a study and those who do not. It is interesting that of the seven sensitivity assessments, the PSSUQ exhibited evidence of sensitivity when the method for parsing the data was to divide the systems into different groups (study, developer, stage of development, type of product, type of evaluation). In contrast, the PSSUQ was insensitive when the basis for parsing the data was to divide the respondents into different groups (gender, completeness of responses). This difference is not strongly compelling evidence of a lack of influence of response style on PSSUQ scores (as suggested by Nunnally, 1978, for tests of sentiments in general), but it is consistent with such a hypothesis.

Finally, it is important to keep in mind that when used as a dependent measure in a standard within- or between-subject experimental design in which cultural differences are not an independent variable, any effect of response style will cancel out across experimental conditions. When used in this way, the presence or absence of effects of response styles on PSSUQ scores is moot.

5. CONCLUSION

The similarity of psychometric properties between the original and current PSSUQ data, despite the passage of time and differences in the types of systems studied, provide evidence of significant generalizability for the questionnaire, supporting its use by practitioners for measuring participant satisfaction with the usability of tested systems.

Due to its generalizability, practitioners can confidently use the PSSUQ when evaluating different types of products and at different times during the development process. Practitioners should be cautious about using the PSSUQ to compare the attitudes of different cultural groups. The PSSUQ can be especially useful in competitive evaluations (see Lewis, 1996) or when tracking changes in usability as a function of design changes made during development (either within a version or across versions).

REFERENCES

Anastasi, A. (1976). *Psychological testing.* New York: Macmillan.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60,* 361–370.

Baumgartner, H., & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38,* 143–156.

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). London: Taylor & Francis.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31,* 187–212.

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

Clarke, I. (2001). Extreme response style in cross-cultural research. *International Marketing Review, 18,* 301–324.

Cliff, N. (1993). *Analyzing multivariate data.* San Diego, CA: Harcourt Brace.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Grimm, S. D., & Church, A. T. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality, 33,* 415–441.

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

Hollemans, G. (1999). User satisfaction measurement methodologies: Extending the user satisfaction questionnaire. In *Proceedings of HCI International '99 8th International Conference on Human–Computer Interaction* (pp. 1008–1012). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ibrahim, A. M. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports, 88,* 497–500.

Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology, 24,* 210–212.

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: Elsevier.

Lewis, J. R. (1991). *User satisfaction questionnaires for usability studies: 1991 manual of directions for the ASQ and PSSUQ* (Tech. Rep. No. 54.609). Boca Raton, FL: International Business Machines Corporation.

Lewis, J. R. (1992a). *Psychometric evaluation of the computer system usability questionnaire: The CSUQ* (Tech. Rep. No. 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. (1993). Multipoint scales: Mean and median differences and observed significance levels. *International Journal of Human–Computer Interaction, 5,* 383–392.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction, 7,* 57–78.

Lewis, J. R. (1996). Reaping the benefits of modern usability evaluation: The Simon story. In G. Salvendy & A. Ozok (Eds.), *Advances in applied ergonomics: Proceedings of the 1st International Conference on Applied Ergonomics—ICAE '96* (pp. 752–757). Istanbul, Turkey: USA Publishing.

Lewis, J. R. (1997). *A general plan for conducting human factors studies of competitive speech dictation accuracy and throughput* (Tech. Rep. No. 29.2246). Raleigh, NC: IBM Corp.

Lewis, J. R. (1999a). *Streamlining a general test plan for competitive evaluation of dictation accuracy and throughput* (Tech. Rep. No. 29.3158). Raleigh, NC: IBM Corp.

Lewis, J. R. (1999b). Tradeoffs in the design of the IBM computer usability satisfaction questionnaires. In *Proceedings of HCI International '99 of the 8th International Conference on Human–Computer Interaction* (pp. 1023–1027). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lewis, J. R., Henry, S. C., & Mack, R. L. (1990). Integrated office software benchmarks: A case study. In *Human–Computer Interaction—INTERACT '90* (pp. 337–343). Cambridge, England: Elsevier.

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

van de Vijver, F. J. R., & Leung, K. (2001). Personality in cultural context: Methodological issues. *Journal of Personality, 69,* 1007–1031.

Zickar, M. J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychological Science, 7,* 104–109.

APPENDIX

The Post-Study System Usability Questionnaire Items

The first item illustrates the item format. The remaining items show only the item text to conserve space. Each item also has an area for comments (not shown).

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

STRONGLY AGREE						STRONGLY DISAGREE	
1	2	3	4	5	6	7	N/A

2. It was simple to use this system.
3. I could effectively complete the tasks and scenarios using this system.
4. I was able to complete the tasks and scenarios quickly using this system.
5. I was able to efficiently complete the tasks and scenarios using this system.
6. I felt comfortable using this system.
7. It was easy to learn to use this system.
8. I believe I could become productive quickly using this system.
9. The system gave error messages that clearly told me how to fix problems.
10. Whenever I made a mistake using the system, I could recover easily and quickly.
11. The information (such as on-line help, on-screen messages and other documentation) provided with this system was clear.
12. It was easy to find the information I needed.
13. The information provided for the system was easy to understand.
14. The information was effective in helping me complete the tasks and scenarios.
15. The organization of information on the system screens was clear.

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 microphone, and the screens (including their use of graphics and language).

16. The interface of this system was pleasant.
17. I liked using the interface of this system.
18. This system has all the functions and capabilities I expect it to have.
19. Overall, I am satisfied with this system.