Tradeoffs in the Design of the IBM Computer Usability Satisfaction Questionnaires

James R. Lewis
International Business Machines Corp.

1 Introduction

Psychometrics is a well-developed field in psychology, and usability researchers began to use psychometric methods to develop and evaluate questionnaires to assess usability a little over ten years ago (Sweeney & Dillon, 1987). 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? Here is a brief review of some basic elements of standard psychometric practice.

2 Brief Review of Psychometric Practice

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).

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 relationship between the measure of interest and a different concurrent or predictive measure). Moderate correlations (with absolute values as small as .30 to .40) are often large enough to justify the use of psychometric instruments (Nunnally, 1978).

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.

Goals of 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.

3 Tradeoffs Considered in the Development of the IBM Questionnaires

Number of scale steps. The more scale steps in a questionnaire 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 most 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 tend to correlate rather highly with one another. Reliability, then, usually is not a problem in the construction of summated attitude scales.

This turned out to be true in the case of the IBM questionnaires (Lewis, 1995). Coefficient alpha exceeded .89 for all instruments using 7-point scales. Coefficient alpha for a questionnaire using 5-point scales ranged from .64 to .93 and averaged .80. A related analysis using the same data (Lewis, 1993) showed that the mean difference of the 7-point scales correlated more strongly than the mean difference of the 5-point scales with the observed significance levels of *t*-tests. For these reasons, we currently use 7-point rather than 5-point scales.

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, making them easier to interpret and compare. For example, with items based on 7-point scales, all the summative scales would also have scores that range from 1 to 7. For these reasons, we average the responses given by a participant across the items for each scale.

Unidimensional or multidimensional instrument. The developer of a questionnaire can have the goal of creating a unidimensional or multidimensional instrument (McIver & Carmines, 1981). A unidimensional instrument will typically require fewer items, so it will take less time to administer and provides a straightforward measurement because it has no subscales. A multidimensional instrument, because it measures several subscales related to the higher-level, overall scale, typically requires more items. For example, the System Usability Scale (Brooke, 199?), a unidimensional instrument, contains ten items. The PSSUQ, a multidimensional instrument that provides measurements for three subscales as well as the overall measurement, contains 19 items.

I actually can't claim that we set out to create a multidimensional instrument when we put together the first version of the PSSUQ. A group of usability evaluators selected the items on the basis of their comprehensive content regarding hypothesized constituents of usability. However, we have found its subscales to be informative and useful. For our purposes, this benefit clearly outweighs its slightly longer administration time relative to shorter instruments.

Control of potential response bias or consistency in item alignment. It is a common practice in questionnaire development to vary the tone of items so that, typically, half of the items elicit agreement and the other half elicit disagreement. The purpose of this is to control potential response bias. An alternative approach, less commonly used, is to align the items consistently.

Probably the most common criticism I've seen of the IBM questionnaires is that they do not use the standard control for potential response bias. 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. Even if this results in some response bias, typical use of the IBM questionnaires is to compare systems or experimental conditions. In this context of use, any systematic response bias will cancel out across comparisons.

I have seen the caution expressed that a frustrated or lazy participant will simply choose one end point or the other and mark all items the same way. With all items aligned in the same way, this could lead to the erroneous conclusion that the participant held a strong belief (either positive or negative) regarding the usability of the system. With items constructed in the standard way, such a set of responses would indicate a neutral opinion. Although this characteristic of the standard approach is appealing, I have seen no evidence of such participant behavior, at least not in the hundreds of PSSUQs that I have personally scored. I am sure it is a valid concern in other areas of psychology – especially some areas of clinical or counseling psychology, where the emphasis is on the individual rather than group comparisons. It is possible that constructing a usability assessment questionnaire in the standard way could lead to more item-marking errors on the part of sincere participants than the approach of consistently aligning items (although I know of no research in this area).

To norm or not to norm. When a questionnaire has norms, data exists that allows researchers to interpret individual and average scores as greater or smaller than the expected norm scores. 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.

When I performed the psychometric qualification of the CSUQ, I acquired a fair amount of data suitable for norms. I never published the norms because they were considered IBM Confidential. Those norms are now about 10 years out of date, and I no longer use them. The only instruments I know of that appear to have useful norms are those created by Kirakowski and his colleagues (Kirakowski & Corbett, 1993; Kirakowski & Dillon, 1988). Researchers should be cautious in the use of such norms, however, because differences between the contexts in which the norms were gathered and the use of the instrument could be misleading. Norms are of clear value in many situations, but it is important not to overgeneralize their applicability in usability evaluation.

4 Advantages of Using Psychometrically Qualified Instruments

Despite any controversies regarding decisions made in the development of such questionnaires, standardized satisfaction measurements (whichever questionnaire you choose to use) offer many advantages to the usability practitioner (Nunnally, 1978). Specifically, standardized measurements (even without norms) 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). The level of measurement of an instrument (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 (Harris, 1985). Measurements using Likert 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 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 B (a ratio claim).

In conclusion, psychometrically qualified, standardized questionnaires can be valuable additions to practitioners' repertoire of usability evaluation techniques.

5 References

Brooke, J. (199?). SUS – A quick and dirty usability scale. Unpublished paper.

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

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.

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. (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.

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.

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