Taylor & Francis
Taylor & Francis Group

International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: https://www.tandfonline.com/loi/hihc20

Measuring Perceived Usability: The CSUQ, SUS, and UMUX

James R. Lewis

To cite this article: James R. Lewis (2018) Measuring Perceived Usability: The CSUQ, SUS, and UMUX, International Journal of Human–Computer Interaction, 34:12, 1148-1156, DOI: 10.1080/10447318.2017.1418805

To link to this article: https://doi.org/10.1080/10447318.2017.1418805

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
2018, VOL. 34, NO. 12, 1148–1156
https://doi.org/10.1080/10447318.2017.1418805

Taylor & Francis
Taylor & Francis Group

Measuring Perceived Usability: The CSUQ, SUS, and UMUX

James R. Lewis

IBM Corporation, Boca Raton, FL, USA

ABSTRACT

The primary purpose of this research was to investigate the relationship between two widely used questionnaires designed to measure perceived usability: the Computer System Usability Questionnaire (CSUQ) and the System Usability Scale (SUS). The correlation between concurrently collected CSUQ and SUS scores was 0.76 (over 50% shared variance). After converting CSUQ scores to a 0–100-point scale (to match the range of the SUS scores), there was a small but statistically significant difference between CSUQ and SUS means. Although this difference (just under 2 scale points out of a possible 100) was statistically significant, it did not appear to be practically significant. Although usability practitioners should be cautious pending additional independent replication, it appears that CSUQ scores, after conversion to a 0–100-point scale, can be interpreted with the Sauro–Lewis curved grading scale. As a secondary research goal, investigation of variations of the Usability Metric for User Experience (UMUX) replicated previous findings that the regression-adjusted version of the UMUX-LITE (UMUX-LITEr) had the closest correspondence with concurrently collected SUS scores. Thus, even though these three standardized questionnaires were independently developed and have different item content and formats, they largely appear to be measuring the same thing, presumably, perceived usability.

1. Introduction

1.1. Perceived usability

An important component of the higher-level construct of usability is perceived usability (Brooke, 2013; Lewis, Utesch, & Maher, 2015; Sauro & Lewis, 2009, 2016). The subjective component of perceived usability, along with the objective components of efficiency and effectiveness, makes up the classical conception of the construct of usability (ISO, 1998), which is in turn a fundamental component of user experience (UX; Diefenbach, Kolb, & Hassenzahl, 2014). Likely driven by the influx of experimental psychologists into the field in the early 1980s, the first standardized usability questionnaires intended for usability testing appeared in the late 1980s (Brooke, 1996; Chin, Diehl, & Norman, 1988; Kirakowski & Dillon, 1988; Lewis, 1990).

Two of the most popular standardized questionnaires used to assess perceived usability are the Computer System Usability Questionnaire (CSUQ; Lewis, 1995) and the System Usability Scale (SUS; Brooke, 1996). They were independently developed in the 1980s at, respectively, IBM and DEC, and published in the mid-1990s. Although both are widely used, over the years the SUS has become the more popular questionnaire. Sauro and Lewis (2009) reported that the SUS accounted for 43% of post-study questionnaire usage in a study of unpublished usability studies while the CSUQ accounted for about 15%. Google Scholar citations (examined 9/12/2017) showed 5015 citations for the paper that introduced the SUS (Brooke, 1996) and 1693 for the paper that introduced the CSUQ (Lewis, 1995). These independent measurements of questionnaire "popularity" show that the CSUQ is popular, but the SUS is almost three times as popular.

The primary goal of this research was to investigate whether the CSUQ and the SUS appear to be measuring the same thing or something different. Despite these questionnaires having been available in the public domain for over 20 years, there has been very little investigation of their psychometric relationship. This is an issue of both theoretical and practical significance.

From a theoretical perspective, if the CSUQ and SUS – possibly the two most popular instruments for the assessment of perceived usability – do not appear to substantially measure the same thing, then what does this mean for the validity of the construct of usability, which has been recently questioned (Tractinsky, 2017)? From a practical perspective, if they do essentially measure the same thing, then practitioners should be able to put data from both questionnaires on a common scale for interpretation as indicating relatively poor, average, or good levels of perceived usability based on grading scale norms developed for the SUS (Sauro & Lewis, 2016).

A secondary goal of this research was to investigate the relationship between the SUS and concurrently collected scores from the Usability Metric for User Experience (UMUX; Finstad, 2010) and metrics derived from the UMUX (UMUX-LITE and UMUX-LITEr; Lewis, Utesch, & Maher, 2013, 2015). Previous research has generally shown that the UMUX and associated metrics provide scores very similar to the SUS, but with fewer items (the SUS has 10 items, the UMUX has 4, and UMUX-LITE has 2). Although not yet as popular as the CSUQ or SUS (144 citations of Finstad, 2010 in Google Scholar on 9/12/2017), the UMUX-related metrics have the potential to become more widely used as researchers and practitioners become more aware of them and, of course, depending on whether research continues to support their use.

CONTACT James R. Lewis jimlewis@us.ibm.com 7329 Serrano Terrace, Delray Beach, FL 33446, USA.

1.2. The CSUQ

The CSUQ is a variant of the Post-Study System Usability Questionnaire (PSSUQ; Lewis, 1995), developed initially for the collection of a large number of completed questionnaires to see if the factor structure found for the PSSUQ in a usability testing setting would stay the same in a mailed survey. The emergence of the same factors demonstrated the potential usefulness of the PSSUQ and CSUQ questionnaires across different user groups and research settings.

The CSUQ is identical to the PSSUQ, with slight changes to the wording due to the change in research context. For example, item 3 of the PSSUQ Version 3 states, "I was able to complete the tasks and scenarios quickly using this system," but item 3 of the CSUQ Version 3 states, "I am able to complete my work quickly using this system." Figure 1 shows the current 16-item version of the CSUQ (which is also available in a Turkish version, Erdinç & Lewis, 2013). Due to their extremely close relationship, it is informative to include research conducted on the PSSUQ in this literature review.

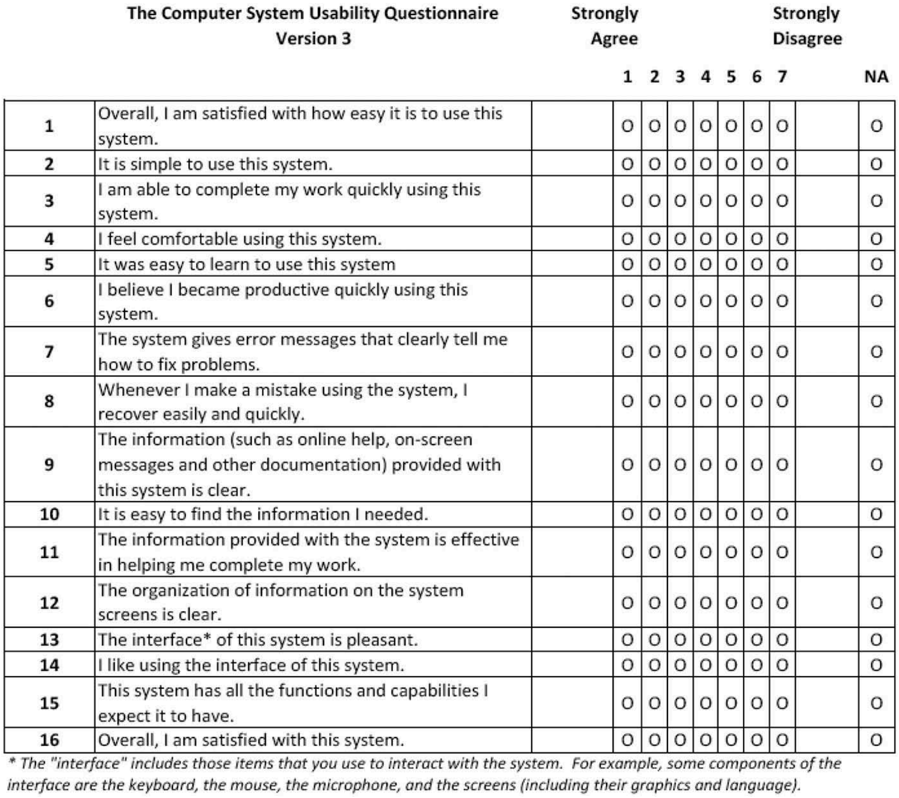
The origin of the PSSUQ was an internal IBM project from the late 1980s called System Usability MetricS (SUMS). The SUMS researchers created a large pool of items based on the contextual usability work of Whiteside, Bennett, and Holtzblatt (1988). After content analysis by that group of human factors engineers and usability specialists, 18 items remained for the first version of the PSSUQ (Lewis, 1990, 1992).

An independent IBM investigation into customer perception of usability of several different user groups indicated a common set of five usability characteristics (Doug Antonelli, personal communication, 5th January 1991). The 18-item version of the PSSUQ had addressed four of those characteristics (quick completion of work, ease of learning, high-quality documentation and online information, and functional adequacy), but had not covered the fifth (rapid acquisition of productivity). The inclusion of an item to address this characteristic led to the second version of the PSSUQ, containing 19 items (Lewis, 1995). After several years' use of the PSSUQ Version 2, item analysis indicated that three questions in that version (3, 5, and 13) contributed relatively little to PSSUQ reliability (Lewis, 2002). Those items were removed to improve the efficiency of the questionnaire in its third version.

The items produce four scores – one overall and three subscales. The rules for computing them are

- Overall: average the responses for items 1–16 (all the items)
- System Usefulness (SysUse): average items 1–6
- Information Quality (InfoQual): average items 7–12
- Interface Quality (IntQual): average items 13–15

The resulting scores can take values between 1 and 7 or not applicable (NA), with lower scores indicating a higher degree of satisfaction. Note that some practitioners prefer higher scores to indicate higher satisfaction and switch the labels for "strongly agree" and "strongly disagree" (e.g., see Tullis & Albert, 2008, p. 140). From a strict interpretation of standardization, it is best to avoid this type of manipulation unless there is evidence that it does not affect the factor structure of the items (e.g., Chen, 1991, found this manipulation to affect means and factor structures of a Mandarin version of the Personal Distress Scale, but these effects did not replicate in



	The Computer System Usability Questionnaire Version 3	Strongly Agree								Strongly Disagree	
			1	2	3	4	5	6	7		NA
1	Overall, I am satisfied with how easy it is to use this system.		O	O	O	O	O	O	O		O
2	It is simple to use this system.		O	O	O	O	O	O	O		O
3	I am able to complete my work quickly using this system.		O	O	O	O	O	O	O		O
4	I feel comfortable using this system.		O	O	O	O	O	O	O		O
5	It was easy to learn to use this system		O	O	O	O	O	O	O		O
6	I believe I became productive quickly using this system.		O	O	O	O	O	O	O		O
7	The system gives error messages that clearly tell me how to fix problems.		O	O	O	O	O	O	O		O
8	Whenever I make a mistake using the system, I recover easily and quickly.		O	O	O	O	O	O	O		O
9	The information (such as online help, on-screen messages and other documentation) provided with this system is clear.		O	O	O	O	O	O	O		O
10	It is easy to find the information I needed.		O	O	O	O	O	O	O		O
11	The information provided with the system is effective in helping me complete my work.		O	O	O	O	O	O	O		O
12	The organization of information on the system screens is clear.		O	O	O	O	O	O	O		O
13	The interface* of this system is pleasant.		O	O	O	O	O	O	O		O
14	I like using the interface of this system.		O	O	O	O	O	O	O		O
15	This system has all the functions and capabilities I expect it to have.		O	O	O	O	O	O	O		O
16	Overall, I am satisfied with this system.		O	O	O	O	O	O	O		O

** 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 graphics and language).*

Figure 1. The CSUQ (Version 3).

Weng & Cheng, 2000). On the other hand, the various psychometric evaluations of the PSSUQ since its initial publication suggest that it should be robust against these types of minor manipulations (Lewis, 2002). If comparing across published studies, however, it is critical to know which item format was in use and, if necessary, to adjust one of the sets of scores. To reverse a 7-point score, subtract it from 8 (e.g., a 1 changes to a 7, a 7 to a 1, and a 4 is unchanged).

Psychometric evaluations of the CSUQ and PSSUQ have consistently shown a factor structure that matches the subscales shown above, providing evidence of construct validity (Berkman & Karahoca, 2016; Lewis, 1995, 2002). The scales have shown high levels of reliability, with values of coefficient alpha (Schmitt, 1996) exceeding 0.80. Significant correlation with other metrics such as task-level satisfaction (Lewis, 1995), percentage of successful task completions (Lewis, 1995), and the SUS (Berkman & Karahoca, 2016) provide evidence of concurrent validity. CSUQ scores have been shown to be sensitive to a number of independent variables, including number of years of experience with the computer system, type of computer used, and range of experience with different computers, and user groups of varying experience (Berkman & Karahoca, 2016; Lewis, 1995, 2002). In a study comparing five different standardized usability questionnaires, the CSUQ was the second fastest to converge on its large-sample mean (Tullis & Stetson, 2004).

Neither the PSSUQ nor the CSUQ require any license fee for their use (see Lewis, 2012, p. 1303). Researchers who use it should cite their source (for Version 3, the best source is Sauro & Lewis, 2016) and should make clear in their method sections which version and item format they used.

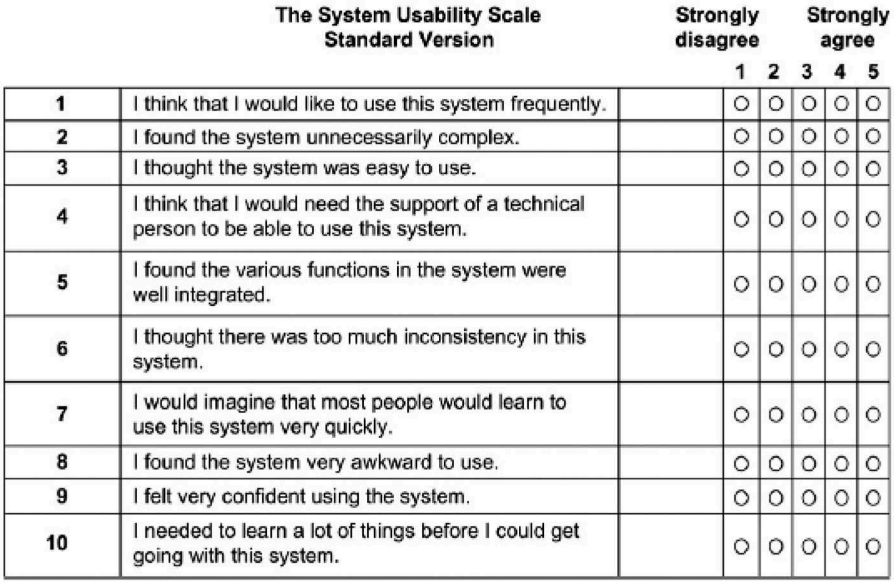
1.3. The SUS

The SUS has become a very popular questionnaire for the assessment of perceived usability, both in usability studies and in surveys (Grier, Bangor, Kortum, & Peres, 2013; Lewis et al., 2013). Like the CSUQ, the SUS is in the public domain, with no license fee required for its use. Research has shown that the SUS has excellent reliability (coefficient alpha typically exceeds 0.90), validity, and sensitivity to a wide variety of independent variables (Sauro & Lewis, 2016), whether used in the lab or in a survey. In a comparison of five different standardized usability questionnaires, the SUS was the fastest to converge on its large-sample mean (Tullis & Stetson, 2004).

Minor wording changes do not appear to affect SUS scores, for example, using "website" or a product name in place of the original "system," or the replacement of the word "cumbersome" with "awkward" (Bangor, Kortum, & Miller, 2008; Finstad, 2006; Lewis & Sauro, 2009). The SUS contains 10 items of mixed tone, with half of the items (the odd numbers) having a positive tone and the other half (the even numbers) having a negative tone, all with a response scale from 1 (strongly disagree) to 5 (strongly agree). Figure 2 shows the standard version of the SUS.

With regard to its construct validity, Brooke (1996) recommended using it as a unidimensional measure of perceived usability, but did not have sufficient data to conduct supporting analyses. Once larger-sample studies were conducted with the SUS, some findings appeared that suggested a bidimensional structure with Usable (items 1, 2, 3, 5, 6, 7, 8, and 9) and Learnable (items 4 and 10) subscales (Borsci, Federici, Gnaldi, Bacci, & Bartolucci, 2015; Borsci, Federici, & Lauriola, 2009; Lewis & Sauro, 2009; which included a reanalysis of the correlation matrix of SUS items provided by Bangor et al., 2008). Other research typically found evidence for two underlying factors, but with inconsistent alignment of items with factors (e.g., Berkman & Karahoca, 2016; Kortum & Sorber, 2015; Lewis, 2014; Lewis, Brown, & Mayes, 2015; Lewis et al., 2013; Sauro & Lewis, 2011) – often indicating a simple two-factor structure with positive-tone items aligning on one factor and negative tone items (which includes items 4 and 10) aligning on the other.

Most recently, Lewis and Sauro (2017) assembled a database of over 9000 completed SUS questionnaires. Exploratory and confirmatory factor analyses showed that the SUS



	The System Usability Scale Standard Version		Strongly disagree				Strongly agree
			1	2	3	4	5
1	I think that I would like to use this system frequently.		○	○	○	○	○
2	I found the system unnecessarily complex.		○	○	○	○	○
3	I thought the system was easy to use.		○	○	○	○	○
4	I think that I would need the support of a technical person to be able to use this system.		○	○	○	○	○
5	I found the various functions in the system were well integrated.		○	○	○	○	○
6	I thought there was too much inconsistency in this system.		○	○	○	○	○
7	I would imagine that most people would learn to use this system very quickly.		○	○	○	○	○
8	I found the system very awkward to use.		○	○	○	○	○
9	I felt very confident using the system.		○	○	○	○	○
10	I needed to learn a lot of things before I could get going with this system.		○	○	○	○	○

Figure 2. The standard SUS. Item 8 shown with "awkward" in place of the original "cumbersome" (Bangor et al., 2008; Finstad, 2006).

Table 1. The Sauro–Lewis curved grading scale.

SUS score range	Grade	Percentile range
84.1– 100	A+	96–100
80.8–84.0	A	90–95
78.9–80.7	A–	85–89
77.2–78.8	B+	80–84
74.1–77.1	B	70–79
72.6–74.0	B–	65–69
71.1–72.5	C+	60–64
65.0–71.0	C	41–59
62.7–64.9	C–	35–40
51.7–62.6	D	15–34
0.0–51.6	F	0–14

appeared to be bidimensional, but not in any interesting or useful way. A comparison of the fit of three confirmatory factor analyses showed that a model in which the SUS's positive tone (odd-numbered) and negative tone (even-numbered) were aligned with two factors had a better fit than a unidimensional model (all items on one factor) or the Usability/Learnability model. Because a distinction based on item tone is of little practical or theoretical interest, they recommended that user experience practitioners and researchers generally treat the SUS as a unidimensional measure of perceived usability.

Another consequence of the large-sample databases assembled for the SUS has been the publication of normative data (Bangor et al., 2008; Sauro & Lewis, 2016). The first approach to developing a grading scale based on this kind of data was by Bangor, Kortum, and Miller (2009), who provided an absolute grading scale with A: >89; B: 80–89; C: 70–79; D: 60–69; and F < 60. Table 1 shows an alternative curved grading scale (CGS) published by Sauro and Lewis (2016), based on data from 446 industrial usability studies (over 5000 completed SUS questionnaires). The CGS provides an empirically grounded approach to the interpretation of mean SUS scores obtained in industrial usability studies and is the grading scale used throughout the rest of this article.

1.4. The UMUX

Even though the SUS is a relatively short questionnaire, there are some situations in which a shorter instrument would be preferable (e.g., when there is a need to measure more attributes than just perceived usability leading to limited "real estate" for any given attribute). The UMUX (Finstad, 2010, 2013) was designed at Intel to get a measurement of perceived usability consistent with the SUS, but using only four (rather than 10) items (see Figure 3).

Like the standard SUS, UMUX items vary in tone but, unlike the SUS, have seven rather than five scale steps from 1 (strongly disagree) to 7 (strongly agree) with the item scores manipulated to obtain an overall score that ranges from 0 to 100. In addition to the initial research by Finstad (2010), other researchers (Berkman & Karahoca, 2016; Borsci et al., 2015; Lewis et al., 2013, 2015) have also reported desirable psychometric properties for the UMUX, including acceptable levels of

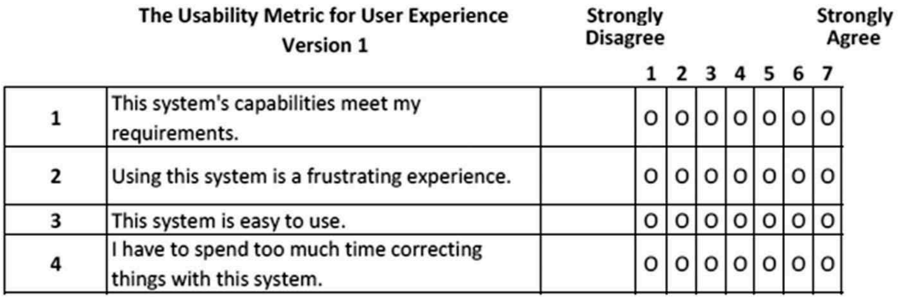
- Reliability (coefficient alpha greater than 0.80).
- Concurrent validity (correlation with SUS greater than 0.55; correlation with CSUQ equal to −0.65).
- Sensitivity to different levels of a variety of independent variables (e.g., discriminating between systems of independently assessed levels of relatively good and poor usability, detecting differences in perceived usability as a function of experience).

Most research in this area has found substantial correspondence between the magnitudes of mean SUS and UMUX. An exception is Borsci et al. (2015), who reported UMUX means that were significantly and markedly higher than concurrently collected SUS means.

Analyses of the factor structure of the UMUX have been inconsistent. With only four items, the most likely structures are one or two factors. Finstad (2010, 2013)) reported a one-factor structure. Lewis et al. (2013, 2015) found a two-factor structure with alignment of positive- and negative-tone items on different factors (like the alignment of SUS items on the basis of tone reported by Berkman & Karahoca, 2016; Lewis & Sauro, 2017). Berkman and Karahoca reported the two-factor positive/negative tone structure when forcing a two-factor solution, but also reported evidence from confirmatory factor analysis suggesting a one-factor structure. Following the same practical reasoning as that for the SUS, it does not matter whether the UMUX has a unidimensional or tone-based bidimensional structure – in either case, practitioners should treat the UMUX as a unidimensional measurement of perceived usability. Like the CSUQ and the SUS, the UMUX and measures derived from it are available for use by researchers or practitioners without a license fee.

1.5. The UMUX-LITE

The UMUX-LITE (Lewis et al., 2013, 2015) is a short version of the UMUX consisting of its positive-tone items (selected on the basis of factor and item analysis), which are



The Usability Metric for User Experience
Version 1

Strongly Disagree — Strongly Agree

			1	2	3	4	5	6	7
1	This system's capabilities meet my requirements.		O	O	O	O	O	O	O
2	Using this system is a frustrating experience.		O	O	O	O	O	O	O
3	This system is easy to use.		O	O	O	O	O	O	O
4	I have to spend too much time correcting things with this system.		O	O	O	O	O	O	O

Figure 3. The standard UMUX.

(1) This system's capabilities meet my requirements.
(2) This system is easy to use.

There are two versions of the UMUX-LITE that usability practitioners should be aware of, and they should also be aware that the UMUX-LITE literature has been inconsistent in its terminology. The formula for computing the standard UMUX-LITE, where x_1 and x_2 are the ratings for Items 1 and 2 using a standard 7-point scale (1–7), is UMUXLITE = $(x_1 + x_2 - 2)(100/12)$. Due to a small but statistically significant difference between the SUS and UMUX-LITE means, Lewis et al. (2013) computed a regression equation to bring the SUS and UMUX-LITE scores into closer correspondence. The formula for this regression-adjusted version is UMUXLITEr = 0.65 (UMUXLITE) + 22.9. Note that because the UMUX-LITEr is a linear adjustment of the UMUX-LITE it has the many of the same statistical properties, such as the magnitude of correlation with other metrics, but can only take values between 22.9 (when UMUX-LITE = 0) and 87.9 (when UMUX-LITE = 100). This range restriction is consistent with the distribution of mean SUS scores reported by Sauro and Lewis (2016), in which the percentile rank for a score of 20 was 1% and for a score of 90 was 99.8%. In this article, we distinguish the two with an "r" to indicate the regression-adjusted version (as in Berkman & Karahoca, 2016).

In addition to the statistical analyses supporting their selection, it is interesting that the content of the two items matches the constructs of the Technology Acceptance Model (TAM) (Davis, 1989), a questionnaire from the market research literature that assesses the usefulness (e.g., capabilities meeting requirements) and ease of use of systems, and has an established relationship to likelihood of future use. According to the TAM, good ratings of usefulness and ease of use (perceived usability) influence the intention to use, which influences the actual likelihood of use.

Research on the UMUX-LITE and UMUX-LITEr (Berkman & Karahoca, 2016; Borsci et al., 2015; Lewis et al., 2013, 2015) has demonstrated acceptable psychometric properties, including

- acceptable reliability (estimates of coefficient alpha ranging from 0.77 to 0.86);
- concurrent validity (correlations with SUS ranging from 0.74 to 0.83; correlation with ratings of likelihood-to-recommend ranging from 0.72 to 0.74); and
- sensitivity (significant differences as a function of respondents' ratings of frequency of use)

Following the study in which it was presented (Lewis et al., 2013), the correspondence in the magnitudes of the UMUX-LITEr and the SUS has been replicated twice (Borsci et al., 2015; Lewis et al., 2015; – scores separated by about 1–3 points on the SUS's 101-point scale). On the other hand, Berkman and Karahoca (2016) found a closer correspondence between the unadjusted UMUX-LITE and the SUS (1 point) than that between the regression-adjusted UMUX-LITEr and the SUS (5 points).

1.6. Research goals

Our research goals were to

- investigate the correlation and correspondence between the CSUQ and the SUS; and
- investigate the correlation and correspondence of the UMUX, UMUX-LITE, and UMUX-LITEr with the SUS.

2. Method

2.1. The survey

Participants completed a Survey Gizmo (surveygizmo.com) designed to recreate the CSUQ survey mailed out to IBM employees for the research reported in Lewis (1995), but with the addition of the SUS and the UMUX. The instructions to participants were

> Thank you for agreeing to participate in this evaluation. It should take about 5-10 minutes to complete this survey. In this survey, you'll use a variety of standardized usability questionnaires to rate the primary computer system (hardware, software, and documentation) that you use for your work at IBM. Full disclosure – I am in no position to help with problems you might be having with your system – you'll need to work with your management and IBM service to resolve anything like that. My primary goal is to understand the statistical relationships among these different questionnaires.
>
> Please keep in mind that you are participating in an evaluation of the usability of computer systems. This is not a test of you – you are helping us to understand the relationships among various ways of measuring perceived usability. Please try to answer all the items in the questionnaires, but don't spend a lot of time on an item – your first impression is fine. The items will differ in whether a low number or a high number indicates a good or poor user experience, so please read each item carefully.
>
> OK, let's get started.

Participants then completed, in this order, the CSUQ (see Figure 1), the SUS (see Figure 2), and the UMUX (see Figure 3), then completed a demographics section that included system questions [type of hardware, operating system (OS), and applications used] and respondent characteristics (length of time at IBM, length of time using the rated system, and frequency of system use).

2.2. Respondents

Respondents were members of the IBM User Experience panel. To form the panel, invitations were emailed to 20,000 randomly selected IBM employees (US only). About 10% (2035) of those invited agreed to join the panel. Of these 2035 members, 746 (36.7%) completed the survey. In accordance with the rules established by their developers, missing data in the CSUQ were left unchanged while missing data in the SUS and UMUX were replaced with the center item of the rating scale (three for the SUS, four for the UMUX).

Referring back to the survey instruction, "The items will differ in whether a low number or a high number indicates a good or poor user experience, so please read each item carefully" – despite providing numerous instructions throughout the survey regarding item format, a comparison of the ratings of a CSUQ item (#2) and a similar SUS item (#3) (both ratings of system ease of use) indicated that 128 (about 17%) of the respondents who completed the survey provided radically different ratings for their systems' ease of use with those two items (more than a 50-point difference after conversion of both items to a common 0–100-point scale). This is a strong indication that they did not notice the shift in item formats from the CSUQ to the SUS, so their data was deleted before conducting any further analyses, leaving a data set with 618 cases. All statistical analyses used SPSS Version 24 (IBM Corporation, Armonk, NY).

3. Results

3.1. Reliability

All of the questionnaires had values of coefficient alpha consistent with the prior literature. A common criterion for acceptable reliability is coefficient alpha equal to or greater than 0.70 (Nunnally, 1978). The values of coefficient alpha computed for the questionnaires were

- CSUQ: 0.97 (with 0.95, 0.93, and 0.91, respectively, for the SysUse, InfoQual, and IntQual subscales)
- SUS: 0.93
- UMUX: 0.88
- UMUX-LITE/UMUX-LITEr: 0.79

3.2. Concurrent validity

A common minimum criterion for evidence of concurrent validity is a correlation between metrics that exceeds 0.3 (Nunnally, 1978). When $r = 0.3$, $R^2 = 0.09$ – in other words, the two measurements have just under 10% of shared variance. Although there are no definitive guidelines, when two metrics measure essentially the same thing, one would expect higher correlations (e.g., correlations of 0.7 have just under 50% shared variance). Because the SUS is the most commonly used measure of perceived usability, the focus of the current study was on the correlations of the other questionnaires with the SUS, which were

- CSUQ: 0.76 (with 0.74, 0.65, and 0.68, respectively, for the SysUse, InfoQual, and IntQual subscales)
- UMUX: 0.79
- UMUX-LITE/UMUX-LITEr: 0.74

3.3. Construct validity

The CSUQ

Analyses of the CSUQ ratings (using unrestricted least-squares factor analysis) were consistent with previous research (Berkman & Karahoca, 2016; Lewis, 1995, 2002). Parallel analysis (O'Connor, 2000) indicated a three-factor solution would be appropriate, and items strongly aligned on factors as expected (1–6, 7–12, and 13–15). As expected when using summative metrics (Cliff, 1987; Nunnally, 1978), there were significant ($p < 0.0001$) correlations among the subscales (SysUse-InfoQual: 0.71; SysUse-IntQual: 0.80; InfoQual-IntQual: 0.71).

The SUS

For the SUS, parallel analysis indicated a two-factor solution. The observed alignment between items and factors generally (but not exactly) matched the expected alignment of odd-numbered (positive tone) items on one factor and even-numbered (negative tone) items on the other (Lewis & Sauro, 2017). Items for which the maximum loading exceeded 0.5 and there was a difference in loadings of at least 0.1 showed clear loading of items 1, 3, 5, 7, and 9, as expected, on the same factor. Using the same criteria, items 2, 4, and 10 clearly aligned on the second factor. Item 8 aligned on the first factor, and the loadings for item 6 were ambiguous (respectively 0.514 and 0.498).

The UMUX

The UMUX literature has been mixed with regard to evidence for unidimensionality and bidimensionality (positive/negative tone). Parallel analysis of the UMUX data in this study (and by extension the UMUX-LITE/UMUX-LITEr) indicated a one-factor solution.

Combined questionnaires

A parallel analysis was conducted to investigate whether the intercorrelations among the CSUQ, SUS, UMUX, and UMUX-LITE were indicative of alignment with one or more than one factor. Comparison of observed with randomly generated eigenvalues indicated a one-factor solution (observed: 3.3, 0.4, 0.2, 0.1; randomly generated: 1.1, 1.0, 1.0, 0.9).

3.4. Correspondence between the CSUQ and SUS

To examine the correspondence between the magnitude of CSUQ and SUS scores, CSUQ scores were converted from their historical scale of a value between 1 and 7 (where a lower number indicates a better experience) to a 0–100-point scale to match the SUS. The formula for this transformation was CSUQ = 100 – (((CSUQ01 + CSUQ02 + CSUQ03 + CSUQ04 + CSUQ05 + CSUQ06 + CSUQ07 + CSUQ08 + CSUQ09 + CSUQ10 + CSUQ11 + CSUQ12 + CSUQ13 + CSUQ14 + CSUQ15 + CSUQ16)/16) – 1)(100/6). Breaking this down, the process of getting from a traditional CSUQ score to one that matches the SUS involves subtracting 1 from the mean of the 16 individual CSUQ items and multiplying that by 100/6 to stretch it out to a 0–100-point scale, then subtracting from 100 to reverse the scale. For example, if the mean CSUQ was 1 (the best possible standard CSUQ mean), the transformed score would be 100 (100 – (1–1)(100/6) = 100 – 0 = 100). If the mean CSUQ was 7 (the worst possible standard CSUQ mean), the transformed score would be 0 (100 – (7–1)(100/6) = 100 – 100 = 0). For a mean CSUQ of 4 (the center of the standard CSUQ 7-point scale), the transformed score would be 50 (100 – (4–1)(100/6) = 100 – 50 = 50).

With this transformation, the overall mean CSUQ was 66.7 and the mean SUS was 68.7 – a difference of 2 points on a 101-point scale. Given the fairly large sample size, this was a statistically significant difference ($t(617) = 3.2$, $p = 0.001$), but matching these scores against the Sauro–Lewis CGS (Table 1) shows that they would both receive a grade of C, so from a user experience perspective there was not a practically significant difference. It is also interesting that the mean SUS of this data set was 68.7 – very close to the center of the CGS (68), which indicates that the average user in this sample reported an average level of perceived usability. The values of the CSUQ subscale means (SysUse: 73.8, InfoQual: 56.4, IntQual: 71.3) suggest that none of them had close enough correspondence with the SUS to warrant interpreting them with the CGS.

One of the system variables for which there was a statistically significant difference was the OS. The mean for respondents using an Apple© system was higher than that for those using a Windows© system for both the CSUQ and the SUS (CSUQ: $t(560) = 6.4$, $p < 0.0001$; SUS: $t(560) = 5.2$, $p < 0.0001$). The means and corresponding CGS grades were

- Apple, CSUQ: 76.6 (B)
- Apple, SUS: 76.8 (B)
- Windows, CSUQ: 64.1 (C–)
- Windows, SUS: 66.9 (C)

The CSUQ and SUS CGS grades were the same for the Apple respondents (both B) and were within one grade boundary for Windows respondents (C– and C).

It would be possible, following the approach of Lewis et al. (2013) with the UMUX-LITE, to compute a regression formula to close this small gap between the CSUQ and the SUS, which for this data set would be SUSPred = 19.178 + 0.742 (CSUQ). This would, however, restrict CSUQ scores to a range less than 0–100, and it is not yet clear whether this discrepancy would be the same in other sets of data.

For example, in Berkman and Karahoca (2016), the reported CSUQ and SUS means for a usability study of a mind-mapping application were, respectively, 2.2 and 79.5. Using the formula to convert CSUQ scores to a SUS-like 0–100-point scale, 2.2 transforms to 80.0, which is just a half-point above the SUS mean of 79.5. Lewis (2002) did not include any SUS data, but did report the grand mean of the PSSUQ collected over 21 studies and 210 participants, getting 2.82, which converts to 69.7, very close to 68 (the center of the CGS).

3.5. Correspondence between the SUS and UMUX-Related Metrics

The grand means for the SUS, UMUX, UMUX-LITE, and UMUX-LITEr for this data set were, respectively (and with CGS grades in parentheses), 68.7 (C), 69.1 (C), 70.7 (C), and 68.9 (C). The only statistically significant difference between a UMUX-related metric and the SUS was for the nonadjusted UMUX-LITE (UMUX: $t(617) = -0.6$, $p = 0.52$; UMUX-LITE: $t(617) = -3.0$, $p = 0.003$; UMUX-LITEr: $t(617) = -0.3$, $p = 0.8$). The closest correspondence was between the SUS and the UMUX-LITEr. As with the CSUQ, even though the difference between SUS and UMUX-LITE grand means was statistically significant, all four means fell into the same grade level on the CGS, suggesting little or no practically significant difference in their measurement of perceived usability.

Table 2. Perceived usability as a function of operating system.

OS	SUS	UMUX	UMUX-LITE	UMUX-LITEr
Apple	76.8 (B)	79.2 (A–)	79.9 (A–)	74.9 (B)
Windows	66.9 (C)	66.5 (C)	68.5 (C)	67.4 (C)

Returning to the analysis of differences between Apple and Windows respondents, Table 2 shows the means and grades for each of the metrics. All Apple vs. Windows comparisons were statistically significant ($p < 0.0001$), but the focus here is on correspondence between CGS grades. For the Windows means, all CGS grades were C. There was a bit more discrepancy for the Apple means, which had CGS grades of B for SUS and UMUX-LITEr, but A- for UMUX and UMUX-LITE. Overall, the regression formula for the UMUX-LITEr seems to have held up well in this independent replication because it had the most consistently close matches to the SUS.

4. Discussion

4.1. Three questionnaires, one underlying construct

The CSUQ and the SUS are two of the most-used standardized questionnaires for the assessment of perceived usability. The UMUX and its variants (UMUX-LITE and UMUX-LITEr) are relative newcomers, but have sparked a considerable amount of research since their initial publication. All three were developed with a common goal, but in different places by different groups of researchers.

On their faces, they are quite different. The CSUQ (Version 3) has 16 items, the SUS has 10, the UMUX has 4, and the UMUX-LITE has 2. Only the CSUQ has well-defined subscales in addition to its overall measurement. The CSUQ items all have a positive tone with 7-point end-anchored scales that include NA outside the scale; the standard SUS and UMUX are mixed-tone questionnaires with, respectively, 5- and 7-point end-anchored scales and no NA response option. The rules for computing CSUQ scores do not require any special treatment of missing values, but the SUS and UMUX developers recommend replacing missing values with the center point of their scales (3 for the SUS, 4 for the UMUX) because the standard method of computing SUS and UMUX scores requires a complete questionnaire.

Despite these structural differences, the CSUQ, SUS, UMUX, and UMUX-LITE correlated highly with one another, and a parallel analysis of their eigenvalues from factor analysis indicated a one-factor solution. In other words, that they appear to be measuring essentially the same thing. Given the goals of their developers and the content of their items, that "thing" is presumably the construct of perceived usability.

In addition to this, the correspondence in the magnitude of their measurements when placed on a consistent 0–100-point scale (standard for the SUS and UMUX-related metrics; conversion required for the CSUQ) was striking. Of the UMUX-related metrics, the UMUX-LITEr appeared to provide the closest match to the SUS when raw scores were converted to grades using the Sauro–Lewis CGS. Despite a statistically significant but small deviation between SUS and CSUQ scores, conversion to letter grades using the CGS showed a remarkable consistency in assigned grades.

One of the most important aspects of using the SUS for practical usability assessment is the development of open-source norms, starting with data published by Bangor et al. (2008) and Tullis and Albert (2008), to which Sauro and Lewis (2016) added additional data to derive their CGS. The finding that transformed CSUQ scores and regression-adjusted UMUX-LITE scores closely correspond to concurrently collected SUS scores means that usability practitioners who use the UMUX-LITEr and CSUQ (and by extension, the PSSUQ) can, with reasonable confidence, use the Sauro–Lewis CGS as an indication of relatively poor, average, or good levels of perceived usability.

4.2. Limits to generalization

Despite these encouraging results, it is important to note some limitations to generalizability. To get large enough sample sizes for psychometric evaluation, most research on standardized usability questionnaires relies on retrospective evaluation using surveys (for exceptions, see Lewis, 2002; Lewis et al., 2015). It would be interesting to see if data collected in traditional usability studies would show similar correspondences between the SUS, CSUQ, and UMUX-LITEr.

Participants in this study completed the questionnaires in the same order: CSUQ, SUS, UMUX. Although it seems unlikely, especially given the close correspondence of the scores in this study, it is possible that this may have had some small effect on the results. In future research, the experimental designs should use a different order or, ideally, alternate the order of presentation.

The body of research on the correspondence between the SUS and UMUX-related metrics has grown over the past few years, but this is the first study to examine the correspondence between scale-matched measurements of the CSUQ and the SUS. Until researchers have replicated the CSUQ-SUS correspondence across a wider variety of systems and research methods, practitioners should be appropriately cautious when using the Sauro–Lewis CGS to interpret overall CSUQ scores.

5. Conclusions

Despite having been independently developed and containing different item content and formats, the SUS, CSUQ, and UMUX-related questionnaires largely appear to be measuring the same thing, presumably, perceived usability. Not only were they strongly correlated, when transformed to a common 0–100-point scale (where 0 is poor and 100 is excellent), the mean scores for the CSUQ, SUS, and UMUX-LITEr had similar magnitudes and similar grades on the Sauro–Lewis CGS for the SUS.

Although usability practitioners should be cautious pending additional independent replication, it appears that CSUQ scores, after conversion to a 0–100-point scale, can be interpreted with the Sauro–Lewis CGS. The UMUX findings from this study replicate the most common finding from the literature that of the various UMUX-related metrics the regression-adjusted version of the UMUX-LITE (UMUX-LITEr) had the closest correspondence with concurrently collected SUS scores, increasing the confidence that usability practitioners can have in using that two-item questionnaire in place of the SUS as a quick measure of perceived usability.

With multiple standardized questionnaires available to researchers and practitioners, it is reasonable to ask which is better to use under which circumstances. For those who have a history with one of these questionnaires, it appears that there is no compelling reason to switch. Given the enormous amount of research conducted with it, the SUS is probably the best default choice. If there would be value in a multidimensional rather than unidimensional instrument, then the CSUQ would be a reasonable choice. If there is limited "real estate" in a survey or for whatever reason it is important to have a few items as possible, then the UMUX-LITEr would be the best choice.

Acknowledgment

Many thanks to members of the IBM User Experience Panel who participated in this study and to the reviewers who volunteered their time to read and comment on the original manuscript.

ORCID

James R. Lewis http://orcid.org/0000-0002-3295-5392

References

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction*, *24*, 574–594.

Bangor, A., Kortum, P. T., & Miller, J. T. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Berkman, M. I., & Karahoca, D. (2016). Re-assessing the usability metric for user experience (UMUX) scale. *Journal of Usability Studies*, *11*(3), 89–109.

Borsci, S., Federici, S., Gnaldi, M., Bacci, S., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: An exploratory analysis of SUS, UMUX and UMUX-LITE. *International Journal of Human-Computer Interaction.*, *31*, 484–495.

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the system usability scale: A test of alternative measurement models. *Cognitive Processes*, *10*, 193–197.

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

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, *8*(2), 29–40.

Chen, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement*, *51*, 531–540.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human–Computer interface. *In* E. Soloway, D. Frye, & S. B. Sheppard (Eds.), *Proceedings of CHI 1988* (pp. 213–218). Washington, DC: Association for Computing Machinery.

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

Davis, D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*, 319–339.

Diefenbach, S., Kolb, N., & Hassenzahl, M. (2014). *The "Hedonic" in human-computer interaction: History, contributions, and future research directions. In* Proceedings of the 2014 Conference on

Designing Interactive Systems - DIS 14 (pp. 305–314), New York, NY: Association for Computing Machinery.

Erdinç, O., & Lewis, J. R. (2013). Psychometric evaluation of the T-CSUQ: The Turkish version of the computer system usability questionnaire. *International Journal of Human-Computer Interaction*, *29* (5), 319–326.

Finstad, K. (2006). The system usability scale and non-native English speakers. *Journal of Usability Studies*, *1*(4), 185–188.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*, 323–327.

Finstad, K. (2013). Response to commentaries on "The usability metric for user experience". *Interacting with Computers*, *25*, 327–330.

Grier, R. A., Bangor, A., Kortum, P. T., & Peres, S. C. (2013). *The system usability scale: Beyond standard usability testing*. In Proceedings of the Human Factors and Ergonomics Society (pp. 187–191), Santa Monica, CA: Human Factors and Ergonomics Society.

ISO. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs), Part 11, Guidance on usability (ISO 9241-11:1998E)*. Geneva, Switzerland: Author.

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.

Kortum, P., & Sorber, M. (2015). Measuring the usability of mobile applications for phones and tablets. *International Journal of Human-Computer Interaction*, *31*(8), 518–529.

Lewis, J. R. (1990). *Psychometric evaluation of a post-study system usability questionnaire: The PSSUQ* (Tech. Report 54.535). Boca Raton, FL: International Business Machines Corp.

Lewis, J. R. (1992). *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. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*, 57–78.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*, 463–488.

Lewis, J. R. (2012). Usability testing. *In* G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (*4th* ed., pp. 1267–1312). New York, NY: John Wiley.

Lewis, J. R. (2014). Usability: Lessons learned… and yet to be learned. *International Journal of Human-Computer Interaction*, *30*, 663–684.

Lewis, J. R., Brown, J., & Mayes, D. K. (2015). Psychometric evaluation of the EMO and the SUS in the context of a large-sample unmoderated usability study. *International Journal of Human-Computer Interaction*, *31*(8), 545–553.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. *In* M. Kurosu (Ed.), *Human centered design* (pp. 94–103). Heidelberg, Germany: Springer-Verlag.

Lewis, J. R., & Sauro, J. (2017). Revisiting the factor structure of the system usability scale. *Journal of Usability Studies*, *12*(4), 183–192.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE—When there's no time for the SUS. *In Proceedings of CHI 2013* (pp. 2099–2102). Paris, France: Association for Computing Machinery.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31*(8), 496–505.

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

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers*, *32*, 396–402.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. *In Proceedings of CHI 2009* (pp. 1609–1618). Boston, MA: Association for Computing Machinery.

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive?. *In Proceedings of CHI 2011* (pp. 2215–2223). Vancouver, Canada: ACM.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research (2nd ed.)*. Cambridge, MA: Morgan Kaufmann.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353.

Tractinsky, N. (2017). The usability construct: A dead end? *Human-Computer Interaction*. doi:10.1080/07370024.2017.1298038

Tullis, T. S., & Albert, B. (2008). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Burlington, MA: Morgan Kaufmann.

Tullis, T. S., & Stetson, J. N. (2004). *A comparison of questionnaires for assessing website usability*. Paper presented at the Usability Professionals Association Annual Conference, UPA, Minneapolis, MN. Retrieved September 13, 2017 from, https://www.researchgate.net/publication/228609327_A_Comparison_of_Questionnaires_for_Assessing_Website_Usability

Weng, L., & Cheng, C. (2000). Effects of response order on Likert-type scales. *Educational and Psychological Measurement*, *60*(6), 908–924.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1988). Usability engineering: Our experience and evolution. *In* M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 791–817). Amsterdam: North-Holland.

About the Author

James R. (Jim) Lewis is a senior human factors engineer (at IBM since 1981), currently focusing on the design/evaluation of speech applications. He has published influential papers in the areas of usability testing and measurement, and is an IBM Master Inventor with 91 US patents issued to date. His books include *Practical Speech User Interface Design* and (with Jeff Sauro) *Quantifying the User Experience*.