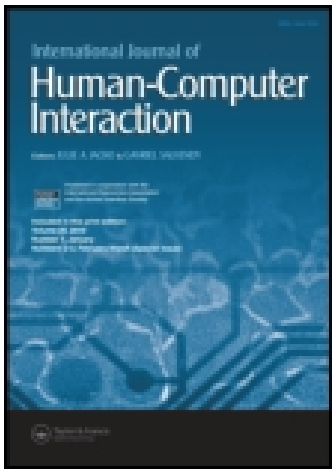

International Journal of Human-Computer Interaction
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hihc20

# Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability

James R. Lewis[a], Brian S. Utesch[a] & Deborah E. Maher[a]
[a] IBM Corporation, Boca Raton, Florida, USA
Accepted author version posted online: 25 Jun 2015.



To cite this article: James R. Lewis, Brian S. Utesch & Deborah E. Maher (2015) Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability, International Journal of Human-Computer Interaction, 31:8, 496-505, DOI: 10.1080/10447318.2015.1064654

To link to this article: http://dx.doi.org/10.1080/10447318.2015.1064654

*Intl. Journal of Human–Computer Interaction*, 31: 496–505, 2015

ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447318.2015.1064654

Taylor & Francis
Taylor & Francis Group

# Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability

**James R. Lewis, Brian S. Utesch, and Deborah E. Maher**
*IBM Corporation, Boca Raton, Florida, USA*



**The purpose of this research was to investigate various measurements of perceived usability, in particular, to assess (a) whether a regression formula developed previously to bring Usability Metric for User Experience LITE (UMUX-LITE) scores into correspondence with System Usability Scale (SUS) scores would continue to do so accurately with an independent set of data; (b) whether additional items covering concepts such as findability, reliability, responsiveness, perceived use by others, effectiveness, and visual appeal would be redundant with the construct of perceived usability or would align with other potential constructs; and (c) the dimensionality of the SUS as a function of self-reported frequency of use and expertise. Given the broad use of and emerging interpretative norms for the SUS, it was encouraging that the regression equation for the UMUX-LITE worked well with this independent set of data, although there is still a need to investigate its efficacy with a broader set of products and methods. Results from a series of principal components analyses indicated that most of the additional concepts, such as findability, familiarity, efficiency, control, and visual appeal covered the same statistical ground as the other more standard metrics for perceived usability. Two of the other items (Reliable and Responsive) made up a reliable construct named System Quality. None of the structural analyses of the SUS as a function of frequency of use or self-reported expertise produced the expected components, indicating the need for additional research in this area and a need to be cautious when using the Usable and Learnable components described in previous research.**

## 1. INTRODUCTION

### 1.1. Perceived Usability

For decades, practitioners and researchers in user-centered design and human–computer interaction (HCI) have had a strong interest in the measurement of perceived usability (Sauro & Lewis, 2012). The subjective component of perceived usability along with the objective components of efficiency and effectiveness make up the classicial conception of the construct of usability (ISO, 1998), which is in turn a fundamental component of user experience (Diefenbach, Kolb, & Hassenzahl, 2014). Although these objective and subjective components tend to be correlated, factor analysis has indicated a distinction between their measurements (Sauro & Lewis, 2009).

The most common approach to the assessment of perceived usability has been through the development and application of standardized questionnaires. A standardized questionnaire is a questionnaire designed for repeated use, typically with a specific set of questions presented in a specified order using a specified format, with specific rules for producing metrics based on the answers of respondents. Developers of standardized questionnaires typically assess the psychometric quality of these types of instruments with measurements of reliability, validity, and sensitivity (Nunnally, 1978).

Earlier standardized scales in HCI, such as the Gallagher Value of MIS Reports Scale and the Hatcher and Diebert Computer Acceptance Scale, were not specifically designed for the assessment of perceived usability (see LaLomia & Sidowski, 1990, for a review of standardized computer satisfaction questionnaires published between 1974 and 1988). 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), likely driven by the influx of experimental psychologists into the field of HCI during the 1980s. Additional standardized usability questionnaires have appeared in the decades since, continuing to the present day (Finstad, 2010; Lewis, 2014; Lewis, Utesch, & Maher, 2013, in press).

At roughly the same time that usability researchers were producing the first standardized usability questionnaires, market researchers were tackling similar issues. Of these, one of the most influential has been the Technology Acceptance Model (TAM; Davis, 1989). According to the TAM, the primary factors that affect a user's intention to use a technology are its perceived usefulness and perceived ease of use (i.e., usability). Actual use of technologies is affected by the intention to use, which is itself affected by the perceived usefulness and usability of the technology. A number of studies support the validity of the TAM and its satisfactory explanation of end-user system usage (Wu, Chen, & Lin, 2007).

Address correspondence to James R. Lewis, 7329 Serrano Terrace, Delray Beach, FL 33446, USA. E-mail: jimlewis@us.ibm.com

### 1.2. The System Usability Scale

The System Usability Scale (SUS), developed in the mid-1980s at Digital Equipment Corporation (Brooke, 1996), has become a very popular questionnaire for the assessment of perceived usability. Sauro and Lewis (2009) reported that the SUS accounted for 43% of poststudy questionnaire usage in a study of unpublished usability studies (compared to 15% each for the other standardized questionnaires in that set of data). In addition to its application in usability studies, the SUS has become a popular questionnaire to include in surveys (Grier, Bangor, Kortum, & Peres, 2013; Lewis et al., 2013). A considerable amount of research has indicated that the SUS has excellent reliability (coefficient alpha typically exceeds .90), validity, and sensitivity to manipulated variables (Sauro & Lewis, 2012), whether used in the lab or in a survey. In a study comparing different standardized usability questionnaires, the SUS was the fastest to converge on its large-sample mean (Tullis & Stetson, 2004).

The SUS is insensitive to minor changes to its wording, 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 original version of 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*). Although it is a standard psychometric practice to vary item tone, this variation can have negative consequences (Lewis, 1999; Sauro & Lewis, 2012). In a study designed to investigate the consequences of replacing the negative-tone items of the SUS with positive-tone items, Sauro and Lewis (2011) found no evidence for differences in response biases or magnitudes of responses between the different versions but did find evidence of an increased number of mistakes on the part of participants and researchers using the standard version, apparently due to tone switching. Table 1 shows the items for the standard and positive versions of the SUS.

There have been some inconsistencies in the reported factor structure of the SUS (Lewis, 2014; Sauro & Lewis, 2012). Although the reported values of coefficient alpha of about .90 for the SUS are evidence of its reliability, it is important to note that high values of coefficient alpha are not proof of unidimensionality (Schmitt, 1996). Originally intended as a unidimensional measure of perceived usability (Brooke, 1996), three independent lines of research published in 2009 converged on solutions with two dimensions, Usable (made up of Items 1, 2, 3, 5, 6, 7, 8, and 9) and Learnable (Items 4 and 10; Borsci, Federici, & Lauriola, 2009; Lewis & Sauro, 2009; a reanalysis of the correlation matrix of SUS items provided by Bangor et al., 2008).

Analyses conducted since 2009 (e.g., Lewis et al., 2013; Sauro & Lewis, 2011), however, have typically resulted in two-factor structures but have not exactly replicated the item-factor alignment that received such support in 2009. Borsci, Federici, Gnaldi, Bacci, and Bartolucci (this issue), analyzing data from early and later use of an educational website, found evidence favoring a unidimensional structure of the SUS for early use and the bidimensional partitioning into Usable and Learnable for later use.

A relatively recent research development for the SUS has been the publication of normative data from fairly large sample databases (Bangor et al., 2008; Sauro & Lewis, 2012). For example, Table 2 shows the curved grading scale (CGS) published by Sauro and Lewis (2012), based on data from 446 industrial usability studies (more than 5,000 completed SUS questionnaires). The CGS provides an empirically grounded approach to the interpretation of mean SUS scores obtained in industrial usability studies. Although we generally

**Table 1.** The System Usability Scale Items (Standard and Positive Versions)

| Item | Standard Version | Positive Version |
|---|---|---|
| 1 | I think that I would like to use this system frequently. | I think that I would like to use this system frequently. |
| 2 | I found the system unnecessarily complex. | I found the system to be simple. |
| 3 | I thought the system was easy to use. | 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. | I think I could use the system without the support of a technical person. |
| 5 | I found the various functions in the system were well integrated. | I found the various functions in the system were well integrated. |
| 6 | I thought there was too much inconsistency in this system. | I thought there was a lot of consistency in the system. |
| 7 | I would imagine that most people would learn to use this system very quickly. | I would imagine that most people would learn to use this system very quickly. |
| 8 | I found the system very cumbersome to use. | I found the system very intuitive. |
| 9 | I felt very confident using the system. | I felt very confident using the system. |
| 10 | I needed to learn a lot of things before I could get going with this system. | I could use the system without having to learn anything new. |

**Table 2.** 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 |

*Note.* SUS = System Usability Scale.



prefer the curved grading scale when interpreting SUS means, the somewhat stricter standard grading scale published by Bangor, Kortum, and Miller (2009) is also noteworthy (A: > 89; B: 80–89; C: 70–79; D: 60–69; F: < 60).

### 1.3. The Usability Metric for User Experience

The Usability Metric for User Experience (UMUX; Finstad, 2010, 2013; Lewis, 2013) was designed to get a measurement of perceived usability consistent with the SUS but using only four (rather than 10) items. The primary purpose for its development was to provide an alternate metric for perceived usability for situations in which it was critical to reduce the number of items while still getting a reliable and valid measurement of perceived usability (e.g., when there is a need to measure more attributes than just perceived usability leading to limited "real estate" for any given attribute).

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*). Finstad (2010) reported desirable psychometric properties for the UMUX, including its discrimination between systems with relatively good and poor usability, high reliability (coefficient alpha of .94), and extremely high correlation with SUS scores ($r = .96$). The four UMUX items are as follows:

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

Lewis et al. (2013) included the UMUX in their study, and found results that generally replicated the findings reported by Finstad (2010). For their two data sets (one using the standard SUS and the other using the positive version), the UMUX correlated significantly with the SUS (standard = .90; positive = .79). Although this is significantly less than Finstad's correlation of .96, it supports his claim of strong concurrent validity. The estimated reliabilities of the UMUX in the two data sets were more than adequate (.87, .81) but, like the correlations with the SUS, a bit less than the originally reported value of .97. For both data sets, there was no significant difference between the mean SUS and mean UMUX scores (extensive overlap between the 99% confidence intervals), consistent with the original data. Note, however, that Borsci et al. (this issue) reported UMUX means that were significally and markedly higher than concurrently collected SUS means.

### 1.4. The UMUX-LITE

The UMUX-LITE (Lewis et al., 2013) is a short version of the UMUX, consisting of its positive-tone items and maintaining the use of 7-point scales. Thus, for the UMUX-LITE, the items are as follows:

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

Factor analysis conducted by Lewis et al. (2013) indicated that the UMUX had a bidimensional structure with item alignment as a function of item tone (positive vs. negative). This, along with additional item analysis, led to the selection of the two items for the UMUX-LITE for the purpose of creating an ultrashort metric for perceived usability. Data from two independent surveys demonstrated adequate psychometric quality of the UMUX-LITE. Estimates of reliability were .82 and .83—excellent for a two-item instrument. Concurrent validity was also high, with significant correlation with standard and positive versions of the SUS (.81, .81) and with likelihood-to-recommend (LTR) scores (.74, .73; Reichheld, 2003, 2006; Sauro & Lewis, 2012). Furthermore, the UMUX-LITE scores were sensitive to respondents' ratings of frequency-of-use. UMUX-LITE score means were slightly lower than those for the SUS but easily adjusted using linear regression to match the SUS scores (Equation 1).

$$\text{UMUX} - \text{LITE} = 0.65 * ((\text{Item1} + \text{Item2} - 2) * (100/12)) + 22.9 \quad (1)$$

Another reason for including the specific two items of the UMUX-LITE was their connection to the content of the items in the 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 TAM, good ratings of usefulness and ease of use (perceived usability) influence the intention to use, which influences the actual likelihood of use.

Applying the regression equation to the UMUX-LITE results in a more constrained possible range than the SUS. Rather than 0 to 100, UMUX-LITE scores can range from 22.9 (when the responses to both items are 1) to 87.9 (when the responses to

both items are 7). When the reponses to both items are 4 (midpoint), the UMUX-LITE score is 55.4. This range restriction is consistent with the distribution of mean SUS scores reported by Sauro and Lewis (2012), in which the percentile rank for a score of 20 was 1% and for a score of 90 was 99.8%.

### 1.5. Research Goals

For a number of years, we have used the SUS to measure perceived usability as part of a larger battery of metrics for the online assessment of software applications. As research has appeared on variants of the SUS and shorter questionnaires such as the UMUX, we have carefully modified our practices by collecting concurrent measurements of the positive and negative versions of the SUS, the UMUX, and the UMUX-LITE. For our standard studies, we currently include the positive version of the SUS and the UMUX-LITE, in addition to the other items in our battery, which are as follows:

- EasyNav: "This system is easy to navigate."
- AbleFind: "I am able to find what I need in this system."
- Familiar: "This system offers capabilities familiar to me."
- Need: "This system does what I need it to do."
- Efficient: "This system helps me to do my job more efficiently."
- SeeData: "I know who can see my data in this system."
- Depend: "I can depend on this system."
- Control: "I feel in control when I work within this system."
- Appeal: "This system is visually appealing."
- OthersUse: "My colleagues use this system."
- MgtUse: "Executive management uses this system."
- Reliable: "This system is very reliable."
- Responsive: "This system responds quickly when I use it."

Our research goals were to:

- Investigate which of these additional items were redundantly measuring the construct of perceived usability.
- Investigate how the additional items interrelated and, in particular, how they might or might not align to assess constructs other than perceived usability.
- Explore the development of an alternative questionnaire (AltUsability) for assessing perceived usability using the additional items that aligned with perceived usability but are not part of the SUS.
- Partially replicate our previous (2013) study to assess whether the regression formula developed using that data would similarly adjust the data from a completely independent set of data to result in close correspondence with the SUS. A successful replication would lead to greater confidence in using the UMUX-LITE in place of the SUS while still using the emerging SUS norms (e.g., Table 2) to interpret the results.
- Investigate the dimensionality of the SUS as a function of experience.
- Study the relationship between constructs and the outcome metrics of overall experience and LTR.

## 2. METHOD

### 2.1. The Surveys

We conducted four different product surveys of IBM products used by IBM employees for a total of 397 cases in which respondents completed the UMUX-LITE, the positive version of the SUS, and the 13 additional items just listed. Each of the 13 additional items were positively worded statements using 7-point scales anchored with 1 (*strongly disagree*) and 7 (*strongly agree*). We used 7 points because, for stand-alone items, psychometric research has shown that 7-point scales are measurably more reliable (Nunnally, 1978) and sensitive to manipulation (Lewis, 1993) than 5-point scales.

### 2.2. Outcome Metrics

Respondents also provided the outcome ratings of LTR and Overall Experience (note that only 239 respondents provided ratings of Overall Experience due to its exclusion from one of the surveys). The LTR scale used the standard format of response options from 0 (*extremely unlikely*) to 10 (*extremely likely*). The Overall Experience item had five response options (1 = *I hate it*; 2 = *I dislike it*; 3 = *I neither like nor dislike it*; 4 = *I like it*; 5 = *I love it*).

### 2.3. Ratings of Frequency of Use and Expertise

Respondents reported their typical frequency of use of the systems using the following response options: 1 (*once every few months or less*), 2 (*once every few weeks*), 3 (*once a week*), 4 (*several times a week*), 5 (*once a day*), or 6 (*more than once a day*). They also provided self-reported assessments of their expertise with the systems as 1 (*novice*), 2 (*intermediate*), 3 (*expert*), or 4 (*evangelist*).

## 3. RESULTS

### 3.1. Reliability and Concurrent Validity of the SUS and the UMUX-LITE

As expected, the SUS and UMUX-LITE had psychometric properties consistent with those reported in previous research. Their respective reliabilities, assessed using coefficient alpha, were .91 and .86. They also correlated significantly with the outcome metrics of Overall Experience and LTR, which is evidence of their concurrent validity: SUS and Overall Experience, $r(395) = .67$; SUS and LTR, $r(395) = .71$; UMUX-LITE and Overall Experience, $r(395) = .72$; UMUX-LITE and LTR,

$r(395) = .72$; all $ps < .01$. The correlation between this version of the SUS and the UMUX-LITE was consistent with the magnitude reported by Lewis et al. (2013), $r(395) = .83$, $p < .01$.

### 3.2. Additional Items and Their Connection to the Construct of Perceived Usability

To investigate the relationship of our additional items to the construct of perceived usability, we conducted a principal components analysis with Varimax rotation of the items plus the SUS and UMUX-LITE overall scores. A discontinuity analysis on the slope of the eigenvalues indicated a three-component solution (eigenvalues = 7.786, 1.197, 1.087, 0.820, 0.640, 0.574, 0.511, 0.445, 0.427, 0.359, 0.329, 0.272, 0.229, 0.200, 0.123). Table 3 shows the resulting structure, with the largest loadings in bold. Most additional items aligned on the first component with the SUS and UMUX-LITE, indicating correspondence with the construct of perceived usability. The exceptions were SeeData, Depend, OthersUse, Reliable, and Responsive.

The next step was to reanalyze the additional items without inclusion of the SUS and UMUX-LITE, as shown in Table 4. The results indicated an Alternate Usability (AltUsability) component consisting of EasyNav, AbleFind, Familiar, Need, Efficient, Control, and Appeal; a System Quality (SysQual) component consisting of SeeData, Depend, Reliable, and Responsive; and an ObservedUsage component made up of OthersUse and MgtUse.

The reliabilities of these new metrics, as measured with coefficient alpha, were as follows:

- AltUsability = .90
- SysQual = .78 (if include only Reliable and Responsive = .87)
- ObservedUsage = .61

The seven-item AltUsability metric had reliability similar to that of the SUS and, from a measurement perspective, probably covers pretty much the same ground, but with a completely different set of items. The difference of .11 for SysQual with and without SeeData and Depend suggests that it would be better to use just the Reliable and Responsive items and to drop the other two, which we have done for the remaining analyses that include this construct.

The reliability of ObservedUsage fell below the typical criterion of .7. This relatively low reliability is likely due, at least in part, to the item loadings on the ObservedUsage component (see Table 3 in which the component was not apparent and Table 4 in which there was almost equal loading of MgtUse on AltUsability and ObservedUsage). Although it does not appear to be a viable construct, it is unique enough that we have included it in some additional analyses, but note that researchers should be cautious regarding its use and interpretation.

### 3.3. Correlations Among the Metrics

Table 5 shows the correlations among the standard, new, and outcome metrics, with the 99% confidence intervals. Although all correlations were significantly greater than 0, many differed significantly in magnitude (when the confidence intervals did not overlap, the difference was statistically significant at $p < .01$). Most notably, the correlations among the three measurements of perceived usability were very high, followed by the correlations among the measurements of perceived

**Table 3.** Assessment of Additional Items Alignment With Construct of Perceived Usability

| Scale/Item | Comp 1 | Comp 2 | Comp 3 |
|---|---|---|---|
| SUS_POS | **0.84** | 0.26 | 0.23 |
| UMUX_LITE | **0.83** | 0.19 | 0.29 |
| EasyNav | **0.76** | 0.10 | 0.37 |
| AbleFind | **0.78** | 0.21 | −0.06 |
| Familiar | **0.61** | 0.30 | −0.20 |
| Need | **0.76** | 0.23 | 0.16 |
| Efficient | **0.66** | 0.38 | 0.20 |
| SeeData | 0.22 | **0.59** | 0.00 |
| Depend | 0.30 | **0.78** | 0.06 |
| Control | **0.77** | 0.33 | −0.06 |
| Appeal | **0.76** | −0.06 | 0.38 |
| OthersUse | 0.16 | 0.16 | **0.74** |
| MgtUse | **0.51** | 0.36 | 0.38 |
| Reliable | 0.04 | **0.69** | 0.57 |
| Responsive | 0.17 | **0.71** | 0.42 |

*Note.* The largest loadings are in bold. SUS = System Usability Scale; UMUX-LITE = Usability Metric for User Experience LITE.

**Table 4.** Item Alignment With AltUsability, SysQual, and ObservedUsage

| Item | AltUsability | SysQual | ObservedUsage |
|---|---|---|---|
| EasyNav | **0.67** | 0.15 | 0.45 |
| AbleFind | **0.81** | 0.15 | 0.03 |
| Familiar | **0.66** | 0.15 | 0.00 |
| Need | **0.75** | 0.23 | 0.24 |
| Efficient | **0.65** | 0.34 | 0.32 |
| SeeData | 0.28 | **0.46** | 0.13 |
| Depend | 0.38 | **0.73** | −0.01 |
| Control | **0.79** | 0.28 | −0.01 |
| Appeal | **0.67** | 0.00 | 0.49 |
| OthersUse | 0.03 | 0.25 | **0.83** |
| MgtUse | 0.47 | 0.35 | **0.50** |
| Reliable | 0.00 | **0.86** | 0.29 |
| Responsive | 0.16 | **0.84** | 0.16 |

*Note.* The largest loadings are in bold. AltUsability = Alternate Usability; SysQual = System Quality.

**Table 5.** Correlations Among the Various Metrics

| | SUS | UMUX-LITE | Alt Usability | SysQual | Observed Usage | Overall Experience | LTR |
|---|---|---|---|---|---|---|---|
| SUS | 1.00 | | | | | | |
| UMUX- LITE | 0.83 [.79, .87] | 1.00 | | | | | |
| AltUsability | 0.86 [.82, .89] | 0.79 [.74, .83] | 1.00 | | | | |
| SysQual | 0.52 [.42, .61] | 0.53 [.43, .62] | 0.51 [.41, .60] | 1.00 | | | |
| ObservedUsage | 0.42 [.31, .52] | 0.36 [.24, .47] | 0.53 [.43, .62] | 0.37 [.25, .48] | 1.00 | | |
| Overall Experience | 0.67 [.57, .75] | 0.72 [.63, .79] | 0.72 [.63, .79] | 0.37 [.22, .51] | 0.30 [.14, .44] | 1.00 | |
| LTR | 0.71 [.64, .77] | 0.72 [.65, .78] | 0.75 [.69, .80] | 0.45 [.34, .55] | 0.40 [.29, .50] | 0.69 [.59, .77] | 1.00 |

*Note.* $df = 395$ for all correlations except those with overall experience, for which $df = 237$, all $ps < .01$. 99% confidence intervals appear in brackets. SUS = System Usability Scale; UMUX-LITE = Usability Metric for User Experience LITE; AltUsability = Alternate Usability; SysQual = System Quality; LTR = likelihood-to-recommend.



usability and the outcome metrics (Overall Experience and LTR). The lowest correlations were those for the constructs of SysQual and ObservedUsage, both with perceived usability and outcomes.

### 3.4. Correspondence Between AltUsability and the SUS

As an exercise in exploring the correspondence between the new AltUsability questionnaire and the SUS, we converted AltUsability scores to a SUS-like measure that could range from 0 to 100 using the following formula (Excel style):

$$\text{AltUsability} = (\text{SUM}(\text{EasyNav, AbleFind, Familiar, Need, Efficient, Control, Appeal}) - 7) * (100/42) \quad (2)$$

The difference between the means for SUS and AltUsability was more than 20 points, so for AltUsability to act as a surrogate for SUS (to take advantage of the published SUS norms), it was necessary to compute a regression equation to align the measurements, as shown in the following formula:

$$\text{AltUsabilityAdj} = 1.257\,(\text{AltUsability}) + 12.585 \quad (3)$$

After adjustment, the means were virtually identical, but it is important to keep in mind that this result is based on the data from which the formula was derived. Furthermore, the regression adjustment changes the possible range from 0–100 to 12.6–138.3. Clearly, there is a need for independent replication of this relationship.

### 3.5. Correspondence Between the UMUX-LITE and the SUS

The overall mean difference between the SUS and regression-adjusted UMUX-LITE scores was just 1.1—only 1% of the range of the values that the SUS and UMUX-LITE can take (0–100). Strictly speaking, that difference was statistically significant, $t(396) = 2.2$, $p = .03$, but for any practical use (such as comparison to norms such as the CGS shown in Table 2), it is essentially no difference, especially for results greater than a point away from the break between grades. When sample sizes are large, it is important not to confuse statistically significant differences with meaningful differences.

### 3.6. The Dimensionality of the SUS as a Function of Product Experience and Expertise

Following the work of Borsci et al. (this issue), we investigated the dimensionality of the SUS, both in its entirety and as a function of product experience. For the entire set of data, analysis of the slope of the eigenvalues indicated a five-component solution, which is an unreasonable result for a 10-item questionnaire. One of those five components matched the hypothesized Learnable construct (Items 4 and 10), but the others were uninterpretable. After setting the number of components to two, the results of a principal components analysis with Varimax rotation found the second component contained Items 1 and 9, a result that was not consistent with the literature (or any that we've seen before).

We also conducted a series of principal components analyses dividing the sample on the basis of reported frequency of use and self-reported expertise, most notably, the two categories of lowest reported frequency (once every few months or less; once

every few weeks), the four categories of higher reported frequency of use (once a week to more than once a day), the two categories of lower self-reported expertise (novice, intermediate), and the two categories of higher self-reported expertise (expert, evangelist). None of the resulting two-component structures matched the previouly reported structure in which the second component contained only Items 4 and 10 and the first contained the other eight items.

### 3.7. Sensitivity Analyses

We conducted a series of analyses of variance on the various metrics to assess their sensitivity to reported frequency of use and self-reported expertise. The results for the three usability metrics (SUS, UMUX-LITE, AltUsability) were virtually identical with respect to their patterns and significance so, unless otherwise specified, statistical test results of perceived usability are those for the SUS.

There was a significant main effect of frequency on Perceived Usability, $F(5, 390) = 8.7$, $p < .0001$, with Bonferroni-corrected multiple comparisons showing significant differences between the two lowest and four highest levels. The main effect of frequency on SysQual was also significant, $F(5, 390) = 5.0$, $p < .0001$, with Bonferroni-corrected multiple comparisons showing significant differences between the first two levels and the third, fourth, and sixth levels—the fifth level was not significantly different from any other level. There was also a significant main effect of frequency on SysQual, $F(5, 390) = 5.0$, $p < .0001$), with Bonferroni-corrected multiple comparisons indicating a significant difference between the least frequent users and more frequent users, but with some complications. Specifically, there was no significant difference between the two lowest frequency groups and the next to highest frequency group, making this a difficult effect to interpret. For the significant main effect of frequency on ObservedUsage, $F(5, 388) = 5.8$, $p < .0001$, the means increased as the reported frequency of use increased. Bonferroni-corrected multiple comparisons indicated no significant difference among the lower three categories or among the higher three categories, but there were significant differences between the lower and higher frequency groups.

There was a significant main effect of expertise on Perceived Usability, $F(3, 393) = 11.5$, $p < .0001$, with Bonferroni-corrected multiple comparisons showing significant differences between the lowest and all other levels of expertise. The main effect of expertise had no significant effect on SysQual, $F(3, 393) = .53$, $p = .66$, or ObservedUsage, $F(3, 391) = 1.1$, $p = .33$.

### 3.8. Contribution of Derived Metrics to Prediction of Outcome Metrics

For overall experience, the stepwise regression of SUS, UMUX-LITE, AltUsability, SysQual, and ObservedUsage retained, in order, AltUsability and UMUX-LITE (adjusted $R^2 = 57.9\%$, $r = .763$). For LTR, the stepwise regression retained the same metrics—AltUsability and UMUX-LITE (adjusted $R^2 = 60.3\%$, $r = .778$).

Because it is unreasonable on the basis of one study to claim a strong superiority of AltUsability over the SUS, we reran the stepwise regressions without AltUsability. For overall experience, the stepwise regression of SUS, UMUX-LITE, SysQual, and ObservedUsage retained, in order, the UMUX-LITE and the SUS (adjusted $R^2 = 53.0\%$, $r = .731$). For LTR, the stepwise regression retained, in order, UMUX-LITE, SUS, and ObservedUsage (adjusted $R^2 = 56.6\%$, $r = .755$).

These analyses indicate that of the variables included in the analysis (SUS, UMUX-LITE, SysQual, and ObservedUsage) perceived usability (as measured using the SUS and UMUX-LITE) was the key driver of ratings of both overall experience and LTR, with some contribution of ObservedUsage to the prediction of LTR.

### 3.9. Reliable, Responsive, or Both?

The data allowed the investigation of the statistical utility of asking just the Reliable question, just the Responsive question, or both. The correlation between Reliable and Responsive was .763 ($n = 397$, $p < .0001$), and their shared variance ($R^2$) was 58% (about 42% unshared variance). A series of regression analyses assessed how well they predicted the outcome variables of overall experience and LTR. They were both significantly predictive individually. When taken together, however, for both outcome metrics Reliable was significant but Responsive was not. This suggests that the variability in Responsive that was predictive was part of the shared variance with Reliable rather than part of the unshared variance, so there was nothing left for Responsive to contribute to the prediction after taking Reliable into account. Table 6 shows the significance of the predictors for the various outcome metrics and models.

**Table 6.** Predictions of Outcome Metrics Using Reliable and Responsive

| Outcome | Model | Reliable | Responsive |
|---|---|---|---|
| Overall experience | Reliable | $p < .0001$ | NA |
| | Responsive | NA | $p < .0001$ |
| | Both | $p < .0001$ | $p = .637$ |
| LTR | Reliable | $p < .0001$ | NA |
| | Responsive | NA | $p < .0001$ |
| | Both | $p < .0001$ | $p = .426$ |

*Note.* LTR = likelihood-to-recommend.

## 4. DISCUSSION

### 4.1. Correspondence Between the UMUX-LITE and the SUS

The broad use of and emerging interpretative norms for the SUS make it an increasingly powerful tool for usability practitioners and researchers. This presents a significant challenge for alternative methods for the assessment of perceived usability. Unless one can establish a correspondence between the alternative metric and the SUS, it may be difficult to justify using the alternative metric, because one would not be able to take advantage of the interpretative norms developed for the SUS.

The research presented in this article provides an important step toward establishing such a correspondence between the UMUX-LITE and the SUS. Using a regression formula derived from an independent set of data, the difference between the overall mean SUS score and overall mean UMUX-LITE score was just 1.1 (on a 0–100-point scale). The linear correlation between the SUS and the UMUX-LITE was not only statistically significant (nonzero) but also of considerable magnitude, $r(395) = .83$.

As in previous research, the UMUX-LITE exhibited excellent psychometric properties. According to Nunnally (1978), for instruments that assess sentiments, the minimum reliability criterion is .70 (typically assessed with coefficient alpha) and the minimum criterion for predictive or concurrent validity is .30 (typically assessed with a correlation coefficient). The UMUX-LITE significantly exceeded these psychometric criteria, and also had the expected sensitivity to self-reported frequency of use and expertise.

Despite these encouraging results, it is important to note some limitations to generalizability. To date, the data we have used for psychometric evaluation of the UMUX-LITE has come from surveys. Indeed, this is the primary intended use of the UMUX-LITE when there is limited survey real estate available for the assessment of perceived usability. It would, however, be interesting to see if data collected in traditional usability studies would show a similar correspondence between the SUS and the UMUX-LITE. Until researchers have validated the UMUX-LITE across a wider variety of systems and research methods, we do not generally recommend its use independent of the SUS.

### 4.2. Investigation of the Additional Items

Analysis of our 13 additional items indicated that seven of them redundantly measured the same construct (perceived usability) as the SUS and the UMUX-LITE. Principal components analysis of the items indicated that in addition to the redundant coverage of perceived usability, there were two additional constructs: System Quality (consisting of the items SeeData, Depend, Reliable, and Responsive) and Observed Usage (made up of the items OthersUse and MgtUse).

Reliability analysis, however, indicated that (a) dropping SeeData and Depend to use just Reliable and Responsive for System Quality substantially improved its reliability and (b) the reliability of Observed Usage did not meet the typical minimum criterion for acceptable reliability. Thus, we recommend combining the ratings of Reliable and Responsive to assess those aspects of System Quality. We also recommend that researchers and practitioners who have an interest in the construct of Observed Usage be cautious in its use due to its relatively low reliability.

The data from the regression analyses of Reliable and Responsive indicated that if one had to select just one of those questions to ask, the better choice appears to be Reliable. It is possible, however, that this could change if a system was unacceptably unreliable. Until more data are available across a broader range of system reliability and responsiveness, we recommend asking both due to the high reliability of the pair of items when used to represent the SysQual construct.

### 4.3. AltUsability

As an exercise in developing an alternative metric for the construct of perceived usability, we created an instrument (AltUsability) using the seven additional items determined to be redundant measures of that construct. First applying a SUS-like formula for computing a score based on these items that can range from 0 to 100 and then applying a UMUX-LITE-like regression equation to improve the correspondence between AltUsability and the SUS led to a metric that, for this data, produced outcomes that almost exactly matched the SUS (keeping in mind that to have confidence in this specific formula, it would be necessary to evaluate its accuracy using an independent set of data).

We are not recommending the replacement of the SUS with AltUsability but find it of interest that a metric using a completely different set of items can be brought into such close correspondence with the SUS. For example, the SUS does not specifically address issues such as navigation, findability, familiarity, efficiency, a feeling of control, or visual appeal, but apparently the items that it does include covers much of the same statistical ground. This means that practitioners who use the SUS can respond to criticisms of its content by pointing out the statistical redundancy with the AltUsability items. Alternatively, practitioners who have practical or theoretical reasons for preferring the content of AltUsability could consider using it in place of or in addition to the SUS.

### 4.4. Dimensionality of the SUS

In 2009, it seemed clear that the SUS was bidimensional, and was bidimensional in a very specific way, with one component being Learnable (Items 4 and 10) and the other being Usable (the other eight items). Research since 2009 has cast some doubt on the universal generalizability of that specific structure. Borsci et al. (this issue) report findings in which the structure of the SUS appears to depend on the users' amount of experience with the product. For relatively new users, the structure of the SUS was unidimensional; for more experienced

users, the structure was the expected bidimensional pattern of Usable and Learnable.

We attempted to replicate the Borsci et al. (this issue) finding by examining the principal components of the items of the SUS as a function of self-reported frequency of use and self-reported levels of expertise. None of the analyses had results matching the hypothesized structure. This suggests that the Learnable component might emerge only in very specific situations such as that studied by Borsci et al. and, for reasons yet unknown, data sets independently examined by Borsci et al. (2009) and Lewis and Sauro (2009).

Given the pattern of results reported to date in various publications, it is unlikely that the differences have anything to do with language differences (e.g., Borsci and his colleagues used Italian versions of the SUS and other questionnaires) or differences between the standard and positive versions of the SUS. Clearly, there is a need for more research in this area. We continue to recommend that researchers who use the SUS and have accumulated sufficiently large databases perform structural analyses and publish their results. We caution researchers and practitioners who have an interest in partitioning SUS scores into Usable and Learnable components to exercise due care given the uncertainty about when these components might or might not exist in a specific set of data.

### 4.5. Predicting Outcome Metrics

LTR and overall experience are two important outcome metrics. In particular, LTR is the basis of the Net Promoter Score (Reichheld, 2003, 2006), which is a popular industrial and market research metric for customer loyalty. Although there are many components to consider when assessing the user experience, the results of the regression analyses predicting these outcome metrics from the derived metrics for the constructs of Perceived Usability, System Quality, and Observed Usage indicated that, among these constructs, perceived usability was the key driver of the outcome metrics.

The observed prominence of perceived usability as a predictor of LTR and overall experience in this study might be due to the context of use, limiting the generalizability of the result. Participants were employees of a major corporation rating software that they used as employees. This is the context in which the TAM was developed, with its emphasis on the dual importance of usefulness and ease-of-use (Davis, 1989). The items of the SUS and AltUsability seem more grounded in the concept of pragmatic usability than that of hedonic usability, where the measurement of hedonic usability attempts to capture aspects of UX that have more to do with noninstrumental qualities such as excitement and enjoyment than the task-oriented attributes of effectiveness and efficiency (Diefenbach et al., 2014).

## 5. CONCLUSIONS

The key findings of this research contribute to an increased understanding of the construct of perceived usability. Verifying that the regression equation that brings the UMUX-LITE into correspondence with the SUS worked well with an independent set of data increases confidence in its use, but it is still important for the foreseeable future for usability practitioners and researchers to continue to investigate the relationship between the SUS and UMUX-LITE over a wider variety of systems and research methods. SysQual, a metric made up of ratings of system reliability and responsiveness, was a statistically reliable component separate from perceived usability but did not appear to have as much impact as perceived usability on outcome metrics such as overall experience and LTR. Assessments of the dimensionality of the SUS did not produce the expected patterns as a function of frequency of use or self-rated expertise, indicating that there is still a need for research investigating the conditions under which the SUS component of Learnable emerges.

## ACKNOWLEDGEMENTS

Some portions of this article originally appeared in the Proceedings of HCII 2015, *Investigating the Correspondence Between UMUX-LITE and SUS Scores.*

## 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*, 114–123.

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.

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

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.

Chin, J. P., Diehl, V. A., Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human–computer interface. In *Proceedings of CHI 1988* (pp. 213–218). Washington, DC: ACM.

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: ACM.

Finstad, K. (2006). The System Usability Scale and non-native English speakers. *Journal of Usability Studies*, *1*, 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*. Human Factors Research Group, University College of Cork, Cork, Ireland.

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.

Lewis, J. R. (1990). *Psychometric evaluation of a post-study system usability questionnaire: The PSSUQ* (Tech. Rep. No. 54.535). Boca Raton, FL: IBM.

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. (1999). Trade-offs in the design of the IBM computer usability satisfaction questionnaires. In H. Bullinger & J. Ziegler (Eds.), *Human–computer interaction: Ergonomics and user interfaces—Vol. I* (pp. 1023–1027). Mahwah, NJ: Erlbaum.

Lewis, J. R. (2013). Critical review of "The Usability Metric for User Experience." *Interacting with Computers*, *25*, 320–324.

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

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., 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: ACM.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (in press). Investigating the correspondence between UMUX-LITE and SUS scores. In *Proceedings of HCII 2015*. Cham, Switzerland: Springer International Publishing.

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

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*, 46–54.

Reichheld, F. F. (2006). *The ultimate question: Driving good profits and true growth*. Boston, MA: Harvard Business School Press.

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: ACM.

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. (2012). *Quantifying the user experience: Practical statistics for user research*. Waltham, MA: Morgan Kaufmann.

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

Tullis, T. S., & Stetson, J. N. (2004). *A comparison of questionnaires for assessing* website *usability*. Paper presented at the Usability Professionals Association Annual Conference, Minneapolis, MN. Available from home.comcast.net/~tomtullis/publications/UPA2004TullisStetson.pdf

Wu, J., Chen, Y., & Lin, L. (2007). Empirical evaluation of the revised end user computing acceptance model. *Computers in Human Behavior*, *23*, 162–174.

## ABOUT THE AUTHORS

**James R. 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. His books include *Practical Speech User Interface Design* and (with Jeff Sauro) *Quantifying the User Experience*.

**Brian S. Utesch** is a Senior Technical Staff Member and User Research Practice Lead in the IBM Enterprise Social Solutions group. Brian and his team are responsible for understanding the voice of the customer across the social solutions product portfolio.

**Deborah E. Maher** is a User Research practitioner in the IBM Enterprise Social Solutions group. She works closely with customers to capture behaviors, expectations and overall ease of use feedback within products such as IBM Connections. (In her spare time she enjoys running around after her two young children.)