# Investigating the Correspondence Between UMUX-LITE and SUS Scores

James R. Lewis[1(✉)], Brian S. Utesch[2], and Deborah E. Maher[3]

[1] IBM Software Group, Boca Raton, FL, USA
jimlewis@us.ibm.com
[2] IBM Software Group, Raleigh, NC, USA
butesch@us.ibm.com
[3] IBM Software Group, Cambridge, MA, USA
debmaher@us.ibm.com

**Abstract.** The UMUX-LITE is a two-item questionnaire that assesses perceived usability. In previous research it correlated highly with the System Usability Scale (SUS) and, with appropriate adjustment using a regression formula, had close correspondence to the magnitude of SUS scores, enabling its comparison with emerging SUS norms. Those results, however, were based on the data used to compute the regression formula. In this paper we describe a study conducted to investigate the quality of the published formula using independent data. The formula worked well. As expected, the correlation between the SUS and UMUX-LITE was significant and substantial, and the overall mean difference between their scores was just 1.1, about 1 % of the range of values the questionnaires can take, verifying the efficacy of the regression formula.

**Keywords:** Perceived usability · System usability scale · SUS · Usability metric for user experience · UMUX-LITE

## 1 Introduction

In 2013, we published the results of an initial investigation into the psychometric properties of the two-item UMUX-LITE questionnaire, and demonstrated that with a regression formula, it was possible to obtain a close correspondence between UMUX-LITE and SUS scores [9]. A limitation of the initial research was that the regression equation was not independent of the data used to evaluate its accuracy. The current study describes a partial replication of the initial research designed to address that limitation.

### 1.1 The System Usability Scale (SUS)

The SUS is questionnaire that uses 10 five-point scales. Item responses are recoded to compute an overall score that ranges from 0 to 100. Although a self-described "quick-and-dirty" questionnaire [3], the SUS appears to have excellent psychometric properties (reliability around 0.9, significant correlation with outcome measures, and sensitivity to


A. Marcus (Ed.): DUXU 2015, Part I, LNCS 9186, pp. 204–211, 2015.
DOI: 10.1007/978-3-319-20886-2_20

variables such as frequency of use and system/product), and accounts for about 43 % of post-study questionnaire usage [1, 2, 8].

The SUS is available in Standard and Positive versions [12]. The Standard (original) version has items with mixed tone – odd items have a positive tone; even items have a negative tone. In the Positive version, all items have a positive tone. Sauro and Lewis [11] found that the Positive version had advantages over the Standard version with regard to reductions in misinterpretation, mistakes, and miscoding. Both versions had high reliability (Standard: 0.92; Positive: 0.96), and had no significant difference in their mean scores. There was no evidence of acquiescence or extreme response biases in the Positive version.

A relatively recent research development for the SUS has been the publication of normative data from fairly large sample databases [1, 12]. For example, Table 1 shows the curved grading scale published by Sauro and Lewis (SL-CGS) [12], based on data from 446 industrial usability studies (over 5000 completed SUS questionnaires). The SL-CGS provides an empirically grounded approach to the interpretation of mean SUS scores obtained in industrial usability studies. Consequently, the adoption of any alternative metric for the assessment of perceived usability would greatly benefit if it were to not only correlate with the SUS, but would also correspond to its magnitude.

**Table 1.** The Sauro/Lewis curved grading scale (SL-CGS)

| 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 |
| 0.0–51.7 | F | 0–14 |

### 1.2 The Usability Metric for User Experience (UMUX)

The Usability Metric for User Experience (UMUX) [5–7] 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 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:

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. [9] included the UMUX in their study, and found results that generally replicated the findings reported by Finstad [5]. For the two datasets (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 datasets 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 datasets, 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.

### 1.3 The UMUX-LITE

The UMUX-LITE is a short version of the UMUX, consisting of its positive-tone (odd-numbered) items (maintaining the use of 7-point scales). Thus, for the UMUX-LITE, the items are:

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

Factor analysis conducted by Lewis et al. [9] 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 ultra-short 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). Furthermore, the scores were sensitive to respondents' 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 (Eq. 1).

$$\text{UMUX-LITE} = .65((\text{Item 1} + \text{Item 2} - 2)(100/12) + 22.9) \quad (1)$$

Another reason for including the specific two items of the UMUX-LITE was their connection to the content of the items in the Technology Acceptance Model (TAM) [4], 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.

### 1.4 Research Goals

Due to its parsimony (two items), reliability, validity, structural basis (usefulness and usability) and, after applying the corrective regression formula, its correspondence to SUS scores, the UMUX-LITE appeared to be a promising alternative to the SUS when it is not desirable to use a 10-item instrument (for example, when the assessment of perceived usability is one part of a larger survey). Our primary goal for the research reported in this paper was to partially replicate our 2013 study so we could investigate 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 norms (e.g., Table 1) to interpret the results.

## 2 Method

To follow up on our initial investigation of the psychometric properties of the UMUX-LITE, especially with regard to how well the regression formula would work with an independent set of data, we combined data from four surveys for a total of 397 cases in which respondents completed the UMUX-LITE and the positive version of the SUS [11, 12] (see Figs. 1 and 2). In addition to collecting SUS and UMUX-LITE ratings, participants also provided ratings of likelihood-to-recommend (LTR).

## 3 Results

### 3.1 Reliability

Consistent with previous research [9], the UMUX-LITE was reliable as assessed using a standard metric of internal consistency (coefficient alpha of .86).

### 3.2 Validity

The UMUX-LITE correlated significantly with ratings of Likelihood-to-Recommend ($r(395) = .72$, $p < .0001$) and SUS scores ($r(395) = .83$, $p < .0001$), providing evidence

| | The System Usability Scale<br>Positive Version | | Strongly Disagree | | | | Strongly Agree |
|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 |
| 1 | I think that I would like to use the website frequently. | | O | O | O | O | O |
| 2 | I found the website to be simple. | | O | O | O | O | O |
| 3 | I thought the website was easy to use. | | O | O | O | O | O |
| 4 | I think that I could use the website without the support of a technical person. | | O | O | O | O | O |
| 5 | I found the various functions in the website were well integrated. | | O | O | O | O | O |
| 6 | I thought there was a lot of consistency in the website. | | O | O | O | O | O |
| 7 | I would imagine that most people would learn to use the website very quickly. | | O | O | O | O | O |
| 8 | I found the website very intuitive. | | O | O | O | O | O |
| 9 | I felt very confident using the website. | | O | O | O | O | O |
| 10 | I could use the website without having to learn anything new. | | O | O | O | O | O |

**Fig. 1.** The system usability scale (positive version)

| | The UMUX-LITE<br>Version 1 | | Strongly Agree | | | | | | Strongly Disagree |
|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | This system's capabilities meet my requirements. | | O | O | O | O | O | O | O |
| 2 | This system is easy to use. | | O | O | O | O | O | O | O |

**Fig. 2.** The UMUX-LITE

of concurrent validity. Note that with a two-item instrument, it is not possible to assess construct validity using analytical methods such as principal components analysis or factor analysis. Thus, the UMUX-LITE is assumed to be a unidimensional metric for perceived usability.

### 3.3 Correspondence

Most importantly, 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 SL-CGS shown in Table 1), it's essentially no difference, especially for results within a point of the break between grades. When sample sizes are large, it's important not to confuse statistically significant differences with meaningful differences.

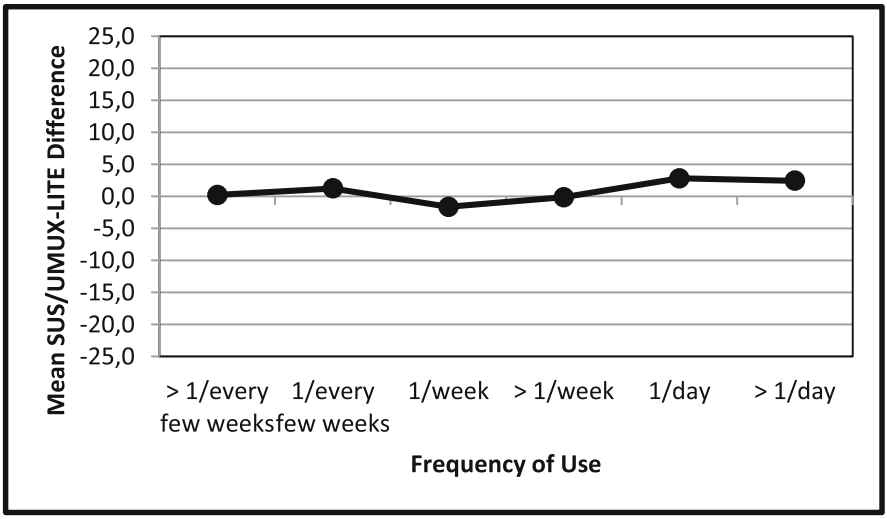


**Fig. 3.** Difference between mean SUS and UMUX-LITE as a function of frequency of use

### 3.4 Sensitivity

To assess sensitivity, an ANOVA was conducted on the main effect of Questionnaire (SUS vs. UMUX-LITE), Frequency of Use (Once every few months or less, Once every few weeks, Once a week, Several times a week, Once a day, More than once a day), and their interaction. Consistent with prior research [9], the effect of Frequency of Use was highly significant ($F(5, 390) = 10.2$, $p < .0001$). The main effect of Questionnaire was not significant ($F(1, 390) = 2.2$, $p = .135$), and the Questionnaire x Frequency of Use interaction was also not significant ($F(5, 390) = 1.6$, $p = .158$). Figure 3 shows the differences between the SUS and UMUX-LITE means as a function of Frequency of Use. As shown in the figure, the mean differences hovered around 0, sometimes slightly positive and sometimes slightly negative, and always with an absolute magnitude less than 3.0.

## 4 Discussion

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 paper is one more 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 was also of considerable magnitude ($r = .83$). The correspondence between the two questionnaires was also evident when assessing the nonsignificant interaction between the questionnaires and reported frequency of use.

As in previous research, the UMUX-LITE exhibited excellent psychometric properties. According to Nunnally [10], 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 reliability of the UMUX-LITE exceeded .70 (coefficient alpha of .86). Two assessments of concurrent validity, correlation with the SUS and correlation with ratings of likelihood-to-recommend, both exceeded .30 (respectively, $r = .83$ and $r = .72$). Finally, the UMUX-LITE had the expected sensitivity to self-reported frequency of use.

Despite these encouraging results, it is important to note some limitations to generalizability. To date, the data 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 recommend its use independent of the SUS.

## 5 Conclusion

These findings add to the emerging literature on the psychometric properties of the UMUX-LITE and increase 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. Researchers who use the SUS should include at least the two UMUX-LITE items in their work (and if possible, the entire UMUX) to build independent databases for future evaluation of its reliability, validity, sensitivity, and correspondence with the SUS.

## References

1. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. Int. J. Hum. Comput. Interact. **6**, 574–594 (2008)
2. Borsci, S., Federici, S., Lauriola, M.: On the dimensionality of the system usability scale: a test of alternative measurement models. Cogn. Process. **10**, 193–197 (2009)
3. Brooke, J.: SUS: a "quick and dirty" usability scale. In: Jordan, P., Thomas, B., Weerdmeester, B. (eds.) Usability Evaluation in Industry, pp. 189–194. Taylor & Francis, London (1996)
4. Davis, D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **13**, 319–339 (1989)
5. Finstad, K.: The usability metric for user experience. Interact. Comput. **22**, 323–327 (2010)

6. Finstad, K.: Response to commentaries on "the usability metric for user experience". Interact. Comput. **25**, 327–330 (2013)
7. Lewis, J.R.: Critical review of "the usability metric for user experience". Interact. Comput. **25**, 320–324 (2013)
8. Lewis, J.R., Sauro, J.: The factor structure of the system usability scale. In: Kurosu, M. (ed.) HCD 2009. LNCS, vol. 5619, pp. 94–103. Springer, Heidelberg (2009)
9. Lewis, J.R., Utesch, B.S., Maher, D.E.: UMUX-LITE—when there's no time for the SUS. In: Proceedings of CHI 2013, pp. 2099–2102. ACM, Paris (2013)
10. Nunnally, J.C.: Psychometric Theory. McGraw-Hill, New York (1978)
11. Sauro, J., Lewis, J.R.: When designing usability questionnaires, does it hurt to be positive? In: Proceedings of CHI 2011, pp. 2215–2223. ACM, Vancouver (2011)
12. Sauro, J., Lewis, J.R.: Quantifying the User Experience: Practical Statistics for User Research. Morgan Kaufmann, Waltham (2012)