Journal of Usability Studies
Vol. 11, Issue 3, May 2016 pp. 89–109

# Re-Assessing the Usability Metric for User Experience (UMUX) Scale

**Mehmet Ilker Berkman**
MSc, MA
Bahcesehir University
Faculty of Communication
Istanbul, Turkey
ilker.berkman@comm.bahcesehir.edu.tr

**Dilek Karahoca**
PhD
Bahcesehir University
Faculty of Engineering
Istanbul, Turkey
dilek.karahoca@bahcesehir.edu.tr

## Abstract

Usability Metric for User Experience (UMUX) and its shorter form variant UMUX-LITE are recent additions to standardized usability questionnaires. UMUX aims to measure perceived usability by employing fewer items that are in closer conformance with the ISO 9241 definition of usability, while UMUX-LITE conforms to the technology acceptance model (TAM). UMUX has been criticized regarding its reliability, validity, and sensitivity, but these criticisms are mostly based on reported findings associated with the data collected by the developer of the questionnaire.

Our study re-evaluates the UMUX and UMUX-LITE scales using psychometric methods with data sets acquired through two usability evaluation studies: an online word processor evaluation survey (n = 405) and a web-based mind map software evaluation survey for three applications (n = 151).

Data sets yielded similar results for indicators of reliability. Both UMUX and UMUX-LITE items were sensitive to the software when the scores for the evaluated software were not very close, but we could not detect a significant difference between the software when the scores were closer.

UMUX and UMUX-LITE items were also sensitive to users' level of experience with the software evaluated in this study. Neither of the scales was found to be sensitive to the participants' age, gender, or whether they were native English speakers. The scales significantly correlated with the System Usability Scale (SUS) and the Computer System Usability Questionnaire (CSUQ), indicating their concurrent validity. The parallel analysis of principal components of UMUX pointed out a single latent variable, which was confirmed through a factor analysis, that showed the data fits better to a single-dimension factor structure.

## Keywords

usability scale, psychometric evaluation, questionnaire, survey

## Introduction

Questionnaires have been widely used to assess the users' subjective attitude related to their experience of using a computer system. Since the 1980s, human-computer interaction (HCI) researchers have developed several standardized usability evaluation instruments. Psychometric methodologies have been employed to develop standardized scales. Nunnally (1978) noted that the benefits of standardization are objectivity, replicability, quantification, economy, communication, and scientific generalization. Using standardized surveys, researchers can verify others' findings by replicating former studies. Powerful statistical methods can be applied to collected quantified data. Standardized questionnaires save the effort required for developing a new research instrument and allow researchers to communicate their results more effectively. After all, a series of research conducted with the same standardized questionnaire is useful for developing generalized findings (Sauro & Lewis, 2012).

The methods for developing a standardized scale (i.e., psychometric methods) are well defined in the literature and are applied to develop many scale instruments in clinical psychology, patient care, education, and marketing. Although these methods are well known, only a few usability scales have been fully evaluated through psychometric methodologies (Sauro & Lewis, 2012).

The Usability Metric for User Experience (UMUX) scale is a new addition to the set of standardized usability questionnaires, and aims to measure perceived usability employing fewer items that are in closer conformity with the ISO 9241 (1998) definition of usability (Finstad, 2010). Although the scale has been developed with a psychometric evaluation approach, it was criticized regarding its validity, reliability, and sensitivity (Bosley, 2013; Cairns, 2013; Lewis, 2013). However, critical statements are based on the results of the original study by Finstad (2010). The only study that has attempted to replicate the results of the original study with different datasets (Lewis, Utesch, & Mahler, 2013) had consistent findings on the validity and reliability of UMUX. In reference to their findings, Lewis et al. (2013) proposed a shorter form of UMUX, called UMUX-LITE, which is a two-item variant that is offered as a time-saving alternative.

Our study contributes to the psychometric evaluation of UMUX and UMUX-LITE, presenting additional evidence for their reliability, validity, and sensitivity by exploring the data collected via UMUX in two different usability evaluation studies. The study also provides a comparison of adjusted (Lewis et al., 2013) and unadjusted UMUX-LITE scores.

### *Psychometric Evaluation*

Psychometrics was intended for studies of individual differences (Nunnally, 1975). Over the decades the methods of psychometrics were intensely used by researchers in the field of psychology and educational sciences. As those disciplines predominantly concentrate on developing standardized scales to identify individual differences, psychometric methods have been of considerable interest in related literature. In the late 1980s, psychometric methods drew attention in HCI as well, because standardized scales became part of the usability testing process to assess the quality of use for a software product from a users' subjective point of view.

Primary measures for a scale's quality are reliability, validity, and sensitivity. Consistency of measurement refers to the reliability of the scale. The extent to which a scale measures what it claims to measure is the validity of a scale. Being reliable and valid, a scale should also be sensitive to experimental manipulations such as changes made in the selection of participants or attributes of the evaluated products.

The reliability of a scale can be assessed by three different approaches: test-retest reliability, alternate-form reliability, and internal consistency reliability. In the test-retest approach, scale items are applied to the same group of participants twice, leaving a time interval between two sessions. Alternate-form questionnaires are intended to measure the same concept with parallel items introducing some changes to the wording and order of items. A high correlation between test-retest or two alternative forms of a questionnaire indicate reliability. However internal consistency, which is typically equated with Cronbach's coefficient alpha, is a widely accepted approach to indicate a scale's reliability because it is easier to obtain from one set of data. The

proportion of a scale's total variance that can be attributed to a latent variable underlying the items is referred as alpha (DeVellis, 2011).

Internal consistency estimates the average correlation among items within a test. If Cronbach's alpha, the indicator of correlation among items, is low, the test is either too short or items have very little in common. Thus, Cronbach's alpha is a highly rated indicator of reliability. It is reported that Cronbach's alpha is remarkably similar to alternate-forms correlation in tests when applied to more than 300 subjects (Nunnally, 1978).

A scale should be examined for criterion validity and construct validity. Criterion validity could be studied in two sub-types: concurrent and predictive. Concurrent validity examines how well one instrument stacks up against another instrument. A scale can be compared with a prior scale, seeking for a correlation between their results. This approach is extensively used in usability scale development studies. Predictive validity, on the other hand, is quite similar to concurrent validity but it is more related to how the instrument's results overlap with future measurements. The Pearson correlation between the instrument and other measurements emphasizes criterion validity.

Construct validation requires the specification of a domain of observables related to the construct at the initial stage. A construct is an abstract and latent structure rather than being concrete and observable. Empirical research and statistical analyses, such as factor analysis, should be made to determine to which extent some of the observables in the domain tend to measure the same construct or several different constructs. Subsequent studies are then conducted to determine the extent to which supposed measures of the construct are consistent with "best guesses" about the construct (Nunnally, 1978).

Another way of examining the constructs is to explore convergent validity, which seeks for the correlation of the instrument with another predictor. From an HCI point of view, survey results can also be compared to the user performance data gathered in usability test sessions. Significant correlations between measurements believed to correspond to the same construct provide evidence of convergent validity. To clarify the constructs, discriminant validity can be explored by checking if there is a mismatch between the instrument's results with those of other instruments that claim to measure a dissimilar construct.

For a researcher in clinical psychology, patient care, education, or marketing, sensitivity is the changes in the responses to a questionnaire across different participants with different attributes. For these disciplines, scales are designed to identify individual differences. However, from an HCI point of view, a usability scale is expected to be sensitive to the differences between systems rather than those between people who use the system. When using the scale to evaluate different systems, one would expect different results, which, in turn, is proof of the scale's sensitivity.

### *Psychometrics of Usability Scales*

Cairns (2013) characterized the evaluation of questionnaires as a series of questions within the context of usability. Validity can be characterized with the question, "Does the questionnaire really measure usability?" When searching for the face validity of a usability questionnaire, Cairns asked, "Do the questions look like sensible questions for measuring usability?" Convergent or concurrent validity seeks the answer to the question, "To what extent does the questionnaire agree with other measures of usability?" Building on convergent validity, the predictive validity of a questionnaire can be assessed by asking, "Does the questionnaire accurately predict the usability of systems?" Discriminant validity is the degree that the questionnaire differentiates "from concepts that are not usability, for example, trust, product support, and so on." Sensitivity, on the other hand, seeks to answer, "To what extent does the measure pick up on differences in usability between systems?" (p. 312).

There a two types of usability questionnaires: post-study and post-task. Post-study questionnaires are administered at the end of a study. Post-task questionnaires, which are shorter in form using three items at most, are applied right after the user completes a task to gather contextual information for each task. We briefly reviewed the post-study questionnaires in the following section because UMUX, which can also be employed to collect data in a field survey, was initially designed to be used as a post-study questionnaire.

***Post-Study Questionnaires***

The Questionnaire for User Interaction Satisfaction (QUIS) was developed as a 27-item, 9-point bipolar scale, representing five latent variables related to the usability construct. Chin, Diehl, and Norman (1988) developed the scale by assessing 150 QUIS forms that were completed for the evaluation of 46 different software programs. The study reported a significant difference in the QUIS results collected for menu-driven applications and command line systems that provided evidence for the scales' sensitivity.

The Software Usability Measurement Inventory (SUMI) consists of 50 items with a 3-point Likert scale representing five latent variables (Kirakowski, 1996). Kirakowski's research provided evidence for construct validity and sensitivity by reporting on the collection of over 1,000 surveys that evaluated 150 different software products. Results affirm that the SUMI is sensitive, as it distinguished two different word processors in work and laboratory settings, while it also produced significantly different scores for two versions of the same product.

The Post-Study System Usability Questionnaire (PSSUQ) initially consisted of 19 items with a 7-point Likert scale and a not applicable (N/A) option. The Computer System Usability Questionnaire (CSUQ) is its variant for field studies (Lewis, 1992; 2002). Three latent variables (subscales), represented by 19 items, are system quality (SysUse), information quality (InfoQual), and interface quality (IntQual). Lewis (2002) offered a 16-item short version that was capable of assessing the same sub-dimensions and used data from 21 different usability studies to evaluate the PSSUQ. He explored the sensitivity of the PSSUQ score for significance of difference to several conditions, such as the study during which the participants completed the PSSUQ, the company that developed the evaluated software, the stage of software development, the type of software product, the type of evaluation, the gender of participants, and the completeness of survey form. As a variant of PSSUQ, CSUQ is designed to assess the usability of a software product without conducting scenario-based usability tests in a laboratory environment (Lewis, 1992; 1995; 2002). Thus, CSUQ is useful across different user groups and research settings.

The System Usability Scale (SUS) was developed for a "quick and dirty" evaluation of usability (Brooke, 1996). Although "it had been developed at the same time period with PSSUQ, it had been less influential since there had been no peer-reviewed research published on its psychometric properties" (Lewis, 2002, p. 464) until the end of the 2000s. After it was evaluated through psychometric methods (Bangor, Kortum, & Miller, 2008; Lewis & Sauro, 2009), it was validated as a unidimensional scale, but some studies suggested that its items represent two constructs: usable and learnable (Borsci, Federici, & Lauriola, 2009; Lewis & Sauro, 2009). SUS consists of 10 items with a 5-point Likert scale. It is reported to provide significantly different scores for different interface types (Bangor et al, 2008) and for different studies (Lewis & Sauro, 2009). Although the SUS score is not affected by gender differences, there is a correlation between the age of participants and the score given to the evaluated applications. It is known that SUS items are not sensitive to participants' native language after a minor change in Item 8, where the word "cumbersome" is replaced with "awkward" (Finstad, 2006).

UMUX has four items with a 7-point Likert scale with a Cronbach's alpha coefficient of .94. Lewis, Utesch, and Maher (2013) reported the coefficient alpha as .87 and .81 for two different surveys. Finstad reported a single underlying construct that conformed to the ISO 9241 definition of usability. However, Lewis et al. (2013) stated that "UMUX had a clear bidimensional structure with positive-tone items aligning with one factor and negative-tone items aligning with the other" (p. 2101). They also reported that UMUX significantly correlated with the standard SUS ($r = .90$, $p < .01$) and another version of SUS in which all items are aligned to have a positive tone ($r = .79$, $p < .01$). These values are lower than the correlation between SUS and UMUX reported in the original study by Finstad (2010; $r = 96$, $p < .01$). However, moderate correlations (with absolute values as small as .30 to .40) are often large enough to justify the use of psychometric instruments (Nunnally, 1978). Accordingly, both studies provided evidence for the concurrent validity of UMUX. To investigate the sensitivity of UMUX to differences between systems, Finstad (2010) conducted a survey study of two systems ($n = 273$; $n = 285$). The *t* tests denoted that both UMUX and SUS produce a significant difference between the scores of the two systems.

The two-item variant of UMUX—UMUX-LITE (Lewis et al., 2013)—is based on the two positive tone items of UMUX, which are items 1 and 3. These items have a connection with the technology acceptance model (TAM) from the market research literature, which assesses usefulness and ease-of-use. UMUX-LITE has a reliability estimate of .82 and .83 on two different surveys, which is excellent for a two-item survey. These items correlated with standard and positive versions of SUS at .81 and .85 ($p < .01$). Correlation of UMUX-LITE with a likelihood-to-recommend (LTR) item was above .7. These findings indicated concurrent validity of UMUX-LITE. On the other hand, Lewis et al. (2013) reported a significant difference between SUS and UMUX-LITE scores that were calculated based on items 1 and 3 of UMUX. For this reason, they have adjusted the UMUX-LITE score with a regression formula to compensate for the difference. A recent study (Lewis, Utesch, & Maher, 2015) confirmed that the suggested formula worked well on an independent data set. Borsci, Federici, Bacci, Gnaldi, and Bartolucci (2015) also replicated previous findings of similar magnitudes for SUS and adjusted UMUX-LITE.

Table 1 gives a quick review of the post-study questionnaires and presents information about their psychometric evaluation.

**Table 1.** Post-Study Questionnaires

| Scale name | No. of items | No. of subscales | Scale type | Reliability (Cronbach's alpha) | Evidence for validity | Evidence for sensitivity | Studies | Number of participants |
|---|---|---|---|---|---|---|---|---|
| QUIS | 27 | 5 | Bipolar (9) | .94 | Yes | Yes | Chin et al., 1988 | 150 |
| SUMI | 50 | 5 | Likert (3) | .92 | Yes | Yes | Kirakowski, 1996 | 1,000+ |
| PSSUQ | 16 | 3 | Likert (7) + N/A option | .94 | Yes | Yes | Lewis, 1992 | 48 |
| | | | | | | | Lewis, 2002 | 210 |
| CSUQ | 16 | 3 | Likert (7) + N/A option | .89 | Yes | Yes | Lewis, 1995 | 377 |
| SUS | 10 | 2 | Likert (5) | .92 | Yes | Yes | Lewis & Sauro, 2009 | 324 |
| | | | | .91 | Yes | Yes | Bangor et al., 2008 | 2,324 |
| UMUX | 4 | 3 | Likert (7) | .94 | Yes | Yes | Finstad, 2010 | 558 |
| | | | | .81 .87 | Yes | - | Lewis et al., 2013 | 402<br>389 |
| UMUX-LITE | 2 | - | Likert (7) | .81 .87 | Yes | Yes | Lewis et al., 2013 | 402<br>389 |

### *Criticism of UMUX and Related Work*

Lewis (2013) criticized Finstad (2010) for his "motivation for scale development," "structural analysis of the scale factors," and "scale validity." Major criticism of structural analysis points out that UMUX should be evaluated not only with a principal component analysis for its factorability, but Finstad should also have conducted "a rotated principal or confirmatory (maximum likelihood) factor analysis to evaluate the possibility that UMUX has two or more underlying factors" (Lewis, 2013, p. 322). Finstad (2013) conducted a maximum-likelihood

analysis to respond to this criticism using the data he collected in the 2010 study. Results verified the one-factor solution.

Criticism of scale validity confirms Finstad's (2013) statement about the limitation of the study and that an industrial study should be conducted to confirm the scale's correlation with "task-based metrics like completion rates, completion times, and error rates" (p. 328). Cairns (2013) also criticized the UMUX study for the same reason in that it does not attempt "to directly position itself in relation to objective measures of usability" (p. 314). He pointed out that the shortness of the questionnaire would also cause participants to provide "socially desirable" responses when they were evaluating two systems. In his response to the Cairns' criticism, Finstad (2013) clarified that these two systems were not compared side by side, but evaluation involved two different groups of participants which resulted with an independent perception about the software—as Cairns suggested.

Although the study provides data that UMUX correlates with SUS, there are some concerns about UMUX's measuring of usability because there is not enough evidence for SUS being capable in that sense. As a response, Finstad (2013) referred to studies on concurrent validity of SUS with other scales, and suggested that this can be extended to UMUX.

Another point of criticism is related with the high Cronbach alpha score obtained by Finstad (2010). Cairns (2013) suggested that UMUX could be "too specific and not measuring usability in a broad sense" (p. 313). Although each UMUX item targets a different aspect of usability that had been found not to correlate strongly on previous studies, the questionnaire items have a strong correlation. He also questioned the item-total correlations. In response, Finstad (2013) presented lower and upper bound values for item-total correlations, illustrating confidence intervals were at the 95% confidence level for correlations of individual UMUX items.

Bosley (2013) criticized Finstad's study for "falling short of producing convincing evidence that UMUX indeed measures 'usability' successfully" (p. 317). Scores provided by UMUX vary substantially with high standard deviation. The systems used for evaluation are reported to have clear differences in usability, one system said to be "low" in usability, and the other claimed to be much better. However, the evaluated *contract worker enterprise application* and an *audio conferencing system* "differ so radically in their functionality" and "such differences may well have subtly influenced usability scores on both UMUX and SUS" (Bosley, 2013, p. 318). Cairns (2013) mentioned the same issue of participants possibly having guessed, "the way they should answer," the question judging from the high face validity of the questionnaire.

Finstad (2013) stated that, even when these groups are separated, correlation between UMUX and SUS for each group "remain strong enough to not be considered problematic for the sensitivity and robustness of the scale" (p. 329). He also presented mean and standard deviation scores for an enhanced version of the "poor usability system" on his original study. The mean score is very close to Sauro's (2011) "average usability score" for SUS, but the standard deviation is higher than scores obtained for each software.

Borsci et al. (2015) explored variation in outcomes of three standardized user satisfaction scales (SUS, UMUX, UMUX-LITE) when completed by users who had spent different amounts of time with a website. Results indicated that users' amount of exposure to the product under evaluation affects the outcomes of each scale. UMUX provided a significant main effect on duration, frequency of use, and interaction of both. As the exposure to the product increased, participants noted higher scores in product evaluation through questionnaires.

Lewis et al. (2015) investigated various measurements of perceived usability, employing SUS and UMUX-LITE. Their study provided evidence on reliability and concurrent validity of UMUX-LITE on an independent dataset.

## Methods

Our study is based on data acquired in two different surveys: (a) a study of two online word processor applications—the word processor (WP) survey (n = 438), and (b) a study of three web-based mind mapping applications—the mind map (MM) survey (n = 153). The systems have similar functionalities, but there is no prior evidence indicating any difference in usability.

***Participants***

Participants were recruited from a business-oriented social web service by sending personal invitations to the members of the online communities related with the evaluated software.

The first part of the WP survey was conducted with 363 participants worldwide, 57 female and 306 male, for system WP01. According to the IP address data of the participants, people from 54 countries participated in the survey. We had 204 participants from Canada, Australia, South Africa, Great Britain, Ireland, and the United States of America; all of these participants were considered to be native English speakers. Participants' age varied from 21 to 73 ($M = 41.1$, $SD = 9.6$).

Another group of 7 female and 35 male participants responded to the questionnaire that was intended to evaluate the system WP02. The majority of the participants were from the country of origin of the software producer, thus, were not native English speakers. Eight of the respondents were from English-speaking countries. Participants' age varied from 21 to 61 ($M = 31.8$, $SD = 8.02$).

A Student's *t* Test revealed a significant difference between the users of WP01 ($M = 6.75$, $SD = .86$) and WP02 ($M = 5.76$, $SD = 1.78$) in terms of their level of experience with the evaluated software, $t(43, 26) = 3.56$, $p < .05$. There was also a significant difference between the two groups in terms of their ages, $t(403) = 5.98$, $p < .05$.

The MM survey involved 151 respondents for three mind map applications: MM01, MM02, and MM03. Demographic details of the participants are given in Table 2. A one-way ANOVA indicated that there were significant differences between the participants' ages, $F(2, 148) = 21.46$, $p < .001$, and their experience with the mind mapping software they evaluated, $F(2, 148) = 15.42$, $p < .001$. A Tukey post-hoc test revealed that the mean score for experience with the software was significantly higher for participants who evaluated MM01 ($M = 6.5$, $SD = 1.18$, $p < .05$) and MM02 ($M = 6.5$, $SD = 1.18$, $p < .05$) compared to the MM03 group ($M = 5$, $SD = 1.85$). There was no difference between the groups MM01 and MM02 ($p > .05$). Participants in MM03 group ($M = 33.8$, $SD = 13.3$) were significantly younger than the participants in MM02 ($M = 49.1$, $SD = 10.9$, $p < .05$) and MM01 groups. ($M = 45.6$, $SD = 9.7$, $p< .05$). We did not observe a significant difference between the mean ages of participants in MM01 and MM02 groups ($p > .05$).

**Table 2.** Participant Demographics for the WP and MM Surveys

| | | WP | | | MM | | | |
|---|---|---|---|---|---|---|---|---|
| | | **WP01** | **WP02** | **Both** | **MM01** | **MM02** | **MM03** | **All** |
| **Gender** | Male | 306 | 35 | 341 | 34 | 62 | 17 | 113 |
| | Female | 57 | 7 | 64 | 11 | 11 | 16 | 38 |
| **English** | Non-native | 167 | 34 | 201 | 23 | 37 | 26 | 86 |
| | Native | 196 | 8 | 204 | 22 | 36 | 7 | 65 |
| **Experience with software** | Mean | 6.8 | 5.8 | 6.8 | 6.5 | 6.5 | 5 | 6.2 |
| | Std. Dev. | .86 | 1.78 | 1.04 | 1.18 | 1.18 | 1.85 | 1.48 |
| **Experience with software** | Tried it once | 2 | 2 | 4 | 1 | - | 2 | 3 |
| | 1-4 times | 8 | 7 | 15 | 1 | 5 | 7 | 13 |
| | 5-10 times | 8 | 2 | 10 | 2 | 3 | 7 | 12 |
| | 11-15 times | 9 | 3 | 12 | 3 | 2 | 3 | 8 |
| | 16-20 times | 6 | 2 | 8 | 2 | 2 | - | 4 |
| | >20 times | 330 | 26 | 356 | 36 | 61 | 14 | 111 |
| **Age** | Mean | 41.1 | 31.8 | 40.1 | 45.6 | 49.1 | 33.8 | 44.7 |
| | Std. Dev. | 9.63 | 8.02 | 9.88 | 9.74 | 10.89 | 13.3 | 12.57 |
| **Age Groups** | < 25 | 11 | 5 | 16 | - | - | 15 | 15 |
| | 25-29 | 33 | 16 | 49 | 1 | 4 | 2 | 7 |
| | 30-34 | 55 | 8 | 63 | 3 | 3 | - | 6 |
| | 35-39 | 64 | 4 | 68 | 12 | 6 | 3 | 21 |
| | 40-44 | 58 | 8 | 66 | 7 | 12 | 4 | 23 |
| | 45-49 | 70 | - | 70 | 5 | 14 | 5 | 24 |
| | 50 to 54 | 41 | - | 41 | 9 | 8 | 2 | 19 |
| | 55-59 | 19 | - | 19 | 3 | 12 | 1 | 16 |
| | 60-64 | 9 | 1 | 10 | 5 | 8 | 1 | 14 |
| | 65-69 | 2 | - | 2 | - | 6 | - | 6 |
| | >69 | 1 | - | 1 | - | - | - | 0 |
| **TOTAL** | | 363 | 42 | 405 | 45 | 73 | 33 | 151 |

### *Procedure*

Participants volunteered for the study by completing an online survey. The survey used in the WP study only contained UMUX items. The MM study survey contained UMUX items as well as SUS and CSUQ items. The survey also contained questions to collect data for the participants' age, gender, and level of experience with the software. Participants were quite familiar with the evaluated systems. Out of 151 MM survey participants, 111 (74%) stated that they had used the application more than 20 times to develop mind maps. Out of 405 WP survey participants, 356 (88%) also stated that their experience with the software exceeded "20 times."

UMUX items were scaled between 1-point for "Strongly Disagree" to 7-point for "Strongly Agree." Participant scores were recoded to maintain a score from 0 to 6, using the method described by Finstad (2010): "Odd items are scored as [score − 1], and even items are scored as [7 – score]" (p. 326; Brackets in original). This method of subtraction is borrowed from the SUS and eliminates the negative/positive keying of the items. The sum of the four UMUX items

was divided by 24, and then multiplied by 100 to achieve parity with the SUS score. Briefly, the following formula is for calculating the UMUX score:

$$UMUX = \big((UMUXitem1 - 1) + (UMUXitem3 - 1) + (7 - UMUXitem2) + (7 - UMUXitem4)\big) x \frac{100}{24}$$

We handled the UMUX-LITE score in adjusted and unadjusted forms. The adjusted score is calculated employing the regression equation established by Lewis et al. (2013). For the remainder of this paper, the adjusted score is going to be referred to as UMUX-LITEr, and the unadjusted score as UMUX-LITE, in cases where it is necessary to distinguish the method of score calculations. Below are the formulas that we used for calculating these scores:

Adjusted score: $UMUX - LITEr = .65\left(\big((UMUXitem1 - 1) + (UMUXitem3 - 1)\big) x \frac{100}{12}\right) + 22.9$

Unadjusted score: $UMUX - LITE = \big((UMUXitem1 - 1) + (UMUXitem3 - 1)\big) x \frac{100}{12}$

Systems evaluated in the word processor survey, namely WP01 and WP02, are web-based online word processors running on a personal cloud service.

The MM survey involved users of three different web-based diagramming tools designed for developing mind maps that are diagrams which represent "semantic or other connections between portions of learned material hierarchically" (Eppler, 2006, p. 203). Besides the UMUX questionnaire, SUS and the 16-item short version of CSUQ questionnaires were also employed in the survey to assess the concurrent validity of UMUX as well as UMUX-LITE and UMUX-LITEr through Pearson correlations between the scale scores and their sub-dimensions. As suggested by Brooke (1996), the SUS score for each participant was equalized within a range of 0 to 100 by scoring odd items as (score − 1) and even items as (5 − score) before multiplying each item by 2.5. CSUQ responses were scored by averaging the items as suggested by Lewis (1995). If an item was not answered or marked as N/A (not applicable), the case was excluded.

Participants' responses to scale items in the WP and MM surveys were evaluated to determine the UMUX and UMUX-LITE scales' sensitivity to the variables of age, language, gender differences, as well as participants' experience with the evaluated systems. A series of multivariate analysis of variance (MANOVA) was employed to observe the effect of each variable on scale items. We also explored the sensitivity of the SUS and CSUQ items based on our MM study data. A significant effect indicated that the scale items were sensitive to the explored variable. Because MANOVA was used for sensitivity analysis, which is based on scale items rather than scale scores, UMUX-LITE and UMUX-LITEr were not evaluated separately.

Cronbach's alpha values were calculated using the merged data acquired in both surveys (n = 556) in order to determine the reliability of UMUX and UMUX-LITE. Corrected item total correlations and the Pearson correlation coefficient of each scale item with the overall scale score were also considered indicators of reliability. UMUX-LITE and UMUX-LITEr are not different in terms of reliability because the Cronbach's alpha is calculated based on items, not the scale scores.

Finally, UMUX was examined for the existence of any possible subscales through applying different factor analysis methods. A principal components based exploratory analysis (PCA) was applied, as well as a principal axis factor analysis with a limited number of factors, to evaluate the possibility that UMUX has two or more underlying factors. A parallel analysis was also conducted to identify the number of components that could be extracted from the available sample. In addition, the single factor and two factor models were tested through a confirmatory factor analysis (CFA) using structural equation modelling (SEM) software. A scree plot was also checked. These methods were applied to the merged data set of word processor and mind map data studies.

Because it has only two items, UMUX-LITE depends on a single latent variable. Thus we did not explore its constructs.

## Results

A discussion of the results related to reliability, sensitivity, and validity of UMUX are discussed in the following sections. We also present a visual abstract of the methodology and results in Figure 1.

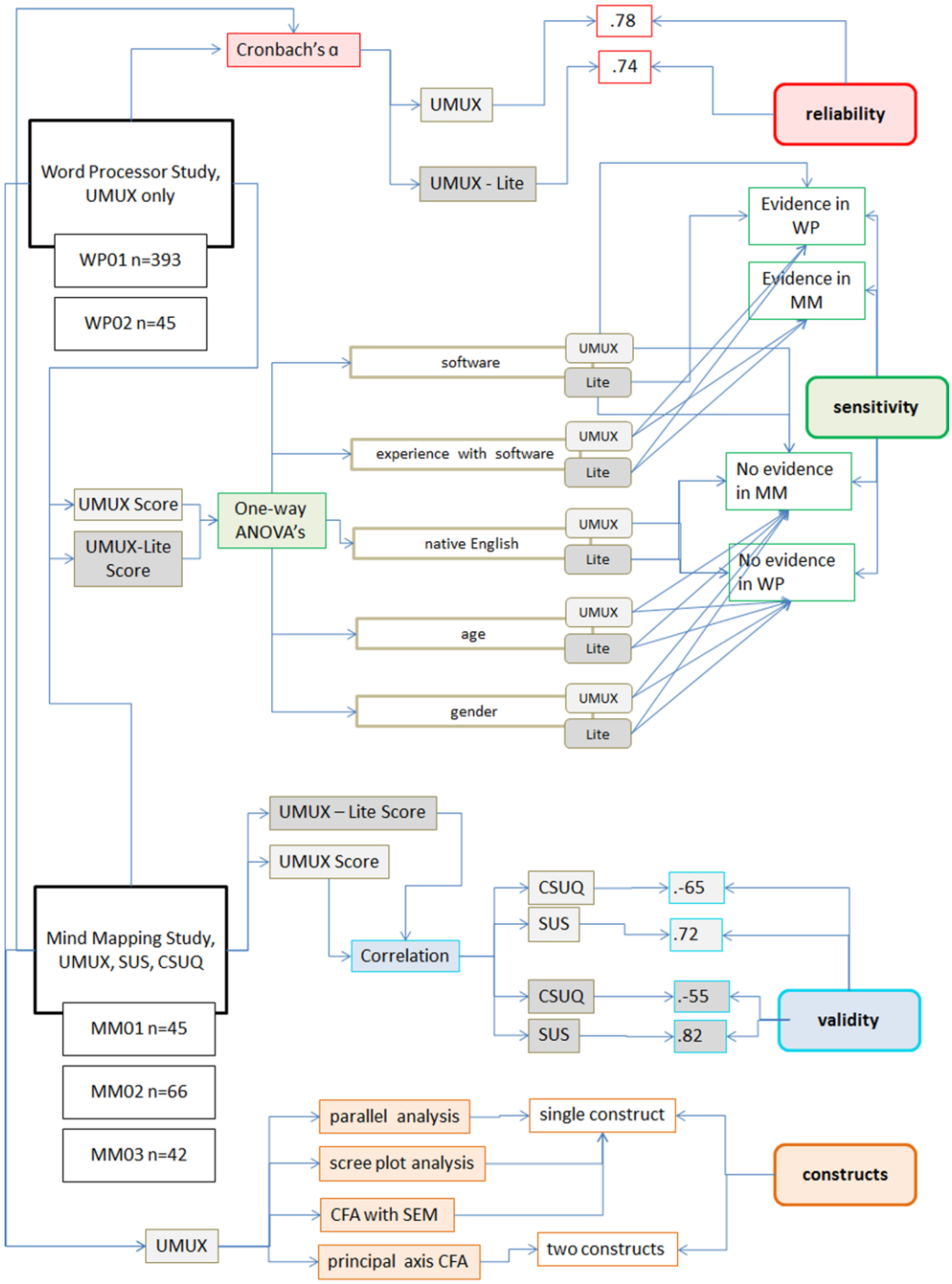


**Figure 1.** Visual abstract of methodology and results.

### *Reliability*

A reliability analysis of the merged data set resulted with a Cronbach's alpha value of .83 for UMUX. When the data sets were examined separately, the same value emerged on each. This score is relatively lower than Finstad's result of .94, but still above the threshold of .7 that is considered the minimum reliability score for a scale which does not measure individual aptitude (Landauer, 1988).

**Table 3.** Reliability Measures of UMUX Items

| | Items | r with score | upper bound | lower bound |
|---|---|---|---|---|
| **UMUX** | UMUX01: This system's capabilities meet my requirements. | .81 | .86 | .74 |
| | UMUX02: Using this system is a frustrating experience. | -.86 | -.90 | -.82 |
| | UMUX03: This system is easy to use. | .78 | .83 | .71 |
| | UMUX04: I have to spend too much time correcting things with this system. | -.82 | -.87 | -.76 |
| **UMUX -LITE** | UMUX01: This system's capabilities meet my requirements. | .91 | .92 | .88 |
| | UMUX03: This system is easy to use. | .90 | .92 | .86 |

UMUX-LITE also yielded an acceptable coefficient alpha (.77). Reliability scores for WP and MM datasets were .77 and .78 respectively. Table 3 shows the correlation of each item with the overall scale score with upper and lower bounds at 99% confidence interval.

SUS provided a Cronbach's alpha value of .83 for the MM dataset. For CSUQ, the reliability score was .95.

For UMUX items, we observed a lower correlation with the overall UMUX score when compared to previous studies (Finstad, 2010; Lewis et al., 2013; 2015). UMUX-LITE items have a higher correspondence with the overall scale score, in consistence with previous work (Lewis et al., 2013; 2015). The results of the study indicate that both UMUX and UMUX-LITE are reliable scales, and their items are significantly correlated at a high level.

### *Sensitivity*

A series of MANOVAs was performed on WP and MM data separately to test the sensitivity of the scale items for the evaluated software; the participants' level of experience with the software; and the differences between participants' age, gender, and native language.

*Gender*

We did not observe a significant effect of gender for participants in the WP study for UMUX items, $F(4, 400) = .14$, $p > .05$; Wilk's $\Lambda = 1$, partial $\eta 2 = .001$, and UMUX-LITE items, $F(2, 402) = .01$, $p > .05$; Wilk's $\Lambda = .1$, partial $\eta 2 = 0$.

In the MM study, results were similar for UMUX items, $F(4, 146) = .145$, $p > .05$; Wilk's $\Lambda = .96$, partial $\eta 2 = .04$, and UMUX-LITE items, $F(2, 148) = 2.5$, $p > .05$; Wilk's $\Lambda = .97$, partial $\eta 2 = .033$.

Recoded SUS items in the MM study emerged to be sensitive to the gender of participants, $F(10, 140) = 2.13$, $p < .05$; Wilk's $\Lambda = .87$, partial $\eta 2 = .13$. The effect of gender was significant for the item "SUS07-I would imagine that most people would learn to use this system very quickly," $F(1, 149) = 1538.8$, $p < .05$). CSUQ items also appeared to be sensitive to gender in our MM study data, $F(16, 134) = 1.94$, $p < .05$; Wilk's $\Lambda = .81$, partial $\eta 2 = .19$.

*Age*

Regarding age groups, UMUX items, $F(40, 1484) = 1.25$, $p > .05$; Wilk's $\Lambda = .88$, partial $\eta 2 = .03$, and UMUX-LITE items, $F(20, 786) = .8$, $p > .05$; Wilk's $\Lambda = .96$, partial $\eta 2 = .032$, did not differ significantly in the WP study. The MM study results also did not reveal a significant

difference for UMUX items, $F(36, 519) = 1.12$, $p > .05$; Wilk's $\Lambda = .76$, partial $\eta 2 = .068$, and UMUX-LITE items, $F(18, 280) = 1.2$, $p > .05$; Wilk's $\Lambda = .86$, partial $\eta 2 = .072$, with regard to age groups. There were no significant differences for the SUS items, $F(90, 905) = 1.01$, $p > .05$; Wilk's $\Lambda = .52$, partial $\eta 2 = .070$, but a significant difference was observed for CSUQ items by age groups, $F(144, 1012) = 1.3$, $p < .05$; Wilk's $\Lambda = .26$, partial $\eta 2 = .014$.

*Native vs. Non-Native English Speakers*

Regardless of whether the participants were native English speakers or not, their responses to items were quite similar for UMUX, $F(4, 400) = 1.8$, $p > .05$; Wilk's $\Lambda = .98$, partial $\eta 2 = .018$, and UMUX-LITE, $F(2, 402) = .65$, $p > .05$; Wilk's $\Lambda = 1$, partial $\eta 2 = .003$, in the WP study.

In the MM study, native English speakers' and non-native English speakers' responses to items were significantly different when they had evaluated the software through UMUX, $F(4, 146) = 3.59$, $p < .05$; Wilk's $\Lambda = .91$, partial $\eta 2 = .09$. The observed effect was statistically significant for items UMUX01, $F(1, 149) = 4.2$, $p < .05$, and UMUX04, $F(1, 149) = 9.6$, $p < .05$. For UMUX01, native English speakers' item mean score ($M = 4.89$, $SD = .99$) was higher than that of non-natives ($M = 4.55$, $SD = 1.06$). Likewise, the UMUX04 item mean score of native speakers ($M = 4.89$, $SD = 1.05$) was higher than that of non-natives ($M = 4.38$, $SD = 1.51$).

MANOVA for UMUX-LITE items also demonstrated a significant difference between native and non-native English speakers, $F(2, 148) = 4.5$, $p < .05$; Wilk's $\Lambda = .94$, partial $\eta 2 = .058$, due to the item UMUX01, $F(1, 148) = 4.8$, $p < .05$). We did not observe a significant effect of native language for any SUS items, $F(10, 140) = 1.17$, $p > .05$; Wilk's $\Lambda = .92$, partial $\eta 2 = .077$, or CSUQ items, $F(16, 134) = 1.4$, $p > .05$; Wilk's $\Lambda = .86$, partial $\eta 2 = .14$.

*Level of Experience with the Evaluated Software*

In the WP study, we observed a significant effect of the participants' level of experience with the software for their responses to UMUX items, $F(20, 1314) = 3.9$, $p < .05$; Wilk's $\Lambda = .83$, partial $\eta 2 = .047$. The effect was observed for all UMUX items. Participants' responses were marginally affected by their level of experience with the software for UMUX01, $F(5, 399) = 2.27$, $p = .047$. The effect of users' experience was significant for UMUX02, $F(5, 399) = 10.2$, $p < .05$; UMUX03, $F(5, 399) = 5.5$, $p < .05$; and UMUX04, $F(5, 399) = 7.8$, $p < .05$. Post-hoc analyses using the Tukey post-hoc criterion for significance indicated that the UMUX01 items score was not significantly different for any participant sub-group regarding the level of experience with the software. For UMUX02, participants who used the software "more than 20 times" scored the item significantly higher than those "who tried it once" or used it "1 to 4 times." These experienced participants also yielded a higher score for UMUX03 compared to participants who used the system "1 to 4 times." They also scored UMUX04 highly compared to those who tried the system once, used it 5 to 10 times, or 11 to 15 times.

MM study data yielded similar results for the effect of the participants' level of experience with the software for their responses to UMUX items, $F(20, 472) = 1.7$, $p < .05$; Wilk's $\Lambda = .79$, partial $\eta 2 = .056$. The effect was significant for items UMUX01, $F(5, 145) = 3.8$, $p < .05$, and UMUX02, $F(5, 145) = 2.4$, $p < .05$. Post-hoc analyses using the Tukey post-hoc criterion for significance indicated that participants who used the software "more than 20 times" scored the item significantly higher than those "who tried it once" or used it "5 to 10 times." The UMUX02 items score was not significantly different for any participant sub-group regarding the level of experience with the software in the MM study.

UMUX-LITE item scores were significantly influenced by participants' level of experience in the WP study, $F(10, 296) = 2.77$, $p < .05$; Wilk's $\Lambda = .93$, partial $\eta 2 = .034$. A marginally significant effect was observed for UMUX01, $F(5, 399) = 2.3$, $p = .047$, and a significant effect was observed for UMUX03, $F(5, 399) = 5.5$, $p < .05$. The mean difference was significantly higher on UMUX03 between the participants who used the system more than 20 times and used the system 1 to 4 times.

MM study results were comparable to those in the WP study, $F(10, 288) = 2.24$, $p < .05$; Wilk's $\Lambda = .86$, partial $\eta 2 = .072$. However, only the UMUX01 item was significantly affected by participants' level of experience with the software, $F(5, 145) = 3.85$, $p < .05$. Post-hoc analyses indicated that participants who used the software more than 20 times scored this item

significantly higher than those who used it 5 to 10 times. We observed a significant difference for SUS items with regard to the participants' level of experience with the software, $F(50, 624) = 1.53$, $p < .05$; Wilk's $\Lambda = .59$, partial $\eta 2 = .1$, due to the items "SUS01-I think that I would like to use this system frequently," $F(5, 145) = 9.57$, $p < .05$, and "SUS05-I found the various functions in this system were well integrated," $F(5, 145) = 2.67$, $p < .05$. On SUS01, participants who used the system more than 20 times sored the item significantly higher than the other groups except those who tried it once. On SUS05, the mean difference was significantly higher for those who used the system more than 20 times compared to participants who used the system 11 to 15 times. CSUQ items also revealed a significant difference regarding the level of experience with the software, $F(80, 630) = 2.4$, $p < .05$; Wilk's $\Lambda = .28$, partial $\eta 2 = .23$. The effect was observed for almost all items, suggesting that as the interaction with the software increased, the participant's score for the item decreased significantly.

*Evaluated Software*

In the WP study, we observed a significant effect of the software evaluated for UMUX, $F(4, 400) = 6$, $p < .05$; Wilk's $\Lambda = .94$, partial $\eta 2 = .057$. The effect was observed for items UMUX02, $F(1, 403) = 8.8$, $p < .05$; UMUX03, $F(1, 403) = 13.9$, $p < .05$; and UMUX04, $F(1, 403) = 3.9$, $p < .05$), but not for item UMUX01, $F(1, 403) = .38$, $p > .05$. Participants who evaluated WP01 scored higher on these three items compared to WP02 evaluators. A similar effect was observed for UMUX-LITE scores, $F(2, 402) = 9.55$, $p < .05$; Wilk's $\Lambda = .80$, partial $\eta 2 = .011$, due to the UMUX03 item, $F(1, 403) = 13.8$, $p < .05$.

On the other hand, we did not observe a significant effect of the software for the UMUX items, $F(8, 290) = 1.6$, $p > .05$; Wilk's $\Lambda = .92$, partial $\eta 2 = .042$, and UMUX-LITE, $F(4, 294) = 1.37$, $p > .05$; Wilk's $\Lambda = .96$, partial $\eta 2 = .018$, for the three mind mapping software subjects in the MM study. When we analyzed the sensitivity of SUS in the MM study, we observed a significant effect of the evaluated software for its items, $F(20, 278) = 1.6$, $p < .05$; Wilk's $\Lambda = .96$, partial $\eta 2 = .018$, due to item SUS07 "I would imagine that most people would learn to use this system very quickly," $F(2, 148) = 8.8$, $p < .05$. Correspondingly, MANOVA for CSUQ items revealed a significant difference between the software evaluated in the MM study, $F(32, 266) = 1.62$, $p < .05$; Wilk's $\Lambda = .70$, partial $\eta 2 = .016$.

Because we provided evidence for the concurrent validity of UMUX and UMUX-LITE with SUS and CSUQ, the results in the MM study did not indicate that UMUX and UMUX-LITE were not sensitive to differences between the software. It is more likely that the evaluated software are not very different in terms of usability. SUS and CSUQ are more sensitive to small differences because they have more items compared to UMUX and its shorter form variant.

**Table 4.** Mean Scores of Scales

| | | WP Study | | | | | | MM Study | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | UMUX | | LITE | | LITEr | | UMUX | | LITE | | LITEr | | SUS | | CSUQ | |
| | | **M** | SD | **M** | SD | **M** | SD | **M** | SD | **M** | SD | **M** | SD | **M** | SD | **M** | SD |
| Gender | Male | **81.2** | 18.7 | **82.5** | 17.9 | **76.5** | 11.7 | **77.7** | 16.4 | **77.6** | 16.5 | **73.3** | 10.7 | **78.6** | 13.9 | **2.2** | 1 |
| | Female | **79.6** | 18.4 | **81.4** | 17.6 | **75.8** | 11.4 | **79.7** | 17.4 | **81.4** | 16 | **75.8** | 10.4 | **82.1** | 14.5 | **2.2** | .8 |
| Age groups | < 25 | **75.3** | 15 | **83.9** | 17.6 | **77.4** | 11.4 | **70.3** | 14.7 | **75** | 15.4 | **71.7** | 10 | **77** | 14.4 | **2.5** | .6 |
| | 25 to 29 | **80.2** | 17.3 | **84.7** | 16.7 | **78** | 10.9 | **81.5** | 11.2 | **84.5** | 11.2 | **77.8** | 7.3 | **75.4** | 12.3 | **2.2** | .7 |
| | 30 to 34 | **82.5** | 20.6 | **83.9** | 18.7 | **77.4** | 12.1 | **80.6** | 23.4 | **83.3** | 16.7 | **77.1** | 10.8 | **85.8** | 13.4 | **2** | .9 |
| | 35 to 39 | **79.3** | 22 | **82.2** | 19.4 | **76.3** | 12.6 | **80.6** | 11.9 | **82.5** | 10.5 | **76.6** | 6.8 | **80.4** | 11 | **2** | .6 |
| | 40 to 44 | **80.9** | 15 | **80.6** | 15.5 | **75.3** | 10.1 | **78.8** | 13.1 | **77.2** | 15.3 | **73.1** | 10 | **78.8** | 12.3 | **2.3** | 1.1 |
| | 45 to 49 | **80.1** | 18 | **79.3** | 18 | **74.4** | 11.7 | **75.7** | 23.1 | **74.3** | 20.8 | **71.2** | 13.5 | **78.4** | 17.7 | **2.4** | 1.1 |
| | 50 to 54 | **83.7** | 17.5 | **83.9** | 17.1 | **77.5** | 11.1 | **78.3** | 18.5 | **76.8** | 19.2 | **72.8** | 12.5 | **80** | 16.5 | **2.1** | 1 |
| | 55 to 59 | **80.9** | 25.1 | **80.3** | 24.1 | **75.1** | 15.7 | **79.4** | 17.1 | **79.2** | 15.2 | **74.4** | 9.9 | **77.5** | 13.8 | **2.1** | .8 |
| | 60 to 64 | **85.4** | 13.2 | **84.2** | 13.9 | **77.6** | 9 | **77.7** | 14.6 | **77.4** | 18.6 | **73.2** | 12.1 | **81.4** | 14.3 | **2.5** | 1.3 |
| | 65 to 69 | **91.7** | 11.8 | **91.7** | 11.8 | **82.5** | 7.7 | **89.6** | 9.4 | **90.3** | 12.3 | **81.6** | 8 | **86.3** | 11.3 | **1.9** | .6 |
| | >69 | **100** | . | **100** | . | **87.9** | . | **.** | . | **.** | . | **.** | . | **.** | . | **.** | . |
| English | Non-native | **80.3** | 17.9 | **81.5** | 17 | **75.9** | 11 | **76.5** | 17.2 | **77.5** | 16.2 | **73.3** | 10.5 | **78** | 14.8 | **2.3** | .9 |
| | Native | **81.5** | 19.4 | **83** | 18.7 | **76.9** | 12.2 | **80.5** | 15.6 | **79.9** | 16.7 | **74.8** | 10.9 | **81.4** | 12.8 | **2.1** | 1 |
| Experience with software | Tried it once | **58.3** | 13.2 | **79.2** | 19.8 | **74.4** | 12.9 | **61.1** | 12.7 | **72.2** | 4.8 | **69.8** | 3.1 | **80** | 5 | **2.2** | .8 |
| | 1-4 times | **65.8** | 16.6 | **69.4** | 19.3 | **68** | 12.6 | **73.7** | 19.7 | **76.3** | 18.6 | **72.5** | 12.1 | **78.7** | 20.5 | **2.5** | .8 |
| | 5-10 times | **67.5** | 12.4 | **76.7** | 16.1 | **72.7** | 10.5 | **69.8** | 13.1 | **66.7** | 17 | **66.2** | 11.1 | **73.8** | 14.2 | **2.8** | 1.4 |
| | 11-15 times | **66.7** | 27.7 | **72.2** | 23.1 | **69.8** | 15 | **68.8** | 15.1 | **67.7** | 18.6 | **66.9** | 12.1 | **66.9** | 17.3 | **3.4** | 1.4 |
| | 16-20 times | **63.5** | 14 | **66.7** | 20.9 | **66.2** | 13.6 | **71.9** | 13.3 | **75** | 11.8 | **71.7** | 7.7 | **72.5** | 7.4 | **2.2** | .7 |
| | >20 times | **83.1** | 17.7 | **83.7** | 17.2 | **77.3** | 11.2 | **81** | 16.1 | **81.2** | 15.6 | **75.7** | 10.1 | **81.3** | 12.8 | **2.1** | .8 |
| Software | WP01 | **81.8** | 18.3 | **83** | 17.3 | **76.8** | 11.3 | **.** | . | **.** | . | **.** | . | **.** | . | **.** | . |
| | WP02 | **73.4** | 20.4 | **76.4** | 21.2 | **72.6** | 13.8 | **.** | . | **.** | . | **.** | . | **.** | . | **.** | . |
| | MM01 | **.** | . | **.** | . | **.** | . | **82.5** | 13.9 | **81.7** | 15.1 | **76** | 9.8 | **82.8** | 10.9 | **2** | .8 |
| | MM02 | **.** | . | **.** | . | **.** | . | **77.5** | 18.3 | **77.4** | 17.4 | **73.2** | 11.3 | **77.5** | 15.5 | **2.3** | 1 |
| | MM03 | **.** | . | **.** | . | **.** | . | **74** | 14.9 | **76.8** | 15.7 | **72.8** | 10.2 | **79.3** | 14.2 | **2.4** | .9 |
| **TOTAL** | | **80.9** | 18.6 | **82.3** | 17.9 | **76.4** | 11.6 | **78.2** | 16.6 | **78.5** | 16.4 | **74** | 10.6 | **79.5** | 14.1 | **2.2** | .9 |

### *Validity*

Our study involves an assessment of the concurrent validity of UMUX. The data obtained in the mind map survey from 151 respondents who evaluated one of the three mind mapping applications include SUS and CSUQ questionnaires in addition to UMUX. We examined the correlations between the scales and subscales to provide evidence for the concurrent validity of UMUX and UMUX-LITE. Results are shown in Table 5, including lower and upper bounds at 99%

confidence intervals. Because both regression-adjusted UMUX-LITEr and unadjusted UMUX-LITE scores correlate identically, the given UMUX-LITE results also cover the UMUX-LITEr.

The UMUX score positively correlates with the total SUS score ($r = .74$, $p < .001$) and CSUQ score ($r = -.65$, $p < .001$). These results are below Finstad's (2010) goal of a correlation higher than 0.8 with SUS. We have detected a moderate correlation between UMUX and SUS, which is lower than the previously reported correlations (Lewis et al., 2013) but higher than results reported by Borsci et al. (2015). However, we observed that both CSUQ and SUS correlated with UMUX at a level that could still be accepted as evidence for the concurrent validity of UMUX. UMUX-LITE yielded similar results.

Before examining the correlation of UMUX with the SUS and CSUQ subscales, we explored our data for underlying factors of SUS and CSUQ. We confirmed the three constructs of CSUQ through an exploratory principal axis analysis for three factors, using a varimax rotation. A similar analysis on SUS for two factors resulted in a factor structure in which negatively and positively keyed items loaded on separate factors. To explore the correlated and uncorrelated factor structures of *usable* and *learnable* constructs suggested by Borsci et al. (2009), we used the Ωnyx structural equation modeling software (von Oertzen, Brandmaier, & Tsang, 2015) for CFA. We checked the root mean square error of approximation (RMSEA), the Tucker–Lewis index (TLI), the comparative fit index (CFI), and the root mean square residual (SRMR) values to identify which model fits the data best (Schreiber, Nora, Stage, Barlow, & King, 2006). Results are presented in Table 5.

In neither of the models were the CFI and TLI indicators above the acceptable value of >.95. We observed that RMSEA values were not within the acceptable range of <.06, and SRMR values were >.08. We found out that neither of the models fits the data. Thus, we have no evidence that our SUS data provides usable and learnable constructs. Consequently, we did not perform any reliability analysis on subscales of SUS in this study.

**Table 5.** Indicators of Model Fit for CFA on SUS Items With SEM Software

| | **Usable and learnable constructs** | |
|---|---|---|
| | **Uncorrelated** | **Correlated** |
| $\chi 2$ | 185 | 143.6 |
| Restricted df | 45 | 44 |
| AIC | 3694 | 3654 |
| BIC | 3695 | 3655 |
| RMSEA | .14 | .12 |
| SRMR (covariances only) | .18 | .1 |
| CFI (to independent model) | .67 | .76 |
| TLI (to independent model) | .52 | .61 |

An examination of the correlations of the UMUX score with the subscales of CSUQ shows that UMUX correlates with the system usefulness (SysUse), information quality (InfoQual), and interface quality (IntQual) constructs.

UMUX-LITE has correlated almost at a similar level with SUS and CSUQ. Compared to UMUX, UMUX-LITE has a slightly higher correlation with CSUQ and its subscales. The overlaps in the 99% confidence intervals were considerable for the correlations between the scales, so it appeared that they had similar magnitudes of association (see Table 6).

It should be noted that UMUX has been developed to conform to the ISO 9241 definition of usability. UMUX items are designed to query the perceived usability of a system based on dimensions of efficiency, effectiveness, and satisfaction. Items related to the SysUse subscale of CSUQ query the ease of use, simplicity, quickness of completing tasks, comfort, and productivity. For this reason, UMUX and UMUX-LITE have higher correlations with the SysUse compared with the InfoQual and IntQual sub-dimensions of CSUQ.

The correlation of UMUX with the latent constructs that can be observed by CSUQ indicates that UMUX has a higher correlation with system usefulness and usability, rather than information quality and interface quality. However, we noticed that in comparison to UMUX-LITE, UMUX presents a slightly closer relation to SUS and CSUQ.

**Table 6.** Correlation Between Scales and Subscales With Confidence Intervals (CI)

| | | UMUX -LITE | SUS | CSUQ | | | |
|---|---|---|---|---|---|---|---|
| | | | Overall | Overall | SysUse | InfoQual | IntQual |
| **UMUX** | | | | | | | |
| | r | .89 | **.74** | **-.65** | -.64 | -.56 | -.56 |
| 99% CI | upper bound | .93 | **.83** | **-.55** | -.52 | -.41 | -.37 |
| 99% CI | lower bound | .81 | **.63** | **-.75** | -.76 | -.66 | -.70 |
| **UMUX-LITE** | | | | | | | |
| | r | | **.72** | **-.69** | -.67 | -.58 | -.59 |
| 99% CI | upper bound | | **.82** | **-.55** | -.54 | -.44 | -.40 |
| 99% CI | lower bound | | **.57** | **-.78** | -.78 | -.72 | -.73 |
| **SUS** | | | | | | | |
| | r | | | **-.72** | -.73 | -.61 | -.58 |
| 99% CI | upper bound | | | **-.40** | -.61 | -.46 | -.39 |
| 99% CI | lower bound | | | **-.81** | -.81 | -.72 | -.71 |

UMUX, UMUX-LITE, UMUX-LITEr, SUS, and CSUQ scores also showed a correspondence in magnitude, with the exception that SUS scores indicated a better usability for MM03 software. As shown in Table 4, mean score magnitudes were consistent for different software for each scale. Participants of the MM01 and MM02 evaluation scored similarly both with UMUX and SUS. Although we used the regression formula to ensure that UMUX-LITEr scores corresponded to SUS scores, the UMUX-LITEr mean score was slightly lower than the SUS mean score. However, the unadjusted UMUX-LITE mean score was almost identical to the SUS mean score even though the purpose of the regression adjustment was to increase the correspondence with SUS.

At this point, we conducted a set of paired sample *t* tests to investigate how these scores differed from each other. There was a significant difference between regression adjusted UMUX-LITEr scores and SUS scores, $t(150) = -6.9$, $p < .05$. The comparison of unadjusted UMUX-LITE scores and SUS scores did not reveal a significant difference, $t(150) = -.97$, $p > .05$; neither did UMUX scores, $t(150) = -1.3$, $p > .05$.

### *Examining the Underlying Constructs of UMUX*

The Kaiser–Meyer–Olkin measure of sampling adequacy of .78 shows that correlations between variables can be explained by other variables. Therefore, the data set has the potential to explain relevant factors. Bartlett's test of sphericity also indicates the existence of correlations between variables, $\chi 2\ (6) = 877.8$, $p < .001$.

An exploratory principal component analysis was performed with the merged data acquired in the word processor and mind map surveys. Four UMUX items revealed only one component that met the Kaiser-Guttman "eigenvalues greater than one" criterion, which explains 67% of the total variance. The "elbow" on the scree plot hints a two-component structure.

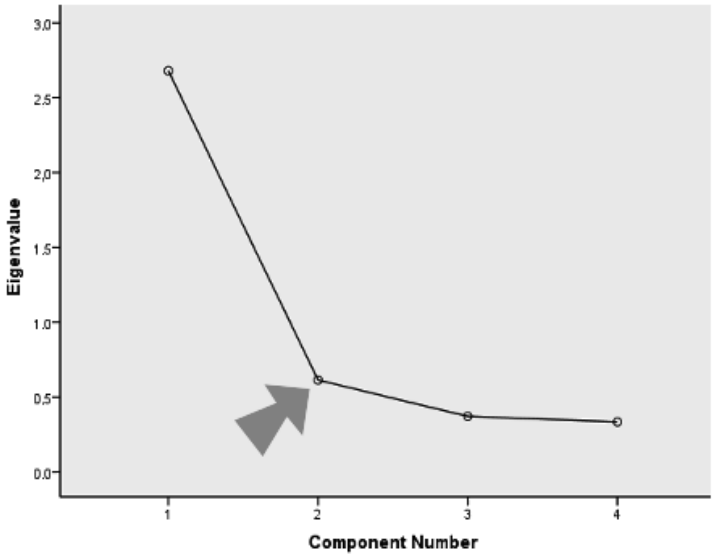


**Figure 2.** Scree plot for UMUX items.

A parallel analysis was also conducted to decide on the number of components to be retained. Results in Table 7 suggest that a single component model is the best fit for our dataset as the actual eigenvalue exceeds the random value generated for four items with 556 participants.

**Table 7.** Actual Eigenvalues and Parallel Analysis Estimations

| Component | Random | | Actual | |
|---|---|---|---|---|
| | Eigenvalue | SD | Eigenvalue | Total variance explained |
| 1 | 1.1 | .030 | 2.68 | 67 |
| 2 | 1.03 | .022 | .51 | 82.3 |
| 3 | .97 | .020 | .37 | 91.6 |
| 4 | .91 | .029 | .33 | 100 |

Although the parallel analysis suggested a single component structure of UMUX, a principal axis analysis for two factors using a varimax rotation was applied to the data to investigate results based on the findings of Lewis et al. (2013). The factor loadings of items are given in Table 8.

**Table 8.** Factors That Emerged by a Varimax Rotated Principal Axis Factoring

| UMUX Item | Factor 1 | Factor 2 |
|---|---|---|
| UMUX01: This system's capabilities meet my requirements. | -.378 | **.717** |
| UMUX03: This system is easy to use. | -.343 | **.968** |
| UMUX02: Using this system is a frustrating experience. | **.724** | .442 |
| UMUX04: I have to spend too much time correcting things with this system. | **.69** | -.312 |

For further investigation of underlying constructs, we used the Ωnyx structural equation modeling software (von Oertzen et al., 2015) for CFA to compare the single construct and two-construct models. The Ωnyx equation affords maximum likelihood parameter estimation. RMSEA, TLI, CFI, and SRMR values were taken as model fit indicators (Schreiber at al., 2006).

For both models, RMSEA values were not within the acceptable range of < .06, but slightly lower for the two-construct model. The SRMR value below .08 indicates a model fit for both models. The CFI value was > .95 for the two-construct model, but not for the single construct model. TLI was not > .95 for both models, but higher for the two-construct when compared to the single construct model.

Those values indicate a better fit between the observed data and the two-construct model. The Akaike information criterion (AIC) as well as the Bayes information criterion (BIC) were slightly lower for the two-construct model, which also indicates a better fit. A comparison of model fit indicators is presented in Table 9.

**Table 9.** Indicators of Model Fit for CFA on UMUX Items With SEM Software

| | **Single construct model** | **Two-construct model** |
|---|---|---|
| $\chi 2$ | 92.8 | 27.2 |
| Restricted df | 5 | 3 |
| AIC | 6755.3 | 6623.8 |
| BIC | 6777 | 6724.1 |
| RMSEA | .18 | .12 |
| SRMR (covariances only) | .07 | .04 |
| CFI (to independent model) | .9 | .97 |
| TLI (to independent model) | .78 | .9 |

Although the parallel analysis results suggested a single principal component indicating that all items are related to a single underlying construct, confirmatory factor analysis provided evidence for a two-construct model. On the other hand, it should be remembered that positive expressions on items UMUX01 and UMUX03 and the negative nature of items UMUX02 and UMUX04 determines this bi-dimensional structure.

## Recommendations

The following are our recommendations for future studies using the scales discussed in this paper:

- Further studies are needed to explore the concurrent validity and sensitivity of UMUX and its variants UMUX-LITE and UMUX-LITEr by way of evaluating different applications that have a similar purpose of use.
- To provide evidence for their convergent validity, such studies should involve usability testing to collect objective metrics such as task time, task success rate and error rate, and assess the correlation of these objective measures with UMUX and its variants.
- Using a within-subject design for the comparison of different software, these studies may also provide evidence for the sensitivity of the scales.
- UMUX and its variants should be investigated for the minimum sample size needed to achieve a significant difference in a comparison of software products in order to fully understand their sensitivity.

## Conclusion

This study provided further evidence on the reliability, validity, sensitivity, and latent variables of UMUX and its shorter variants.

Data suggests that UMUX and its variants are reliable scales with items that correlate significantly. Compared with previous studies, lower internal reliabilities observed in our study respond to critiques that suggest UMUX has a narrow sense of usability.

Regarding sensitivity, our data partially confirmed previous studies, providing evidence that UMUX distinguishes between different software. Although we observed a significant difference for UMUX and UMUX-LITE in the WP study, neither of the scales was able to distinguish the

software evaluated in the MM study at a significant level. On the other hand, SUS and CSUQ items provided a significant difference in the MM study. However, it should be noted that the differences between the scale scores for each software were quite low, which may indicate that the three software evaluated in the mind mapping study were quite similar in terms of usability.

There was evidence that UMUX and UMUX-LITE were highly sensitive to differences in users' level of experience with the evaluated software. For this reason, participants with a wide range of experience levels could cause results predominantly based on the level of experience rather than the properties of the evaluated software.

Neither scales' items were affected by the gender of the participants or whether they were native English speakers. The age of the participants did not affect scores on items of UMUX-LITE and UMUX either. Our data suggest that it may not be necessary to form homogenous groups in terms of age, gender, and native language while recruiting participants for a study that employs UMUX or UMUX LITE.

Results for the concurrent validity demonstrated that UMUX and UMUX-LITE are capable of measuring usability in conformance with the ISO 9241 definition, rather than in relation to concepts of information quality and interface quality. Although they correlate significantly with their predecessors, the focal point of the relation here is efficiency, ease of use, and satisfaction.

Although previous studies suggest using a regression formula to increase the correspondence of the UMUX-LITE score with SUS, this approach was futile with our data. On the contrary, compared to the adjusted UMUX-LITEr score, the unadjusted UMUX-LITE score had a higher correspondence with SUS.

Our results fall short in suggesting a construct structure for UMUX. UMUX provides information on the users' subjective attitude towards the system use, and this subjective attitude was observed as a single dimension in our data, through a parallel analysis approach. However, we also observed that items tended to load on two different factors depending on their negative and positive keying, and there is evidence that this factor structure model fits the data.

## Tips for Usability Practitioners

The following are tips for usability practitioners using the scales discussed in this paper:

- Researchers and practitioners could benefit from UMUX and UMUX-LITE as lightweight tools to measure the perceived usability of a software system. Although they are short and simple, there is evidence for their reliability, sensitivity, and concurrent validity.
- For a comparative study of highly similar systems, practitioners may consider using SUS and CSUQ in addition to UMUX because we observed that UMUX and UMUX-LITE may not be sensitive to differences between the software when the scale scores were not very different.
- While recruiting participants, researchers should consider that both UMUX and UMUX-LITE were found to be sensitive to participants' level of experience with the software.
- Participants' native language, age, and gender were not identified as sensitivity issues in either scale.
- UMUX can be evaluated as a single-dimension construct, but it should be considered that the negative/positive wording of items causes items to load on two different factors.
- Practitioners should employ some other tools if research requires further information for additional latent constructs related to usability such as information quality or interface quality.
- Practitioners should avoid a direct conclusion that the UMUX score indicates the user performance in actual use. UMUX has been evaluated for reliability, sensitivity, construct validity, and concurrent validity. Yet there is no evidence on its convergent validity with objective metrics such as task time, task success rate, and error rate.

## 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*(6), 574–594.

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International Journal of Human-Computer Interaction, 31*(8), 484–495.

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the System Usability Scale: A test of alternative measurement models. *Cognitive processing, 10*(3), 193–197.

Bosley, J. J. (2013). Creating a short usability metric for user experience (UMUX) scale. *Interacting with Computers, 25*(4), 317–319.

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

Cairns, P. (2013). A commentary on short questionnaires for assessing usability. *Interacting with Computers, 25*(4), 312–316.

Chin, J. P., Diehl, V. A., & Norman, K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *In Proceedings of CHI '88 Conference on Human Factors in Computing Systems* (pp. 213–218). New York, NY: ACM.

DeVellis, R. F. (2011). *Scale development: Theory and applications* (3rd Ed., Vol. 26).Los Angeles, CA: Sage Publications.

Eppler, M. J. (2006). A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing. *Information Visualization, 5*(3), 202–210.

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*(5), 323–327.

Finstad, K. (2013). Response to commentaries on "the usability metric for user experience." *Interacting with Computers, 25*(4), 327–330.

International Organizational for Standardization (ISO; 1998). ISO: Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on Usability, ISO9241. Genève, Switzerland: International Organization for Standardization.

Kirakowski, J. (1996). The software usability measurement inventory: Background and usage. In P. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 169–178). London, UK: Taylor & Francis.

Landauer, T. K. (1988). Research methods in human–computer interaction. In M. Helander (Ed.), *Handbook of human–computer interaction* (pp. 905–928). New York: Elsevier.

Lewis, J. R. (1992, October). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 36(16), 1259–1260. Sage Publications.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7*(1), 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*(3-4), 463–488.

Lewis, J. R. (2013). Critical review of "the usability metric for user experience." *Interacting with Computers, 25*(4), 320–324.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In *Human Centered Design* (pp. 94–103). Berlin Heidelberg: Springer.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013, April). UMUX-LITE: When there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099–2102). New York, NY: ACM.

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. (1975). Psychometric theory: 25 Years Ago and Now. *Educational Researcher, 4(1*0), 7–14, 19–21.

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

Sauro, J. (2011). *A practical guide to the system usability scale: Background, benchmarks & best practices*. Denver, CO: Measuring Usability LLC.

Sauro, J., & Lewis, J. R. (2012). Standardized usability questionnaires. In *Quantifying the user experience* (pp. 185-240). Boston, MA: Morgan Kaufmann.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323–338.

von Oertzen, T., Brandmaier, A. M., & Tsang, S. (2015). Structural equation modeling with Ωnyx. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(1), 148–161. doi: 10.1080/10705511.2014.935842

## About the Authors

**Mehmet Ilker Berkman**
Berkman teaches on CAD applications for visual design, user-centered design methods, and interactive media design in the Communication Design department at Bahcesehir University. Holding an MA on interactive media design and MSc on information technologies, he is a PhD candidate at the Bahcesehir University Computer Engineering Programme.

**Dilek Karahoca**
Karahoca is a social anthropologist. Holding a PhD in Computer Education and Instructional Technologies, she is interested in human-computer interaction, web based education systems, and blended learning methodologies. She has published several articles about use of information systems in health, tourism, and education.