Interacting with Computers 22 (2010) 323–327

Contents lists available at ScienceDirect

Interacting with Computers

journal homepage: www.elsevier.com/locate/intcom

# The Usability Metric for User Experience

Kraig Finstad

*Intel® Corporation, 2501 NW 229th Ave., M/S RA1-222, Hillsboro, OR 97124, United States*

ARTICLE INFO

*Article history:*
Received 21 September 2009
Accepted 6 April 2010
Available online 6 May 2010

*Keywords:*
Usability
User experience
Scale
Metric

ABSTRACT

The Usability Metric for User Experience (UMUX) is a four-item Likert scale used for the subjective assessment of an application's perceived usability. It is designed to provide results similar to those obtained with the 10-item System Usability Scale, and is organized around the ISO 9241-11 definition of usability. A pilot version was assembled from candidate items, which was then tested alongside the System Usability Scale during usability testing. It was shown that the two scales correlate well, are reliable, and both align on one underlying usability factor. In addition, the Usability Metric for User Experience is compact enough to serve as a usability module in a broader user experience metric.



## 1. Introduction

Measuring and tracking usability is an ongoing challenge for organizations that are concerned with improving user experience. A popular and cost-effective approach to usability measurement is the use of standardized surveys. When the Information Technology (IT) department at Intel® decided to standardize on a usability inventory, it selected the System Usability Scale (SUS). The SUS is a 10-item, five-point Likert scale with a weighted scoring range of 0–100 and which has been shown to be a reliable measure of usability. It is anchored with one as Strongly Disagree and five as Strongly Agree. According to Holyer (1993), it correlates at 0.86 with the 50-item Software Usability Measurement Inventory (Kirakowski et al., 1992). Tullis and Stetson (2004) found the SUS to outperform the Questionnaire for User Interface Satisfaction (Chin et al., 1988) and the Computer System Usability Questionnaire (Lewis, 1995) at assessing website usability. The SUS was adopted as a standard usability measure because of these performance characteristics, in addition to being free and relatively compact. It proved to be easy for project teams to understand, but several issues emerged. As IT at Intel® began to pursue a more comprehensive approach to user experience, the SUS was originally considered as a usability module for a more comprehensive index of user experience. This definition describes user experience as a lifecycle consisting of: Marketing and Brand Awareness, Acquisition and Installation, Product or Service Use, Product Support, and Removal/End of Life (Sward and Macarthur, 2007). However, it became apparent that simply adapting the SUS to work as a Product Use component was not feasible. Early trials with internal project teams showed that a 10-item Product Use module would be too large when other elements such as Product Support were factored in and required their own additional scales. The concept of user experience covers a lot of ground: any Product Use or usability component of a larger user experience index would have to be much more compact than 10 items. Also, in its original form, the SUS did not lend itself well to electronic distribution in a global environment due to non-native English speakers not understanding the word "cumbersome" in SUS Item 8 (Finstad, 2006), and it used a five-point Likert scale which has been shown to be inadequate in many cases. Diefenbach et al. (1993) found that seven-point scales outperformed five-point scales in reliability, accuracy, and ease of use, while Cox's (1980) review of Likert scales found the optimal number of alternatives to be seven. Finstad (in press) found that respondents were more likely to provide non-integer interpolations (e.g., saying "three and a half" instead of "three" or "four") in the five-point SUS than in a seven-point alternate version of the same instrument. These interpolations indicate a mismatch between the scale and a user's actual evaluation. From a more theoretical standpoint, the SUS items did not map well onto the concepts that comprise usability according to ISO 9241-11 (1998), namely effectiveness, efficiency, and satisfaction. These mappings are important because the SUS is not a diagnostic tool; it can indicate whether there is s a problem with a system's usability but not what those problems actually are. It is often used as a starting point in usability efforts, but an alignment with known usability factors can provide a stronger foundation for user experience efforts.

These issues with the SUS motivated a research program aimed at developing a replacement. The goal was to provide an inventory that was substantially shorter than the SUS and therefore appropriate as the usability component of a larger user experience index. An early attempt at item set reduction aimed to leverage a single

*E-mail address:* kraig.a.finstad@intel.com

0953-5438/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.intcom.2010.04.004

ease of use item from the SUS. A SUS survey with 43 responses was conducted on an enterprise portal product. It was found that Item 3 in the SUS, "I thought [the system] was easy to use" correlated with the final SUS score at $r = 0.89$, $p < 0.01$; the strongest correlation in the set of 10 items. This result was not surprising in light of recent findings. Sauro and Dumas (2009) have demonstrated the utility of a general ease of use Likert item, and have also shown a promising alternative in the Subjective Mental Effort Questionnaire (SMEQ). Tedesco and Tullis (2006) found that a single "Overall this task was: Very Difficult. . .Very Easy" Likert item correlated significantly with usability test performance. This direction motivated further analysis of SUS surveys and SUS Item 3 with other systems, but no consistent pattern emerged. In some cases SUS Item 3 correlated most strongly with the final SUS score, and in others it did not. The idea of reducing an instrument to one general ease of use item was abandoned. Instead, a new direction was taken – the development of a concise scale that would more closely conform to the ISO 9241-11 (1998) definition of usability, would minimize bias and language issues, and would still perform as well as the baseline it was intended to replace. In this case the baseline was the updated, internationally-appropriate SUS (with "cumbersome" clarified as "awkward") and the performance goal of total SUS score to the total score of the new scale was set at a correlation of 0.80 or better. The resulting instrument is the Usability Metric for User Experience (UMUX), and this paper outlines the research and development of this usability component of a more general user experience measurement model.

## 2. Pilot study

A pilot study was developed to explore these possibilities. The end goal of the pilot study was the determination of how candidate Likert items would fare in an analysis of actual responses to items.

### 2.1. Method

#### 2.1.1. Participants

A total of 42 Intel® employees were recruited as part of a larger worldwide usability test. As a control for cultural and language factors in both the usability task and the candidate Likert items, participants were recruited worldwide. Users from the United States, Germany, Ireland, the Netherlands, China, the Philippines, Malaysia, and Israel participated in this study.

#### 2.1.2. Materials

A pool of candidate Likert items was developed that were related to the ISO 9241-11 (1998) definition of usability. A total of 12 such items were developed, four each for effectiveness, efficiency, and satisfaction. Some were intentionally generic, while others were behavior-based (e.g., "I don't make many errors with this system") or emotion-based (e.g., "I would prefer to use something other than this system"). These candidate items used a five-point scale so they could be used alongside the SUS in an actual post-deployment usability survey. Also like the SUS, they used an alternating positive/negative keying to control for acquiescence bias. These 12 candidate items and their usability factors are listed in Table 1.

#### 2.1.3. Design and procedure

Participants first engaged in a usability test of an enterprise software prototype involving the selection of contract workers and adding them to a database. After completing the usability test, participants received a modified version of the SUS. The first three items were candidate items, followed by the SUS, which was then followed by three more candidate items. This format presented the SUS as an intact instrument in order to achieve a valid final score for analysis. Each participant therefore responded to six candidate items, two per usability component (effectiveness, efficiency, and satisfaction), in addition to the SUS. This allowed a direct per-participant comparison of candidate item responses with a final SUS score. Presentation of candidate items was counterbalanced across participants. Response to the Likert items was verbal, with the entire items read aloud to help ensure comprehension of the scale. The facilitator recorded responses manually. After completion of the composite survey, participants were thanked for their time, debriefed, and excused.

### 2.2. Results

#### 2.2.1. Item correlations

The odd items in the SUS were scored as [score – 1], and the even items were scored as [5 – score]. This aligned all scores in one direction, removing the positive/negative keying of the language in the instrument. It also allowed zeroes at the bottom of the range. The ten rescored items were summed and multiplied by 2.5, providing a range of 0–100 (Brooke, 1996). The critical measure of this study was the correlation of UMUX candidate items (scored similarly to the SUS) with the final SUS score. A high correlation coefficient indicated that the candidate item was in line with the total SUS score, regardless of direction. That is, a good candidate item would correlate highly with the SUS regardless of whether the SUS itself was indicating good or poor usability. This is a different approach from that used in developing the original SUS, which selected candidates based on their tendencies toward extreme (non-neutral) responses (Brooke, 1996). The UMUX is intended to match the performance of the SUS, so alignment with existing measures is more important.

As the UMUX was being designed to reflect the ISO 9241-11 (1998) definition of usability with as few items as possible, the highest-correlating candidate items for each usability component were chosen for further study. Table 2 below summarizes these results.

All the correlations in this table were negative due to the negative keying of the candidates; for instance, if the application was usable then the participants disagreed on the item. The more general items with language like "I am satisfied. . ." tended to correlate poorly. As a point of comparison, the correlations of the items in the SUS to the SUS score itself varied from $r = 0.36$ to $r = 0.78$.

No participants required assistance with the terminology or phrasing of the UMUX candidate items. This was taken as evidence

**Table 1**
Candidate items used (pilot study).

| Usability component | Candidate item |
|---|---|
| Efficiency | [This system] saves me time. |
| | I tend to make a lot of mistakes with [this system]. |
| | I don't make many errors with [this system]. |
| | I have to spend a lot of time correcting things with [this system]. |
| Effectiveness | [This system] allows me to accomplish my tasks. |
| | I think I would need a system with more features for my tasks. |
| | I would not need to supplement [this system] with an additional one. |
| | [This system's] capabilities would not meet my requirements. |
| Satisfaction | I am satisfied with [this system]. |
| | I would prefer to use something other than [this system]. |
| | Given a choice, I would choose [this system] over others. |
| | Using [this system] was a frustrating experience. |

Note: Bracketed text is custom-replaced by relevant system.

**Table 2**
Items having highest correlation with overall SUS score (pilot study).

| Usability component | Candidate UMUX Item | *r* |
|---|---|---|
| Efficiency | I have to spend a lot of time correcting things with [this system]. | −0.48* |
| Effectiveness | [This system's] capabilities would not meet my requirements. | −0.50* |
| Satisfaction | Using [this system] was a frustrating experience. | −0.76* |

Note: Bracketed text is custom-replaced by relevant system.
* $p < 0.05$.

that the items were appropriate for an international English-speaking audience.

*2.2.2. Analysis of preliminary instrument*

These results motivated an analysis to determine how the best candidate items would perform if they comprised an actual instrument that yielded a SUS-like usability score. If candidate item data from the pilot study could produce a result comparable to the SUS, those items would be subjected to a wider scale validation study with new participants. The preliminary UMUX was comprised of the three highest-correlating candidate items from each ISO (1998) usability factor shown in Table 2, plus the overall ease of use from the SUS ("I thought the system was easy to use"), which had shown earlier to be promising as a general question with $r = 0.89$, $p < 0.01$ (see Section 1).

Data for the analysis consisted of the 21 response sets from the pilot study that included the candidate items. These four candidate items from a five-point scale were used with a 2.5 multiplier, providing a score range of 0–40 (compared to 0–100 for the SUS). The preliminary UMUX attained a mean score of 24 out of 40, and the SUS for the same participants attained a mean score of 60 out of 100. Both of these scores were 60% of their respective maximums. The preliminary UMUX correlated with the SUS at $r = 0.81$, $p < 01$.

*2.3. Discussion*

The pilot study identified the three most promising candidates to be included in a measurement instrument along with an additional ease of use item. The results for the candidate items were in line with correlations achieved by the SUS itself. When combined into a preliminary user experience inventory, the four candidate items met the research program's goal of a correlation higher than 0.80 with the SUS.

## 3. Survey study

The next step was to design an experiment directly comparing the SUS with the new UMUX instrument.

*3.1. Method*

*3.1.1. Participants*

Participants consisted of two groups of users of enterprise software at Intel®. System 1 was a contract worker enterprise application that had been rated as having poor usability, and System 2 was an audio-conferencing application that had been rated as having good usability. Valid responses received from survey requests resulted in System 1 with $n = 273$ and System 2 with $n = 285$.

*3.1.2. Materials*

Some minor changes were made to the candidate items to build the experimental UMUX to better balance the positive/negative keying and to clear up some potential confusion with the Efficiency item. For comparison with the original items in Table 2, see the completed UMUX in Table 3.

**Table 3**
Usability components and scale items (survey study).

| Usability component | Candidate UMUX item |
|---|---|
| Effectiveness | [This system's] capabilities meet my requirements. |
| Satisfaction | Using [this system] is a frustrating experience. |
| Overall | [This system] is easy to use. |
| Efficiency | I have to spend too much time correcting things with [this system]. |

Note: Bracketed text is custom-replaced by relevant system.

The UMUX used in this survey study was a seven-point Likert scale, anchored with one as Strongly Disagree and seven as Strongly Agree. Like the SUS, all other response options were numbered but otherwise unlabeled. This move to a seven-point scale gave the UMUX an initial range of 0–60, after applying the 2.5 multiplier from the SUS. These UMUX scores could be presented as a percentage of the maximum (60) to provide a final range comparable to that found in the SUS (0–100). The SUS was also modified as per Finstad (2006), clarifying "cumbersome" as "awkward".

*3.1.3. Design and procedure*

This study used a between-subjects design, with participants using one of two systems (having poor usability or good usability) and then responding to both the UMUX and the SUS. Presentation of the instruments was counterbalanced so that half the participants responded to the UMUX first, and the other half responded to the SUS first. These instruments were administered electronically through a combination of email invitation and online survey tool.

*3.2. Results*

*3.2.1. Principal components*

A common first step in validating instruments is through principal components analysis (Tabachnik and Fidell, 1989). The results from the initial principal component extraction are shown below in Table 4.

The strength of the first principal component led to the conclusion that UMUX items were aligning along one usability component. This perspective is supported by the scree plot of the components, shown in Fig. 1 below.

Tabachnik and Fidell (1989) recommend the point where the scree plot line changes direction as a determinant of the number of components; this plot's direction drops off dramatically after the first component. This is strong evidence for the scale measuring one "usability" component. Because no secondary components

**Table 4**
Principal components (survey study).

| Principal component | Eigenvalue | Percent of variance explained |
|---|---|---|
| 1 | 3.37 | 84.37 |
| 2 | 0.31 | 7.83 |
| 3 | 0.20 | 4.88 |
| 4 | 0.12 | 2.92 |

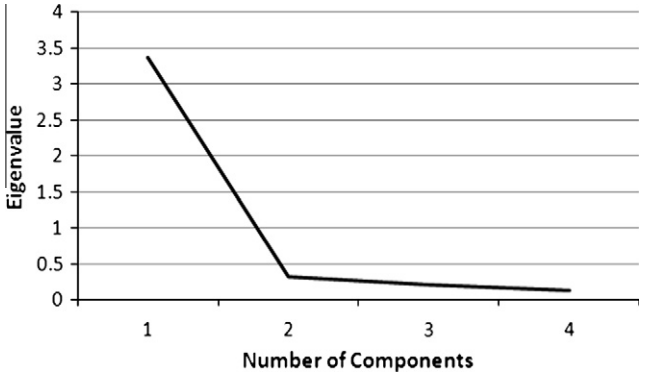


**Fig. 1.** Scree plot of principal components (survey study).

emerged from the analysis, no attempts at further extractions or rotations were performed. The SUS provided a similar one-component extraction, with no additional elements emerging. For a more thorough treatment of factoring in the SUS, see Lewis and Sauro (2009), who found evidence that the SUS may be comprised of two factors (usability and learnability). The conclusion from this analysis is that both instruments were unidimensional and align on just one component (usability) rather than several.

*3.2.2. Reliability*

Instruments need to measure an underlying construct consistently. At the early stages of a metric's development, one way to establish this is through reliability estimation. Cronbach's alpha is a correlation coefficient that indicates how well a factor is being measured. The rule of thumb for Cronbach's alpha is that a coefficient of higher than an absolute value of 0.70 indicates a high degree of internal reliability. Instruments farther along in their development are subjected to more longitudinal reliability measures. The Cronbach's alpha for both instruments indicated high reliability: 0.94 for the UMUX and 0.97 for the SUS. Therefore, both instruments were reliable.

*3.2.3. Validity and sensitivity*

The overall correlation of UMUX with the SUS, across both system conditions, was $r = 0.96$, $p < 0.001$. These results exceed the goal criterion of $r > 0.80$, providing evidence of validity. *T*-tests demonstrated that System 2 was more usable than System 1, $t(533) = -39.04$, $r = 0.86$, $p < 0.01$ for UMUX, $t(556) = -44.47$, $r = 0.89$, $p < 0.01$ for SUS, thereby providing evidence for sensitivity. The breakdown of usability inventory scores and correlations is shown in Table 5.

*3.2.4. Item correlations*

After the UMUX had been developed and finalized, the performance of its individual items was examined in two applied situations. All final UMUX items were analyzed for their contribution to the overall UMUX score, both as a post-usability test questionnaire ($n = 45$) and in the first seven internal usability projects completed with the new scale as a standard instrument ($n = 272$). The results shown in Table 6 demonstrate significant item-total correlations. It was therefore concluded that all UMUX items were valid contributors to the overall score.

**Table 5**
Means, standard deviations, and correlation (survey study).

| System | UMUX (0–100) | SUS (0–100) | *r* |
|---|---|---|---|
| System 1 | 27.66 (20.54) | 28.77 (18.19) | 0.84* |
| System 2 | 87.91 (15.98) | 88.39 (13.18) | 0.81* |

* $p < 0.001$.

**Table 6**
Correlations of UMUX items with overall score (survey study).

| Scale item | Post-test *r* | Surveys *r* |
|---|---|---|
| 1. [This system's] capabilities meet my requirements. | 0.78* | 0.85* |
| 2. Using [this system] is a frustrating experience. | −0.76* | −0.89* |
| 3. [This system] is easy to use. | 0.76* | 0.87* |
| 4. I have to spend too much time correcting things with [this system]. | −0.69* | −0.81* |

* $p < 0.05$.

## 4. Discussion

### 4.1. Implementation

The UMUX can be administered electronically as a survey, or as a follow-up in usability testing. It is simple to administer, as it requires no branching or reordering of items. The UMUX is implemented as shown below, where bracketed text is custom-replaced by the relevant system.

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

### 4.2. Analysis

Once data are collected, they need to be properly recoded, with a method that borrows from the SUS. Odd items are scored as [score – 1], and even items are scored as [7 – score]. As with the SUS, this removes the positive/negative keying of the items and allows a minimum score of zero. Each individual UMUX item has a range of 0 – 6 after recoding, giving the entire four-item scale a preliminary maximum of 24. To achieve parity with the 0–100 range provided by the SUS, a participant's UMUX score is the sum of the four items divided by 24, and then multiplied by 100. This calculation replaces the earlier methodology of weighting items by a 2.5 multiplier. These scores across participants are then averaged to find a mean UMUX score. It is this mean score and its confidence interval that become the application's UMUX metrics for a system's usability tracking and goal-setting.

### 4.3. Limitations

The UMUX, like the SUS, provides a subjective evaluation of a system's usability. Its scoring has yet to be compared to objective

metrics, such as error rates and task timings, in a full experiment. Additionally, as it is currently the first module in a planned series of user experience measures, it only measures usability.

As the UMUX consists of only four Likert items, it has fewer total data points available to respondents than the SUS, although the move to a seven-point scale does provide some mitigation. It has four seven-point items for a total of 28 data points, while the SUS has ten five-point items for a total of 50 data points. By comparison, a singular ease of use item like that used in Tedesco and Tullis (2006) may have only five data points. Reducing the total information capacity of a survey effort can result in a less sensitive measure. Once validity and reliability are established, there is still a potential risk of application beyond the metric's scope. For example, a simple ease of use item may do an exemplary job of measuring ease of use, but a particular user experience professional needs to determine whether that information is sufficient as a usability metric.

### 4.4. Conclusion

It can be concluded that the Usability Metric for User Experience is a reliable, valid, and sensitive alternative to the System Usability Scale. It correlates with the SUS at a rate higher than 0.80, its items align on one usability factor, and it is fully capable as a standalone subjective usability metric. It is also aligned to a fundamental learning for the user experience community: in order to measure user experience effectively, its components need to be measured efficiently. The compact size of the UMUX is suited to a more fully realized measurement model of user experience. Such a model would go beyond Product Use, and would include other product lifecycle stages such as Brand Awareness and Installation (Sward and Macarthur, 2007). Sward (personal communication, August 5, 2009) indicates significant progress in this area. The UMUX is well-positioned as a foundation for developing future instruments, with the ultimate goal of metrics that can target any of a product's user experience aspects in a way that is concise and cross-validated.

## Acknowledgements

Thanks to David Sward of Symantec™ for his work on the user experience lifecycle model, Pete Lockhart of Intel® for item language suggestions, Charles Lambdin of Intel® for statistical assistance, and Linda Wooding of Intel® for management support in implementing this research.

## References

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

Chin, J.P., Diehl, V.A., Norman, K., 1988. Development of an instrument measuring user satisfaction of the human–computer interface. In: Proceedings of ACM CHI '88, Washington, DC, pp. 213–218.

Cox III, E.P., 1980. The optimal number of response alternatives for a scale: a review. Journal of Marketing Research 17, 407–422.

Diefenbach, M.A., Weinstein, N.D., O'Reilly, J., 1993. Scales for assessing perceptions of health hazard susceptibility. Health Education Research 8, 181–192.

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

Finstad, K., in press. Response interpolation and scale sensitivity: evidence against five-point scales. Journal of Usability Studies.

Holyer, A., 1993. Methods for Evaluating user Interfaces. Cognitive Science Research Paper No. 301. School of Cognitive and Computing Sciences, University of Sussex.

ISO 9241-11 (1998). Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 11: Guidance on Usability.

Kirakowski, J., Porteous, M., Corbett, M., 1992. How to use the software usability measurement inventory: the users' view of software quality. In: Proceedings European Conference on Software Quality, Madrid.

Lewis, J., 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., Sauro, J., 2009. The factor structure of the System Usability Scale. In: Proceedings of the Human–Computer Interaction International Conference (HCII 2009), San Diego CA, USA.

Sauro, J., Dumas, J.S., 2009. Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems. Boston.

Sward, D., Macarthur, G., 2007. Making user experience design a business strategy. Towards a UX Manifesto; SIGCHI Workshop. Lancaster, UK, September 3–4.

Tabachnik, B.G., Fidell, L.S., 1989. Using Multivariate Statistics, 2nd ed. Harper Collins, New York.

Tedesco, D., Tullis, T., 2006. A comparison of methods for eliciting post-task subjective ratings in usability testing. Usability Professionals Association (UPA) 2006, 1–9.

Tullis, T.S., Stetson, J.N., 2004. A comparison of questionnaires for assessing website usability. In: Proceedings of UPA 2004, June 7–11.