Comments on the reliability and validity of UMUX and UMUX-LITE short scales

Costin Pribeanu
National Institute for Research and Development in Informatics - ICI Bucharest
Blvd. Maresal Averescu, nr.8-10, Bucharest, Romania
pribeanu@ici.ro

ABSTRACT
Recent work on usability and user experience shows several concerns on the validity of evaluation instruments. There is a debate on the use of standardized scales versus short scales, such as UMUX and UMUX-LITE or even a single-item measure of usability. Nevertheless, there are relatively few papers reporting the testing of these scales together for reliability and validity. This paper aims at discussing the UMUX and UMUX-LITE scales that have been tested in the context of Facebook use by university students. From a theoretical point of view, both scales are questionable. From an empirical point of view, the testing results confirmed a lack of unidimensionality as well as a poor reliability and convergent validity of these scales.

Author Keywords
Usability scales, UMUX, UMUX-LITE, factor analysis, validity, Facebook use.

INTRODUCTION
Recent work on usability and user experience shows several concerns on the reliability and validity of the scales used to measure the perceived usability. There is currently a debate on the use of standardized scales versus short scales, such as UMUX [6] and UMUX-LITE [14] or even a single-item measure of usability [13]. There are many pros and cons as regards the reliability and validity as well as the practical benefits. Nevertheless, there are relatively few papers reporting the testing of these scales for reliability and validity.

This paper aims at discussing the UMUX and UMUX-LITE scales from both a theoretical and an empirical point of view. Additionally, the perceived ease of use (PEU) is analyzed that is a widely used concept in the context of technology acceptance [5]. Since PEU is a short scale taping on several key usability aspects, it could be a better alternative than UMUX and UMUX-LITE.

Theoretically, the analysis is following the scale development recommendations. Empirically, the analysis is focused on the scale testing that has been carried on by using two samples collected during a larger study on Facebook use by university students [9]. The first sample is from the pilot study and is used to assess the UMUX and UMUX-LITE scales. The second sample is from a subsequent study using a revised evaluation instrument and is used to assess the UMUX-LITE and PEU scales.

The rest of this paper is organized as follows. The following section briefly presents recent approaches in the area of scale development with an emphasis on the scale development process and the existing usability scales. In section 3, the analysis of UMUX and UMUX-LITE is presented based on two empirical studies. The same assessment criteria are used to analyze the PEOU scale. The paper ends with conclusion and future research directions.

RELATED WORK

Scale development
The interest in developing rating scales increased after the adoption of the ISO 9241-11 standard that included satisfaction as a key usability aspect. As Lindgaard & Kirakowski [15] pointed out, the landscape of scale development in HCI shows many usability scales, many approaches, as well as many opinions as regards the scale reliability and validity.

When analyzing the reliability and validity of the usability scales, two aspects are usually neglected: the theoretical meaning and the multidimensional nature of the usability concept. Psychometrics is not a favorite area of expertise in HCI [15]. As such, the scale development process is not well understood as regards both the ordering of steps to be carried on and the suitable techniques that should be used in each step.

Long time ago, Gerbing & Anderson [7] outlined an updated paradigm for scale development that includes a confirmatory factor analysis (CFA) to assess the scale unidimensionality. They underlined that, only after achieving an acceptable unidimensionality level, the reliability could be assessed. This precondition is usually ignored in the existing papers reporting the development and testing of usability scales. The authors rely on the traditional approach that only includes the Cronbach's alpha coefficient, the item-to-total correlations, and the exploratory factor analysis (EFA).

More recently, MacKenzie et al. [16] emphasized the importance of the conceptualization as a first step in the scale development process. They noticed that an adequate conceptualization is difficult and requires a review of the literature on the meaning of related constructs, aspects these constructs refer to, dimensionality and preliminary research with domain experts or practitioners. Another important issue is to specify if the construct is measured reflectively or formatively.

Usability in the ISO standards
ISO 9241-11 standard defined usability as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use [11]. The ISO standard 9126-1 defined usability as the capability of a

software product to be understood, learned, used, and attractive to the user, when used under specified conditions [10]. Later on, both definitions were integrated in the ISO standard 25010 [12], under two key terms:

- Quality in use: the degree to which a product used by specific users meets their needs to achieve specific goals with effectiveness, efficiency, safety, and satisfaction in specific contexts of use.
- Usability: the degree to which a software product is able to satisfy the following needs when used under specified conditions: appropriateness recognizability, learnability, operability, error protection, user interface aesthetics, and accessibility.

As pointed out by Bevan et al. [1], the quality in use defines usability as a high level concept, focusing on the outcomes of the interaction rather than on the characteristics that make a product usable. Unfortunately, this distinction is rarely made in the mainstream of HCI literature.

Measurement scales for the perceived usability

A well-known usability scale is System Usability Scale (SUS) that has been developed by Brooke as a simple, "quick and dirty" scale [3]. SUS has been widely used and is considered an industrial standard [2, 14].

Several authors noticed that SUS lacks unidimensionality. For example, Borsci et al. [2] found that a learnability dimension of SUS might emerge under certain conditions (when administrated to experienced users). SUS has been also criticized for using both positive and negative wording, since this may lead to mistakes (made by respondents) and mis-coding (made by researchers) [17].

More recently, Finstad proposed the UMUX (Usability Metrics for User Experience) as a shorter alternative to SUS. UMUX have been criticized for dimensionality [4, 14] and for using negative wording [14, 17]. Lewis et al. [14] found that UMUX has a bi-factorial structure with positive tones aligning with one factor and negative tones with the second factor.

An even shorter scale that is based on UMUX has been proposed by Lewis et al. [14]. UMUX-LITE was intended as a very quick, two-item scale, that uses the first and the third item from UMUX. The authors found that UMUX-LITE is unidimensional and has acceptable reliability. However, they recommended using this scale with caution until it will be validated across a wider variety of systems.

A well-known scale measuring the perceived ease of use (PEU) has been developed and tested in the context of technology acceptance studies. The technology acceptance model (TAM) has been developed by Davis et al. [5], in order to explain and predict the technology acceptance on a large variety of technologies. Although it is a short scale with a widely recognized psychometric quality, PEU has been rarely used in the HCI studies.

ANALYSIS OF UMUX AND UMUX-LITE

Method

The analysis follows the recommendations in the literature for scale development and assessment of dimensionality, reliability, and validity [7, 8, 16]. The first step is to analyze the conceptualization based on the definition of concepts in the literature. Then, the dimensionality is assessed via exploratory and confirmatory factor analysis. After demonstrating that the construct is unidimensional, the reliability could be analyzed checking the magnitude of Cronbach's alpha and the item-to-total correlations. The convergent validity is assessed by the examination of the composite reliability (CR), and average variance extracted (AVE).

Empirical studies

UMUX and UMUX-LITE have been tested in a larger study on the use of Facebook (FB) by university students. Two samples collected during these studies are used for the analysis of the psychometric quality of UMUX and UMUX-LITE. The respondents were asked to answer questions related to demographics, enrollment, FB usage (the size of their FB network, frequency of use, minutes per day), and to evaluate items on a 7-point Likert scale.

The first sample has been collected in 2014 and consists of 152 students (110 female, 42 male) from two universities in Lithuania. The negatively worded items in Table 1were recoded. The first and third items in Table 1 represent the UMUX-LITE scale.

Item	Statement	M	SD
U1	FB's capabilities meet my requirements	4.02	1.38
U2	Using FB is a frustrating experience	3.44	1.64
U3	FB is easy to use	4.91	1.15
U4	I have to spend too much time correcting things with FB	4.12	1.63

Table 1. Descriptives for UMUX scale (N=152).

The second sample has been collected in 2015 and consists of 414 students (258 female, 156 male) from a Romanian university. Since the testing results from the first study revealed poor psychometric properties of UMUX, the scale has been removed from the evaluation instrument and replaced with PEU. However, the first item has been preserved in order to test again UMUX-LITE.

Item	Statement	M	SD
U1	FB's capabilities meet my requirements	4.36	1.51
PEU1	It is easy to learn how to use FB	6.10	1.27
PEU2 / U3	FB is easy to use	6.21	1.17
PEU3	My interaction with FB is clear and understandable	5.69	1.38

Table 2. Descriptives for UMUX-LITE and PEU (N=414).

The PEU scale has been developed by adapting items from the existing scales. The three items tap on three usability attributes: understandability, learnability, and operability. As it could be observed, the item U3 in Table 1 is identical with the item PEU2 in Table 2.

Conceptualization

In most studies discussing UMUX and UMUX-LITE the reliability and validity is limited at dimensionality which is assessed with exploratory factor analysis techniques, reliability, and correlation with other usability scales. The main shortcoming of UMUX (and, consequently, of UMUX-LITE) is the poor conceptualization which is due, on one side, to the underlying usability definition and, on another side, to the misunderstanding of the nature of the measurement model.

Both scales take the roots from the quality in use concept and try to measure the user's subjective satisfaction. Nevertheless, the measured variable is not the satisfaction (there is no item saying "I am satisfied with…". Rather, the operationalization of the construct is done with a mix of items measuring various aspects of user experience and usability. As such, the conceptualization is ambiguous and is not clear what UMUX is actually measuring.

The first item is the most ambiguous and undermines the conceptualization of both UMUX and UMUX-LITE. The fit between the user's needs and requirements could refer to anything: ease of use, aesthetics, flexibility, robustness, safety, usefulness, enjoyment, etc.

The second shortcoming is the lack of a clear definition of the nature of the measurement model. The measurement model describes the relationship between a construct and its measures [7, 16]. According to the direction of the causal relationship, the constructs could be reflective (from the construct to its measures) or formative (from measures to construct). It is also possible to define multidimensional constructs where the dimensions are specified as first order constructs.

Failure to adequately specify the measurement model leads to a poor operationalization and a lack of validity. The point is that the conceptualization and validation recommendations are different for reflective and formative measurement models. Unidimensionality and inter-item correlation are required for reflective constructs, since all items are supposed to measure the same thing. For a more detailed discussion on the scale development of reflective and formative constructs, see MacKenzie et al. [16].

Therefore, neither UMUX nor UMUX-LITE could be adequately assessed as measurement scales, since both of them suffer from a lack of clear definition of the construct domain. As regards the measurement model, although is not specified, it is assumed to be reflective according to the assessment techniques used by the authors.

Dimensionality

The dimensionality of UMUX and UMUX-LITE has been analyzed within the two empirical studies. The first study enabled testing of UMUX and UMUX-LITE.

The principal component analysis with Varimax rotation for UMUX resulted in two factors explaining 34.56%, respectively 26.69% of the variance. The same analysis for UMUX-LITE resulted in one factor explaining 60.06 % of the variance.

A confirmatory factor analysis has been then carried on. The results revealed the lack of dimensionality for both constructs. The loadings of the underlying construct on its measures (α regression coefficients) is below the cutoff value of 0.60 [8]. The results are presented in Table 3.

Item	UMUX			UMUX-LITE	
	EFA		CFA (α)	EFA	CFA (α)
	1	2		1	
U1	.707	-.364	0.66	.775	0.65
U2	.700	.332	-0.38		
U3	.609		0.31	.775	0.34
U4		.919	0.32		

Table 3. Dimensionality of UMUX and UMUX-LITE (N=152).

The second study enabled the analysis of dimensionality of UMUX-LITE and PEU. The principal component analysis with Varimax rotation for UMUX-LITE resulted in one factor explaining 68% of the variance. The same analysis for PEU resulted in one factor explaining 80.61% of the variance.

The confirmatory factor analysis for UMUX-LITE revealed a low item loading of U1 (α=0.49). The same analysis for PEU confirmed its unidimensionality (item loadings: 0.86, 0.92, and 0.73).

The results of the two empirical studies demonstrate the lack of dimensionality for the UMUX and UMUX-LITE short scales as well as the limitations of the exploratory factor analysis for testing the dimensionality.

Reliability

The Cronbach's alpha was unacceptable low in the first study: 0.213 for UMUX and 0.331 for UMUX-LITE. The item-to-total correlations were in the range of 0.05-0.23for UMUX, respectively 0.20 for UMUX-LITE.

In the second study, the Cronbach's alpha for UMUX-LITE was low (0.517) and the item-to-total correlation also low (0.36). Cronbach's alpha for PEU was 0.874 and the item-to-total correlation in the range of 0.68-0.82.

Convergent validity

Convergent validity refers to the degree to which the measures of a construct that are supposed to be related, are in fact related.

In the first study, the low item loadings make no sense to test the convergent validity of UMUX. For UMUX-LITE, the composite reliability of 0.412 and the average variance extracted of 0.279 demonstrate the lack of convergent validity. The second study confirmed the poor convergent validity of UMUX-LITE (CR=0.552, AVE=0.390). The

convergent validity for PEU was very good (CR=0.885, AVE=0.722).

Interpretation of scores

The final step in scale development is to provide the prospective researchers with some recommendations for the interpretation of scores. Since this step is beyond the purpose of this study, it will not be discussed. However, is important to note that the correlation of a scale under consideration with other existing scales does not ensure the scale validity. It is expected that two scales pointing to similar usability aspects correlate. The problem is that if the scale under consideration and the reference scale are not unidimensional, then a comparison leads to ambiguous if not erroneous conclusions.

DISCUSSION AND CONCLUSION

In the area of HCI, several misconceptions exist as regards the scale development and the validity criteria. The two short scales analyzed in this paper suffer from an ambiguous definition of the construct. An adequate conceptualization should include the specification of the dimensionality and the nature of the construct (reflectively vs. formatively measured).

It seems that the relationship between the definition of the target construct and the criteria for its assessment are not well understood: if a usability scale is not unidimensional, then more than one thing (e.g. usability) is measured. In other words, what is actually measured is not what has been supposed to be measured.

A problem with the conceptualization of existing short scales in HCI, such as SUS, UMUX, and UMUX-LITE is the confusion between usability and quality in use. The quality in use is a multidimensional construct since it taps on different concepts. Another problem is the overlapping between two HCI concepts: usability and user experience.

It is advisable to keep apart the scales measuring the pragmatic and hedonic aspects (each scale should undergo a separate validation procedure). For example, the perceived ease of use refers to pragmatic aspects, while the perceived enjoyment refers to hedonic aspects. Both scales have been widely used and validated in technology acceptance studies.

The empirical studies confirmed the recommendation of Gerbing and Anderson to use confirmatory factor analysis to assess the dimensionality [7]. The exploratory factor analysis is clearly not enough.

The two empirical studies show that both UMUX and its shorter version, UMUX-LITE, suffer from poor reliability, lack of unidimensionality, and poor convergent validity. The correlation with other usability scales, which is frequently mentioned as an argument for reliability, is a poor surrogate when the candidate scale and the reference scale does not measure the same thing.

This paper does not deny the practical value of the short questionnaires which are less expensive and could provide a useful feedback for the developers. However, these should not be referred as usability or UX scales.

As it was shown, the PEU scale is unidimensional and reliable. PEU provides with a useful feedback on some key aspects of usability and could be combined with other short scales pointing to other usability / user experience aspects. This approach enables a step-by-step development of valid and reliable evaluation instruments and a flexible choice of scales, according to the objectives of the evaluation.

ACKNOWLEDGMENTS

This work was supported by the Romanian grant financed by ANCS under COGNOTIC 1609 0101 / 2016.

REFERENCES

1. Bevan, N., Carter, J. and Harker, S. ISO 2041-11 revised: What have we learned about usability since 1998? *Human-Computer Interaction Design and Evaluation, Proc. HCI International Conference, LNCS 9169*, Springer, (2015), 143-151.
2. Borsci, S., Federici, S., Bacci, S., Gnaldi, M. and Bartolucci, F. 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), (2015), 484-495.
3. Brooke, J. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189 (194), (1996), 4-7.
4. Cairns, P. A Commentary on short questionnaires for assessing usability. *Interacting with Computers* 25(4), (2013), 312-316.
5. Davis, F.D., Bagozzi, R.P. and Warshaw, P.R. User acceptance of computer technology: A comparison of two theoretical models, *Management Science*, 35 (8), (1989), 982-1003.
6. Finstad, K. The usability metric for user experience. *Interacting with Computers* 22(5), (2010), 323-327.
7. Gerbing, D.W. and Anderson, J.C. An updated paradigm for scale development incorporating unidimensionality and its assesment. *Journal of Marketing Research* 25(2), (1988), 86-192.
8. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. and Tatham, R.L. *Multivariate Data Analysis*. 6th ed., Prentice Hall, 2006.
9. Iordache, D.D., Pribeanu, C., Lamanauskas, V. and Raguliene, L. Usage of Facebook by university students in Romania and Lithuania: a comparative study. *Informatica Economica* 19(1), (2015), 46-54.
10. ISO/IEC 9126-1, Software Engineering –Product quality. Part I: Quality Model, 2001.
11. ISO 9241-11, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11 *Guidance on usability*, 1998.
12. ISO/IEC 25010, Systems and software engineering – Systems and software product Quality Requirements and Evaluation (SQuaRE) – System and software quality models, 2011.

13. Konradt, U., Wandke, H., Balazs, B. and Christophersen, T. Usability in online shops: scale construction, validation and the influence on the buyers' intention and decision. *Behaviour and Information Technology* 22, (2003), 165–174.

14. Lewis, J. R., Utesch, B. S. and Maher, D. E. UMUX-LITE: when there's no time for the SUS. Proceedings of *CHI 2013*, ACM, (2013), 2099-2102.

15. Lindgaard, G. and Kirakowski, J. Introduction to the Special Issue: The tricky landscape of developing rating scales in HCI. *Interacting with Computers* 25 (4), (2013), 271-277.

16. MacKenzie, S. B., Podsakoff, P. M., and Podsakoff, N. P. Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS quarterly*, 35(2), (2011), 293-334.

17. Sauro, J. and Lewis, J. When designing usability questionnaires, does it hurt to be positive? Proceedings of the *CHI 2011*, ACM, (2011), 2215-2223.