Cogn Process (2009) 10:193–197
DOI 10.1007/s10339-009-0268-9

LETTER TO THE EDITOR

On the dimensionality of the System Usability Scale: a test of alternative measurement models

Simone Borsci · Stefano Federici · Marco Lauriola

Received: 30 May 2009 / Revised: 12 June 2009 / Accepted: 15 June 2009 / Published online: 30 June 2009
© Marta Olivetti Belardinelli and Springer-Verlag 2009

Abstract The System Usability Scale (SUS), developed by Brooke (Usability evaluation in industry, Taylor & Francis, London, pp 189–194, 1996), had a great success among usability practitioners since it is a quick and easy to use measure for collecting users' usability evaluation of a system. Recently, Lewis and Sauro (Proceedings of the human computer interaction international conference (HCII 2009), San Diego CA, USA, 2009) have proposed a two-factor structure—Usability (8 items) and Learnability (2 items)—suggesting that practitioners might take advantage of these new factors to extract additional information from SUS data. In order to verify the dimensionality in the SUS' two-component structure, we estimated the parameters and tested with a structural equation model the SUS structure on a sample of 196 university users. Our data indicated that both the unidimensional model and the two-factor model with uncorrelated factors proposed by Lewis and Sauro (Proceedings of the human computer interaction international conference (HCII 2009), San Diego CA, USA, 2009) had a not satisfactory fit to the data. We thus released the hypothesis that Usability and Learnability are independent components of SUS ratings and tested a less restrictive model with correlated factors. This model not only yielded a good fit to the data, but it was also significantly more appropriate to represent the structure of SUS ratings.

Keywords Questionnaire · Usability evaluation · System Usability Scale

S. Borsci (✉)
ECoNA, Interuniversity Centre for Research on Cognitive Processing in Natural and Artificial Systems,
University of Rome 'La Sapienza', Rome, Italy
e-mail: simone.borsci@uniroma1.it; siomone.bo21@alice.it

S. Federici
Department of Human and Educational Sciences,
University of Perugia, Perugia, Italy

M. Lauriola
Department of Psychology of Socialization and Development Processes, University of Rome 'La Sapienza', Rome, Italy

Introduction

The System Usability Scale (SUS) developed in 1986 by Digital Equipment Corporation© is a ten-item scale giving a global assessment of Usability, operatively defined as the subjective perception of interaction with a system (Brooke 1996). The SUS items have been developed according to the three usability criteria defined by the ISO 9241-11: (1) the ability of users to complete tasks using the system, and the quality of the output of those tasks (i.e., effectiveness), (2) the level of resource consumed in performing tasks (i.e., efficiency), and (3) the users' subjective reactions using the system (i.e., satisfaction).

Practitioners have considered the SUS as unidimensional (Brooke 1996; Kirakowski 1994) since the scoring system of this scale results in a single summated rating of overall usability. Such scoring procedure is strongly based on the assumption that a single latent factor loads on all items. So far this assumption has been tested with inconsistent results. Whereas Bangor et al. (2008) retrieved a single principal component of SUS items, Lewis and Sauro (2009) suggested a two-factor orthogonal structure, which practitioners may use to score the SUS on independent Usability and Learnability dimensions. This latter finding is very inconsistent with the unidimensional SUS scoring system as items loading on independent factors of Usability and Learnability cannot be summated according to the classical test theory (Carmines and Zeller 1992). Furthermore, these factor analyses of the SUS have been carried out by exploratory techniques, nevertheless

these techniques lack of the necessary formal developments to test which of the two proposed factor solutions is the best account of collected data.

Unlike exploratory factor analysis, confirmatory factor analysis (CFA) is a theory-driven approach who needs a priori specification of the number of latent variables (i.e., the factors), of the observed-latent variables correlations (i.e., the factor loadings) as well as of the correlations among latent variables (Fabrigar et al. 1999). Once the model's parameters have been estimated, the hypothesized model is evaluated according to its ability to replicate sample's data. These features make the CFA approach the state of the art most accurate methodology to compare alternative factorial structures and eventually decide which is the best one.

Purpose

In the present study, we aim at comparing three alternative factor models of the SUS items: the one-factor solution with an overall usability factor (overall SUS) resulting from Bangor et al. (2008) (Fig. 1a); the two-factor solution resulting from Lewis and Sauro (2009) with uncorrelated Usability and Learnability factors (Fig. 1b) and its less restrictive alternative assuming Usability and Learnability as correlated factors (Fig. 1c).

Methods

Procedure

One hundred and ninety-six Italian students of University of Rome "La Sapienza" (28 males, 168 females, age mean = 21) were asked to navigate a website (http://www.serviziocivile.it) in three consecutive sections (all the students declared they never had previous surfing experience with the website):

1. In the first *20-min pre-experimental training section*, the participants were asked to navigate the website freely in order to learn features, graphic layouts, information structures and lays of the interface.
2. Afterwards, in the second *no-time-limit-scenario-based navigation section*, the participants were asked to navigate the website following four scenario targets.
3. Finally, in the third *usability evaluation section*, the SUS-Italian version was administered to the participants (Table 1).

Statistical analyses

All models were estimated by the Maximum Likelihood Robust Method as the data were not normally distributed (Mardia's normalized coefficient = 10.72). This method provided us with the Satorra–Bentler scaled chi-square statistic (S–Bχ^2), which is an adjusted measure of fit for non-normal data that is more accurate than the standard ML statistic (Satorra and Bentler 2001). According to the inspection of the model's χ^2, virtually any factor model can be rejected if the sample size is large enough, therefore many authors (McDonald and Ho 2002; Widaman and Thompson 2003) recommended to supplement the evaluation of the model's fit by some more "practical" indices. The so-called Comparative Fit Index (Bentler 1990) was purposefully designed to take sample size into account, as

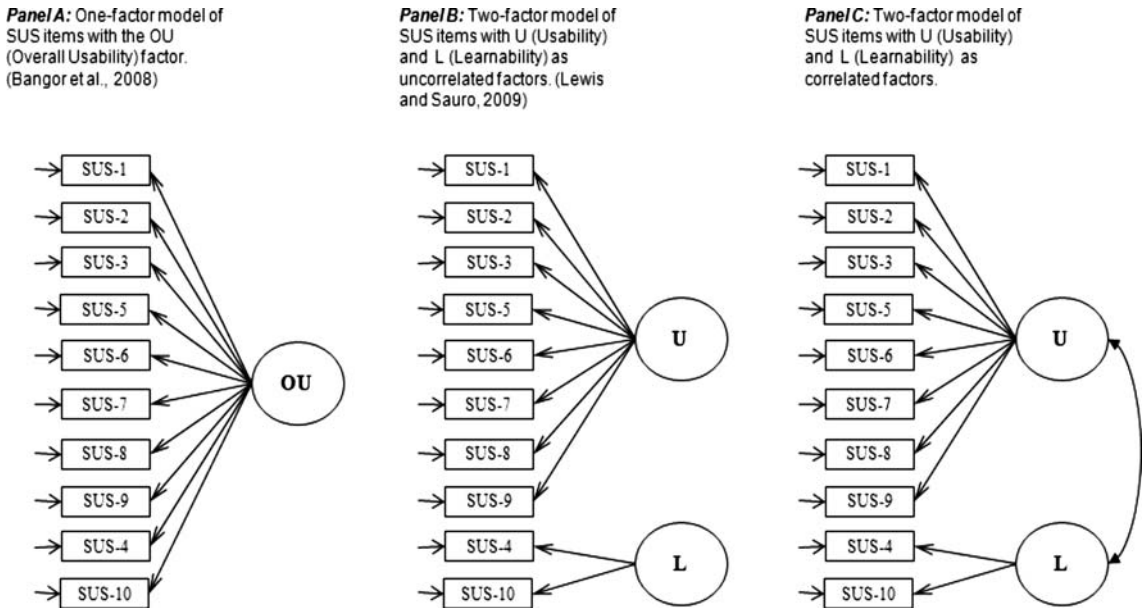

Fig. 1 SUS models tested: one-factor (**a**), two uncorrelated factors (**b**), two correlated factors (**c**)

Table 1 Synoptical table of the English and Italian versions of the SUS

Original English version	Italian version
1. I think I would like to use this system frequently	1. Penso che mi piacerebbe utilizzare questo sistema frequentemente
2. I found the system unnecessarily complex.	2. Ho trovato il sistema complesso senza che ce ne fosse bisogno
3. I thought the system was easy to use	3. Ho trovato il sistema molto semplice da usare
4. I think I would need the support of a technical person to be able to use this system	4. Penso che avrei bisogno del supporto di una persona già in grado di utilizzare il sistema
5. I found the various functions in this system were well integrated	5. Ho trovato le varie funzionalità del sistema bene integrate
6. I thought there was too much inconsistency in this system	6. Ho trovato incoerenze tra le varie funzionalità del sistema
7. I would imagine that most people would learn to use this system very quickly	7. Penso che la maggior parte delle persone potrebbero imparare ad utilizzare il sistema facilmente
8. I found the system very cumbersome to use	8. Ho trovato il sistema molto macchinoso da utilizzare
9. I felt very confident using the system	9. Ho avuto molta confidenza con il sistema durante l'uso
10. I needed to learn a lot of things before I could get going with this system	10. Ho avuto bisogno di imparare molti processi prima di riuscire ad utilizzare al meglio il sistema

it compares the hypothesized model's χ^2 with the null model's χ^2. By convention (Hu and Bentler 2004), a CFI greater than 0.90 indicates an acceptable fit to the data, with values greater 0.95 being strongly recommended. A second suggested index is the Root Mean Square Error of Approximation (Browne and Cudeck 1993). Like the CFI, the RMSEA is relatively insensitive to sample size, as it measures the difference between the reproduced covariance matrix and the population covariance matrix. Unlike the CFI, the RMSEA is a "badness of fit" index as a value of 0 indicates perfect fit and the greater the RMSEA the worse the model's fit. By convention (Hu and Bentler 2004), a RMSEA less than 0.05 corresponds to a "good" fit and an RMSEA less than 0.08 corresponds to an "acceptable" fit.

Results

Table 2 shows that the S–Bχ^2 was statistically significant for all the models we tested regardless of the number of factors and of whether the factors were correlated or not (Bentler 2004). The inspection of the CFI and RMSEA fit indexes indicated, however, that the less restrictive model assuming Usability and Learnability as correlated factors (Fig. 1c) resulted in a good fit (i.e., CFI > 0.95 and RMSEA < 0.06), whereas the unidimensional factor model (Fig. 1a) proposed by Bangor et al. (2008) resulted only in an acceptable fit (i.e., CFI > 0.90 and RMSEA < 0.00). Differently, the two-factor model proposed by Lewis and Sauro (2009) with uncorrelated factors (Fig. 1b) did not meet with any of the recommended fit indexes.

Since both the Bangor's and the Lewis and Sauro's factor models are nested within the less restrictive and best fitting model (i.e., the model with Usability and Learnability as correlated factors) we could formally compare the fit of each of the model proposed in the literature to the fit of the model which they were nested in. Nevertheless, given that we used the Satorra–Bentler scaled χ^2 measure for not multivariate normal data, we could not merely assess the χ^2 difference of two nested models. Rather we have assessed the scaled S–Bχ^2 difference according to the procedures devised by Satorra and Bentler (2001). The first contrast, which involved the comparison of the Lewis and

Table 2 Exact and close fit confirmatory factor analysis statistics/indices maximum likelihood estimation for the system usability scale

Model	S–Bχ^2 (*df*)	CFI	RMSEA	RMSEA CI
One-factor, overall usability	76.50 (35)	0.921	0.079	0.054–0.103
Two-factor, usability and learnability, uncorrelated	108.58 (35)	0.857	0.105	0.083–0.127
Two-factor, usability and learnability, correlated	54.81 (34)	0.959	0.057	0.026–0.083

All χ^2 measures were statistically significant at the 0.001 level

Table 3 Maximum likelihood standardized solution for the two-factor model of the system usability scale

Item	$\lambda_{\text{Usability}}$	$\lambda_{\text{Learnability}}$	Var ε
Q1	0.440		0.898
Q2	−0.737		0.676
Q3	0.750		0.662
Q4		0.752	0.660
Q5	0.629		0.777
Q6	−0.578		0.816
Q7	0.670		0.742
Q8	−0.600		0.800
Q9	0.681		0.732
Q10		0.712	0.702

Sauro's (2009) model (Fig. 1b) to the less restrictive two-factor model with correlated factors (Fig. 1c), was statistically significant (ΔS–B$\chi^2 = 30.17$; $df = 1$; $p < 0.001$). Likewise, the second contrast, which involved the comparison of the unidimensional model (Bangor et al. 2008) (Fig. 1a) to the less restrictive two-factor model with correlated factors (Fig. 1c), was also statistically significant (ΔS–B$\chi^2 = 28.54$; $df = 1$; $p < 0.001$). Based on the inspection of absolute and relative fit indexes as well as on the results of formal tests of χ^2 differences, we may conclude that the two-factor model with correlated factors outperformed both the factor models proposed in the literature to account for the measurement model of the SUS.

The inspection of model parameters assessed for the best fitting model (Table 3) indicated that all the SUS items significantly loaded on the appropriate factor, with factor loadings ranging from |0.44| to |0.74| for Usability and greater than 0.70 for Learnability. Accordingly, the factor reliability assessed by the ω coefficient[1] yielded fairly high values, such as 0.81 and 0.76, respectively, for Usability and Learnability factors. The correlation of Usability and Learnability was positive and significant ($r = 0.70$) thus showing that the greater the perceived Usability the greater the perceived Learnability.

Conclusions

Despite the SUS is one of the most used questionnaires to evaluate usability of systems, recent contributions have provided inconsistent results regarding the factorial structure of its items, which in turn has important consequences in determining the most appropriate scoring system of this scale for practitioners and researchers. The traditional unidimensional structure (Brooke 1996; Kirakowski 1994; Bangor et al. 2008) has been challenged by the more recent view of Lewis and Sauro (2009), assuming Learnability and Usability as independent factors. Based on a relatively large sample of users' evaluations of an existing website, we tested which of the two alternative models was the best for SUS ratings. Our data indicated that both the proposed models had a not satisfactory fit to the data with the unidimensional model—being too narrow to represent the contents of all SUS items—and with the two-factor model with uncorrelated factors—being too restrictive for its psychometric assumptions. We thus released the hypothesis that Usability and Learnability are independent components of SUS ratings and tested a less restrictive model with correlated factors. This model not only yielded a good fit to the data, but it was also significantly more appropriate to represent the structure of SUS ratings. Albeit the literature reported greater reliability coefficients (e.g., >0.80) of the Overall SUS scale, the reliability of the two Learnability and Usability factors was in keeping with required psychometric standards for short scales (Carmines and Zeller 1992). Thus, we propose that future usability studies may evaluate systems according to the scoring rule suggested by Lewis and Sauro (2009) which is very consistent with the bidimensional and best fitting model we have retrieved in this study. However, since we have found a relative correlation of Usability factors with Learnability ones, future studies should clarify under which circumstances researchers may expect to obtain Usability scores dissociated from Learnability (e.g., systems with high Learnability but low Usability). In the present study, users evaluated a single system (i.e., the serviziocivile.it website) and this might have boosted up the association of the two factors. Alternatively, our sample of users, who is comprised of college students, might be considered a sample with high computer skills compared to the general population and this might have also boosted up the factor correlation. Other studies of the SUS should, then, consider different combinations of systems and users to test the generality of the correlation of the two factors.

References

Bangor A, Kortum PT, Miller JT (2008) An empirical evaluation of the system usability scale. Int J Hum Comp Interact 24:574–594

Bentler PM (1990) Comparative fit indexes in structural models. Psychol Bull 107:238–246

Bentler PM (2004) EQS structural equations modeling software (Version 6.1) (Computer software). Multivariate Software, Encino

[1] $\omega = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum \text{Var}(\varepsilon_i)}$ where λ_i the standardized factor loadings for the factor and Var(ε_i) the error variance associated with the individual indicator variables (both reported in Table 3).

Brooke J (1996) SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL (eds) Usability evaluation in industry. Taylor & Francis, London, pp 189–194

Browne MW, Cudeck R (1993) Alternative ways of assessing model fit. In: Bollen KA, Long JS (eds) Testing structural equation models. Sage, Beverly Hills, pp 136–162

Carmines EG, Zeller RA (1992) Reliability and validity assessment. SAGE, Beverly Hills

Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ (1999) Evaluating the use of exploratory factor analysis in psychological research. Psychol Meth 4:272–299

Hu L, Bentler PM (2004) Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model 6:1–55

Kirakowski J (1994) The use of questionnaire methods for usability assessment (unpublished manuscript). http://sumi.ucc.ie/sumipapp.html

Lewis JR, 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

McDonald RP, Ho MR (2002) Principles and practice in reporting structural equation analyses. Psychol Meth 7:64–82

Satorra A, Bentler PM (2001) A scaled difference chi-square test statistic for moment structure analysis. Psychometrika 66:507–514

Widaman KF, Thompson JS (2003) On specifying the null model for incremental fit indices in structural equation modeling. Psychol Meth 8:16–37