The Factor Structure of the System Usability Scale

James R. Lewis [1] and Jeff Sauro [2]

[1] IBM Software Group, 8051 Congress Ave, Suite 2227
Boca Raton, FL 33487, jimlewis@us.ibm.com

[2] Oracle, 1 Technology Way, Denver, CO 80237, jeff.sauro@oracle.com

Abstract. Since its introduction in 1986, the 10-item System Usability Scale (SUS) has been assumed to be unidimensional. Factor analysis of two independent SUS data sets reveals that the SUS actually has two factors – Usability (8 items) and Learnability (2 items). These new scales have reasonable reliability (coefficient alpha of .91 and .70, respectively). They correlate highly with the overall SUS (r = .985 and .784, respectively) and correlate significantly with one another (r = .664), but at a low enough level to use as separate scales. A sensitivity analysis using data from 19 tests had a significant Test by Scale interaction, providing additional evidence of the differential utility of the new scales. Practitioners can continue to use the current SUS as is, but, at no extra cost, can also take advantage of these new scales to extract additional information from their SUS data.

Keywords: System Usability Scale, SUS, factor analysis, psychometric evaluation, subjective usability measurement

1 Introduction

In 1986, John Brooke, then working at DEC, developed the System Usability Scale (SUS) [1]. The SUS consists of 10 items, with odd-numbered items worded positively and even-numbered items worded negatively.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

To use the SUS, present the items to participants as 5-point scales numbered from 1 (anchored with "Strongly disagree") to 5 (anchored with "Strongly agree"). If a participant fails to respond to an item, assign it a 3 (the center of the rating scale). After completion, determine each item's score contribution, which will range from 0 to 4. For positively-worded items (1, 3, 5, 7 and 9), the score contribution is the scale position minus 1. For negatively-worded items (2, 4, 6, 8 and 10), it is 5 minus the scale position. To get the overall SUS score, multiply the sum of the item score contributions by 2.5. Thus, SUS scores range from 0 to 100 in 2.5-point increments.

The ten SUS items were selected from a pool of 50 potential items, based on the responses of 20 people who used the full set of items to rate two software systems, one of which was relatively easy to use, and the other relatively difficult. The items selected for the SUS were those that provided the strongest discrimination between the systems. In the original paper by Brooke [1], he reported strong correlations among the selected items (absolute values of *r* ranging from .7 to .9), but he did not report any measures of reliability or validity, referring to the SUS as a quick and dirty usability scale. For these reasons, he cautioned against assuming that the SUS was any more than a unidimensional measure of usability (p. 193): "SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own." Given data from only 20 participants, this caution was appropriate.

1.1 Psychometric Qualification of the SUS

Despite being a self-described "quick and dirty" usability scale, the SUS has become a popular questionnaire for end-of-test subjective assessments of usability [2, 3]. Research conducted on the SUS has shown that although it is fairly quick, it is probably not all that dirty. The typical minimum reliability goal for questionnaires used in research and evaluation is .70 [4, 5]. An early assessment of the reliability of the SUS based on 77 cases indicated a value of .85 for coefficient alpha (a measure of internal consistency often used to estimate reliability of multi-item scales) [6, 7]. More recently, Bangor, Kortum, and Miller [8], in a study of 2324 cases, found the coefficient alpha of the SUS to be .91. Bangor et al. also provided some evidence of the validity of the SUS, both in the form of sensitivity (detecting significant differences among types of interfaces and as a function of changes made to a product) and concurrent validity (a significant correlation of .806 between the SUS and a single 7-point adjective rating question for an overall rating of "user friendliness").

Although not directly measuring reliability, Tullis and Stetson [9] provided additional evidence of the reliability of the SUS. They conducted a study with 123 participants in which the participants used one of five standard usability questionnaires to rate the usability of two websites. With the entire sample size, all five questionnaires indicated superior usability for the same website. Because no practical usability test would have such a large number of participants, they conducted a Monte Carlo simulation to see, as the sample size increased from 6 to 14, which of the questionnaires would converge most quickly to the "correct" conclusion regarding the difference between the websites' usability, where "correct" meant a significant *t*-test consistent with the decision reached using the total sample size. They found that

two of the questionnaires, the SUS and the CSUQ [10, 11] met this goal the most quickly, making the correct decision over 90% of the time when $n = 12$. This result is implicit evidence of reliability, and also suggests that comparative within-subject summative usability studies using the SUS should have sample sizes of at least 12 participants.

1.2 The Assumption of SUS Unidimensionality

As previously mentioned, there has been a long-standing assumption that the SUS assesses the single construct of usability. In the most ambitious investigation of the psychometric properties of the SUS to date, Bangor et al. [8] conducted a factor analysis of their 2324 SUS questionnaires and concluded, on the basis of examining the eigenvalues and factor loadings for a one-factor solution, that there was only one significant factor, consistent with prevailing practitioner belief and practice.

The problem with this conclusion is that Bangor et al. [8] did not explore the possibility of a multifactor solution, especially, the possibility of a two-factor solution. The mechanics of factor analysis virtually guarantee high loadings for all items on the first unrotated factor, so although this finding supports the use of an overall SUS measure, it does not exclude the possibility of additional structure. Examination of the scree plot (see their Figure 5) shows the expected very high value for the first eigenvalue, but also a fairly high value for the second eigenvalue – a value just under 1.0. There is a rule-of-thumb used by some practitioners and computer programs to set the appropriate number of factors to the number of eigenvalues greater than 1, but this rule-of-thumb has been discredited because it is often the case that the appropriate number of factors is more than the number of eigenvalues greater than 1 [12, 13].

1.3 Goals of the Current Study

The primary purpose of the current study was to conduct factor analyses to explore the factor structure of the SUS, using data published by Bangor et al. [8] and an independent set of data we collected as part of a larger data collection and analysis program [14] that included 324 complete SUS questionnaires. Secondary goals were to use the new data to assess the reliability and, to as great an extent as possible, the validity of the SUS.

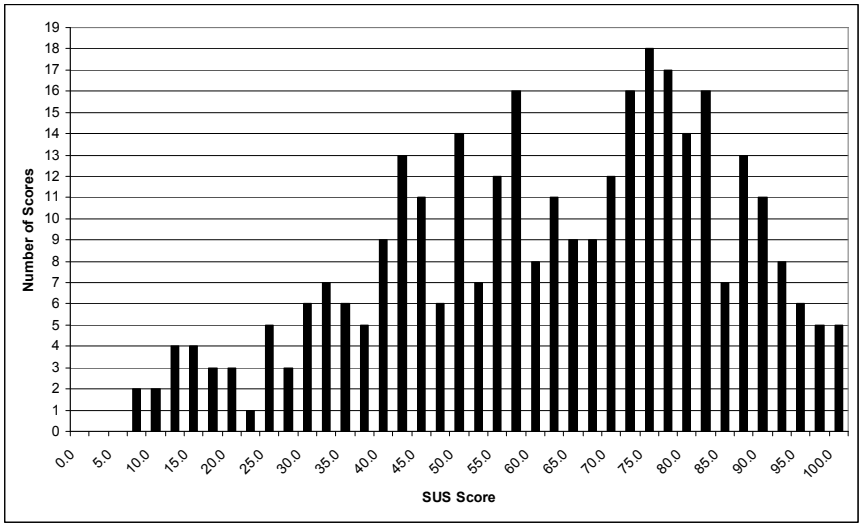
2 The Distribution of SUS Scores

Bangor et al. [8] provided some information about the distribution of SUS scores in their data. Table 1 shows basic statistical information about their distribution and the distribution of our new data. Figure 1 shows a graph of the distribution of our SUS scores (for comparison with Figure 2 of Bangor et al.).

Table 1. Basic statistical information about the SUS distributions

	Bangor et al.	Sauro & Lewis
N	2324	324
Minimum	0.0	7.5
Maximum	100.0	100.0
Mean	70.1	62.1
Variance	471.32	494.38
Standard Deviation	21.7	22.2
Standard Error	0.45	1.24
1st Quartile	55.0	45.0
Median	75.0	65.0
3rd Quartile	87.5	75.0
Interquartile Range	32.5	30.0
Critical Z (99.9%)	3.09	3.09
Critical d (99.9%)	1.39	3.82
99.9% CI Upper Limit	71.5	65.9
99.9% CI Lower Limit	68.7	58.3

Fig. 1. Distribution of the SUS scores from the current data set



Of particular interest is that the central tendencies of the distributions were not identical, with a mean difference of 8.0. The mean of the Bangor et al. distribution was 70.1, with a 99.9% confidence interval ranging from 68.7 to 71.5 [8]. The mean

of our data was 62.1, with a 99.9% confidence interval ranging from 58.3 to 65.9. Because the confidence intervals did not overlap, this difference in central tendency as measured by the mean was statistically significant ($p < .001$). There were similar differences (with the Bangor et al. scores higher) for the 1st quartile (10 points), median (10 points), and 3rd quartile (12.5 points). The distributions' measures of dispersion (variance, standard deviation, and interquartile range) were close in value.

3 Factor Analysis of the SUS

At the time of this study, we had collected 324 completed SUS questionnaires from the usability data for 19 usability studies, which was an adequate number for investigating the factor structure of the SUS [5]. Fortunately, Bangor et al. [8] published the correlation matrix of the SUS items from their studies (see their Table 5). It is possible to use an item correlation matrix as the input for a factor analysis, which meant that data were available for two independent sets of solutions – one using the Bangor et al. correlation matrix, and another using the 324 cases from Sauro and Lewis [14].

Having two independent data sources for a factor analysis of the SUS afforded a unique method for assessing the factor structure. It takes at least two items to form a scale, which makes it very unlikely that the 10-item SUS would have a structure with more than four factors. Table 2 shows side-by-side solutions for both sets of data for four, three, and two factors. Our strategy was to start with the four-factor solution (using common factor analysis with varimax rotation), then work our way down until we obtained similar item-to-factor loading for both data sets. The failure of this approach would be evidence in favor of the unidimensionality of the SUS.

As Table 2 shows, however, the results converged for the two-factor solution. Indeed, given the differences in the distributions and the differences in the four- and three-factor solutions, the extent of convergence at the two-factor solution was striking, with the solutions accounting for 56-58% of the total variance. For both two-factor solutions, Items 1, 2, 3, 5, 6, 7, 8, and 9 aligned with the first factor, and Items 4 and 10 aligned with the second factor. Given 8 items in common between the Overall SUS and the first factor, we named the first new scale Usability. Based on the content of Items 4 and 10 ("I think I would need the support of a technical person to be able to use this system" and "I needed to learn a lot of things before I could get going with this system"), we named the second new scale Learnability. It was surprising that Item 7 ("I would imagine that most people would learn to use this system very quickly") did not also align with this factor, but its non-alignment was consistent for both data sets, possibly due to its focus on considering the skills of others rather than the rater's own skills.

Table 2. Four-, three-, and two-factor solutions for the two independent data sets

Bangor et al.

Item	1	2	3	4
Q1	**0.64**	0.19	0.31	0.04
Q2	0.38	0.30	**0.53**	0.25
Q3	**0.66**	0.42	0.31	0.22
Q4	0.22	**0.67**	0.22	0.03
Q5	**0.61**	0.20	0.38	0.00
Q6	0.37	0.32	**0.58**	-0.04
Q7	**0.59**	0.33	0.30	-0.01
Q8	0.41	0.35	**0.52**	0.03
Q9	**0.61**	0.52	0.20	0.10
Q10	0.25	**0.66**	0.25	0.05

Item	1	2	3
Q1	**0.63**	0.19	0.33
Q2	0.41	0.32	**0.49**
Q3	**0.66**	0.42	0.33
Q4	0.22	**0.67**	0.23
Q5	**0.60**	0.19	0.40
Q6	0.35	0.31	**0.59**
Q7	**0.58**	0.33	0.31
Q8	0.40	0.35	**0.54**
Q9	**0.62**	0.52	0.20
Q10	0.25	**0.67**	0.26

Item	1	2
Q1	**0.70**	0.22
Q2	**0.59**	0.38
Q3	**0.71**	0.45
Q4	0.27	**0.69**
Q5	**0.71**	0.23
Q6	**0.58**	0.39
Q7	**0.64**	0.36
Q8	**0.60**	0.41
Q9	**0.60**	0.52
Q10	0.31	**0.69**

Var	3.46	2.12	**Total**
% Var	34.63	21.18	55.81

Current

Item	1	2	3	4
Q1	**0.65**	0.17	0.19	0.29
Q2	**0.59**	0.43	0.20	0.25
Q3	**0.50**	0.39	0.18	0.47
Q4	0.25	**0.64**	0.07	0.14
Q5	0.32	0.16	0.18	**0.64**
Q6	**0.46**	0.36	0.16	0.35
Q7	0.49	0.28	**0.58**	0.31
Q8	**0.67**	0.34	0.22	0.37
Q9	0.46	0.45	0.14	**0.47**
Q10	0.18	**0.68**	0.45	0.24

Item	1	2	3
Q1	**0.69**	0.20	0.26
Q2	**0.60**	0.46	0.23
Q3	**0.54**	0.43	0.43
Q4	0.27	**0.58**	0.12
Q5	0.33	0.20	**0.71**
Q6	**0.47**	0.38	0.33
Q7	**0.52**	0.45	0.35
Q8	**0.69**	0.38	0.35
Q9	**0.50**	0.46	0.40
Q10	0.24	**0.78**	0.24

Item	1	2
Q1	**0.71**	0.21
Q2	**0.62**	0.46
Q3	**0.69**	0.43
Q4	0.28	**0.58**
Q5	**0.60**	0.26
Q6	**0.58**	0.39
Q7	**0.62**	0.46
Q8	**0.77**	0.38
Q9	**0.64**	0.47
Q10	0.32	**0.79**

Var	3.61	2.20	**Total**
% Var	36.07	21.95	58.01

4 Additional Psychometric Analyses

4.1 Item Weighting

Rather than weighting each scale item the same (unit weighting), it can be tempting to use the factor loadings to weight items differentially. Such a practice is, however, rarely worth the effort and increased complexity of measurement. Nunnally [5] pointed out that such weighting schemes usually produce a measurement that is highly correlated with the unweighted measurement, so there is no statistical advantage to the weighting. That was the case with these new Usability and Learnability scales, which had, respectively, weighted-unweighted correlations of .993 and .997 (both $p < .0001$), supporting the use of unit weighting for these scales.

4.2 Scale Correlations

The correlations between the new scales and the Overall SUS were .985 for Usability and .784 for Learnability (both $p < .0001$). Because each of the new scales had items in common with the Overall SUS, this is an expectedly high level of correlation. The correlation between Usability and Learnability was .664 ($p < .0001$). They are not completely independent factors, but neither are they completely dependent, with shared variance (R^2) of about 44%. Consistent with the interpretation of the factor analyses, this finding supports both the use of an Overall SUS score and the decomposition of that score into Usability and Learnability components.

4.3 Reliability

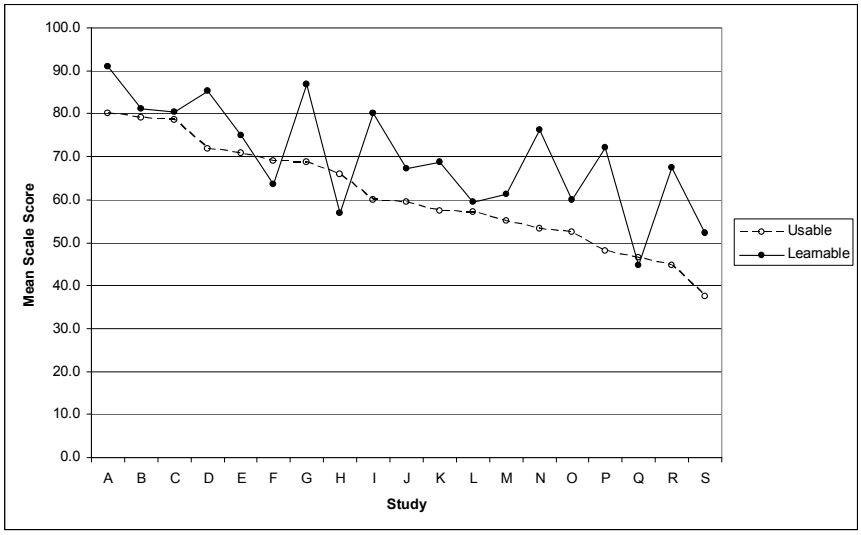
For our 324 cases, coefficient alpha for Overall SUS was .92, a finding consistent with the value of .91 reported by Bangor et al. [8]. Coefficient alphas for Usability and Learnability were, respectively, .91 and .70. Even though only two items contributed to Learnability, the scale had sufficient reliability to meet the typical minimum standard of .70 for this type of measurement [4, 5].

4.4 Sensitivity

To assess scale sensitivity, we conducted an ANOVA with Test as an independent variable with 19 levels (for the 19 tests from which the SUS scores came) and Scale as a dependent variable with 2 levels (Usability and Learnability). To make the Usability and Learnability scores comparable with the Overall SUS score (ranging from 0 to 100), we multiplied their summed score contributions by 3.125 and 12.5, respectively. The resulting scale score for Usability ranged from 0 to 100 in 32 increments of 3.125, and for Learnability ranged from 0 to 100 in eight increments of 12.5. The ANOVA had a significant main effect of Test ($F(18, 305) = 7.73$, $p <$

.0001), a significant main effect of Scale ($F(1, 305) = 47.6$, $p < .0001$), and a significant Test by Scale interaction ($F(18, 305) = 3.81$, $p < .0001$). In particular, the significant Test by Scale interaction provided evidence of the sensitivity of the Scale variable. If there had been no interaction, then this would have been evidence that Usability and Learnability were contributing the same information to the analysis. As expected from the factor and correlation analyses, however, the results confirmed the differential information provided by the two scales, as shown in Figure 2 (with the tests ordered by decreasing value of Usability). As expected due to the moderate correlation between Usable and Learnable, when the value of Usable declined, the value of Learnable also tended to decline, but with a different pattern. In most of the studies (except for three cases), the value of Learnable tended to be greater than the value of Usable, but to varying degrees as a function of Test.

Fig. 2. The Test by Scale interaction



5 Discussion

5.1 Benefit of an Improved Understanding of the Factor Structure of the SUS – A Cleaner and Possibly Quicker Usability Scale

In the 23 years since the introduction of the SUS, it has certainly stood the test of time. The results of the current research show that it would be possible to use the new Usability scale in place of the Overall SUS. The scales have an extremely high correlation (.985), and the reduction in reliability in moving from the 10-item Overall

SUS to the 8-item Usability scale is negligible (.92 to .91). The time saved by dropping Items 4 and 10, however, would be of relatively little benefit compared to the advantage of getting an estimate of perceived Learnability along with a cleaner estimate of perceived Usability. For this reason, we encourage practitioners who use the SUS to continue doing so, but to recognize that in addition to working with the standard Overall SUS score, they can decompose the Overall SUS score into its Usability and Learnability components, extracting additional information from their SUS data with very little additional effort.

5.2 Implications for SUS Item Wording

Psychometric findings for one version of a questionnaire do not necessarily generalize to other versions. Research on the SUS and similar questionnaires has shown, however, that slight changes to item wording most often lead to no detectable differences in factor structure or reliability.

For example, in a study of the interpretation of the SUS by non-native English speakers, Finstad [15] found that in Item 8 ("I found the system very cumbersome to use"), all native English speakers claimed to understand the term, but half of the non-English speakers asked for clarification. When told that "cumbersome" meant "awkward", the non-English speakers indicated that this was sufficient clarification.

Bangor et al. [8] also reported some confusion (about 10% of participants) with the word "cumbersome", and replaced it with "awkward" early in their use of the SUS. They also replaced the word "system" with "product" in all items. Consequently, about 90% of their 2324 cases used the modified version of the SUS. Our 324 cases, however, used the original SUS item wording. Despite these differences in item wording, estimates of reliability and the two-factor solutions for the two data sets were almost identical, which leads to the following two guidelines for practitioners.

- For Item 8, use "awkward" rather than "cumbersome".
- Use either "system" or "product" depending on which seems more appropriate for a given test, but for consistency of presentation, use the same term in all items for any given test or across a related series of tests.

References

1. Brooke, J.: SUS: A "quick and dirty" usability scale. In: Jordan, P. W., Thomas, B., Weerdmeester, B. A., McClelland (eds.) Usability Evaluation in Industry pp. 189--194. Taylor & Francis, London, UK (1996)
2. Lewis, J. R.: Usability Testing. In: Salvendy, G. (ed.) Handbook of Human Factors and Ergonomics pp. 1275--1316. John Wiley, New York, NY (2006)
3. Zviran, M., Glezer, C., Avni, I.: User Satisfaction from Commercial Web Sites: The Effect of Design and Use. Information & Management. 43, 157--178 (2006)

4. Landauer, T. K.: Behavioral Research Methods in Human-Computer Interaction. In: Helander, M., Landauer, T., Prabhu, P. (eds.) Handbook of Human-Computer Interaction pp. 203--227. Elsevier, Amsterdam, Netherlands (1997)
5. Nunnally, J. C.: Psychometric Theory. McGraw-Hill, New York, NY (1978)
6. Lucey, N. M.: More than Meets the I: User-Satisfaction of Computer Systems. Unpublished thesis for Diploma in Applied Psychology, University College Cork, Cork, Ireland (1991)
7. Kirakowski, J.: The use of questionnaire methods for usability assessment (1994), http://sumi.ucc.ie/sumipapp.html
8. Bangor, A., Kortum, P. T., Miller, J. T.: An Empirical Evaluation of the System Usability Scale. International Journal of Human-Computer Interaction. 24, 574--594 (2008)
9. Tullis, T. S., Stetson, J. N.: A Comparison of Questionnaires for Assessing Website Usability (2004), unpublished presentation given at the UPA Annual Conference, http://home.comcast.net/~tomtullis/publications/UPA2004TullisStetson.pdf
10. Lewis, J. R.: IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. International Journal of Human-Computer Interaction. 7, 57--78 (1995)
11. Lewis, J. R.: Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. International Journal of Human-Computer Interaction. 14, 463--488 (2002)
12. Cliff, N.: Analyzing Multivariate Data. Harcourt Brace Jovanovich, San Diego, CA, (1987)
13. Coovert, M. D., McNelis, K.: Determining the Number of Common Factors in Factor Analysis: A Review and Program. Educational and Psychological Measurement. 48, 687--693 (1988)
14. Sauro, J., Lewis, J. R.: Correlations among Prototypical Usability Metrics: Evidence for the Construct of Usability. To appear in the Proceedings of CHI 2009.
15. Finstad, K.: The System Usability Scale and Non-Native English Speakers. Journal of Usability Studies. 1, 185--188 (2006)