SUS 2.0: Updating the System Usability Scale to conform with insights from questionnaire design research

René F. Kizilcec
Department of Communication
Stanford University
kizilcec@stanford.edu

Hendrik Mueller
UX Research
Google
hendrikm@google.com

ABSTRACT

The System Usability Scale (SUS) is probably the most widely employed measure of usability. Numerous studies have assessed its psychometric properties and used it as a 'gold standard' in the development of alternative scales. From a questionnaire design perspective, however, the SUS has striking deficiencies which lead to biased measures of usability. First, we review the literature on survey biases, inspect each SUS item for issues and show that the SUS is vulnerable to significant acquiescence bias using a survey experiment. We then propose the SUS 2.0, an updated version of the SUS, which conforms to insights from questionnaire design research, and present its favorable psychometric properties. We present evidence from a second study that suggests that the SUS 2.0 is a more sensitive usability measure than the original SUS.

Author Keywords

SUS; questionnaire; surveys

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

With the increased focus on developing products that have a high level of usability, it has become important to quantify the perceived usability of a product or system. Especially during usability studies, large or small, standardized questionnaires are being used widely for such measurement. Commonly used such questionnaires include the Questionnaire for User Interface Satisfaction (QUIS) [4], the Software Usability Measurement Inventory (SUMI) [6], the Computer System Usability Questionnaires (CSUQ) [7], and the System Usability Scale (SUS) [3], among many others. Out of all usability questionnaires developed, SUS has received the highest level of adoption in both industry and academia, with hundreds of references in various publications alone.

The SUS was developed in 1986 by John Brooke while working at Digital Equipment Corporation in the UK. It was used as a quick and dirty scale to administer after usability studies on electronic office systems, such as the VT100, a text-based terminal system. SUS measures attitudes and perceptions regarding the effectiveness, efficiency, and satisfaction with a system (in accordance with the measures of usability defined in ISO 9241-11), yielding a single score to represent the global usability assessment for that system and to enable cross-system comparisons. To measure the systems usability on these dimensions, SUS is comprised of ten statements (see 3 for their exact wording) which the respondent is asked to rate on a five-point Likert scale. The Likert scale, established by Rensis Likert in 1932, allows questionnaire respondents to specify their level of agreement or disagreement on a symmetric agreement scale for a series of statements [10]. SUS Likert scale ranges from Strongly disagree to Strongly agree, with only its endpoints labeled, while additionally all five scale items are numbered from 1 to 5. As already noted in Brookes initial work [3], the phrasing of the statement heavily influences the respondents level of agreement or disagreement for it. During the development of the SUS, statements that received the most extreme responses were selected. Finally, when analyzing the SUS responses, the individual responses are added up using a particular scheme. To ensure that this summation is possible, all items of the SUS need to be evaluated by the respondent.

Over the years, SUS has been used across a variety of different systems, including hardware, software, websites, and applications. However, since the time of its development in 1986, research regarding questionnaire design has advanced significantly, with several insights that now challenge some of the foundations of the SUS. The remainder of this note will explain relevant advances in questionnaire design research, evaluate the original SUS in the context of those, propose an updated version to conform with these insights, and finally offer a comparison between the original and the updated SUS. Note that our intention is not to reduce the number of statements asked about in the SUS, contrary to recent work that attempts to create a shorter usability measure to save time [9, 5]. Our goal is to merely update the original questionnaire.

RELATED WORK

[Maybe some background on the development and testing of the SUS. Describe the original SUS and a few of its incarnations until today.]

Survey Biases

Satisficing
Acquiescence
Question order
Social Desirability
Answer Options
Hypotheticals
Leading Information

SUS EVALUATION

Heuristic review

In this section, we inspect the SUS, one item at a time, for deficiencies in its question design. The item numbers that are refered to correspond to those in Table 3. [Go through each question and describe biases, with references to articles that described them]

Experiment setup

Participants of a massive open online course offered by Stanford University were asked to complete an optional post-course survey. The survey received 1746 responses. At the beginning of the survey respondents were asked to rate their overall experience with the course, their likelihood of taking another course with the same format, their satisfaction with the amount they learnt, and the difficulty of the course. Respondents were then randomly assigned to one of three weighted groups: 25% were presented with the original SUS (n=439), 25% with the reversed SUS (n=438), and 50% received the SUS 2.0 proposed in this paper (n=869). The system that respondents were asked to evaluate comprised of the course sites for browsing and watching lecture videos. The rest of the survey was the same for all respondents and contained typical course assessment questions.

Psychometric Properties of the SUS

As the psychometric propoerties of the SUS have been studied extensively [1, 8, 2], a brief evaluation of key statistics is sufficient for this paper. Table 1 provides basic statistics that describe the distribution of the SUS scores.

Concurrent Validity

We expect the original SUS to have high concurrent validity in the form of strong associations with related and weak associations with unrelated constructs. We find that the original SUS correlates weakly to moderately albeit significantly with the following related constructs: respondents' overall course experience (r=0.34, $t(437)$=8, p<0.001), their likelihood to take another course with the same format (r=0.19, $t(437)$=4, p<0.001), and their satisfaction with how much they learnt (r=0.27, $t(437)$=6, p<0.001). However, there we find no significant association between the SUS and the difficulty of the course (r=0.04, $t(437)$=0.9, p=0.4).

Table 1. Statistical information on the SUS distributions (short SUS 2.0 scores scaled to 0-100 for comparison)

SUS	N	Min	Max	Mean	SD	Median	IQR
Original	439	38	100	84.5	12.9	88	16
Reversed	438	26	100	82.3	11.4	84	14
Short 2.0	869	24.5	84.5	69.7	10.4	71	13

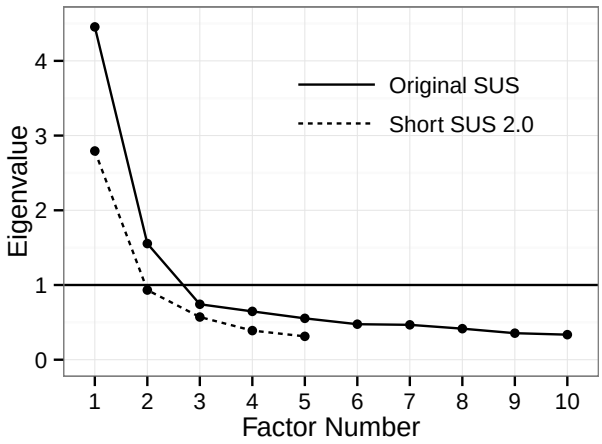

Figure 1. Scree plot for original SUS and SUS 2.0 showing that a single-factor solution is viable, although the original SUS has two factors with eigenvalue greater one

Internal consistency

For the 439 responses to the original SUS, Cronbach's α is 0.86 and the correlation of each item with the total score lies between 0.53 and 0.81 with 95% confidence. Although the SUS is frequently reported to have a higher coefficient alpha [1, 8, 2], an alpha of 0.86 reflects a high degree of interrelatedness.

Factor analysis

A factor analysis of the 439 responses to the original SUS suggests that the scale has a two-factor structure. A scree plot (Figure 1) illustrates that two factors have eigenvalues greater than one. This is consistent with previous work on the SUS factor structure [8].

Acquiescence Bias in SUS

A comparison of original SUS scores with scores from the reversed SUS provides strong evidence for acquiescence bias in the SUS. Without acquiescence bias, the average for each item on the original SUS would not be significantly different from the reverse-coded average for each item on the reversed SUS. However, if acquiescence bias exists, respondents would tend to agree with statements independent of the statement's tone, which would be reflected in a significant differenece between the original SUS average and reversed SUS reverse-coded average.

Table 2 provides means, standard deviations, and p-values from non-parametric Mann-Whitney tests of the hypothesis that there is no location shift (this non-parametric alternative of the t-test is used as scores are

not normally distributed). We find highly significant differences at 99% significance in all but two items and the overall SUS score. This is very strong evidence for the claim that the original SUS induces acquiescence bias.

Table 2. Strong evidence for acquiescence bias in the SUS for individual items and the overall score

	Original		Reversed		
#	M	SD	M	SD	p value
1	8.52	2.12	6.69	2.98	<0.001
2	9.09	1.85	8.15	2.47	<0.001
3	8.58	1.79	8.90	1.83	<0.001
4	8.19	2.36	8.66	1.68	0.17
5	8.24	1.81	8.18	1.90	0.86
6	8.55	2.12	8.46	1.73	0.006
7	8.61	1.68	8.90	1.79	<0.001
8	7.99	1.85	8.19	2.02	0.01
9	8.52	2.04	7.77	2.29	<0.001
10	8.17	1.80	8.42	2.01	0.002
	84.46	12.91	82.33	11.39	<0.001

SUS 2.0 PROPOSAL

Following the heuristic evaluation of the original SUS items, we propose an updated set of items that, at their core, are equivalent to the SUS, but reduce vulnerability to survey biases, like acquiescence bias. The first step was to change the questionnaire items from being statements to questions in an effort to reduce acquiescence bias. Moreover, statements that were phrased as hypotheticals, such as item 5 in Table 3, were rephrased as concrete questions about the underlying construct.

The second step was to change the scale from an agree-disagree Likert scale to scales that reflect the relevant constructs, such as confidence, learnability, or complexity. Following question design research, unipolar scales were presented as 5-point scales, while bipolar scales were presented as 7-point scales.

Our goal was to update the SUS items for future use given that the SUS is clearly the most established usability scale. These updated items with corresponding answer scales are presented in Table 4. For our evaluation of the SUS 2.0, we were unable to use the full 10 items and opted for using a reduced number of items which we refer to as the short SUS 2.0. The short SUS 2.0 consists of five items marked with asterisks in Table 4 and covers the key facets of the SUS (confidence, ease of use, consistency, leanability).

The score calculation for the SUS 2.0 is different to that of the original SUS, because the SUS 2.0 consists of six items with 5-point unipolar answer scales and four items with 7-point bipolar answer scales, instead of ten items on a 5-point scale. The SUS 2.0 can be scored as a proportion of the maximum score that is achievable. First, assign values from 0–4 or 0–6 depending on the number of scale points such that 0 reflects the worse response and 4 or 6 the best. Second, sum up the values for all ten responses; the sum should be an integer between 0–48. Third, divide by 48 to get the SUS 2.0 score. For the short SUS 2.0, follow the same steps except that the sum of response values will lie between 0–24 and you divide by 24, not 48.

Psychometric Properties of the SUS 2.0

Table 1 provides basic statistics on the distribution of short SUS 2.0 scores scaled to the same 0-100 score range as the original SUS to facilitate a comparison. Most notably, the SUS 2.0 scores range from about 25 to 85 for a system with good usability, while the original and reversed SUS suffer from ceiling effects which artificially constrain the distribution of scores.

Concurrent Validity

We start to establish the SUS 2.0's concurrent validity by investigating its association with related and unrelated constructs. We find that the short SUS 2.0 correlates moderately and significantly with respondents' overall course experience (r=0.31, $t(866)$=10, p<0.001), their likelihood to take another course with the same format (r=0.22, $t(867)$=7, p<0.001), and their satisfaction with how much they learnt (r=0.25, $t(867)$=8, p<0.001); however, the short SUS 2.0 is only marignally associated with the difficulty of the course (r=0.08, $t(867)$=2.5, p=0.01).

Internal consistency

Cronbach's α for the short SUS 2.0 is 0.77 based on 869 responses and individual item correlations with the total score vary between 0.47 and 0.86 with 95% confidence. Coefficient alpha is smaller for the short SUS 2.0, but given that coefficient alpha increases with the number of items, the ten-item SUS 2.0 will have higher reliability thant the five-item short SUS 2.0.

Factor analysis

A factor analysis of 869 responses to the short SUS 2.0 suggests that a single-factor structure is most appropriate for the measure. As shown in the scree plot (Figure 1), only one factor has an eigenvalues greater than one and the slope changes considerably at the two factor point.

Sensitivity of the SUS 2.0

A good usability scale should exhibit a high level of sensitivity to reflect subtle differences in usability. We conducted a second survey study in a post-course survey of an online course that ran on a different system (web interface) to investigate how the SUS 2.0 compares to the original SUS in terms of sensitivity.

The two systems offered the same basic features, browsing and playing video lectures, but differed considerably in their design. We employed Molich and Nielsen's heuristic evaluation criteria [11] to informally establish which system has better usability. While both systems showed generally high usability, one system was deemed superior in these four categories: match between system and the real world, consistency and standards, aesthetic and minimalist design, and help and documentation. This informal usability comparison was the basis

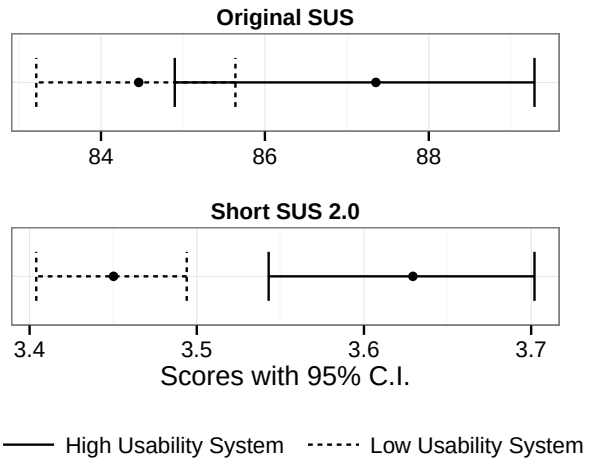

Figure 2. Evaluation of scale sensitivity for original SUS and SUS 2.0 showing that the SUS 2.0 has high enough sensitivity to distinguish between a high and low usability system

for labeling one system as having 'high usability' and the other 'low usability'.

Figure 2 illustrates usability ratings using the original SUS and the short SUS 2.0 for the high-usability and the low-usability system. As the usability scores from both scales were not normally distributed, 95% confidence intervals were computed from 10,000 bootstrap replicates using the adjusted bootstrap percentile method. While the original SUS is not sensitive enough to differentiate the usability of the two interfaces with 95% confidence, the short SUS 2.0 exhibits good sensitivity. A Mann-Whitney test of the difference between the usability estimates for the two systems for each scale supports this result (W=18585, p=0.07 for the original SUS; W=74598, p<0.001 for the short SUS 2.0).

DISCUSSION

The paper's unique contribution is to analyze the SUS from a questionnaire design perspective which has not been done before. We provide strong evidence that the original SUS induces acquiescence bias and address this and other potential issues by proposing the SUS 2.0: an updated version of the original SUS with statements rephrased as questions with relevant answer scales. The SUS 2.0 is less likely to induce response biases, because it conforms to insights from questionnaire design research. Moreover, we find that the short SUS 2.0, with only five items, is a more sensitive measure of usability than the original, ten-item SUS.

Future work could investigate the strength of association between the original SUS and the SUS 2.0, which was not possible in these studies as each respondent saw just one usability scale. Moreover, our findings are limited in that we could only administer the short SUS 2.0, which constraint us from performing an investigation of the SUS 2.0's factor structure.

CONCLUSION

This paper presents compelling evidence that highlights big questionnaire design deficits in the SUS. The authors propose a redesign of the scale (the SUS 2.0) and provide evidence that a sub-scale of the SUS 2.0 has high reliability, concurrent validity, an single-factor structure, and better sensitivity than the orignal SUS. While more work on evaluating the SUS 2.0 is needed, the authors strongly recommend the use of the updated scale for measuring usability.

ACKNOWLEDGMENTS

We are grateful to the instructors of the two Stanford courses for allowing us to conduct survey experiments in their post-course surveys.

REFERENCES

1. Bangor, A., Kortum, P. T., and Miller, J. T. An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction 24*, 6 (2008), 574–594.
2. Borsci, S., Federici, S., and Lauriola, M. On the dimensionality of the system usability scale: a test of alternative measurement models. *Cognitive processing 10*, 3 (2009), 193–197.
3. Brooke, J. Sus: A quick and dirty usability scale. *Usability evaluation in industry 189* (1996), 194.
4. Chin, J. P., Diehl, V. A., and Norman, K. L. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, ACM (New York, NY, USA, 1988), 213–218.
5. Finstad, K. The usability metric for user experience. *Interacting with Computers 22*, 5 (2010), 323–327.
6. Kirakowski, J., and Corbett, M. Sumi: The software usability measurement inventory. *British journal of educational technology 24*, 3 (1993), 210–212.
7. Lewis, J. R. Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int. J. Hum.-Comput. Interact. 7*, 1 (Jan. 1995), 57–78.
8. Lewis, J. R., and Sauro, J. The factor structure of the system usability scale. In *Human Centered Design*. Springer, 2009, 94–103.
9. Lewis, J. R., Utesch, B. S., and Maher, D. E. Umux-lite: when there's no time for the sus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, ACM (New York, NY, USA, 2013), 2099–2102.

Table 3. Items from the original and reversed SUS

#	**Original SUS***	**Reversed SUS***
1	I needed to learn a lot of things before I could get going with this system	I needed to learn very few things before I could get going with this system
2	I think that I would need the support of a technical person to be able to use this system	I think that I would not need any support of a technical person to be able to use this system
3	I felt very confident using the system	I did not feel very confident using the system
4	I found the system very cumbersome to use	I found the system very manageable to use
5	I would imagine that most people would learn to use this system very quickly	I would imagine that most people would learn to use this system very slowly
6	I found the system unnecessarily complex	I found the system appropriately simple
7	I thought the system was easy to use	I thought the system was hard to use
8	I found the various functions in this system were well integrated	I found the various functions in this system were not well integrated
9	I thought there was too much inconsistency in this system	I did not think there was too much inconsistency in this system
10	I think that I would like to use this system frequently	I do not think that I would like to use this system frequently

*Items were presented in a matrix with a 5-point Likert scale: Strongly disagree (1), (2), (3), (4), Strongly agree (5)

Table 4. SUS 2.0 questions with answer scales

#	**Questions**	**Answer Scales**
1⋆	How much more is there to learn about the system?	Nothing at all, A little, A moderate amount, A lot, A great deal
2	How likely are you to need the support of a technical person to be able to use the system?	{Extremely, Very, Somewhat} unlikely, Neither likely nor unlikely, {Somewhat, Very, Extremely} likely
3⋆	How confident are you using the system?	{Not at all, Slightly, Moderately, Very, Extremely} confident
4	How cumbersome is it to use the system?	{Not at all, Slightly, Moderately, Very, Extremely} cumbersome
5⋆	How easy or difficult is it to learn how to use the system?	{Extremely, Moderately, Slightly} difficult, Neither difficult nor easy, {Slightly, Moderately, Extremely} easy
6	How complex is the system?	{Not at all, Slightly, Moderately, Very, Extremely} complex
7⋆	How easy or difficult is it to use the system?	{Extremely, Moderately, Slightly} difficult, Neither difficult nor easy, {Slightly, Moderately, Extremely} easy
8	How integrated are the systems various functions?	{Not at all, Slightly, Moderately, Very, Extremely} integrated
9⋆	How consistent is the system?	{Not at all, Slightly, Moderately, Very, Extremely} consistent
10	How much do you like or dislike the system?	{Extremely, Moderately, Slightly} dislike, Neither like nor dislike, {Slightly, Moderately, Extremely} like

*Items included in the Short SUS 2.0

10. Likert, R. A technique for the measurement of attitudes. *Archives of Psychology 22*, 140 (1932), 1–55.

11. Molich, R., and Nielsen, J. Improving a human-computer dialogue. *Communications of the ACM 33*, 3 (1990), 338–348.