

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 publications have been devoted to assessing its psychometric properties and using it as a gold standard in the development of alternative metrics. 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.

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 2 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 end-points 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.]

Paste the appropriate copyright statement here. ACM now supports three different copyright statements:

- ACM copyright: ACM holds the copyright on the work. This is the historical approach.
- License: The author(s) retain copyright, but ACM receives an exclusive publication license.
- Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.

This text field is large enough to hold the appropriate release statement assuming it is single spaced.

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 referred to correspond to those in Table 2. [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 properties of the SUS have been studied extensively [1, 8, 2], a brief evaluation of key statistics is sufficient for this paper. Table 4 provides basic statistics that describe the distribution of the SUS scores.

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 our 439 responses suggest that a single-factor solution is viable, even though two factors have eigenvalues greater than one. A scree plot (Figure 2) illustrates this result.

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 difference between the original SUS average and reversed SUS reverse-coded average.

Table 1 provides means, standard deviations, and significance levels from Welch two-sample tests (which does not assume equal sample variances like the t-test). Given the much larger sample size of responses on system A than B, we observe more significant differences for comparisons for system A than B. Most notably, highly significant acquiescence bias at $p < 0.05$ was found for 6 out of 10 items and the overall SUS score for system A. This result clearly shows that the SUS is significantly biased.

Table 1. Strong evidence for acquiescence bias in the SUS for individual items and the overall score

#	Original		Reversed		p value
	M	SD	M	SD	
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	0.008
4	8.19	2.36	8.66	1.68	<0.001
5	8.24	1.81	8.18	1.90	0.67
6	8.55	2.12	8.46	1.73	0.49
7	8.61	1.68	8.90	1.79	0.012
8	7.99	1.85	8.19	2.02	0.13
9	8.52	2.04	7.77	2.29	<0.001
10	8.17	1.80	8.42	2.01	0.056
	84.46	12.91	82.33	11.39	0.010

SUS 2.0 PROPOSAL

[Describe the proposal and refer to table with the updated questions. Mention that it was a non-goal to reduce the number of questions, etc.]

[We gotta talk about how this impacts the SUS score calculations, comparison to past scores, etc.]

Psychometric Properties of the SUS 2.0

Distribution in Table 4

Internal consistency

Coefficient α is 0.77

Factor analysis

Figure 2

Sensitivity of the SUS 2.0

[Introduce second data set] The two online courses were offered on two distinct online platforms that shared the same core features but differed considerably in design. Based on a design heuristic evaluation of the two systems, it was determined that B had more usability problems than A. [Need to talk about how this was evaluated. E.g. http://en.wikipedia.org/wiki/Heuristic_evaluation]

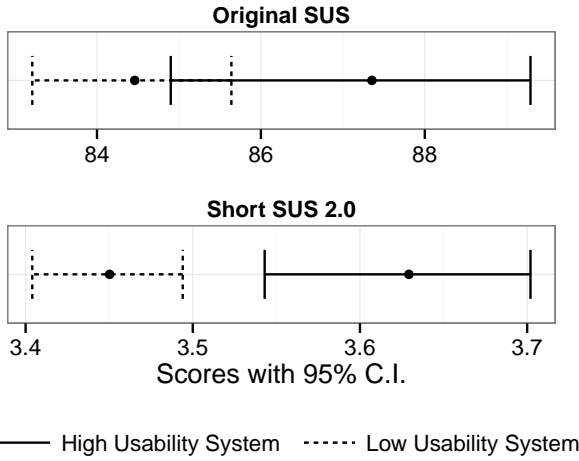


Figure 1. Evaluation of scale sensitivity for original SUS and SUS 2.0 showing that the SUS 2.0 has high enough sensitivity to distinguish between a high and low usability system

Table 4. Statistical information on the SUS distributions (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 illustrates usability ratings using the original SUS and the short SUS 2.0 for two online course interfaces where one interface has an objectively higher usability level than the other. 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 non-parametric Wilcoxon rank sum test for each scale supports this result ($W=18585$ and $p=0.07$ for the original SUS, and $W=74598$ and $p<0.001$ for the short SUS 2.0)

DISCUSSION

CONCLUSION

ACKNOWLEDGMENTS

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.

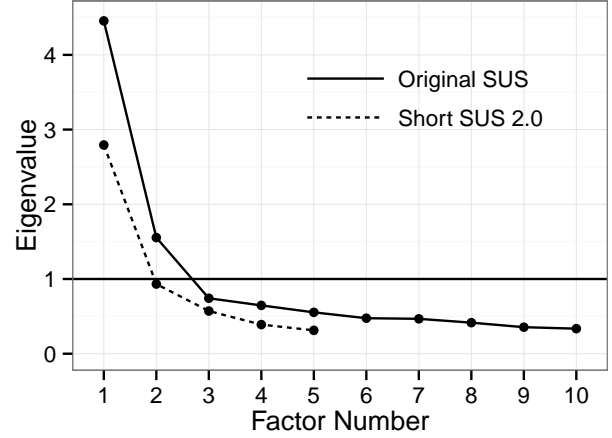


Figure 2. Scree plot for original SUS and SUS 2.0 showing that a single-factor solution is viable, although the original SUS has two factors with eigenvalue greater one

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

Table 2. 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 3. Items from the proposed SUS 2.0 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