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

Author Keywords
SUS; questionnaire; surveys

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

INTRODUCTION

Some of the most used such questionnaires are QUIS [3], SUMI [4], CSUQ [5], and SUS [2].

The SUS is used extensively to evaluate systems' usability and it is thus important that it provides accurate results. Since its introduction in 19XX, there have been critical advancement in the literature on questionnaire and question design regarding the identification of question biases and how to avoid them. As a matter of fact, the SUS is vulnerable to several well-known biases as a result of how the items and scale are constructed.

[Insert high-level summary of SUS]

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 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

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, 6], an alpha of 0.86 reflects a high degree of interrelatedness.

Factor analysis

Lewis & Sauro [6] investigated the factor structure of the SUS and found two moderately correlated factors: Usability (items 2-4, 6-10) and Learnability (items 1, 5). 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 1) 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

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.

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 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		
#	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	0.008
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 1

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`]

DISCUSSION

CONCLUSION

ACKNOWLEDGMENTS

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
2.0	869	24.5	84.5	69.7	10.4	71	13

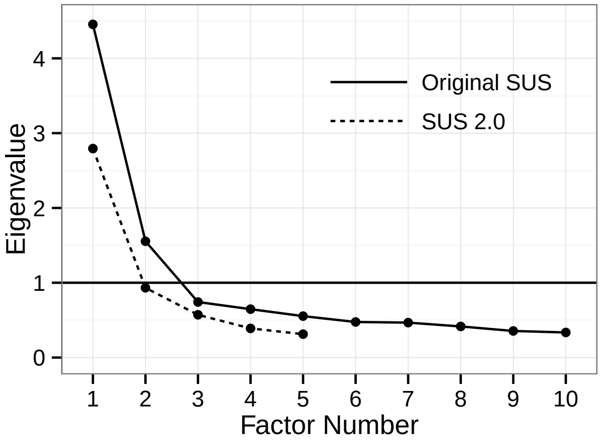

Figure 1. Scree plot for original SUS and SUS 2.0 showing that a single-factor solution is viable, although the original SUS has two factors with eigenvalue greater one

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. Brooke, J. Sus: A quick and dirty usability scale. *Usability evaluation in industry 189* (1996), 194.

3. 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.

4. Kirakowski, J., and Corbett, M. Sumi: The software usability measurement inventory. *British journal of educational technology 24*, 3 (1993), 210–212.

5. 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.

6. Lewis, J. R., and Sauro, J. The factor structure of the system usability scale. In *Human Centered Design*. Springer, 2009, 94–103.

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