Critiquing the System Usability Scale from a Questionnaire Design Perspective: Beware of Acquiescence Bias

1st Author Name
Affiliation
Street
City
E-mail address

2nd Author Name
Affiliation
Street
City
E-mail address

ABSTRACT
The System Usability Scale (SUS) is probably the most widely employed measure of usability today. Numerous studies have assessed its psychometric properties and used it as a "gold standard" in the development of alternative scales. Recent advances in questionnaire design research on satisficing, acquiescence, and other biases, however, now challenge some of the foundations of the SUS. In this note, we review literature on relevant survey biases, inspect each SUS item for such biases, and using a survey experiment, show that the SUS is vulnerable to significant acquiescence bias. We then propose a more robust scale, rooted in the SUS, which conforms with recent insights from questionnaire design research, and provide an example of how the proposed scale outperforms the SUS in terms of measurement sensitivity.

Author Keywords
Usability, questionnaire, survey, System Usability Scale, SUS

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

INTRODUCTION
With the increased focus on developing highly usable products, it has become important to actually quantify the perceived usability of product or systems. Especially during usability studies, large or small, standardized questionnaires are being used widely for such measurement. Commonly used such questionnaires, in order of its first publication, include the Computer User Satisfaction Inventory (CUSI) [8], the Questionnaire for User Interface Satisfaction (QUIS) [4], the After Scenario Questionnaire (ASQ) [14], the Software Usability Measurement Inventory (SUMI) [7], the Computer System Usability Questionnaires (CSUQ) [15], and the System Usability Scale (SUS) [3], among many others.

Out of all usability measurement questionnaires, the SUS has received the highest level of adoption in both industry and academia, with hundreds of references across publications alone. The SUS is comprised of ten statements measured on a 5-point agreement scale, yielding a single score summarising the usability assessment of the evaluated system. 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 inception in 1986, research regarding the design of valid and reliable questionnaires has advanced significantly, with several insights that challenge some of the foundations of the SUS. Most notably are research insights related to satisficing [23, 9, 10], acquiescence bias [22, 13, 20], optimal scale lengths [12, 6], social desirability, question and response order biases, the use of hypothetical questions, among others.

The remainder of this note will explain such advances in questionnaire design research relevant to the original SUS, evaluate the SUS in the context of those, propose an updated version to conform with these insights, and finally show the increased quality of 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 identify distinct factor and to create a shorter usability measure to save time [5, 17]. The goal of this note is to critique the wording of the questions and response options of the SUS on theoretical grounds with empirical evidence and propose a more robust alternative scale.

RELATED WORK
Even though first published in 1996, the SUS was developed in 1986 by John Brooke while working at Digital Equipment Corporation (DEC) in the UK. It was used as a "quick and dirty" scale to be administered after usability studies on electronic office systems, such as DEC's VT100, a text-based terminal system. The 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). To measure a system's usability on these dimensions, the SUS is comprised of ten statements (see 3 for their exact wording) which the respondent is asked to rate individually. The SUS uses a Likert scale, established by Rensis Likert in 1932 [18], which allows questionnaire respondents to specify their level

of agreement or disagreement with each of the statements on a symmetric five-point agreement scale, ranging from "Strongly disagree" to "Strongly agree". Note that only its endpoints are labeled, while additionally all five scale items numbered from 1 to 5. As already noted in Brooke's initial work [3], the phrasing of the statement strongly influences the expressed level of agreement; hence, during the development of the SUS, statements that received the most extreme responses were selected. When analyzing responses to the SUS, the individual responses are consolidated into a single score to represent the global usability assessment for that system and to enable cross-system comparisons. To ensure that this summation is possible, all items of the SUS need to be evaluated by each respondent.

One of the questionnaire biases that has been researched thoroughly is that of acquiescence bias. *acquiescence bias* is the tendency of a respondent to be more likely to agree with a given statement independent of its substance [22]. The respondent only thinks of reasons why the statement is true, rather than expending cognitive effort to consider reasons for disagreement [9]. On top of that, respondents with lower self-perceived status assume the survey administrator agrees with the posed statement, resulting in deferential agreement bias [20]. Acquiescence bias is the strongest when presented with agree/disagree, yes/no, or true/false answer options [22]; however, the same effects have been shown for agreement scales (such as the Likert scale) [20]. To minimize this bias, questions should refer to the underlying construct in a neutral, non-leading way and offer a neutral scale [20].

Another effect that often leads to less reliable questionnaire responses is that of satisficing. *satisficing* occurs when respondents compromise one or more of the four cognitive steps required for reliably responding to a question: comprehension, retrieval, judgement, and mapping [23]. Satisficing can take weak and strong forms; while weak satisficers make an attempt to answer correctly yet less than thorough, strong satisficers typically skip entire steps and typically pick what they consider to be the first acceptable response alternative [21, 9, 10]. Satisficing often also leads to straight-lining when subsequent questions are presented with the same answer options [11].

Picking the most valid and reliable *rating scale* has been another heavily researched topic in the field of questionnaire design. First, research has shown that scale points that are fully labeled as compared to those that just use numbers optimize reliability and minimize bias [6]. Second, rating scales may either use a unipolar or bipolar scale, depending on the nature of the construct being measured. Unipolar constructs range from zero to an extreme amount and are best measured on a five-point scale, optimizing reliability while minimizing respondent burden [12]. Bipolar constructs, on the other hand, range from an extreme negative to an extreme positive with a natural midpoint, which are best measured with a 7-point rating scale to maximize reliability and data differentiation [12].

Finally, we would like to mention the use of *hypotheticals* and questions that ask the respondent to predict future behavior or attitudes for themselves or other people. Even though the respondent may have a rational answer to such questions, their response does not predict actual future behavior, neither theirs nor others.

EVALUATION OF THE ORIGINAL SUS

This section evaluates the SUS in regards to latest advances in questionnaire design research. It also includes an experiment through which acquiescence bias for the SUS statements has been evaluated.

Heuristic review

In this section, we inspect the original SUS for weaknesses in its questionnaire design, through the consideration of SUS in relation to recent questionnaire design research mentioned in the previous section. The item numbers of the SUS referred to in the remainder of this note correspond to those in Table 3.

Each of the ten items of the SUS is constructed of a non-neutral statement and an agreement scale. For each item, this particular design encourages the effects of acquiescence bias, i.e., the likelihood for the respondents to agree with the statement provided [22]. Exemplary for all other items, let's review item #3 of the SUS: "I thought the system was easy to use" In this case, respondents will expend unproportional effort on finding reasons that confirm that the system is actually complex, instead of reasonably simple. As a result, respondents are more likely to think of aspects that make the system appear more complex (instead of simple), hence, leading to a more agreeable response for the statement. On top of that, the nature of the agreement scale leads respondents further towards agreement as disagreeing with anyone requires courage and cognitive effort [20]. This effect is especially strong when the survey is not administered anonymously and when the survey administrator is known to the respondent, as it is true for most applications of the SUS, i.e., its administration at the end of usability studies. To minimize such acquiescence bias, each of the statements should be transformed into a construct-specific, neutral question with similarly neutral answer options matching the question construct. Item #3 may then result in less such bias if it is instead asked as "How easy or difficult was it to use the system?" on a seven-point, fully labeled scale from "Extremely difficult" to "Extremely easy". By giving "easy" and "difficult" equal weight in both the question and the answer options, respondents are less led into either direction. Note, as items #3 and #8 ask about the same underlying construct of ease/difficulty, they would result in the same reworded question, and hence should only be included once.

The original SUS uses the exact same scale for all of its items. While one may interpret this as a way to minimize the required effort by the respondents, as they do not have to relearn a new scale for each item, this design is subject to satisficing. With the same scale repeated for subsequent questions, respondents are likely to straight-line, i.e., select the same answer option for items without thoroughly considering the question [11]. Furthermore, satisficing may be further encouraged as the design of the SUS requires the respondent to answer all of the items. Respondents may randomly select an answer for questions that they simply cannot respond to, as they may not fully understand the question wording or as they may not have experienced the system sufficiently to reliably respond. For example, a respondent may not fully understand the term "well integrated" in item #5 and would prefer to skip this question, however, is now forced to leave an answer. As a result of these different tendencies towards satisficing, results for the SUS may be less reliable overall. To minimize satisficing, none of the questions should be required and different (but construct-appropriate) answer scales should be used. The length of the different scales needs to depend on the identified construct and if it is unipolar or bipolar in nature [12]. As demonstrated earlier for item #3, its construct is bipolar in nature, as ease/difficulty can have an extreme negative, an extreme positive, with a neutral midpoint. In this case, a seven-point scale is used to achieve the highest reliability while minimizing the respondent effort low. For item #5 on the other hand, a five-point scale may be used as the construct of "complexity" is unipolar in nature.

Finally, the SUS uses several hypothetical questions, in particular item #1 ("I think that I would like to use this system frequently"), item #4 ("I think I would need the support of a technical person to be able to use this system"), and item #7 ("I would imagine that most people would learn to use this system very quickly"). All of these three items have in common that they ask the respondent to make predictions about their future behavior, needs, and attitudes, in the case of item #7 even predict attitudes of other people. Even though respondents may have a rational response to these questions, the relationship between predictions and actual future behavior is weak. Instead, past behavior is a much better predictor of future behavior. These items need to be entirely rephrased to refer back to the system just used before measuring its usability on the SUS.

Experimental Evaluation of Acquiescence Bias

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. Respondents were 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 an example of a more robust scale proposed in this note (n=869) (see Table 3 for scale details). 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, 16, 2], a brief evaluation of key statistics should be sufficient here. Table 1 provides an overview of statistics and psychometric properties of the original and reversed SUS, and the more robust example scale. Although the SUS is frequently reported to have a higher coefficient α [1, 16, 2], an α of 0.86 reflects a good level of internal consistency. Moreover, the factor analysis yields two eigenvalues greater than one, which is consistent with previous work on the SUS's factor structure [16].

Acquiescence Bias Conclusion

A comparison between scores from the original and reversed SUS provides strong evidence that the SUS induces acquiescence bias. 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 2 provides means, standard deviations, and p-values from non-parametric Mann-Whitney tests of the hypothesis that there is no location shift (a non-parametric alternative of the t-test is used as scores are not normally distributed). We find highly significant differences with at least 99% confidence 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.

ALTERNATIVE SCALE PROPOSAL

In light of the SUS's deficiencies uncovered in the previous section, 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, satisficing, and hypothetical projections. The first step was to change the questionnaire items from being statements

Table 1. Statistical and psychometric scale properties

Statistic	SUS	Rev. SUS	Robust Example
N	439	438	869
Range	[23, 100]	[8, 100]	[22, 100]
Mean	80.6	77.9	76.7
SD	16.1	14.2	14.7
Median	85	80	78
IQR	20	18	22
Cronbach α	0.86	0.73	0.67
$\lvert\lambda > 1\rvert^*$	2	2	1

*number of eigenvalues greater than one in factor analysis

Table 2. Means, standard deviations, and p values from Mann-Whitney tests for each item and the overall score of the original and reversed SUS providing strong evidence that the SUS induces acquiescence bias

	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	0.173
5	8.24	1.81	8.18	1.90	0.868
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	0.013
9	8.52	2.04	7.77	2.29	<0.001
10	8.17	1.80	8.42	2.01	0.002
	80.58	16.14	77.92	14.23	<0.001

to questions in an effort to reduce acquiescence bias. Moreover, statements that were phrased as hypotheticals, such as item #7 in Table 3, were rephrased as concrete questions about the underlying construct referring back to the system that was just used by the respondent.

The second step was to change the scale from a Likert agreement scale to scales that reflect the relevant construct, such as ease/difficulty (item X), confidence (item X), learnability (item X), or complexity (item X). Following recommendations from question design research [12], unipolar constructs were presented as five-point scales, while bipolar constructs were presented as seven-point scales.

Our goal was to update the SUS items instead of creating a new measure, given that the SUS is probably the most established usability scale. These updated items with corresponding answer scales are presented in Table 3. For our evaluation of the SUS 2.0, we were unable to use all ten 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 four items marked with asterisks in Table 3 and covers the key dimensions of the SUS (confidence, ease of use, consistency, learnability).

The score calculation for the SUS 2.0 is different to that of the 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. To calculate the SUS 2.0 score, first assign values between 0 – 4 or 0 – 6 depending on the number of scale points such that 0 reflects the worse usability response and 4 or 6 the best. Second, sum up the values for all ten responses to obtain an integer between 0 – 48. Third, divide by 0.48 to obtain the SUS 2.0 score. For the short SUS 2.0, follow the same steps except that the sum of response values lies between 0 – 18 and is divided by 0.18. In contrast to the SUS, non-responses are permitted, albeit not encouraged, in the SUS 2.0 and are accounted for by adjusting the denominator accordingly.

Comparing Scale Sensitivity

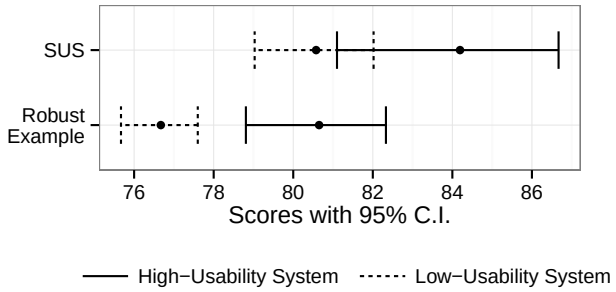

Figure 1. Evaluation of scale sensitivity for SUS and short SUS 2.0 showing that the latter has high enough sensitivity to distinguish between a high and low usability system, while SUS scores are not significantly different

A good usability scale should exhibit a high level of sensitivity to reflect even 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 provide an example of how a more robust usability scale compares to the SUS in terms of sensitivity.

The two systems offered the same basic features, i.e. browsing and playing video lectures, but differed considerably in their design. We employed Molich and Nielsen's heuristic evaluation criteria [19] to informally establish which system has better usability. While both systems showed generally high usability, one system was deemed superior in four evaluation 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 for labeling one system as having 'high usability' and the other 'low usability'.

Figure 1 illustrates usability ratings on the SUS and the alternative example scale 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 SUS is not sensitive enough to differentiate the usability of the two interfaces with 95% confidence, the alternative scale 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 SUS; W=74598, p<0.001 for the alternative scale).

CONCLUSION

This note's unique contribution is to critique the SUS from a questionnaire design perspective which has not been done before. We find strong evidence that the SUS induces acquiescence bias and show how a robust alternative scale with statements rephrased as questions and relevant answer scales achieves higher sensitivity in measuring usability than the SUS.

Future work should investigate the strength of association between the SUS and the proposed alternative scale, and establish the latter's reliability, validity, and factor structure. While more work on evaluating the proposed alternative scale is needed, the authors recommend using a more robust 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. Groves, R. M., Singer, E., Lepkowski, J. M., Heeringa, S. G., and Alwin, D. F. Survey methodology.
7. Kirakowski, J., and Corbett, M. Sumi: The software usability measurement inventory. *British journal of educational technology 24*, 3 (1993), 210–212.
8. Kirakowski, J., and Dillion, A. The computer user satisfaction inventory. *Proceedings from the IEE: Evaluation Techniques for Interactive System Design, London, England* (1987).
9. Krosnick, J. A. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology 5* (1991), 213–236.
10. Krosnick, J. A. Survey research. *Annual review of psychology 50*, 1 (1999), 537–567.
11. Krosnick, J. A., and Alwin, D. F. A test of the form-resistant correlation hypothesis ratings, rankings, and the measurement of values. *Public Opinion Quarterly 52*, 4 (1988), 526–538.

Table 3. Original and reversed SUS items and updated SUS items with corresponding answer scales

#	Original SUS*	Reversed SUS*	Proposed Alternative	Proposed Answer Scale
1	I think that I would like to use this system frequentl	I do not think that I would like to use this system frequently	How much do you like or dislike the system?	{Extremely, Moderately, Slightly} dislike, Neither like nor dislike, {Slightly, Moderately, Extremely} like
2	I found the system unnecessarily complex	I found the system appropriately simple	How complex is the system?	{Not at all, Slightly, Moderately, Very, Extremely} complex
3	I thought the system was easy to use	I thought the system was hard to use	How easy or difficult is it to use the system?	{Extremely, Moderately, Slightly} difficult, Neither difficult nor easy, {Slightly, Moderately, Extremely} easy
4	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	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
5	I found the various functions in this system were well integrated	I found the various functions in this system were not well integrated	How integrated are the system's various functions?	{Not at all, Slightly, Moderately, Very, Extremely} integrated
6	I thought there was too much inconsistency in this system	I did not think there was too much inconsistency in this system	How consistent is the system?	{Not at all, Slightly, Moderately, Very, Extremely} consistent
7	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	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
8	I found the system very cumbersome to use	I found the system very manageable to use	How cumbersome is it to use the system?	{Not at all, Slightly, Moderately, Very, Extremely} cumbersome
9	I felt very confident using the system	I did not feel very confident using the system	How confident are you using the system?	{Not at all, Slightly, Moderately, Very, Extremely} confident
10	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	How much more is there to learn about the system?	Nothing at all, A little, A moderate amount, A lot, A great deal

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

12. Krosnick, J. A., and Fabrigar, L. R. Designing rating scales for effective measurement in surveys. *Survey measurement and process quality* (1997), 141–164.

13. Krosnick, J. A., and Presser, S. Question and questionnaire design. *Handbook of Survey Research. 2nd edition. Bingley, UK: Emerald* (2010), 263–314.

14. Lewis, J. R. Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the asq. *SIGCHI Bull. 23*, 1 (Jan. 1991), 78–81.

15. 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.

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

17. 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.

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

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

20. Saris, W. E., Krosnick, J. A., and Shaeffer, E. M. Comparing questions with agree/disagree response options to questions with construct-specific response options. *Unpublished manuscript, Political, Social, Cultural Sciences, University of Amsterdam* (2005).

21. Simon, H. A. Rational choice and the structure of the environment. *Psychological review 63*, 2 (1956), 129.

22. Smith, D. H. Correcting for social desirability response sets in opinion-attitude survey research. *The Public Opinion Quarterly 31*, 1 (1967), 87–94.

23. Tourangeau, R. *Cognitive science and survey methods*, vol. 73. National Academy Press Washington, 1984.