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 undoubtedly 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 scale proposals. Recent advances in questionnaire design research on acquiescence and other biases, however, now challenge some of the foundations of the SUS. In this note, we review literature on relevant questionnaire design aspects, inspect the SUS for biases, and using a survey experiment, show that the SUS is vulnerable to significant acquiescence bias. We then propose an alternative scale, strongly 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
System Usability Scale, SUS; response biases; usability evaluation; standardized questionnaires

ACM Classification Keywords
H.5.2. Information Interfaces and Presentation (e.g. HCI): User InterfacesEvaluation/Methodology

INTRODUCTION

With the increased focus on developing highly usable products, it has become highly important to actually measure the perceived usability of products 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) [13], the Software Usability Measurement Inventory (SUMI) [7], the Computer System Usability Questionnaires (CSUQ) [14], and the System Usability Scale (SUS) [3], among several 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 summarizing the usability assessment of the evaluated system. Over the years, the SUS has been used across a variety of different systems, including hardware, software, websites, applications, and even non-electronic products. 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 now challenge some of the foundations on which the SUS was developed. Most notably are research insights related to acquiescence bias [20, 12, 19], its relationship to satisficing [9, 10], and the use of scales with optimal lengths [11, 6].

The remainder of this note outlines such advances in questionnaire design research relevant to the original SUS, review the SUS in the context of those, propose an alternative version to conform with these insights, and finally compare the original SUS to the proposed alternative. Note that this note's intention is not to reduce the number of statements asked about in the SUS (as attempted by others [5, 16]), nor the identification and hence elimination of biases that do not relate to acquiescence. Instead, the goal of this note is to critique the wording of the statements and response options of the SUS on theoretical grounds with empirical evidence and propose an 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 [17], 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" (see Figure 1). 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.

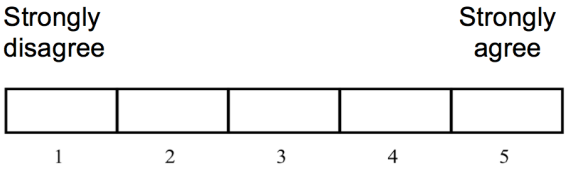

Figure 1. 5-point agreement scale as used in the SUS

One of the questionnaire biases that has been researched thoroughly is that of acquiescence bias, the insight from which are also highly relevant to the SUS. Acquiescence bias is the tendency of a respondent to be more likely to agree with a given statement independent of its substance [20]. Several aspects of questionnaire design contribute to the likelihood for acquiescence bias when responding to a questionnaire. First, when given a non-neutral statement, respondents are more likely to think of reasons why the statement is true, rather than expending cognitive effort to consider reasons for disagreement. This form of shortcutting the question answer process is often referred to as satisficing [9]. Second, respondents with lower self-perceived status assume the questionnaire administrator agrees with the posed statement, resulting in deferential agreement bias [19]. Similarly, respondents whose personality naturally skews towards agreeableness are more likely to suffer from acquiescence bias [19]. Finally, if the respondent's cognitive ability or motivation is lower, acquiescence bias is more likely [**?**].

Acquiescence bias is the strongest when presented with binary agree/disagree, yes/no, or true/false answer options [20]; however, similar effects have been shown for agreement scales (such as the Likert scale) [19]. To minimize this bias, research has proposed to refer to the underlying construct in a neutral, non-leading question and offer a neutral scale, instead of using statements with agreement scales [19]. Another suggestion to minimize this bias has been to use reverse-keyed constructs, i.e., the same construct is asked positively and negatively in the same survey, the raw scores of which are then combined to correct for acquiescence bias.

Picking the most valid and reliable rating scale for a given question has been another heavily researched topic, especially when moving away from agreement scales. 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, the scale length and its items depends on the nature of the construct being measured, i.e., if the construct is unipolar or bipolar in nature. Unipolar constructs range from zero to an extreme amount and are best measured on a 5-point scale, optimizing reliability while minimizing respondent burden [11]. 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 [11].

ACQUIESCENCE IN THE SUS

In this section, we inspect the original SUS for weaknesses in regards to acquiescence bias, based on questionnaire design research insights mentioned in the previous section. It also describes an experimental evaluation to identify the effects of acquiescence bias in the SUS. The item numbers of the SUS referred to in the remainder of this note correspond to those in Table 3.

Heuristic review

Each of the ten items of the SUS is constructed of a non-neutral statement and an agreement scale. In accordance with questionnaire design research [20], this particular design encourages the effects of acquiescence bias, i.e., the likelihood for the respondents to simply agree with the statement provided. As an example, let's review item 3 of the SUS: "I thought the system was easy to use." In this case, respondents are likely to expend unproportional effort on finding reasons that confirm that the system is actually easy to use, instead of thinking about aspects that would contribute to the system being interpreted as difficult to use. As a result, respondents are led towards a more agreeable response in line with "easy to use" as expressed in the given statement. As another example, let's take a look at item 9 in the original SUS ("I felt very confident using the system"). This statement leads the respondent towards higher confidence, as that's what the statement suggests and since the agreement scale increases the likelihood to agree that the system actually made them feel confident. Additionally, the nature of the agreement scale leads respondents further towards agreement as disagreeing with anyone requires courage and cognitive effort [19]. As the SUS is often administered at the end of usability studies, hence, not anonymously and with the questionnaire administrator present, this effect may be especially strong. As all other items of the SUS are phrased in the exact same way, this consideration applies equally across all of the SUS.

Notably, some of the items in the SUS have been reverse-keyed to ask about the same construct from opposite directions, as for example in items 3 ("easy to use") and 8 ("cumbersome to use"). This suggests that potential acquiescence bias effects may cancel each other out; however, this claim needs further evaluation. Nevertheless,

this reverse-keyed approach is used only for a subset of the items in the SUS, hence, all other items (1, 2, 4, 5, 6, 7, 9, 10) continue to suffer from significant acquiescence biases overall. While some of the items (e.g., 4 and 10) are reasonably similar, they are still asking about a slightly different underlying constructs and cannot simply used to cancel out biases.

The SUS' agreement scale is offered as a bipolar scale with 5 items. As agreement/disgreement is a bipolar scale in its nature, the most appropriate scale length should be 7 points, to maximize reliability and data differentiation [11]. However, when transforming the current agreement scale into construct-specific, neutral scales, the different scales and their lengths then depend on the nature of the underlying construct being measured, i.e., if the construct is unipolar or bipolar in nature [11].

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, 15, 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 alternative scale proposed below. A Cronbach's α of 0.86 reflects a good level of internal consistency, though the SUS is typically reported to have higher reliability [1, 15, 2]. A quick factor analysis yields two eigenvalues greater than one, which is consistent with previous work on the SUS's factor structure [15].

Acquiescence Bias Result

Statistic	Original SUS	Reversed SUS	Alternative Scale†
N	439	438	869
Range	[23, 100]	[8, 100]	[22, 100]
Mean (SD)	80.6 (16.1)	77.9 (14.2)	76.7 (14.7)
Median (IQR)	85 (20)	80 (18)	78 (22)
Cronbach α	0.86	0.73	0.67
$\lvert\lambda > 1\rvert$*	2	2	1

*number of eigenvalues greater than 1 in factor analysis
†proposed alternative items 3, 6, 9, 10 from Table 3

Table 1. Statistical and psychometric scale properties.

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.

	Original		Reversed		
Item #	M	SD	M	SD	p value
1	3.09	0.90	3.21	1	0.002
2	3.28	1.06	3.23	0.86	0.006
3	3.30	0.84	3.45	0.89	<0.001
4	3.54	0.92	3.08	1.23	<0.001
5	3.00	0.92	3.10	1.01	0.013
6	3.26	1.02	2.89	1.14	<0.001
7	3.12	0.90	3.09	0.95	0.868
8	3.10	1.18	3.33	0.84	0.173
9	3.29	0.89	3.45	0.91	<0.001
10	3.26	1.06	2.35	1.49	<0.001
overall	80.58	16.14	77.92	14.23	<0.001

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.

ALTERNATIVE PROPOSAL

Based on the review and evaluation of the original SUS presented above, this section now discusses a proposal for an alternative scale. While reasoning based on recent questionnaire design research is used to change the wording of the different SUS items as well as the response scales, an experiment discussed its quality and sensitivity as compared to SUS.

Proposed Wording Changes

To minimize acquiescence bias in the SUS, each of the statements may be transformed into a construct-specific, neutral question with similarly neutral answer options matching the question construct. The appropriate scale and its length will then depend on the nature of the construct, i.e., if it is unipolar or bipolar in nature. To explain the reasoning that let to the proposed changes, this section now discusses a few exemplary items from the SUS. The intention is that the same consideration can then be applied to all other items, as summarized in Table 3.

For example, to minimize acquiescence bias for item 9, we would first need to identify the underlying construct being asked about, which in this case is likely "confidence". The leading statement can then easily be transformed into a construct-specific, neutral question, such as "How confident were you using the system?". To determine the appropriate scale, it is critical to consider the polarity of "confidence." As confidence naturally starts from a zero point (the absence of confidence), spans to a positive extreme (high confidence), without a natural negative extreme (i.e., several levels of negative confidence), this construct is unipolar in nature. As a result, it should then be measured on a 5-point, fully-labeled scale from "Not at all confident" to "Extremely confident" (see full scale in Table 3).

As another example, item 3 refers to the bipolar construct of ease/difficulty. Bias may then be minimized by changing its wording to "How easy or difficult was it to use the system?", giving "easy" and "difficult" equal weight so respondents are less led into either direction. Due to the bipolarity of this construct, a 7-point, fully labeled scale from "Extremely difficult" to "Extremely easy" may be used (see full scale in Table 3). Note, as items 3 and 8 in the original SUS ask about the same underlying construct of ease/difficulty, they would result in the same reworded question, and hence, should only be included once in the full questionnaire.

Our goal was to propose alternative wording for the SUS items instead of creating a new measure, given that the SUS is probably the most established usability scale to date. Note that the purpose of this note is not an evaluation of the summarization of the questionnaire responses into a single score, hence, it is excluded from this discussion.

Experimental Evaluation

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 investigate how an example of the proposed alternative scale (items 3, 6, 9, 10 from Table 3) 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 [18] 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".

Psychometric Properties of the Alternative Scale

This note does not provide a thorough evaluation of the proposed alternative scale. Instead, Table 1 provides statistical and psychometric information from a subset of items of the alternative scale (items 3, 6, 9, 10) based on 869 responses for comparison with the SUS. Notably, the scales share similar distributional characteristics, but the alternative scale has lower reliability than the SUS and one rather than two-factor structure.

Scale Sensitivity Result

A Mann-Whitney test of the difference between the usability ratings for the two systems for each scale suggests that one the alternative scale example is more sensitive than the SUS (W=18585, p=0.069 for the SUS; W=17744, p=0.014 for the alternative scale). This result is based on 439 usability ratings of the low-usability system and 96 of the high-usability system. Although this comparison uses relatively small sample sizes and only four items from the proposed alternative scale, this finding could point at further potential benefits of using a more robust scale than the SUS.

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*

Table 3. Original and reversed SUS items and proposed alternative scale with corresponding answer scales.

#	Original SUS*	Reversed SUS*	Proposed Alternative	Proposed Answer Scale
1	I think that I would like to use this system frequently	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 was 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 was 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 was 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 were 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)

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

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

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

14. 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.

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

16. 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.

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

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

19. 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).

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