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 used 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 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 an experiment, show that it 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 measure their perceived usability. Especially during usability studies, large or small, standardized questionnaires are being used widely for such measurement. Commonly used such questionnaires include the Computer User Satisfaction Inventory (CUSI) [6], the Questionnaire for User Interface Satisfaction (QUIS) [4], the After Scenario Questionnaire (ASQ) [11], and the System Usability Scale (SUS) [3].

The SUS has received the highest level of adoption, 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 systems. However, since its inception in 1986, research regarding the design of valid and reliable questionnaires has advanced significantly, with insights that now challenge some of the foundations of the SUS. Most notably are insights related to acquiescence bias [16, 15], satisficing [7, 8], and scales lengths [9, 5].

The remainder of this note outlines such advances in questionnaire design relevant to the SUS, reviews the original SUS in the context of those, proposes an alternative version to conform with these insights, and finally compares the SUS to the proposed alternative. Note that this note's intention is not to reduce the number of statements asked about in the SUS, nor the identification and hence elimination of biases that do not relate to acquiescence. Instead, the goal 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

Although first published in 1996, the SUS was developed in 1986 by John Brooke at Digital Equipment Corporation (DEC). It was used as a "quick and dirty" scale administered after usability studies on electronic office systems, such as DEC's VT100, a text-based terminal system. The SUS measures attitudes regarding the effectiveness, efficiency, and satisfaction with a system. The SUS is comprised of ten statements (see 3 for their wording) which the respondent is asked to specify their level of agreement or disagreement on a symmetric 5-point Likert scale (as established by Rensis Likert in 1932 [13]). Its endpoints are labeled with "Strongly disagree" and "Strongly agree", while additionally all five scale items are numbered from 1 to 5 (see Figure 1). All statements of the SUS need to be evaluated by respondents. Responses are consolidated into a single score to represent the global usability assessment for that system and to enable cross-system comparisons.

One of the questionnaire biases researched extensively is that of acquiescence response bias: the tendency of a respondent to agree with a given statement independent of its substance [16]. When given a non-neutral statement, respondents are more likely to think of reasons why the statement is true, rather than affording cognitive effort

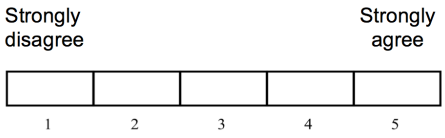

Figure 1. 5-point agreement scale as used in the SUS.

to consider reasons for disagreement. This form of short-cutting the question answer process is often referred to as satisficing [7]. Furthermore, respondents with lower self-perceived status assume the questionnaire administrator agrees with the posed statement, resulting in deferential agreement bias [15]. Similarly, respondents whose personality naturally skews towards agreeableness [15] and those with lower cognitive ability [10] are more likely to suffer from this bias. Acquiescence bias is the strongest when presented with binary agree/disagree, yes/no, or true/false answer options [16]; however, similar effects have been shown for agreement scales (such as the Likert scale) [15]. To minimize this bias, research has proposed to refer to the underlying construct in a neutral, non-leading question and offer a neutral scale [15].

Determining the current scale and the appropriate length has been another aspect subject to much research. First, fully labeled scale points compared to those that just use numbers improve reliability and reduce bias [5]. Second, the choice of scale length and labels depends on the nature of the construct being measured, i.e., if it is unipolar or bipolar in nature. Unipolar constructs range from zero to an extreme amount and are best measured on a 5-point scale, while bipolar constructs range from an extreme negative to an extreme positive with a natural midpoint and are best measured on a 7-point scale. These guidelines improve a scales reliability and maximizes data differentiation while minimizing the cognitive burden on respondents [9].

ACQUIESCENCE BIAS IN THE SUS

In this section, we review the SUS in regards to acquiescence bias, first on theoretical grounds and then through empirical evaluation.

Heuristic review

Each of the ten items of the SUS is constructed of a non-neutral statement and an agreement scale. In accordance with research [16], this particular design induces acquiescence. For instance, consider item 3 of the SUS (see Table 3): "I thought the system was easy to use." Respondents are likely to spend relatively more effort on finding reasons that confirm rather than contradict the statement. As a result, respondents tend to be more agreeable towards the "easy to use" description in the statement. Consider item 9 from the SUS as another example: "I felt very confident using the system". The loaded phrasing of this statement in combination with the endpoint-labeled agreement scale induces respondents to report higher levels of confidence than they would otherwise. Note also that this item confounds two dimensions by containing "very", because disagreement with feeling very confident is not equivalent to feeling unconfident, but could theoretically indicate extreme confidence. Additionally, the nature of the agreement scale leads respondents further towards agreement as disagreeing with anyone requires courage and cognitive effort [15]. This effect may be especially strong, as the SUS is frequently administered at the end of usability studies without anonymity and in the presence of the questionnaire administrator. These issues apply to all SUS items as they are phrased in the exact same way.

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"). It may be that potential acquiescence bias effects cancel each other out; however, this claim requires empirical 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) remain vulnerable to significant acquiescence bias. While some items (e.g., 4 and 10) are reasonably similar, they effectively tap into different underlying constructs and are thus unlikely to mitigate any biases.

The SUS' agreement scale is offered as a bipolar scale with five items. As agreement/disagreement is bipolar, the appropriate scale length would be seven points to maximize reliability and data differentiation [9].

Experimental Evaluation

To provide empirical support for the principal claim that the SUS is vulnerable to acquiescence bias in particular, we ran an online survey experiment with 877 participants who evaluated an online learning system.

Participants of a massive open online course were asked to complete an optional post-course survey. Respondents were randomly assigned to one of two groups: one received the original SUS (n=439) and the other the reversed SUS (n=438) (see Table 3 for scale details). The system respondents were asked to evaluate comprised of the course sites for browsing and watching lecture videos.

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 1 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	1.00	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 1. Original and reversed SUS scores providing strong evidence for 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 of Scale Sensitivity

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 post-course surveys of two online courses that ran on different systems (Web interfaces) 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 [14] to informally establish which system has better usability. While both systems showed generally high usability, one system was deemed superior in four evaluation categories. This informal usability comparison was the basis for labeling one system as having "high usability" and the other "low usability". We collected 439 alternative scale responses for the low usability system, and 96 for the high usability system.

A Mann-Whitney test of the difference between the usability ratings for the two systems for each scale suggests that the alternative scale example is more sensitive than the SUS (SUS: W=18585, p=0.069; alternative scale: W=17744, p=0.014). 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 neutral scale than the SUS.

Comparing Psychometric Properties

As the psychometric properties of the SUS have been studied extensively [1, 12, 2], a brief evaluation of key statistics should be sufficient here. Table 2 provides an overview of statistics and psychometric properties of the original and reversed SUS, and an example (items 3, 6, 9, 10) of the alternative scale proposed above. The scales share similar distributional characteristics, but the SUS has higher internal consistency than the alternative scale in terms of Cronbach's α. However, the SUS is 2.5 times longer than the example of the alternative scale which

increases reliability. A factor analysis yields two eigenvalues greater than one for the SUS, consistent with previous work on the SUS's factor structure [12].

Statistic	Original SUS	Reversed SUS	Alternative Scale$^{\text{item } \#3,6,9,10}$
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

Table 2. Statistical and psychometric scale properties.

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 an alternative scale with statements rephrased as questions and construct-specific, neutral answer scales achieve higher sensitivity in measuring usability than the SUS.

Future work could evaluate the SUS with respect to other biases, such as satisficing and the use of hypotheticals. Moreover, a thorough psychometric evaluation of the proposed alternative scale using a variety of systems is needed before we can recommend the adoption of this new scale. This will also require addressing the compatibility issue between old and new usability scores for the transition phase.

ACKNOWLEDGMENTS

We are grateful to the instructors of the two online 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. Groves, R. M., Singer, E., Lepkowski, J. M., Heeringa, S. G., and Alwin, D. F. Survey methodology.

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)

6. Kirakowski, J., and Dillion, A. The computer user satisfaction inventory. *Proceedings from the IEE: Evaluation Techniques for Interactive System Design, London, England* (1987).

7. Krosnick, J. A. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology 5* (1991), 213–236.

8. Krosnick, J. A. Survey research. *Annual review of psychology 50*, 1 (1999), 537–567.

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

10. Krosnick, J. A., Narayan, S., and Smith, W. R. Satisficing in surveys: Initial evidence. *New Directions for Evaluation 1996*, 70 (1996), 29–44.

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

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

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

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

15. 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).

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