Sample Paper for "how to read a scientific paper" (with no knowledge of statistics)

When Designing Usability Questionnaires, Does It Hurt to Be Positive?

Jeff Sauro
Oracle Corporation
1 Technology Way, Denver, CO 80237
jeff@measuringusability.com

James R. Lewis
IBM Software Group
8051 Congress Ave, Suite 2227
Boca Raton, FL 33487
jimlewis@us.ibm.com

ABSTRACT

When designing questionnaires there is a tradition of including items with both positive and negative wording to minimize acquiescence and extreme response biases. Two disadvantages of this approach are respondents accidentally agreeing with negative items (mistakes) and researchers forgetting to reverse the scales (miscoding).

The original System Usability Scale (SUS) and an all positively worded version were administered in two experiments (n=161 and n=213) across eleven websites. There was no evidence for differences in the response biases between the different versions. A review of 27 SUS datasets found 3 (11%) were miscoded by researchers and 21 out of 158 questionnaires (13%) contained mistakes from users.

Central research question

Is it true that there is an...

We found no evidence that the purported advantages of including negative and positive items in usability questionnaires outweigh the disadvantages of mistakes and miscoding. It is recommended that researchers using the standard SUS verify the proper coding of scores and include procedural steps to ensure error-free completion of the SUS by users.

Researchers can use the all positive version with confidence because respondents are less likely to make mistakes when responding, researchers are less likely to make errors in coding, and the scores will be similar to the standard SUS.

Author Keywords

Usability evaluation, standardized questionnaires, satisfaction measures, acquiescent bias, System Usability Scale (SUS)

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): User Interfaces–Evaluation/Methodology.

General Terms

Experimentation, Human Factors, Measurement, Reliability, Standardization, Theory

INTRODUCTION

Designers of attitudinal questionnaires (of which questionnaires that measure satisfaction with usability are one type) are trained to consider questionnaire response styles such as extreme response bias and acquiescence bias [17]. In acquiescence bias, respondents tend to agree with all or almost all statements in a questionnaire. The extreme response bias is the tendency to mark the extremes of rating scales rather than points near the middle of the scale. To the extent that these biases exist, the affected responses do not provide a true measure of an attitude. Acquiescence bias is of particular concern because it leads to an upward error in measurement, giving researchers too sanguine a picture of whatever attitude they are measuring.

A strategy commonly employed to reduce the acquiescent response bias is the inclusion of negatively worded items in a questionnaire [1], [2], [17]. Questionnaires with a mix of positive and negatively worded statements force attentive respondents to disagree with some statements. Under the assumption that negative and positive items are essentially equivalent and by reverse scoring the negative items, the resulting composite score should have reduced acquiescence bias.

More recently, however, there is evidence that the strategy of including a mix of positively and negatively worded items creates more problems than it solves [4]. Such problems include lowering the internal reliability [25], distorting the factor structure [19], [23], [22] and increasing interpretation problems with cross-cultural use [29].

The strategy of alternating item wording has been applied in the construction of most of the popular usability questionnaires, including the System Usability Scale (SUS) [6], SUMI [11], and QUIS [7]. The ASQ, PSSUQ and CSUQ [12], [13], [14] are exceptions, with all positive items.

The System Usability Scale is likely the most popular questionnaire for measuring attitudes toward system usability [14], [30]. Its popularity is due to it being free and short—with 10 items that alternate between positive and negative statements about usability (odd items positive, even items negative). It has also been the subject of some recent investigations [3], [15] [8], [9], which makes it a good candidate to manipulate to study whether the benefits outweigh the potential negatives of alternating item wording. The ten traditional SUS items are shown in Exhibit 1.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The response options, arranged from the left to right, are Strongly Disagree (1) to Strongly Agree (5).

Exhibit 1: The System Usability Scale (SUS).

The proper scoring of the SUS has two stages:

1. Subtract one from the odd numbered items and subtract the even numbered responses from 5. This scales all values from 0 to 4 (with four being the positive response).
2. Add up the scaled items and multiply by 2.5 (to convert the range of possible values from 0 to 100 instead of from 0 to 40).

Previous research on the SUS

Much of the research applied to the design of rating scales for usability attitudes comes from disciplines other than usability, typically marketing and psychology. In other fields, items can include more controversial or ambiguous topics than is typical of system usability. Although many findings should still apply to usability questionnaire design, it is of value to the design of future usability questionnaires to review research related specifically to the analysis of rating scales used in usability—especially the SUS.

Bangor et al. [3] analyzed a large database of SUS questionnaires (over 2300) and found that participants tended to give slightly higher than average ratings to most of the even numbered statements (negatively phrased items 4, 6, 8 and 10), and also tended to give slightly lower than average ratings to most of the odd numbered statements (positively phrased items: 1, 2, 3, 5 and 9). This suggests participants tended to agree slightly more with negatively worded items and to disagree slightly more with positively worded items (from this point on, referred to as positive and negative items). The magnitude of the difference was modest, with the average absolute deviation from the average score of .19 of a point and the highest deviation on item 4 ("I think that I would need the support of a technical person to be able to use this system.") – a deviation of .47 of a point.

Finstad [9] compared a 7-point version to the original 5-point version of SUS and found users of enterprise systems "interpolated" significantly more on the 5-point version than on the 7-point version; however, there was no investigation on the effects of changing item wording. Based on difficulties observed with non-native speakers completing the SUS, Finstad [8] recommended changing the word "cumbersome" in Item 8 to "awkward" – a recommendation echoed in [3] and [15].

In 2008 a panel at the annual Usability Professionals Association conference entitled "Subjective ratings of usability: Reliable or ridiculous?" was held [10]. On the panel were two of the originators of the QUIS and SUMI questionnaires. As part of the panel presentation, an experiment was conducted on the effects of question wording on SUS scores to investigate two variables: item intensity and item direction (for details see [21]). For example, the extreme negative version of the SUS Item 4 was "I think that I would need a permanent hot-line to the help desk to be able to use the web site."

Volunteer participants reviewed the UPA website. After the review, participants completed one of five SUS questionnaires -- an all positive extreme, all negative extreme, one of two versions of an extreme mix (half positive and half negative extreme), or the standard SUS questionnaire (as a baseline). Sixty-two people in total participated, providing between 10-14 responses per condition. Even with this relatively small sample size the extreme positive and extreme negative items were significantly different from the original SUS ($F(4,57) = 6.90$, $p < .001$) and represented a large effect size (Cohen d >1.1).

These results were consistent with one of the earliest reports of attitudes in scale construction [27]. Research has shown that people tend to agree with statements that are close to their attitude and disagree with all other statements [24].

By rephrasing items to extremes, only respondents who passionately favored the usability of the UPA website tended to agree with the extremely phrased positive statements—resulting in a significantly **lower** average

score. Likewise, only respondents who passionately disfavored the usability agreed with the extremely negatively questions—resulting in a significant **higher** average score.

The results of this experiment confirmed that extreme intensity can affect item-responses towards attitudes of usability, so designers of usability questionnaires should avoid such extreme items. Due to the confounding of item intensity and direction, however, the results do not permit making claims about the effects of alternating positive and negatively worded items.

Advantages for alternating question items

The major impetus for alternating scales is to control acquiescent response bias (including the potential impression that having only positive items may lead respondents to think you want them to like the system under evaluation – John Brooke, personal communication, 8/2010). The alternation of positive and negative items also provides protection against serial extreme responders – participants who provide all high or all low ratings – a situation that could be especially problematic for remote usability testing. For example, when items alternate, responses of all 1's make no sense. When items do not alternate, responses of all 1's could represent a legitimate set of responses.

empathy with designer

Disadvantages for alternating question items

Despite the potential advantages of alternation, we consider three major potential disadvantages.

1. **Misinterpret**: Users may respond differently to negatively worded items such that reversing responses from negative to positive doesn't account for the difference. As discussed in the previous section, problems with misinterpreting negative items include creating an artificial two-factor structure and lowering internal reliability, especially in cross-cultural contexts.

2. **Mistake**: Users might not intend to respond differently, but may forget to reverse their score, accidentally agreeing with a negative statement when they meant to disagree. We have been with participants who have acknowledged either forgetting to reverse their score or commenting that they had to correct some scores because they forgot to adjust their score.

3. **Miscode**: Researchers might forget to reverse the scales when scoring, and would consequently report incorrect data. Despite there being software to easily record user input, researchers still have to remember to reverse the scales. Forgetting to reverse the scales is not an obvious error. The improperly scaled scores are still acceptable values, especially when the system being tested is of moderate usability (in which case many responses will be neutral or close to neutral).

A researcher may only become aware of coding errors after subjecting the results to internal reliability testing and obtaining a negative Cronbach's alpha – a procedure that few usability practitioners routinely practice. In fact, this problem is prevalent enough in the general practice of questionnaire development that the makers of statistical software (SPSS) have included it as a warning "The [Cronbach's alpha] value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings" [26].

We are able to estimate the prevalence of the miscoding error from two sources. First, in 2009, eight of 15 teams used the SUS as part of the Comparative Usability Evaluation-8 (CUE-8) workshop at the Usability Professionals Association annual conference [18]. Of the eight teams, one team improperly coded their SUS results. Second, as part of an earlier analysis of SUS, we [15] examined 19 contributed SUS datasets. Two were improperly coded and needed to be recoded prior to analysis.

Considering these two sources, three out of 27 SUS datasets (11.1%) had negative items that weren't reversed. Assuming this to be a reasonably representative selection of the larger population of SUS questionnaires, we can be 95% confident that miscoding affects between 3% and 28% of SUS datasets. Hopefully, future research will shed light on whether this assumption is correct.

Research gap

Despite published concerns about acquiescence bias, there is little evidence that the "common-wisdom" of including both positive and negatively worded items solves the problem. To our knowledge there is no research documenting the magnitude of acquiescence bias in general, or whether it specifically affects the measurement of attitudes toward usability.

H1

H2

The goals of this paper are to determine whether an acquiescence bias exists in responses to the SUS, and if so, how large is it? Does the alternating wording of the SUS provide protection against acquiescence and extreme response biases? Further, does its alternating item wording outweigh the negatives of misinterpreting, mistaking and miscoding?

links to central research question

METHODS

We conducted two experiments to look for potential advantages and disadvantages of reversing items in questionnaires.

Experiment 1

In April 2010, 51 users (recruited using Amazon's Mechanical Turk) performed two representative tasks on one of four websites (Budget.com, Travelocity.com, Sears.com and Bellco.com). Examples of the tasks include making reservations for a car or flight, locating items to purchase, finding the best interest rate and locating store hours and locations.

At the end of the test users answered the standard 10 item SUS questionnaire. There were between 12 and 15 users for each website.

In August 2010, a new set of 110 users (again recruited using Amazon's Mechanical Turk) performed the same tasks on one of four websites tested four months earlier. There were between 20 and 30 users for each website. The testing protocol was the same except the new set of users completed a positive-only version of the SUS as shown in Exhibit 2. Note that other than replacing "system" with "website", the odd items are identical to those of the standard SUS but the even items are different – worded positively rather than negatively.

1. I think that I would like to use the website frequently.
2. I found the website to be simple.
3. I thought the website was easy to use.
4. I think that I could use the website without the support of a technical person.
5. I found the various functions in the website were well integrated.
6. I thought there was a lot of consistency in the website.
7. I would imagine that most people would learn to use the website very quickly.
8. I found the website very intuitive.
9. I felt very confident using the website.
10. I could use the website without having to learn anything new.

Response options appeared from the left to right anchored with Strongly Disagree 1 to Strongly Agree 5.

Exhibit 2: A positively worded SUS questionnaire.

Results of Experiment 1

Both samples contained only respondents from the US, with no significant differences in average age (32.3 and 32.2; $t(81) = .04$, $p > .95$), gender (57% and 56% female; $\chi2(1) = .003$, $p > .95$) or highest degree obtained (63% and 59% with college degrees $\chi2(3) = 1.54$, $p > .67$) and prior experience with the sites (63% and 54% had no prior experience $\chi2(1) = 1.17$, $p > .27$).

The internal reliability of both versions of the questionnaires was high and nearly identical. For the original SUS questions with 51 cases Cronbach's alpha was .91. For the positively worded SUS with 110 cases Cronbach's alpha was .92.

To look for an overall effect between questionnaire types, we conducted a t-test using the scaled SUS scores, the average of the evenly numbered items, and the average of the odd-numbered items. The means and standard deviations appear in Tables 1-3.

Questionnaire	Mean	SD	N
Normal SUS	73.4	17.6	51
Positive SUS	77.1	17.1	110

Table 1: SUS Scaled scores for four websites ($p > .20$).

Questionnaire	Mean	SD	N
Normal SUS	3.25	.70	51
Positive SUS	3.21	.66	110

Table 2: Even number items for four websites ($p > .74$).

Questionnaire	Mean	SD	N
Normal SUS	2.62	.79	51
Positive SUS	2.97	.75	110

Table 3: Odd numbered items ($p < .02$).

The difference between questionnaires was not statistically significant for scaled SUS scores ($t(95) = -1.25$, $p > .20$) or for the average of the even items ($t(92) = 0.33$, $p > .74$). There was a significant difference for the odd items ($t(93) = -2.61$ $p < .02$) – the items not changed between versions of the questionnaire.

There was a statistically significant difference between the odd and even-numbered items for the original SUS questionnaire ($t(98) = 4.26$, $p < .001$) and the all positive SUS questionnaire ($t(214) = 2.60$, $p < .02$), suggesting the even items elicit different responses than the odd items in both questionnaires. Furthermore, a repeated-measures ANOVA with odd/even as a within subjects variable and questionnaire type as a between subjects variable revealed a

significant interaction between odd/even questions and questionnaire type ($F(1,159) = 32.4$, $p < .01$), as shown in Figure 1.

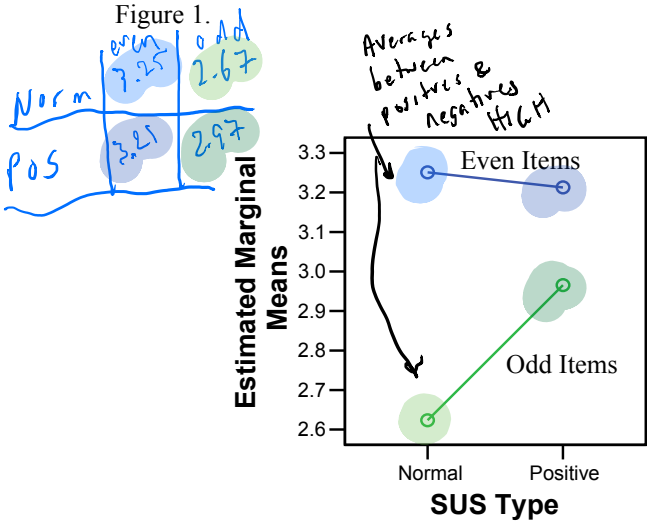

Figure 1: Even/odd by n/positive interaction from Experiment 1 (asynchronous data collection).

H1: Is there a difference in acquiescence bias between groups?

Acquiescence Bias

Operational (active) variable

To assess acquiescence bias, we counted the number of agreement responses (4 or 5) to the odd numbered (consistently and positively worded) items in both questionnaires. The mean number of agreement responses was 3.2 per questionnaire for the standard SUS ($SD = 1.67$, $n = 51$) and 3.69 for the positive version ($SD = 1.46$, $n = 110$). The positive questionnaire had a slightly higher average number of agreements than the standard, although the difference was only marginally significant ($t(86) = -1.82$, $p > .07$.

H2: Is there a difference in extreme response bias between groups?

Extreme Response Bias

active variable

To measure extreme response bias, we counted the number of times respondents provided either the highest or lowest response option (1 or 5) for both questionnaire types for all items. The mean number of extreme responses was 3.45 for the standard SUS ($SD = 2.86$, $n = 51$) and 4.09 for the positive version ($SD = 3.29$, $n = 110$), a nonsignificant difference ($t(111) = -1.26$, $p > .21$).

Mistakes

We used two approaches to assess the magnitude of the potential mistake problem. First, we looked for internal inconsistencies within questionnaires by comparing the number of times respondents agreed (responses of 4 and 5) to negatively worded items and also agreed to positively worded items (responses of 4 and 5)—an approach similar to [28]. We considered a questionnaire to contain mistakes if there were at least 3 responses indicating agreement to positively and negatively worded items or 3 responses with disagreement to positively and negatively worded items. We found 3 such cases (5.8%, 95% CI ranging from 1.4% to 16.5%).

Our second approach was to examine responses to the most highly correlated negative and positive item which, according to [3]'s large SUS dataset were items 2 and 3 ($r = -.593$). The correlation between those items from this experiment was also high and significant ($r = -.683$, $p < .01$, $n = 51$). For this assessment, we counted the number of times respondents provided a response of a 4 or 5 to both items 2 and 3. There were 18 such cases (35.3%, 95% CI ranging from 23.6% to 49.1%).

Experiment 1 Discussion

The overall SUS scores between the standard and all positive versions of the SUS were not significantly different, which suggests that changing the wording of the items in this way does not appear to have a strong effect on the resulting SUS measurements. There was no difference in the even numbered item averages (the ones changed in the positive only questionnaire). However, the odd-numbered item averages (the ones NOT changed in this experiment) were significantly different, with slightly lower scores for positive and slightly higher scores for the negative versions of the items.

To say the least, we did not expect this result, and found it difficult to explain. In examining the difference by website, the bulk of the differences came from two of the four websites (Travelocity.com and Sears.com). Investigating systems in the wild can be tricky because researchers have no control over the timing of system changes (for example, see [20], reanalyzed in [16]). It is possible that changes to the websites somehow affected only the odd numbered questions, but that is pure speculation. To minimize the potential confounding effects of website changes and item wording, we conducted another experiment with the questionnaires tested simultaneously rather than asynchronously. We also planned for a larger sample size to increase the power of the experiment.

Experiment 2 Mostly, repetition of first experiment

In August and September 2010, 213 users (recruited using Amazon's Mechanical Turk) performed two representative tasks on one of seven websites (third party automotive or primary financial services websites: Cars.com, Autotrader.com, Edmunds.com, KBB.com, Vanguard.com, Fidelity.com and TDAmeritrade.com). The tasks included finding the best price for a new car, estimating the trade-in value of a used-car and finding information about mutual funds and minimum required investments. At the end of the study users randomly completed either the standard or the positively worded SUS. There were between 15 and 17 users for each website and questionnaire type. Why so few?

Results of Experiment 2

Both samples contained only respondents from the US. There were no significant differences in average age (32.3 and 31.9; $t(210) = .26$, $p > .79$), gender (62% and 58%

female; $\chi 2(1) = .38$, $p > .53$) or highest degree obtained (58% and 63% with college degrees $\chi 2(3) = 4.96$, $p > .17$) and prior experience with the sites (64% and 66% had no prior experience, $\chi 2(1) = .144$, $p > .70$).

The internal reliability of both questionnaires was high – Cronbach's alpha of .92 (n = 107) for the original and .96 (n = 106) for the positive.

To look for an overall effect between questionnaire types, we conducted a t-test using the scaled SUS scores, the average of the evenly numbered items and the average of the odd-numbered items. The means and standard deviations appear in Tables 4-6.

Questionnaire	Mean	SD	N
Normal SUS	52.2	23.3	107
Positive SUS	49.3	26.8	106

Table 4: SUS Scaled scores for seven websites (p > .39).

Questionnaire	Mean	SD	N
Normal SUS	2.30	1.04	107
Positive SUS	2.15	1.09	106

Table 5: Even number items for four websites (p>.27).

Questionnaire	Mean	SD	N
Normal SUS	1.88	.97	107
Positive SUS	1.79	1.11	106

Table 6: Odd numbered items (p > .54).

The questionnaires were not statistically significant for scaled SUS scores ($t(206) = 0.85$, $p > .39$), the average of the even items ($t(210) = 1.09$, $p > .27$), or the average of the odd items ($t(206) = 0.60$, $p > .54$).

There continued to be a statistically significant difference between the odd and even-numbered items for the original SUS questionnaire ($t(210) = 3.09$, $p < .01$) and the all positive SUS questionnaire ($t(209) = 2.32$, $p < .03$).

In contrast to Experiment 1, a repeated-measures ANOVA with odd/even as a within subjects variable and questionnaire type as a between subjects variable indicated no significant interaction between odd/even questions and questionnaire type ($F(1, 211) = .770$, $p > .38$), as shown in Figure 2.

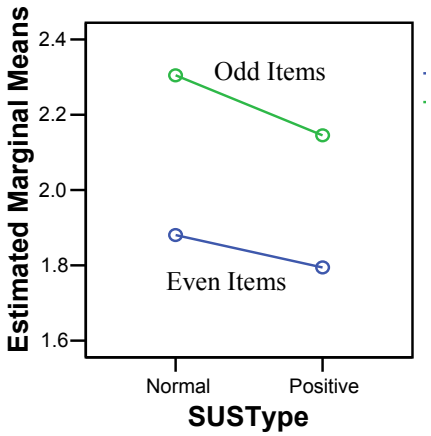

Figure 2: Interaction plot between odd/even questions and questionnaire type show no significant difference by questionnaire type (synchronous data collection).

Mistakes

As in Experiment 1, we assessed (1) the magnitude of mistaken responses: internal inconsistencies in at least 3 questions, and (2) the consistency of responses to items 2 and 3. We found 18 of the 107 original SUS questionnaires contained at least 3 internal inconsistencies (16.8%, 95% CI between 10.8% and 25.1%) and 53 questionnaires with inconsistent responses for items 2 and 3 (49.5%, 95% CI between 40.2% and 58.9%).

Acquiescence Bias

To assess acquiescence bias, we counted the number of agreement responses (4 or 5) to the odd numbered (consistently and positively worded) items in both questionnaires. The mean number of agreement responses was 1.64 per questionnaire for the standard SUS (SD = 1.86, n = 107) and 1.66 for the positive version (SD = 1.87, n = 106). There was no significant difference between the questionnaire versions ($t(210) = -.06$, $p > .95$).

Extreme Response Bias

To measure extreme response bias, we counted the number of times respondents provided either the highest or lowest response option (1 or 5) for both questionnaire types for all items. The mean number of extreme responses was 1.68 for the standard SUS (SD = 2.37, n = 107) and 1.36 for the positive version (SD = 2.23, n = 106), a nonsignificant difference ($t(210) = 1.03$, $p > .30$).

Factor Analysis of the Questionnaires

Finally, we compared the factor structures of the two versions of the SUS with the SUS factor structure reported in [15], based on the large sample of SUS questionnaires collected by [3] (and replicated by [5]). The key finding from the prior factor analytic work on the SUS was that the SUS items clustered into two factors, with one factor containing items 1, 2, 3, 5, 6, 7, 8, and 9, and the other factor containing items 4 and 10.

As shown in Table 7, neither of the resulting alignments of items with factors exactly duplicated the findings with the large samples of the SUS, and neither were they exactly consistent with each other, with discrepancies occurring on items 6, 8, and 9. Both the original and positive versions were consistent with the large-sample finding of including items 4 and 10 in the second factor. The original deviated slightly more than the positive from the large-sample factor structure (original items 6 and 8 aligned with the second rather than the first factor; positive item 9 aligned with the second rather than the first factor).

Items	Original Factor 1	Positive Factor 1	Original Factor 2	Positive Factor 2
1	**.784**	**.668**	.127	.300
2	**.594**	**.832**	.555	.437
3	**.802**	**.834**	.375	.488
4	.194	.301	**.812**	**.872**
5	**.783**	**.826**	.243	.362
6	.319	**.815**	**.698**	.230
7	**.763**	**.734**	.322	.467
8	.501	**.776**	**.688**	.404
9	**.599**	.619	.518	**.661**
10	.193	.419	**.865**	**.811**
% Var	35.9%	47.8%	32.7%	29.4%

Table 7: Two-factor structures for the standard and positive versions of the SUS (synchronous data collection)

DISCUSSION

The results of Experiments 1 and 2 were reasonably consistent, other than the bizarre outcome in Experiment 1 in which the unchanged items had significantly different values as a function of the SUS version (standard vs. positive). Because that finding did not replicate in Experiment 2, it was very likely a consequence of having collected the data asynchronously. It could be that the websites changed or the type of users who participated were in some way different.

In almost every way, the data collected in Experiment 2 with the standard and positive versions of the SUS were similar, indeed, almost identical. There were no significant differences in total SUS scores or the odd or even averages. Cronbach's alpha for both versions was high (> .90). The differences in the factor structures (both with each other and in comparison to published large-sample evaluations) were very likely due to the relatively small sample sizes. There was little evidence of any differences in acquiescence or extreme response biases between the original SUS questionnaire and the all positive version in either experiment.

Using the more conservative of the two estimation methods for mistaken responses, there were 3 out of 51 from Experiment 1 and 18 out of 107 in Experiment 2 for a total of 21 out of 158 questionnaires which contained at least 3 internal inconsistencies. This would suggest 13.3% (95% CI between 8.8% and 19.5%) of SUS questionnaires administered remotely contain mistakes. For miscoding errors, three out of 27 SUS datasets (11%; 95% CI between 3.0% and 28.8%) were improperly coded resulting in incorrect scoring.

We did not find any evidence for a strong acquiescence or extreme response bias. Even if strong evidence were found, the recommendation by [4] to reverse scale responses instead of item wording would not correct the problems of mistakes and miscoding. The data presented here suggest the problem of users making mistakes and researchers miscoding questionnaires is both real and much more detrimental than response biases.

Interesting finding: Do any later papers dispute or contain evidence to the contrary?

CONCLUSION

There is little evidence that the purported advantages of including negative and positive items in usability questionnaires outweigh the disadvantages. This finding certainly applies to the SUS when evaluating websites using remote-unmoderated tests. It also likely applies to usability questionnaires with similar designs in unmoderated testing of any application. Future research with a similar experimental setup should be conducted using a moderated setting to confirm whether these findings also apply to tests when users are more closely monitored.

Key finding

Researchers interested in designing new questionnaires for use in usability evaluations should avoid the inclusion of negative items.

Researchers who use the standard SUS have no need to change to the all positive version provided that they verify the proper coding of scores. In moderated testing,

researchers should include procedural steps to ensure error-free completion of the SUS (such as when debriefing the user).

In unmoderated testing, it is more difficult to correct the mistakes respondents make, although it is reassuring that despite these inevitable errors, the effect is unlikely to have a major impact on overall SUS scores.

Researchers who do not have a current investment in the standard SUS can use the all positive version with confidence because respondents are less likely to make mistakes when responding, researchers are less likely to make errors in coding, and the scores will be similar to the standard SUS.

REFERENCES

1. Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.
2. Anderson, A. B., Basilevsky, A., & Hum, D. P. J. (1983). Measurement: Theory and techniques. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of Survey Research* (pp. 231-187). New York, NY: Academic Press.
3. Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, *6*, 574-594.
4. Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60(3)*, 361–370.
5. Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the System Usability Scale: A test of alternative measurement models. *Cognitive Processes*, *10*, 193-197.
6. Brooke, J. (1996). SUS: A quick and dirty usability scale. In P.W. Jordan, B. Thomas, B.A. Weerdmeester & I.L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189-194). London, UK: Taylor & Francis.
7. Chin, J. P, Diehl, V. A., & Norman, K. (1988). Development of an instrument measuring user satisfaction of the human–computer interface. In *Conference on Human Factors in Computing Systems* (pp. 213–218). New York, NY: Association for Computing Machinery.
8. Finstad, K. (2006). The System Usability Scale and non-native English speakers. *Journal of Usability Studies*, *1*, 185–188.
9. Finstad, K. (2010). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies*, *5(3)*, 104-110.
10. Karn, K., Little, A., Nelson, G., Sauro, J. Kirakowski, J., Albert, W. & Norman, K., (2008) Subjective Ratings of Usability: Reliable or Ridiculous? Panel Presentation at the Usability Professionals Association (UPA 2008) Conference Baltimore, MD.
11. Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology, 24,* 210–212.
12. Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction*, *7*, 57-78.
13. Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14(3&4)*, 463-488.
14. Lewis, J. R (2006). Usability testing. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (3rd ed.) (pp. 1275-1316). New York, NY: John Wiley.
15. Lewis, J. R., & Sauro, J. (2009). The factor structure of the System Usability Scale. In M. Kurosu (Ed.), *Human Centered Design, HCII 2009* (pp. 94–103). Berlin, Germany: Springer-Verlag.
16. Lewis, J. R. (2011). *Practical speech user interface design*. Boca Raton, FL: Taylor & Francis.
17. Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
18. Molich, R., Kirakowski, J., Sauro, J., & Tullis, T., (2009) Comparative Usability Task Measurement Workshop (CUE-8) at the UPA 2009 Conference in Portland, OR.
19. Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement*, *50*, 603–610.
20. Ramos, L. (1993). The effects of on-hold telephone music on the number of premature disconnections to a statewide protective services abuse hot line. *Journal of Music Therapy*, *30(2)*, 119-129.
21. Sauro, J. (2010). That's the worst website ever!: Effects of extreme survey items. www.measuringusability.com/blog/extreme-items.php (last viewed 9/23/2010).
22. Schmitt N., & Stuits, D. (1985) Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9(4)*, 367-373.

23. Schriesheim, C.A., & Hill, K.D. (1981). Controlling acquiescence response bias by item reversals: the effect on questionnaire validity. *Educational and Psychological Measurement*, *41(4)*, 1101–1114.
24. Spector, P., Van Katwyk, P., Brannick, M., & Chen, P. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management*, *23(5)*, 659-677.
25. Stewart, T. J., & Frye, A. W. (2004). Investigating the use of negatively-phrased survey items in medical education settings: Common wisdom or common mistake? *Academic Medicine*, *79(10 Supplement)*, S1–S3.
26. SPSS. (2003). *SPSS for Windows, Rel. 12.0.1*. (2003). Chicago, IL: SPSS Inc.
27. Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529-554.
28. Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: the number of response categories and response category labels. *International Journal of Research in Marketing*, 27, 236-247.
29. Wong, N., Rindfleisch, A., & Burroughs, J. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the materia**l** values scale. *Journal of Consumer Research*, *30*, 72-91.
30. Zviran, M., Glezer, C., & Avni, I. (2006). User satisfaction from commercial web sites: The effect of design and use. *Information & Management*. *43*, 157-178.