Journal of Usability Studies

Vol. 13, Issue 1, November 2017 pp. 38–46

Can I Leave This One Out? The Effect of Dropping an Item From the SUS

James R. Lewis
Senior HF Engineer
IBM Corp.
5901 Broken Sound Parkway
Suite 514C
Boca Raton, FL 33487
USA
jimlewis@us.ibm.com

Jeff Sauro
MeasuringU
Principal
jeff@measuringu.com

Abstract

There are times when user experience practitioners might consider using the System Usability Scale (SUS), but there is an item that just doesn't work in their context of measurement. For example, the first item is "I think I would like to use this system frequently." If the system under study is one that would only be used infrequently, then there is a concern that including this item would distort the scores, or at best, distract the participant. The results of the current research show that the mean scores of all 10 possible nine-item variants of the SUS are within one point (out of a hundred) of the mean of the standard SUS. Thus, practitioners can leave out any one of the SUS items without having a practically significant effect on the resulting scores, as long as an appropriate adjustment is made to the multiplier (specifically, multiply the sum of the adjusted item scores by 100/36 instead of the standard 100/40, or 2.5, to compensate for the dropped item).

Keywords

System Usability Scale, SUS, custom nine-item version

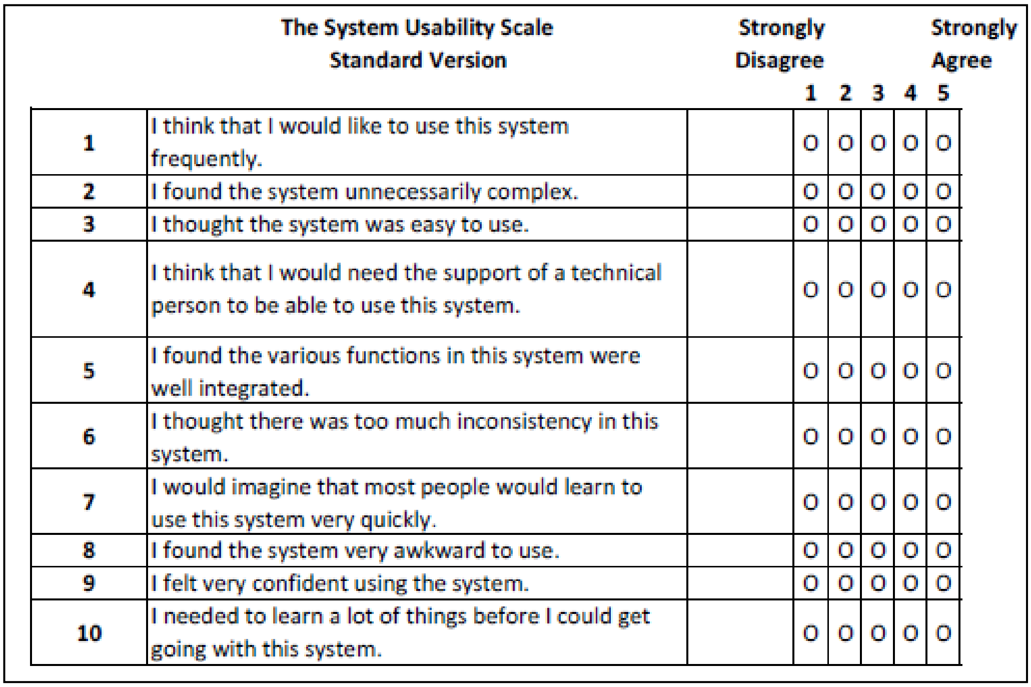
Introduction

In this section, we give a brief overview of the SUS as well as its psychometric properties and flexibility. We also discuss an issue practitioners have encountered when deciding how to use the SUS for their study and a possible solution for that issue.

What Is the SUS?

The System Usability Scale (SUS; Brooke, 1996) is a very popular (if not the most popular) standardized questionnaire for the assessment of perceived usability. Sauro and Lewis (2009), in a study of unpublished industrial usability studies, found that the SUS accounted for 43% of post-test questionnaire usage. It has been cited in over 1,200 publications (Brooke, 2013).

As shown in Figure 1, the standard version of the SUS has 10 items, each with five steps anchored with "Strongly Disagree" and "Strongly Agree." It is a mixed-tone questionnaire in which the odd-numbered items have a positive tone and the even-numbered items have a negative tone. The first step in scoring a SUS is to determine each item's score contribution, which will range from 0 (a poor experience) to 4 (a good experience). For positively-worded items (odd numbers), the score contribution is the scale position minus 1 (for example, a rating of 4 on an odd-numbered item would have a score contribution of 3). For negatively-worded items (even numbers), the score contribution is 5 minus the scale position (for example, a rating of 2 on an even-numbered item would have a score contribution of 3). To get the overall SUS score, multiply the sum of the item score contributions by 2.5, which produces a score that can range from 0 (very poor perceived usability) to 100 (excellent perceived usability) in 2.5-point increments.



	The System Usability Scale Standard Version	Strongly Disagree					Strongly Agree
			1	2	3	4	5
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 awkward 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.		○	○	○	○	○

Figure 1. The standard System Usability Scale. Note: Item 8 shows "awkward" in place of the original "cumbersome."

Psychometric Properties

The SUS has excellent psychometric properties. Research has consistently shown the SUS to have reliabilities at or just over 0.90 (Bangor, Kortum, & Miller, 2008; Lewis, Brown, & Mayes, 2015; Lewis & Sauro, 2009; Lewis, Utesch, & Maher, 2015), far above the minimum criterion of 0.70 for measurements of sentiments (Nunnally, 1978). The SUS has also been shown to have acceptable levels of concurrent validity (Bangor, Joseph, Sweeney-Dillon, Stettler, & Pratt, 2013; Bangor et al., 2008; Kortum & Peres, 2014; Lewis, Brown, et al., 2015; Peres, Pham, Philips, 2013) and sensitivity (Kortum & Bangor, 2013; Kortum & Sorber, 2015; Lewis & Sauro, 2009; Tullis & Stetson, 2004). Norms (tables of percentiles based on a large number of user research studies that included the SUS) are available to guide the interpretation of the SUS (Bangor, Kortum, & Miller, 2008, 2009; Sauro, 2011; Sauro & Lewis, 2016).

Flexibility

The SUS has turned out to be a very flexible questionnaire. Since its initial publication, some researchers have proposed minor changes to the wording of the items. For example, Finstad (2006) and Bangor et al. (2008) recommended replacing "cumbersome" with "awkward" in Item 8. The original SUS items refer to "system," but substituting the word "website" or "product," or using the actual website or product name seems to have no effect on the resulting scores (Lewis & Sauro, 2009; these types of substitutions should be consistent across the items within a study). SUS scores did not appear to be significantly affected even when the even items were rewritten with a positive tone (Sauro & Lewis, 2011). In addition to its use as a post-test questionnaire for the assessment of perceived usability, the SUS is also useful for the retrospective evaluation of products and services using surveys (Grier, Bangor, Kortum, & Peres, 2013).

What Is the Problem?

There are times when user experience practitioners might consider using the System Usability Scale (SUS), but there is an item that just doesn't work in their context of measurement. For example, the first item is "I think I would like to use this system frequently." If the system under study is one that would only be used infrequently, then there is a concern that including this item would at worst distort the scores and at best would confuse participants. We have both been asked by practitioners if it would be acceptable to replace a problematic SUS item with a different item (Sauro, 2016) or to just leave it out.

However, when you use a standardized questionnaire created and assessed using classical test theory, you are, strictly speaking, not supposed to modify the questionnaire, at least, not without conducting the necessary research to investigate how the changes affect its properties. As mentioned above, there is evidence that the SUS is a flexible measuring instrument, but there has been no systematic research on the effect of substituting or dropping SUS items.

A Potential Solution

Even though there has not yet been any systematic research on the effect of substituting or dropping SUS items, some research has found that missing data from standardized usability questionnaires had little or no effect on the resulting scores (Lewis, 2002; Lah & Lewis, 2016). It would be challenging to substitute new items for existing SUS items because each change would require extensive research with new data collected with the substituted item. A more promising approach would be to systematically examine the effect of dropping a problematic item—an approach that can be performed using historical SUS data without the need to collect new data.

In his original paper, Brooke (1996) reported strong correlations among the SUS items (which had absolute values of *r* ranging from 0.7 to 0.9). Given such high correlations among the 10 SUS items, it seems likely that dropping any individual item should have little effect on the resulting score. The resulting 10 different nine-item versions of the SUS would be expected to correlate almost perfectly with the standard 10-item version. With an appropriate adjustment to the SUS multiplier, it would be reasonable to expect very little deviation from the overall mean of the standard SUS.

To understand how to adjust the SUS multiplier, consider how the standard multiplier works. The process of determining score contributions described above results in a score that, without multiplication, would range from 0 to 40 (a maximum score contribution of 4 multiplied by 10 items). To stretch that out so it ranges from 0 to 100, you need to multiply the sum of the score contributions by 100/40, which is where the "2.5" multiplier comes from.

If you keep the scoring system the same but drop one item, the score contributions can range from 0 to 36. To stretch this out so it ranges from 0 to 100, you need to multiply the sum of the score contributions by 100/36 (which resolves to a repeating decimal of 2.7777…).

Objective of the Current Study

The objective of the current study was to determine the magnitude of the scoring discrepancies between the standard SUS and the 10 possible nine-item versions of the SUS that are created by dropping one item. If the discrepancies are small, then UX researchers and practitioners who object to the inclusion of a single problematic SUS item would be able to drop that item and still use the norms that have been developed for the interpretation of SUS scores.

Method

In this section, we discuss the data sets that were assembled for this study and the analyses of the data.

Data Set

For this study, we assembled a data set of 9,156 completed SUS questionnaires from 112 unpublished industrial usability studies and surveys. Note that with *n* = 9,156, the study has the power to reliably detect very small differences and to precisely compute confidence intervals around estimated means, allowing us to focus on differences that have practical rather than simply statistical significance (which only supports claims that differences are not plausibly 0).

Analyses

For comparison with standard SUS scores, we computed the 10 possible nine-item scores that are possible when leaving one SUS item out. We followed the standard scheme for computing these SUS scores but multiplied the sum of the score contributions by 100/36 instead of 2.5 to compensate for the missing item. For each nine-item variant of the SUS, we assessed scale reliability using coefficient alpha, the correlation with the standard SUS, and the magnitude of the mean difference.

Results

Tables 1 and 2 show the results of the various analyses. In the tables and figures, "SUS" by itself indicates the standard SUS. "SUS-01" indicates the SUS without Item 1, "SUS-02" indicates the SUS without Item 2, and so on.

Table 1. Correlations, Mean SUS With 95% Confidence Intervals, and Coefficient Alphas

Metric	**SUS**	**SUS-01**	**SUS-02**	**SUS-03**	**SUS-04**	**SUS-05**	**SUS-06**	**SUS-07**	**SUS-08**	**SUS-09**	**SUS-10**
Mean	74.1	74.8	74.5	74.1	73.2	74.6	74.1	74.2	73.9	73.9	73.7
Std dev	20.2	20.7	19.9	20.2	20.5	20.5	20.1	20.5	19.8	20.3	20.2
n	9156	9156	9156	9156	9156	9156	9156	9156	9156	9156	9156
Std error	0.21	0.22	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
df	9155	9155	9155	9155	9155	9155	9155	9155	9155	9155	9155
Critical *t*	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
Margin of error	0.41	0.42	0.41	0.41	0.42	0.42	0.41	0.42	0.40	0.42	0.41
Upper limit	74.5	75.2	74.9	74.5	73.6	75.0	74.5	74.6	74.3	74.3	74.1
Mean	74.1	74.8	74.5	74.1	73.2	74.6	74.1	74.2	73.9	73.9	73.7
Lower limit	73.7	74.3	74.1	73.7	72.8	74.2	73.6	73.8	73.5	73.5	73.3
Correlation with SUS	1.000	0.992	0.995	0.997	0.994	0.996	0.996	0.996	0.996	0.997	0.994
Coefficient alpha	0.91	0.91	0.90	0.90	0.91	0.90	0.90	0.90	0.90	0.90	0.90

The key findings from Table 1 were the following:

- Dropping an item did not have an adverse effect on scale reliability: All values of coefficient alpha were at or just above 0.90, consistent with the previous literature.
- As expected, all nine-item versions correlated highly, almost perfectly (*r* > 0.99), with the standard SUS.

Table 2. Mean Differences With 95% Confidence Intervals Around the Difference

Metric	**SUS-01**	**SUS-02**	**SUS-03**	**SUS-04**	**SUS-05**	**SUS-06**	**SUS-07**	**SUS-08**	**SUS-09**	**SUS-10**
Upper limit	-0.60	-0.34	0.06	0.92	-0.47	0.08	-0.05	0.20	0.21	0.41
Mean difference	**-0.66**	**-0.39**	**0.03**	**0.87**	**-0.51**	**0.04**	**-0.09**	**0.16**	**0.17**	**0.37**
Lower limit	-0.71	-0.43	-0.01	0.83	-0.55	0.00	-0.12	0.13	0.14	0.32
Margin of error	0.05	0.04	0.03	0.05	0.04	0.04	0.04	0.04	0.03	0.04
t	-24.5	-17.8	1.6	37.0	-25.3	2.0	-4.5	8.6	9.7	16.2
df	9155	9155	9155	9155	9155	9155	9155	9155	9155	9155
p	0.000	0.000	0.11	0.000	0.000	0.05	0.000	0.000	0.000	0.000

The key findings from Table 2 were the following:

- Most of the observed differences were statistically significant (except for the version without Item 3), but that is influenced by the high power of the analysis given *n* = 9,156.
- The estimates of the mean differences were very precise (due to the power of having *n* = 9,156), with margins of error ranging from 0.03 to 0.05.
- The magnitudes of the mean differences between the standard SUS and all the nine-item versions were small—all less than 1 SUS point.

Figure 2 provides a graphic illustration of the magnitudes of the differences between the nine-item versions and the standard SUS. All of the mean differences (including the boundaries of the 95% confidence intervals) fell within ±1.0 SUS points, with most of them very close to a mean difference of 0.

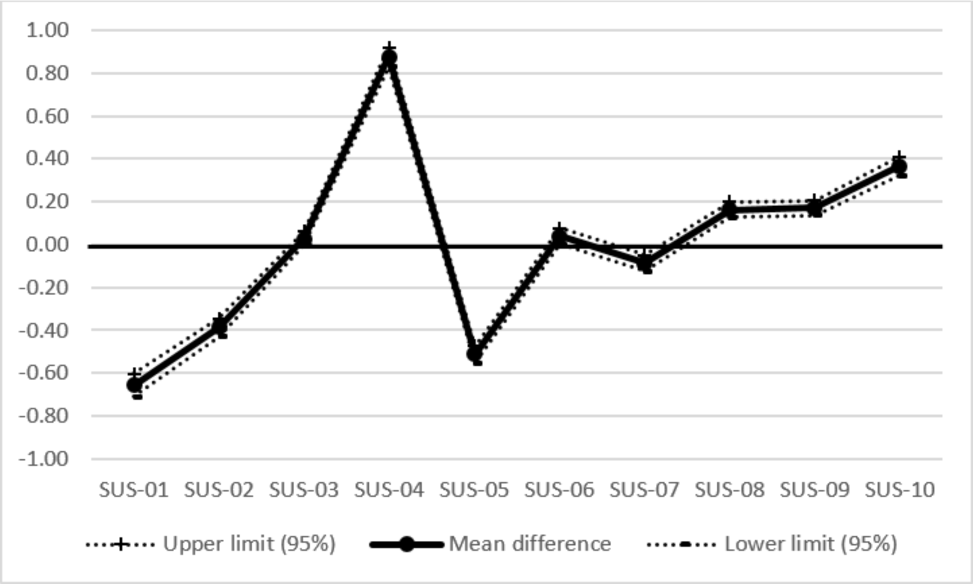


Figure 2. Mean differences between the standard SUS and the nine-item versions.

Recommendations

Future research on this topic could take a couple of different directions:

- Other researchers who have large samples of completed SUS questionnaires should consider conducting a similar analysis to explore the extent to which their results match the current results.
- It would be interesting to see what happens when systematically removing two items from the SUS, which would result in 45 different eight-item versions.
- Future research could determine the smallest version of the SUS (the one with the fewest items) that still produces scores that are close in magnitude to the standard SUS.
- A more complicated research design would be needed to investigate if there is any effect of asking the missing question to see if the SUS means for the nine-item variants are the same when (a) respondents answer all 10 items but the computation only uses nine (as in the current study) versus (b) respondents only see nine items. We suspect that any difference in these two conditions would be negligible, but it would be interesting to see what happens.

Conclusion

The results show that although removing an item from the SUS generally leads to a measurable deviation from the full score, the magnitude of the deviation is small. For all nine-item versions of the SUS, we found the deviation to be less than ±1.0 points on the 0–100 SUS scale. This means that the deviation of the nine-item variants of the SUS from the standard SUS were consistently less than 1%.

To help assess the practical consequences of a 1% "error," consider Table 3 that shows the Sauro-Lewis curved grading scale for the SUS (Sauro & Lewis, 2016, p. 204; based on the means from 241 industrial usability studies, both tests and surveys). Due to the distribution of the mean SUS scores, some of the grade ranges are wide and some are narrow. The narrowest are for C+ and B- (both 1.4). Thus, an error of 1% in this range could, for example, lead to some discrepancy in the assigned grade. Also, note that when using percentile ranks instead of raw scores or letter grades, an error of 1 point results in an approximately 3 percentile point difference when the SUS score falls between 60 and 80. For example, a raw score of 68 has a percentile rank of 50%, whereas scores of 67 and 69 have percentile ranks of 47% and 53% respectively (Sauro, 2011).

As shown in Table 1, the mean of the standard SUS for this sample of 9,156 individual SUS questionnaires was 74.1, the lowest point of the range for B. Six of the nine-item versions were also in the B range; four were in the B- range. Thus, there is always some risk that using a nine-item version of the SUS could move a mean over a grade boundary. The error associated with the use of a nine-item version of the SUS would not, however, be enough to move a mean over two grade boundaries.

Table 3. Curved Grading Scale for the SUS

Grade	SUS	Max-Min SUS	Percentile Range
A+	84.1 - 100	15.9	96 - 100
A	80.8 - 84.0	3.2	90 - 95
A-	78.9 - 80.7	1.8	85 - 89
B+	77.2 - 78.8	1.6	80 - 84
B	74.1 - 77.1	3	70 - 79
B-	72.6 - 74.0	1.4	65 - 69
C+	71.1 - 72.5	1.4	60 - 64
C	65.0 - 71.0	6	41 - 59
C-	62.7 - 64.9	2.2	35 - 40
D	51.7 - 62.6	10.9	15 - 34
F	0 - 51.6	51.6	0 - 14

Most user research has sample sizes much smaller than *n* = 9,156. When samples are smaller, the standard error of the mean is larger, with correspondingly larger margins of error for associated confidence intervals. In most cases, this increase in random error will completely wash out the small systematic error associated with dropping one SUS item. For example, Adell, Várhelyi, and dalla Fontana (2011) reported a mean SUS of 81.6 (A) with a standard deviation of 10.5, based on a sample of 20 drivers. The lower and upper limits of the associated 95% confidence interval ranged from 77.0 (B) to 86.8 (A+), a margin of error of ±4.9, which spanned over four grade boundaries (and 19 percentile ranks—80% to 99%).

Thus, the primary conclusion is that practitioners can leave out any one of the SUS items without having a practically significant effect on the resulting scores, as long as an appropriate adjustment is made to the multiplier (specifically, multiply the sum of the adjusted item scores by 100/36 instead of the standard 100/40, or 2.5, to compensate for the dropped item).

Tips for Usability Practitioners

Here are tips for practitioners to consider when using the SUS:

- Unless there is a strong reason to do otherwise, use the full 10-item version of the SUS.
- When there is one problematic item, you can still use the SUS. Remove the item and follow the standard procedures for computing the score contributions, but remember to multiply the sum of the score contributions by 100/36, not 2.5.
- There are a number of SUS variants that are currently available to practitioners, and the current research adds 10 more to the practitioner’s toolkit. Whenever you use the SUS, be sure to carefully document which one you used.

- When reporting a mean SUS, also report the lower and upper limits of a reasonable confidence interval (usually 95% confidence unless there is a reason to do otherwise).
- If using a guide like the Sauro-Lewis curved grading scale for interpreting SUS means, include grades for the lower and upper limits of the confidence interval in addition to the grade for the mean to report the range of plausible SUS grades.

References

Adell, E., Várhelyi, A., & dalla Fontana, M. (2011). The effects of a driver assistance system for safe speed and safe distance–A real-life field study. *Transportation Research Part C*, *19*, 145–155.

Bangor, A., Joseph, K., Sweeney-Dillon, M., Stettler, G., & Pratt, J. (2013). Using the SUS to help demonstrate usability's value to business goals. In *Proceedings of the Human Factors Society and Ergonomics Society Annual Meeting* (pp. 202–205). Santa Monica, CA: HFES.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, *24*, 574–594.

Bangor, A., Kortum, P. T., & Miller, J. T. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*. *4*(3), 114–123.

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 189–194). London, UK: Taylor & Francis.

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, *8*(2), 29–40.

Finstad, K. (2006). The System Usability Scale and non-native English speakers. *Journal of Usability Studies*, *1*(4), 185–188.

Grier, R. A., Bangor, A., Kortum, P., & Peres, S. C. (2013). The System Usability Scale: Beyond standard usability testing. In *Proceedings of the Human Factors Society and Ergonomics Society Annual Meeting* (pp. 187–191). Santa Monica, CA: HFES.

Kortum, P., & Bangor, A. (2013). Usability ratings for everyday products measured with the System Usability Scale. *International Journal of Human-Computer Interaction*, *29*, 67–76.

Kortum, P., & Peres, S. C. (2014). The relationship between system effectiveness and subjective usability scores using the System Usability Scale. *International Journal of Human-Computer Interaction*, *30*, 575–584.

Kortum, P., & Sorber, M. (2015). Measuring the usability of mobile applications for phones and tablets. *International Journal of Human-Computer Interaction*, *31*, 518–529.

Lah, U., & Lewis, J. R. (2016). How expertise affects a digital-rights-management-sharing application's usability. *IEEE Interactions*, *33*(3), 76–82.

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.

Lewis, J. R., Brown, J., & Mayes, D. K. (2015). Psychometric evaluation of the EMO and the SUS in the context of a large-sample unmoderated usability study. *International Journal of Human-Computer Interaction*, *31*, 545–553.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the System Usability Scale. In Kurosu, M. (Ed.), *Human Centered Design, HCII 2009* (pp. 94–103). Heidelberg, Germany: Springer-Verlag.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31*, 496–505.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Peres, S. C., Pham, T., & Phillips, R. (2013). Validation of the System Usability Scale (SUS): SUS in the wild. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 192–196). Santa Monica, CA: HFES.

Sauro, J. (2011). *A practical guide to the System Usability Scale*. Denver, CO: Measuring Usability.

Sauro, J. (2016, August 30). *Can you change a standardized questionnaire*? MeasuringU. Retrieved from http://www.measuringu.com/blog/change-standardized.php

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of CHI 2009* (pp. 1609–1618). Boston, MA: ACM.

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? In *Proceedings of CHI 2011* (pp. 2215–2223). Vancouver, Canada: ACM.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.). Cambridge, MA: Morgan-Kaufmann.

Tullis, T. S., & Stetson, J. N. (2004). *A comparison of questionnaires for assessing website usability*. Paper presented at the Usability Professionals Association Annual Conference, June. Minneapolis, MN: UPA.

About the Authors

James R. (Jim) Lewis

Dr. Lewis is a senior human factors engineer (at IBM since 1981). He has published influential papers in the areas of usability testing and measurement. His books include *Practical Speech User Interface Design* and (with Jeff Sauro) *Quantifying the User Experience* (now in its second edition).

Jeff Sauro

Dr. Sauro is a six-sigma trained statistical analyst and founding principal of MeasuringU, a customer experience research firm based in Denver. He has conducted usability tests and statistical analysis for companies such as Google, eBay, Walmart, Autodesk, Lenovo, and Dropbox, and has published over 20 peer-reviewed research articles and 5 books, including *Customer Analytics for Dummies*.