International Journal of Human–Computer Interaction

Intl. Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: http://www.tandfonline.com/loi/hihc20

An Empirical Evaluation of the System Usability Scale

Aaron Bangor , Philip T. Kortum & James T. Miller

To cite this article: Aaron Bangor , Philip T. Kortum & James T. Miller (2008) An Empirical Evaluation of the System Usability Scale, Intl. Journal of Human–Computer Interaction, 24:6, 574-594, DOI: 10.1080/10447310802205776

To link to this article: https://doi.org/10.1080/10447310802205776

Published online: 30 Jul 2008.

INTL. JOURNAL OF HUMAN–COMPUTER INTERACTION, *24*(6), 574–594, 2008
Copyright © Taylor & Francis Group, LLC
ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447310802205776

An Empirical Evaluation of the System Usability Scale

Aaron Bangor[1], Philip T. Kortum[2], and James T. Miller[3]
[1]AT&T Labs, Inc.
[2]Rice University
[3]AT&T Labs, Inc.

This article presents nearly 10 year's worth of System Usability Scale (SUS) data collected on numerous products in all phases of the development lifecycle. The SUS, developed by Brooke (1996), reflected a strong need in the usability community for a tool that could quickly and easily collect a user's subjective rating of a product's usability. The data in this study indicate that the SUS fulfills that need. Results from the analysis of this large number of SUS scores show that the SUS is a highly robust and versatile tool for usability professionals. The article presents these results and discusses their implications, describes nontraditional uses of the SUS, explains a proposed modification to the SUS to provide an adjective rating that correlates with a given score, and provides details of what constitutes an acceptable SUS score.

1. INTRODUCTION

The System Usability Scale (SUS) was developed by Brooke (1996) as a "quick and dirty" survey scale that would allow the usability practitioner to quickly and easily assess the usability of a given product or service. Although there are a number of other excellent alternatives (see Table 1) the SUS has several attributes that make it a good choice for general usability practitioners. Chief among these is that the survey is technology agnostic, making it flexible enough to assess a wide range of interface technologies, from interactive voice response systems (IVRs) and novel hardware platforms to the more traditional computer interfaces and Web sites. Second, the survey is relatively quick and easy to use by both study participants and administrators. Third, the survey provides a single score on a scale that is easily understood by the wide range of people (from project managers to computer programmers) who are typically involved in the development of products and services and who may have little or no experience in human factors and usability. Finally, the survey is nonproprietary, making it a cost effective tool as well.

Correspondence should be addressed to Philip T. Kortum, Rice University, Department of Psychology, MS25, 6100 Main Street, Houston, TX 77005. E-mail: pkortum@rice.edu

Table 1: Summary of Examined Usability Surveys

Survey Name	*Abbreviation*	*Developer*	*Survey Length (Questions)*	*Availability*	*Interface Measured*	*Reliability*
After Scenario Questionnaire	ASQ	IBM	3	Nonproprietary	Any	0.93[a]
Computer System Usability Questionnaire	CSUQ	IBM	19	Nonproprietary	Computer based	0.95[b]
Poststudy System Usability Questionnaire	PSSUQ	IBM	19	Nonproprietary	Computer based	0.96[b]
Software Usability Measurement Inventory	SUMI[c]	HFRG	50	Proprietary	Software	0.89[d]
System Usability Scale	SUS	DEC	10	Nonproprietary	Any	0.85[e]
Usefulness, Satisfaction and Ease of Use	USE	Lund	30	Nonproprietary	Any	Unreported[f]
Web Site Analysis and Measurement Inventory	WAMMI	HFRG	20	Proprietary	Web based	0.96[g]

[a]Lewis (1995). [b]Lewis (2002). [c]Kirakowski and Corbett (1993). [d]Igbaria and Nachman (1991). [e]Kirakowski (1994). [f]Lund (2001). [g]Kirakowski , Claridge, and Whitehand (1998).

The primary goal of this article is to share nearly a decade's worth of SUS data—more than 2,300 individual surveys collected while conducting more than 200 studies—to provide a benchmark that can be used by other usability professionals who are also using the SUS or who are considering adopting it. One common question that is seen on usability newsgroups is one concerning the "goodness" of a single SUS score. Typically, a researcher has conducted a test, used the SUS to help measure the usability and customer satisfaction of a particular system, but is now unsure how to interpret the resulting SUS scores. The data contained in this article helps those professionals better assess their SUS score and determine how their product's usability fits into a larger universe of SUS scores for similar products.

As part of the analysis of the SUS data from these studies, we also address some of the common questions that are asked by consumers of the SUS data (e.g., managers, clients, programmers, fellow practitioners, etc.), including questions about the meaningfulness of individual questions in the SUS, whether different interface types are intrinsically more usable than others, and whether the age and gender of the test participants makes a difference in how they rate usability. Finally, we introduce an adjective description measure to the SUS that might aid practitioners in the interpretation of SUS scores.

2. DESCRIPTION OF THE SUS DATA

2.1. The SUS Instrument

The original SUS instrument (Brooke, 1996), is composed of 10 statements that are scored on a 5-point scale of strength of agreement. Final scores for the SUS can range from 0 to 100, where higher scores indicate better usability. Because the statements alternate between the positive and negative, care must be taken when scoring the survey. See Brooke (1996) for the details of scoring the SUS.

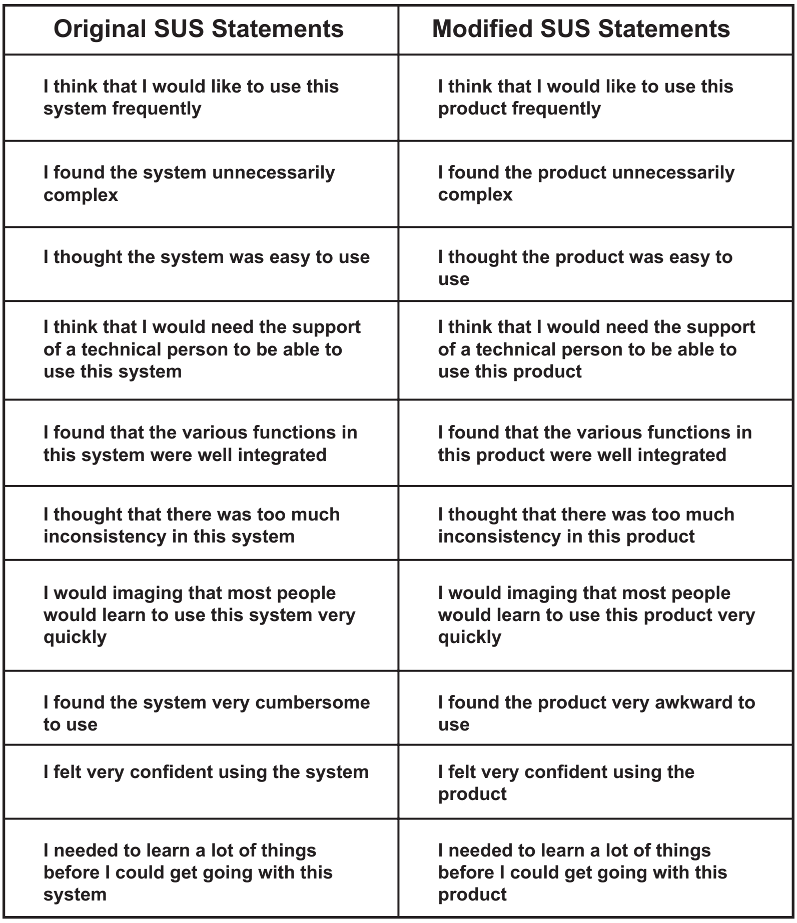
It is important to note that approximately 90% of the SUS data presented in this article was collected using a slightly modified form of the original SUS. In our initial use of the SUS, it was noted that some participants (about 10%) had a question about the word *cumbersome* in Statement 8—"I found the system very cumbersome to use." We grew concerned that perhaps a larger portion of the tested population also did not know what the word meant but did not ask the test administrator for clarification. Consequently, we decided it was necessary to replace *cumbersome* with a more recognizable synonym. We eventually settled on *awkward* because *awkward* is a much more commonly used word in English than *cumbersome* (Oxford University Computing Service, 2001) and most test administrators reported using *awkward* in their explanation of the definition of *cumbersome* to a participant, with good success. Other researchers (Finstad, 2006) have independently reached a similar conclusion. We also replaced the word 'system' with the word 'product' throughout the survey, based on user feedback. The original SUS statements from Brooke (1996) and the modified statements used in these studies are shown in Figure 1.

2.2. Aggregate Data

Since we began collecting data with the SUS in 1996, 2,324 surveys have been completed over the course of 206 studies. The mean SUS score for all surveys is 70.14 (s = 21.71) with a median of 75 and a range from 0 to 100. Analyzing these data on a per *study* basis (M number of surveys per study = 11.3) reveals a mean SUS of 69.69 (s = 11.87) with a median of 70.91 and a range from 30.00 to 93.93. The mean of all study sample standard deviations is 18.00. Table 2 summarizes these data.

These data indicate that there is a negative skew to the distribution of survey scores, which is not unexpected for this type of data. Figure 2 shows the histogram of individual SUS scores. A histogram of study means in Figure 3 reveals a slightly more normal distribution.

One interesting characteristic of the SUS scores that has been observed is a rather robust range restriction for study means. Fewer than 6% of the mean study scores fall below 50, and there are no group scores below 30. This is a rather surprising result, as usability task success rates have been recorded across the entire range (0–100%), and we initially believed that SUS scores would track the success metric very closely. However, as can be seen in Figure 4, this is clearly not the case. This figure shows the results of successive tests with the same user interface during the design lifecycle. In the early tests, success rates were in the 20 to 30% range and,



Original SUS Statements	Modified SUS Statements
I think that I would like to use this system frequently	I think that I would like to use this product frequently
I found the system unnecessarily complex	I found the product unnecessarily complex
I thought the system was easy to use	I thought the product was easy to use
I think that I would need the support of a technical person to be able to use this system	I think that I would need the support of a technical person to be able to use this product
I found that the various functions in this system were well integrated	I found that the various functions in this product were well integrated
I thought that there was too much inconsistency in this system	I thought that there was too much inconsistency in this product
I would imaging that most people would learn to use this system very quickly	I would imaging that most people would learn to use this product very quickly
I found the system very cumbersome to use	I found the product very awkward to use
I felt very confident using the system	I felt very confident using the product
I needed to learn a lot of things before I could get going with this system	I needed to learn a lot of things before I could get going with this product

FIGURE 1 The original SUS statements (Brooke, 1996) and the modified statements used in these studies.

Table 2: Summary Statistics for System Usability Scale Scores

	Data for Individual Surveys	*Data for Multisurvey Studies*
Count	2,324	206
M	70.14	69.69
Mdn	75.00	70.91
SD	21.71	11.87
Range	0.00–100.0	30.00–93.93

although the SUS scores reflect the problems encountered by the participants, they do not mirror the significant decline seen in the success metrics. Unfortunately, we have insufficient paired SUS/success data to build a model of this relationship, but further data collection toward this goal is currently taking place.

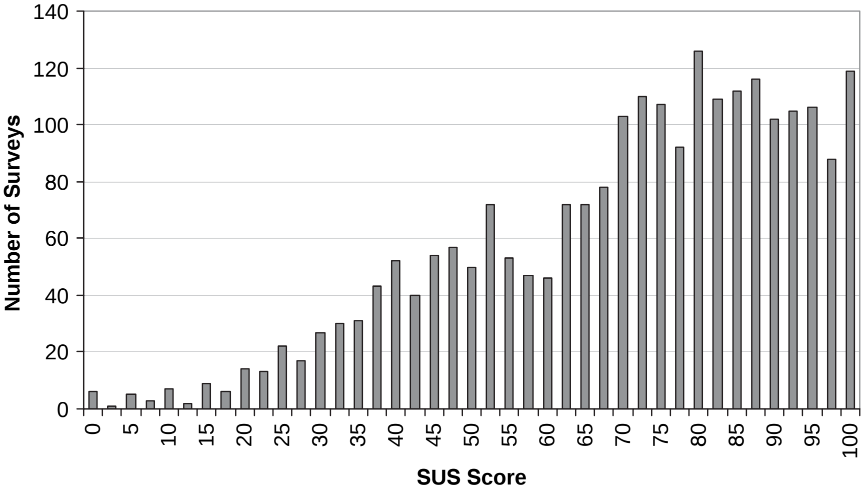


FIGURE 2 A histogram of individual System Usability Scale (SUS) scores ($n = 2{,}324$).

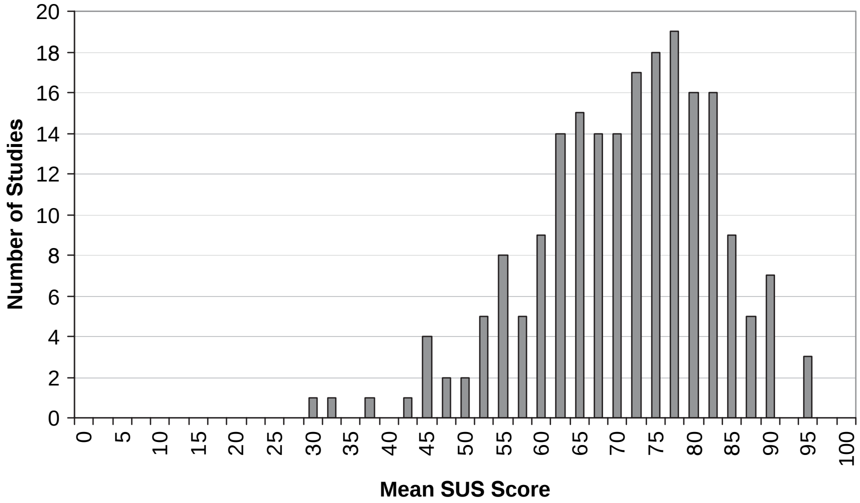


FIGURE 3 A histogram of *study* means of System Usability Scale (SUS) scores ($n = 206$).

A quartile breakdown of the individual surveys and the study means is shown in Table 3. One fourth of individual survey responses fall at 55.0 and below, whereas the worst quarter of overall study outcomes falls at 62.26 and below. To be in the top fourth of responses, a participant must have given a rating of 87.5 to 100.0, whereas the best quarter of studies range from 78.51 to 93.93.

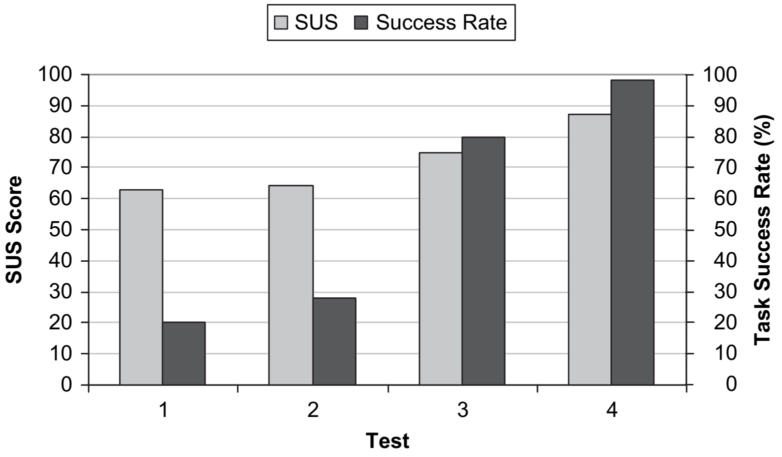


FIGURE 4 An illustration of range restriction for System Usability Scale (SUS) means for four tests where the success rates varied from 20% to nearly 100%.

Table 3: Quartile Breakdown of Surveys and Study Means

Quartile	*Survey Means*	*Study Means*
1	55.0	62.26
2	75.0	70.91
3	87.5	78.51
4	100.0	93.93

2.3. Statement-by-Statement Analysis

The principal value of the SUS is that it provides a single reference score for participants' view of a product's usability. As such, the individual statements that compose the SUS are secondary to the discussion of the instrument, in favor of the emergent score. In fact, Brooke (1996, p. 194) cautioned that "scores for individual items are not meaningful on their own." Our experience, however, indicates that clients and practitioners alike tend to ignore this admonition and frequently seek to gain further information about the goodness of specific aspects of the interface by looking at individual statements. Because of this tendency, a close examination of participants' scores for each statement is warranted so that the usability practitioner can speak knowledgeably in these situations.

Table 4 presents each statement, the mean and standard deviation for the raw scores, and an absolute score that adjusts the scores of Statements 2, 4, 6, 8, and 10 so that positive responses are associated with a larger number, like the other five statements. Individual responses have been standardized on a per survey basis, with the mean and standard deviation for the transformed data also given.

Table 4: Responses to Individual System Usability Scale Statements

Statement	*M*	*SD*	*Absolute*	*Standardized*	*Standard SD*
1. I think that I would like to use this product frequently.	3.68	1.15	3.68	–0.18	0.99
2. I found the product unnecessarily complex.	2.34	1.21	3.66	–0.16	0.89
3. I thought the product was easy to use.	3.69	1.15	3.69	–0.13	0.73
4. I think that I would need the support of a technical person to be able to use this product.	1.83	1.16	4.17	0.47	0.94
5. I found the various functions in the product were well integrated.	3.62	1.05	3.62	–0.28	0.90
6. I thought there was too much inconsistency in this product.	2.12	1.12	3.88	0.10	0.88
7. I imagine that most people would learn to use this product very quickly.	3.82	1.15	3.82	–0.01	0.92
8. I found the product very awkward to use.	2.09	1.22	3.91	0.16	0.85
9. I felt very confident using the product.	3.64	1.19	3.64	–0.20	0.83
10. I needed to learn a lot of things before I could get going with this product.	2.03	1.24	3.97	0.23	0.97

Performing a *t* test on the standardized results shows that all but Statement 7 differ significantly from the average (absolute) response. In other words, relative to all responses a participant gave on their survey, Statements 4, 6, 8, and 10 tended to have more of a positive rating than the participant's average rating, whereas Statements 1, 2, 3, 5, and 9 tended to have a more negative rating relative to their other ratings. Participants tended to give Statement 7 a rating that was about average with respect to their other ratings.

This finding could mean that there are characteristics of the statements that tend to result in generally positive or negative ratings. It could also be that participants tend to more strongly disagree with negative statements than they agree with positive statements. A third possible explanation is that participants had a bias toward the left side of the survey, as suggested by Friedman and Amoo (1999).

By inspection, Statement 4 ("I think that I would need the support of a technical person to be able to use this product") received the most extreme of all responses, with a mean standardized rating of 0.47. Statement 3 ("I thought the product was easy to use") had the lowest standard deviation of its standardized scores, at 0.73, well below the others, which ranged from 0.83 to 0.99. The most varied responses were to Statement 1 ("I think that I would like to use this product frequently"), with a standard deviation of 0.99. This is to be expected because the range of products tested included those that would be used on a daily basis, like a cell phone, and those that would only be completed once, for example the initial setup of a service.

To further analyze the individual statements, a correlational analysis was performed. Table 5 presents the results of correlating all 10 statements with one another. The responses to all 10 statements are very highly correlated with one another with every correlation achieving Pearson product moments that are significant at $\alpha = .01$ or better.

Table 5: Pearson Correlations for all 10 Statements

Statement	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	1									
2	.466	1								
3	.616	.593	1							
4	.339	.407	.509	1						
5	.542	.497	.596	.347	1					
6	.482	.528	.553	.427	.518	1				
7	.526	.476	.615	.426	.543	.487	1			
8	.502	.552	.581	.448	.507	.566	.531	1		
9	.553	.517	.694	.524	.555	.497	.591	.543	1	
10	.367	.438	.526	.558	.381	.455	.433	.464	.555	1

Note. All correlations are significant at the α = .01 level (two-tailed).

Although examining individual statements is appealing, the statement-by-statement analysis shows that there is not much differentiation among the statements. The differences that are found among the statements may have more to do with the use of positive and negative statements and/or a left-hand scale bias, rather than the individual statements.

2.4. Factor Analysis

Because the SUS was originally designed to give a single score, a factor analysis was performed to determine if the SUS statements address different dimensions of the participant's experience or just one dimension—usability—as intended. The Eigenvalues resulting from the analysis are shown in Figure 5 and the factor loading matrix in Figure 6. The results show that there was only one significant factor for the ten SUS statements. These results indicate that the SUS questionnaire, as a whole, reflects participants' estimates of the overall usability of an interface, regardless of the type of interface.

Given the results that indicate that all statements are significantly correlated and the factor analysis finding that there is only one factor, practitioners should heed Brooke's (1996) original admonition and avoid the temptation to analyze the scores to individual questions and report only the overall SUS scores instead.

2.5. Reliability Analysis

To determine how well the ten statements used in the SUS correlate with the hypothetical population of statements regarding the concept of usability, a reliability analysis was performed. The absolute ratings (i.e., transformed responses for Statements 2, 4, 6, 8, and 10 so all scales have 1 as the negative and 5 as the positive) for the 10 statements were used to compute Cronbach's alpha. The result of the test was a result of 0.911. This is slightly higher than the value of 0.85 reported in Table 1 from previous data for the SUS. The current reliability value obtained from these data compares favorably with the reliability values of the other instruments presented in Table 1.

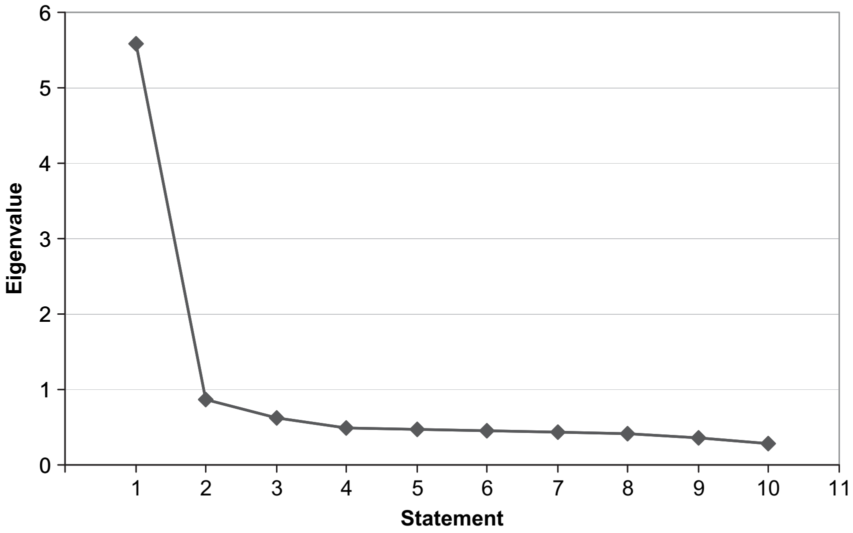


FIGURE 5 Eigenvalue output from the factor analysis of System Usability Scale statements.

Statement	*Component*
S3	.849
S9	.813
S8	−.764
S7	.757
S6	−.737
S5	.737
S2	−.734
S1	.725
S10	−.687
S4	−.658

FIGURE 6 Factor loading matrix for the SUS statements factor analysis. Positive and negative loadings correspond to the positive and negative wording of the statements.

2.6. SUS Scores by User Interface Type

Many factors affect a participant's ratings of usability. Obviously, the specific user interface design that study participants use plays a highly significant part in those ratings. To determine if the specific type of interface has a significant impact on the SUS score, each survey response has been categorized by type and analyzed (see Table 6 and Figure 7). Six different classes of interface types were identified in the data used for this article. These include cellular phone interfaces, equipment that is used in a customer's premise like landline telephones, non-Web graphical user interfaces, IVR (i.e., automated telephone) interfaces, Web-based

Table 6: System Usability Scale Scores by User Interface Type

Interface Type	*Count*	*M*	*SD*
Cell	189	66.55	19.84
CPE	219	71.60	21.60
GUI	208	75.24	20.77
IVR	401	73.84	22.15
Web	1180	68.05	21.56
Web/IVR	50	59.45	19.19

Note. An interface type was required to have at least 50 surveys to be included in the analysis. Cell = cell phone equipment; CPE = customer premise equipment (e.g., phones, modems, etc.); GUI = graphical user interface for OS-based computer interfaces; IVR = interactive voice response phone systems, including speech based; Web = Internet-based Web pages and applications.

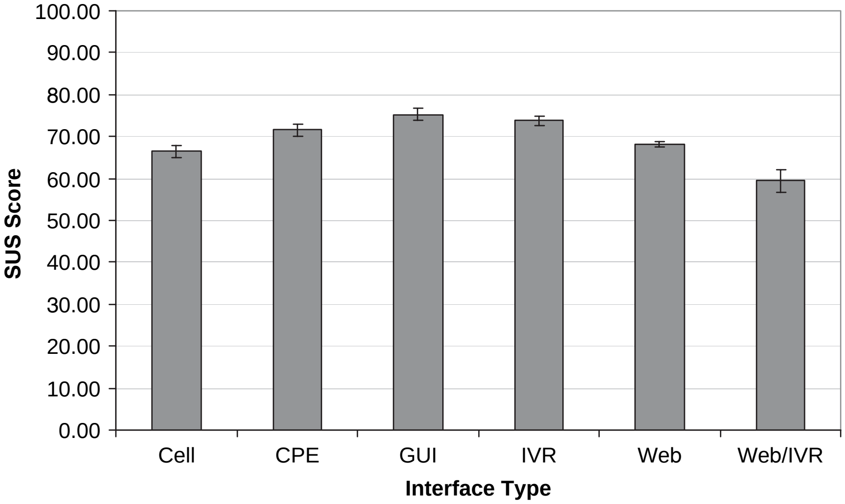
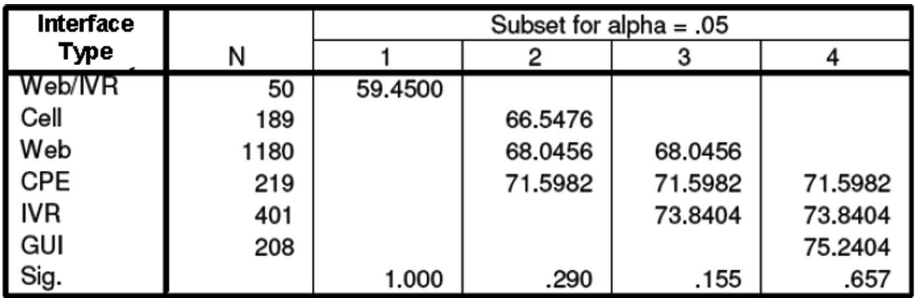


FIGURE 7 The mean System Usability Scale (SUS) scores for six user interface types. *Note.* Cell = cell phone equipment; CPE = customer premise equipment (e.g., phones, modems, etc.); GUI = graphical user interface for OS-based computer interfaces; IVR = interactive voice response phone systems, including speech based; Web = Internet-based Web pages and applications.

interfaces, and applications that combine a web and IVR interface. Each of the six user interface types was required to have at least 50 surveys to be included in the analysis (error bars are ± one standard error of the mean).

A one-way analysis of variance on these data showed that SUS scores do vary significantly by the type of interface being tested ($\alpha = .05$, $p < .001$). Post hoc analysis showed that there were homogeneous variances (Levene = 1.02, $p = .405$) and so a Tukey's Honestly Significance Difference test was conducted to identify level



Interface Type	N	Subset for alpha = .05			
		1	2	3	4
Web/IVR	50	59.4500			
Cell	189		66.5476		
Web	1180		68.0456	68.0456	
CPE	219		71.5982	71.5982	71.5982
IVR	401			73.8404	73.8404
GUI	208				75.2404
Sig.		1.000	.290	.155	.657

FIGURE 8 Tukey Honestly Significant Difference test results for the means of different interface types. *Note.* Cell = cell phone equipment; CPE = customer premise equipment (e.g., phones, modems, etc.); GUI = graphical user interface for OS-based computer interfaces; IVR = interactive voice response phone systems, including speech-based; Web = Internet-based Web pages and applications.

differences. Four levels were found, with graphical user interfaces being only in the top level, IVRs spanning the top two, customer premise equipment spanning the top three, Web user interfaces spanning the second and third levels, cell phone user interfaces only being in the second lowest level, and applications with Web and IVR user interfaces alone in the lowest level. See Figure 8 for a depiction of these results.

To explore further how participants have responded to each statement, Figure 9 presents the standardized (per survey) scores for each statement, broken down by user interface modality. It should be noted that, for the graph in Figure 9, data to the right of the vertical axis are a relatively positive rating. The graphs illustrate the generally positive ratings that Statements 4, 6, 8, and 10 received and the generally negative ratings received by Statements 1, 2, 3, 5, and 9, as discussed in section 2.3.

The graph in Figure 9 also shows some differentiation for particular interface types. For example, Statement 1 appears to show that daily use products like cell phones and the messaging product associated with the Web/IVR modality are rated highly, whereas the infrequently used products like IVR systems were rated rather low. Note the variance in response to this question previously reported in section 2.3.

The other statements did not show much difference due to interface type. Although there are a few instances where modalities may differ on a per statement basis, the vast majority do not, reinforcing the results that the SUS is a robust instrument for generating a single score of usability.

2.7. SUS Scores by Age and Gender

The effect of age and gender on the resulting SUS score was also examined. This analysis was carried out on a subset of the data (213 surveys) that was coded with age and gender attributes.

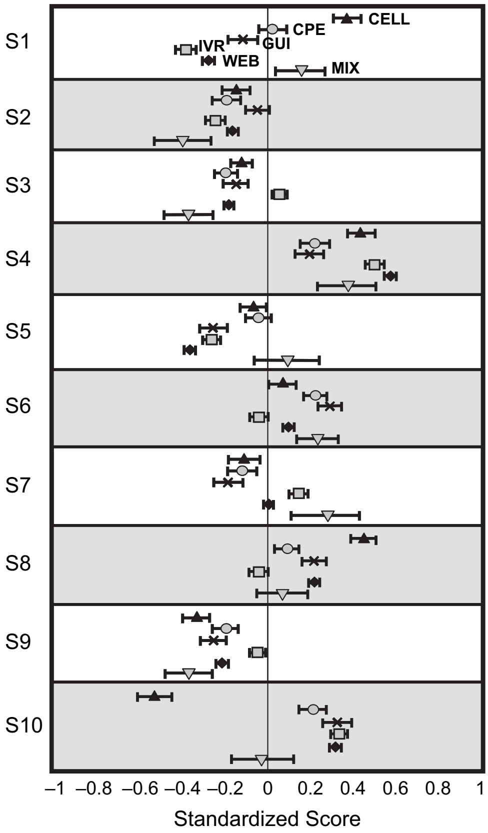


FIGURE 9 Mean standardized System Usability Scale (SUS) scores (with standard error of the mean bars shown) for each statement on the SUS (S1, S2, etc.), broken down by interface type (positive ratings to the right, negative to the left, for all statements). *Note.* Cell Phone equipment (▲), Customer Premise Equipment (e.g., phones, modems, etc.) (●), OS-based Graphical User Interfaces (X), Interactive Voice Response Systems, including speech systems (■), Web Interfaces (◆), Mixed Web and IVR systems (▼).

The correlation between SUS score and age is significant at $r = -.203$ ($p = .003$). The slope is not very strong at –.311. Although not strong, the significant relationship between age and SUS scores may indicate that the age of the user does have some negative impact on the usability score that is given to an interface. We believe that much more data need to be collected with the age attribute to make any definitive claims about its impact on usability assessment or to offer an interpretation about that result.

The mean SUS score for male participants in this subset of data was 70.2 ($s = 19.2$) and 71.6 ($s = 17.4$) for female participants. There is no significant difference between the mean SUS scores between women and men (two-tailed t test, $p = .586$). An F test to determine if the variances between genders were different was not significant at $\alpha = .05$ ($p = .338$).

3. ADDING AN ADJECTIVE RATING SCALE

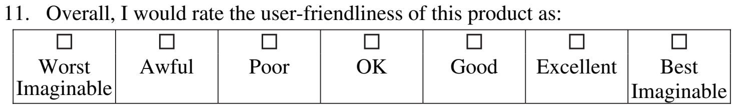
Although the single number generated by the SUS is very useful for relative judgments (e.g., comparing competitive alternatives, iterative versions, etc.), a valid assessment of what the absolute numerical score means is another matter. Explaining what a specific mean score means (e.g., 68.5) to project managers and design teams is a common and frustrating experience. As more and more data were collected, a pattern that was familiar to engineers and product managers alike began to take form and became the standard rule of thumb. This rule judged the SUS scores based on the typical grading scale used in most schools and became known as the "university grade analog." This is logical because both range from 0 to 100, with 100 being the absolute best and 0 being the worst. Managers and engineers also saw an anecdotal correlation between task success rates, subjective user comments, and the value of the SUS scores that were being reported.

Consequently, this "university grade analog" rule-of-thumb judgment meant that a SUS score of 90 to 100 was an A, 80 to 89 a B, and so on. Practically speaking, this was an extremely useful and powerful notion to foster, even though there was no basis in the original survey development that suggests or supports this. Nevertheless, the rule of thumb is immediately clear to project teams and metrics of improvement are intuitive as well, with no special interpretation required by the human factors researchers.

As useful as this concept is, however, it has not been validated. To address this need, a pilot program was begun to determine if adjective descriptors could be associated with intervals of SUS scores to give more of an absolute rating to the SUS scores. A 7-point adjective rating scale (a modified version of the scale used by Bangor, 2000, and Olacsi, 1998, for subjective image quality ratings) was used. The modified adjective rating scale was added at the bottom of the page with the 10 SUS statements. Figure 10 shows the adjective rating scale used in the pilot program.

The phrasing of the statement for the scale has three major components. First, the "overall" modifier comes from the original subjective quality statement used by Bangor (2000) and Olacsi (1998) and is meant to inquire about the summative experience of the participant. Second, the term "user-friendliness" is used because it is one of the most widely known synonyms for usability and one that participants were likely to immediately grasp. Third, "product" is the same term used in the other SUS statements to refer to the object of the study.

The intent of this question is to provide a qualitative answer that can be used in conjunction with a SUS score to better explain the overall experience when using the SUS to summarize a user interface's usability.



11. Overall, I would rate the user-friendliness of this product as:

□ Worst Imaginable	□ Awful	□ Poor	□ OK	□ Good	□ Excellent	□ Best Imaginable

FIGURE 10 The adjective rating statement appended to the System Usability Scale.

To date, 212 participants have completed a SUS that includes the 11th adjective rating question. In the analysis of this pilot data, a numeric value of 1 was assigned to "Worst Imaginable" and 7 to "Best Imaginable." The correlation between the adjective rating and SUS score turns out to be very high at $r = .806$. The slope of a regression line drawn through the scatter plot of SUS scores versus adjective rating values is m = 15.62. Table 7 presents the summary statistics for the SUS scores, and Figure 11 depicts the means for each adjective with error bars (± one standard error of the mean).

There is only one data point for each of the two extremes and there are no data points for "Awful"; otherwise this distribution is generally consistent with the slope of about 16. In addition to the consistency, an examination of the standard error of the means in Figure 11 suggests that the descriptors are different.

Table 7: Summary Statistics for Adjective Ratings

No.	*Rating*	*Count*	*M*	*SD*
7	Best imaginable	1	100	NA
6	Excellent	69	85.58	9.473
5	Good	90	72.75	10.56
4	OK	36	52.01	12.13
3	Poor	15	39.17	12.38
2	Awful	0	NA	NA
1	Worst imaginable	1	25	NA

Note. NA= not applicable.

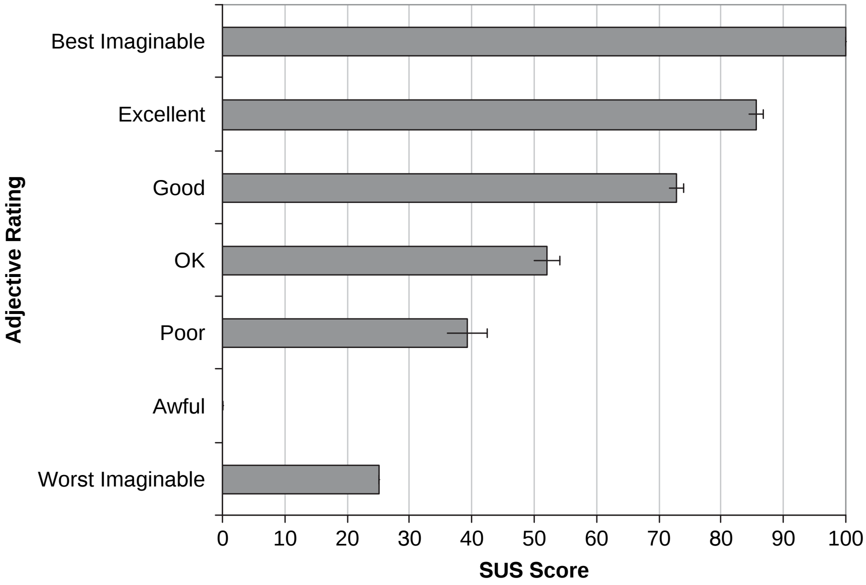


FIGURE 11 Mean System Usability Scale (SUS) scores for each adjective rating.

Table 8: Mapping Adjective Ratings to Study Mean Quartiles

No.	*Adjective*	*M*	*Upper Bound*	*Quartile*
7	Best imaginable	100	100	1
6	Excellent	85.58	87.5	2
5	Good	72.75	75.0	3
4	OK	52.01	55.0	4
3	Poor	39.17		
2	Awful	NA		
1	Worst imaginable	25.00		

Note. NA = not applicable.

At this point, the data used to validate the adjective rating scale represent less than 10% of all of the surveys discussed in this study. Furthermore, the sample is not fully representational of the overall population of SUS surveys, given the fact that it has not yet been used in all of the interface modalities discussed earlier. We believe that these preliminary data suggest that an associated adjective rating scale is a legitimate complement to the SUS statements and overall SUS score. Of interest, the adjective ranges closely match those of the quartiles, as can be seen in Table 8.

4. A CASE STUDY OF ITERATIVE TESTING OF A PRODUCT USING THE SUS

This section describes the use of the SUS in an iterative design process and underscores the strong face validity of the instrument that has caused us to so fully embrace the SUS as a standard instrument for use in nearly all of the usability studies we conducted.

Early in 1999, a broadband Internet self-installation kit was developed and deployed that had significant software and hardware components. The hardware components were essentially a collection of original equipment manufacturer vendor parts in their original packaging (e.g., modems, cables, Ethernet cards, etc.), with each component having its own software and attendant instructions. Predictably, the kit was not particularly easy to use. Over the course of the next 2 years development progressed and substantial changes were made to the kit as more data were collected from the field and the usability lab. With each version of the self-installation kit, SUS scores were collected as part of successive usability tests to assess the usability at each iteration of the kit.

Plotting SUS scores for this product over an extended period and relating these scores to the events surrounding the product clearly demonstrates the SUS's strong face validity and its ability to measure the progress (or regression) in usability as the self-installation kit proceeded through its iterative design lifecycle.

The 12 laboratory-based usability tests of this product are shown in Figure 12. The first recorded SUS score was a weak 62, which mirrored our heuristic assessment of the first iteration of the self-installation kit. By the 2nd iteration, the SUS

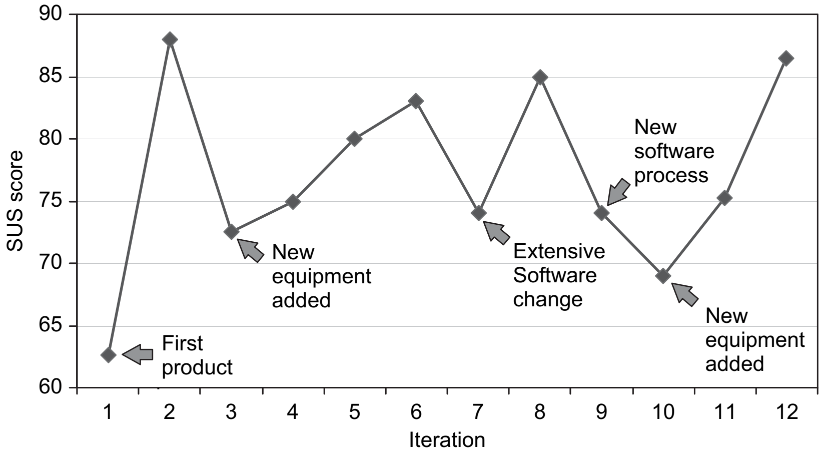


FIGURE 12 System Usability Scale (SUS) scores and their relationship to critical events in the product lifecycle process.

score improved drastically because most of the recommendations from the first test had been incorporated. The effect of adding new (untested) equipment was evaluated in the 3rd iteration and the SUS scores dropped substantially. Over the next 3 iterations, SUS scores slowly improved, as minor improvements were implemented, until Iteration 7, when a significant software change was put into service. This software change resulted in a substantial decrease in the SUS scores. Iteration 8 saw the score dramatically increase after improvements to the user interface were made, only to be followed by several decreases because of the introduction of more software and hardware changes. Following Iteration 10, the hardware and software began to stabilize and SUS scores improved through Iterations 11 and 12. The mean SUS scores for each iteration track very closely with changes—both positive and negative—made to the product.

Employing the SUS in these kinds of iterative tests provided us with significant confidence that the instrument was providing valuable benchmark information that we could share with our product design colleagues, secure in our belief that the SUS score was indeed measuring important aspects of the product's usability.

5. DISCUSSION

The SUS has been used across a wide range of user interfaces, including standard OS-based software interfaces, Web pages and Web applications, cell phones, landline phones, modem and networking equipment, pagers, IVRs, speech systems, and video delivery hardware and software.

Originally, the SUS was used primarily in one-time, isolated tests to determine a single usability and satisfaction score for a given product or service. As its use within the human factors group became more frequent, more uses became evident and were adopted. Over the last 10 years we have identified six major ways that the SUS (or any good technology agnostic survey, for that matter) can be used to positively supplement a usability testing and evaluation program:

1. **Providing a point estimate measure of usability and customer satisfaction:** The SUS provides a single score that estimates the overall usability of a given product or service. With a sufficient history of test scores (like the one provided in this article) for a basis of comparison, these data provide the usability practitioner with sufficient evidence that a given interface is—or is not—sufficiently usable in its current form.
2. **Comparing different tasks within the same interface:** Complex interfaces can be difficult to assess because the more difficult, but less frequently performed tasks (e.g., installation), can overshadow simpler tasks that are done more frequently. The ease of administering the SUS allows the investigator to use the SUS after completion of a difficult, low-frequency task (e.g., installation) and then after the participant completes the easier, more frequent tasks. This flexibility allows the investigator to measure two very different aspects of an interface's usability and satisfaction, so they can be analyzed and weighted appropriately.
3. **Comparing iterative versions of the same system:** Throughout the development and testing process, interfaces typically undergo a large number of changes as the development process progresses. Obtaining a SUS score at each step in the process allows the experimenter to compare each iteration using a standard tool. This provides a convenient and powerful way of determining if the interface is becoming more or less usable with each testable iteration. Negative trends in usability can be identified early so that corrective action can be taken, whereas positive trends indicate that the interface is headed in the right direction. This method was used in the case study described earlier in the article.
4. **Comparing competing implementations of a system:** Frequently, competing designs of an interface need to be compared to determine which design has superior usability and customer satisfaction. Quite often, these different alternatives are not available at the same time, so standard head-to-head comparison testing is not practical. Having a common measure that allows direct comparison of competing interfaces over an extended period of time offers significant value to the design and testing process.
5. **Competitive assessment of comparable user interfaces:** The SUS provides yet another tool for assessing competing products that may have the same functional properties implemented with very different interfaces. Although the typical measures of success/failure, error rate, time to complete, and so on, are used to provide objective measures of competitors' products, the SUS provides another way to assess the overall usability and satisfaction of the interfaces of the competing products.
6. **Comparing different interface technologies:** One question that frequently comes up concerns general usability and customer satisfaction across different interface technologies. A product or service can be fielded using more than one interface technology (e.g., interactive voice response systems and the Web) and assessing the overall usability and satisfaction of each interface

> technology when users are performing the same task is desirable. In fact, it has become clear that a generalized assessment of the usability and customer satisfaction for different delivery types of interfaces is valuable information to have when trying to determine which technology would be best suited for different deployments. Indeed, the ease and technology-agnostic nature of the SUS has led to its recent adoption by other researchers in novel interface studies, including voting technologies (including lever machines, paper ballots, and e-voting; Everett, Byrne, & Greene, 2006) and wearable interfaces (Kostov, Ozawa, & Matsuura, 2004).

Although we have extolled the many benefits of the SUS, we believe that the modifications we made to the instrument early in its adoption has improved the SUS. The user population tested at our labs is drawn from the entire range of socioeconomic demographics, giving us a somewhat wider background of participants than one might find in other settings. The minor modification of substituting the word *awkward* for the word *cumbersome* in Statement 8 eliminated most of the questions that were received during the administration of the SUS. Although other modest changes have been debated over the years, the overall consensus has been to leave the modified SUS in its current form.

Comparing iterative versions of the same system is one of the important uses of the SUS, as we described earlier. The case study showing SUS score changes as they related to different product changes demonstrates that the instrument is sensitive to differences in a product as it evolves. This use of the SUS is especially important in environments where a product goes through a lengthy development lifecycle and undergoes significant revisions. The use of the SUS in an iterative development environment has also made the intuitive nature of the instrument extremely valuable when communicating usability trends and comparisons to clients. As such, the great value of the SUS is not just to the researcher who uses it in a study but to the recipients of the results of user research.

5.1. What is an Acceptable SUS Score?

As with any metric, the SUS score should not be used in isolation to make absolute judgments about the "goodness" of a given product. Factors such as success rate and the nature of the failures observed when the system was tested with representative users should play a large part in determining how usable a product is (ISO, 1998).

That said, the SUS score can and does provide a very useful metric for overall product usability, with the fundamental question of "What constitutes an acceptable SUS score?" Certainly the operational definition one assigns to the term "acceptable" makes a large difference in the answer. It also seems clear that although the operational definition will certainly vary across industries, as well as the point at which a product is in its lifecycle, each usability lab will have some sense of what the SUS scores mean for them if that lab consistently collects SUS data with each usability test.

We believe that the rule-of-thumb standard discussed earlier, the so-called university grade analog, is well supported by both the adjective rating scale data and our own collective experience that takes into account the nature of the failures. This means that products which are at least passable have SUS scores above 70, with better products scoring in the high 70s to upper 80s. Truly superior products score better than 90. Products with scores of less than 70 should be considered candidates for increased scrutiny and continued improvement and should be judged to be marginal at best.

We can make a further distinction in the marginal scores, by dividing them into "low marginal" and "high marginal." This break occurs roughly at the beginning of the second quartile range. Products with scores less than 50 should be cause for significant concern and are judged to be unacceptable. This structure is consistent with the quartile ranges found for study averages and with the adjective rating scales. Figure 13 shows a comparison of acceptability score, quartile ranges, and the adjective rating scale.

The restricted range seen for the per study SUS scores suggests that it is important to run enough participants to avoid misinterpreting the SUS data because of insufficient sample size. Any interpretation of a SUS score also needs to take into account that the range of scores is essentially half of the nominal value. Thus, a score of 50 does not represent a product that is "half as good" as a product that scores 100 but rather is likely an indication of a serious usability failure for that product.

Although we have insufficient data linking individual success metrics with their associated SUS scores, anecdotal evidence suggests a participant's SUS score may be correlated to their individual task performance (i.e., participants who perform poorly on objective measures tend to have low SUS scores and vice versa). A noteworthy caveat to this, however, is for participants who think they performed well on tasks but did not because they did not receive proper feedback. In these cases, SUS scores may very well be inflated, representing perceived success on the part of the user even though they actually failed.

Of course, as Lewis (1996) pointed out, it is easier to show that a product is unacceptable than it is to show that it is acceptable—if one wishes to make the claim with a high degree of confidence. This must be kept in mind when determining the acceptability of a product based on SUS scores, unless uncommonly large numbers of participants are tested. With SUS scores below 50, one can be

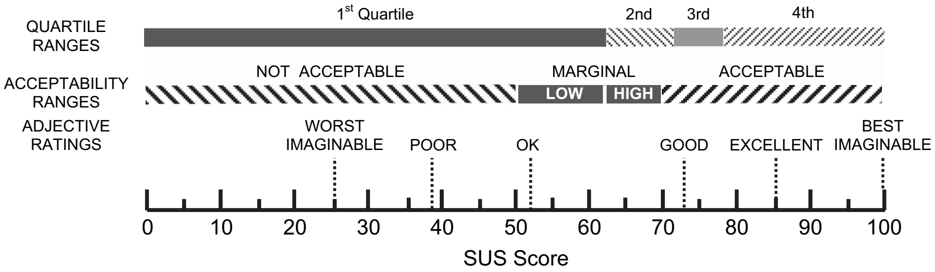


FIGURE 13 A comparison of mean System Usability Scale (SUS) scores by quartile, adjective ratings, and the acceptability of the overall SUS score.

almost certain that the product will have usability difficulties in the field, whereas scores in the 70s and 80s, although promising, do not guarantee high acceptability in the field.

6. CONCLUSIONS AND FUTURE DIRECTIONS

The primary objective of this article has been to share nearly a decade's worth of SUS data to allow other researchers and practitioners in the field who are using the SUS to compare and better understand their data, relative to an independent source of SUS data representing a range of user interfaces. The establishment of baseline "acceptability" scores is one important aspect that we believe will help further the ability of practitioners to score and report the SUS with greater confidence. To further understand the SUS, we are collecting more data about study success rates to match to SUS scores. This will give a better understanding of the sensitivity of the SUS to quantitative measures of performance. Also, more data regarding adjective ratings and the age and gender of participants are being collected to come to firmer conclusions about their relationship to SUS scores.

Several potentially interesting alternatives have been developed in the last several years that might serve as good alternates for the SUS. Notable among these is the Usability Magnitude Estimation metric (Rich & McGee, 2004), which includes users' expectations of the usability in the estimation of an overall usability metric. One of the most promising alternatives, however, is the Summated Usability Metric described by Sauro and Kindlund (2005). Like the SUS, the Summated Usability Metric also produces a single score, but it does so by including actual performance data in addition to the subjective data provided by the user. Given the evidence that the three usability measures recommended by ISO 9241-11 (1998) (efficiency, effectiveness, and satisfaction) may not be highly correlated (Frøkjær, Hertzum, & Hornbæk, 2000) the method may provide additional and powerful information about product usability. Although these new metrics hold significant promise, the SUS has proven itself a valuable and robust tool in helping assess the quality of a broad spectrum of user interfaces.

REFERENCES

Bangor, A. W. (2000). *Display technology and ambient illumination influences on visual fatigue at VDT workstations*. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

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: Taylor & Francis.

Everett, S. P., Byrne, M. D., & Greene, K. K. (2006). Measuring the usability of paper ballots: Efficiency, effectiveness and satisfaction. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, pp. 2547–2551. Santa Monica, CA: Human Factors and Ergonomics Society.

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

Friedman, H. H., & Amoo, T. (1999). "Rating the rating scales." *Journal of Marketing Management, 9*(3), 114–123.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency and satisfaction really correlated? *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computer Systems,* 345–352.

Igbaria, M., & Nachman, S.A. (1991). Correlates of user satisfaction with end user computing: An exploratory study. *Information and Management, 19,* 73–82.

ISO. (1998). *Ergonomic requirements for office work with visual display terminal (VDT's)–Part 11: Guidance on usability* (ISO 9241-11(E)). Geneva, Switzerland: Author.

Kostov, V., Ozawa, J., & Matsuura, S. (2004). Analysis of wearable interface factors for appropriate information notification. *Proceedings of the Eighth International Symposium on Wearable Computer*, 102–109.

Kirakowski, J. (1994). *The use of questionnaire methods for usability assessment*. Unpublished manuscript. Retreived from http://sumi.ucc.ie/sumipapp.html

Kirakowski, J., Claridge, N., & Whitehand, R. (1998). Human centered measures of success in Web design. *Conference Proceedings of the 4th Conference on Human Factors and the Web*. Retrieved from http://www.research.att.com/conf/hfweb/proceedings/kirakowski/index.html

Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Measurement Inventory. *British Journal of Educational Technology, 24,* 210–212.

Lewis, J. (1995). IBM Computer Satisfaction Questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7*(1), 57–78.

Lewis, J. (1996). Binomial confidence intervals for small sample usability studies. In A. F. Ozok & G. Salvendy (Eds.), *Advances in applied ergonomics: Proceedings of the 1st International Conference on Applied Ergonomics* (pp. 732–737). West Lafayette, IN: USA Publishing.

Lewis, J. (2002). Psychometirc evaluation of the PSSUQ using data from 5 years of usability studies. *International Journal of Human-Computer Interaction, 14*(3&4), 463–488.

Lund, A. (2001). Measuring usability with the USE Questionnaire. *Usability Interface: The usability SIG newsletter of the Society for Technical Communications*, *8*(2). Retrieved from http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html

Olacsi, G. S. (1998). *Subjective image quality of CRT displays under ambient glare: Assessing the ISO 9241-7 Ergonomic Technical Standard*. Unpublished master's thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Oxford University Computing Service. (2001). *The British national corpus* (version 2). Oxford, UK: Author. Available from http://www.natcorp.ox.ac.uk/

Rich, A., & McGee, M. (2004). Expected usability magnitude estimation. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 912–916). Santa Monica, CA: Human Factors and Ergonomics Society.

Sauro, J., & Kindlun, E. (2005). A method to standardize usability metrics into a single score. *Proceedings of CHI 2005: Technology, Safety, Community*, 401–409.