The Relationship Between Task-level and Test-level System Usability Scale Scores

Amber Callan and Philip Kortum
Rice University, Houston, TX

This study examined whether the average usability score for a series of tasks was the same as the usability score for the product if usability was measured only after all the tasks had been completed. Fifty participants completed a set of tasks for five websites and fourteen mock voting ballots. Subjective usability assessment was made with the System Usability Scale (SUS). Participants completed the SUS either after each task (five or fourteen SUS administrations, respectively) or after completing the entire set of tasks (one SUS). The results show that the average SUS scores for the task-level assessments were significantly higher than the SUS scores for the test-level assessments. Results were similar for the ballot and website conditions. Task-level SUS scores on the Honda websites ($M = 65.5$) were significantly higher than the test-level SUS scores ($M = 42.8$), $p < 0.0001$. Similar results were observed in the ballot condition, where task-level usability assessments were higher ($M = 59.5$) than test-level assessments ($M = 38.5$), $p < 0.0001$. Practitioners and those interpreting SUS scores need to be aware of how these experimental differences can lead to different assessment metrics.

INTRODUCTION

The System Usability Scale (SUS), developed by Brooke (1996), is a commonly used ten-item Likert scale which assesses the usability of a product or service. It yields a score from 0-100 that allows researchers to make basic judgments about a product’s subjective usability, based on the product’s effectiveness, efficiency and satisfaction as suggested by ISO 9241-11. Because it can assess a wide variety of systems and is easy to use, score, and interpret, the SUS has become one of more widely used usability assessments in testing (Bangor, Kortum & Miller, 2008). The SUS has been well-studied since its creation, and so practitioners have a fairly good idea of what a SUS score means.

Understanding subjective usability scores can help researchers interpret their results and analyze the usability of the product they are testing. In most basic terms, a low score on the SUS indicates a low usability while a high score shows high usability.

While the SUS is commonly used to measure the subjective usability of all manners of products and services, its exact application in the testing environment is not specified. Some practitioners are interested in the summative information of a single SUS score for the entire product as a whole, and in this case a single SUS is administered after the participant has completed all of the tasks of interest. This method of measure is often called ‘test-level’ measurement, and is meant to capture the overall usability of the product or service. Other practitioners administer the SUS after each task, which gives them insight into the usability of specific features or task-based elements of the product. This form of administration is called ‘task-level’ assessment. While this allows a practitioner to get task-level information, there is often still a need for a task-level measure. In this case, these individual task SUS scores are often averaged together in order to obtain a summative SUS score for the overall product.

Clearly, there are benefits to collecting subjective usability scores with each of these procedures, but the impact of using these two different administration schemes on the final summative SUS score is not well understood.

We know from previous research that task-level and test-level measures are not perfectly correlated (Sauro and Lewis, 2009; Hornbæk & Law, 2007), with correlations ranging between .38 and .70. This means that practitioners who average the scores of their task-level metrics can garner different summative scores than practitioners who take a single test-level summative measure. This previous research collapsed the analysis across multiple forms of subjective usability assessment questionnaires, and so the direction of the difference and the specific magnitude of that difference for the System Usability Scale are unknown.

Because this information could have significant implications for determining the best and most accurate way to administer the SUS, we explore it in this study.
Knowing exactly how the administration method affects overall scores would strengthen our ability to more accurately interpret scores from studies regardless of the method used and would help improve our confidence in the reliability of the measure itself.

**METHODS**

**Participants**

Fifty participants were recruited from the Rice University undergraduate research pool. Of the fifty students, 19 were male and 31 were female with ages ranging from 18-22 years (average age = 19.0).

**Materials**

This study was comprised of two parts. Part one was the assessment of a set of websites and part two was an assessment of a set of voting ballots.

In condition one of the study, subjects navigated through five different Honda websites (Powersports, Racing, Power Equipment, Marine, and Automobile) which could all be accessed through Honda’s main site (www.honda.com). The task list (Table 1) was created so each task had an intermediate difficulty and the information could be located within four direct clicks. The tasks served as a way to get participants to navigate through the websites and experience them with a purpose in mind as an individual might while naturally visiting the websites.

In condition two of the study, subjects voted with fourteen different mock ballots where each differed in layout and question subject matter. The ballots used were those created by Kortum and Acymyan (2013), and are shown in Figure 1. These mock ballots represent exemplars of different forms of ballots that have been used in real elections, or have been proposed for use. Some ballots represent combinations of different ballot forms into a single ballot. The ballots were designed to contain non-controversial items so as to not evoke emotional responses. As such, no party labels were used.

Each of these ballots was compiled into a packet where participants viewed a single ballot at a time and then voted according to the method required by the ballot form. Participants then wrote their selected ballot choice on the following page, which allowed us to verify the intent of the participant in their ballot selection.

<table>
<thead>
<tr>
<th>Honda Website</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Powersports</td>
<td>When does the Vikings Grand Prix take place?</td>
</tr>
<tr>
<td>Racing</td>
<td>Performance wheels have one critical difference when compared to standard wheels. What is on the standard wheels, but not on performance wheels?</td>
</tr>
<tr>
<td>Power Equipment</td>
<td>When did Honda copyright the WSP50 manual?</td>
</tr>
<tr>
<td>Marine</td>
<td>What is the tilt range of the 15 HP motor?</td>
</tr>
<tr>
<td>Automobiles</td>
<td>What do you need to do in order to locate the navigation serial number on a 2004 Accord?</td>
</tr>
</tbody>
</table>

**Procedures**

Participants were randomly assigned to either the task-level assessment condition, where they would evaluate the usability of the product after each and every task, or the test-level assessment condition, in which they would evaluate the usability of the product after completing all tasks. Both test-level and task-level participants completed tasks on the Honda websites and the mock ballots. Presentation order of the website or ballot condition was counter-balanced.

**Condition one.** Participants completed an IRB approved consent form before proceeding. Participants were then directed to an internet connected computer that had one of the Honda family websites pre-loaded. Presentation of the 5 different websites was randomized. Participants were informed that for each website they would be given a task and they were to search the website to try and find relevant information in order to answer the task question. Participants were informed that they were not allowed to use the search bar and that the information required might or might not be on the website.

Participants were given a maximum of 5 minutes to complete a task. However, if at any point, the participant felt that the information was not available on the site, they could let the experimenter know and they would be allowed to move on to the next task. The experimenter would read the first task from the list.
Figure 1: Ballots used in condition two of the study
(see Table 1) and then let the participant find the information. Once found, participants verbally reported the answer to the experimenter who verified the information. Depending on condition, the experiment administrator would either give participants the SUS or proceed to the next task.

Participants would repeat this procedure until they had visited all five Honda websites and taken either one overarching SUS test or five individual SUS tests.

**Condition Two:** If they had not already done so, participants completed an IRB approved consent form before proceeding. Participants were then given a packet that contained 14 different ballots, each on its own page. They were instructed to look over the first ballot, cast their desired vote and then indicate what that vote was on a separate page. Then, after each ballot or after all fourteen ballots, participants would assess the usability with the SUS. Participants would repeat this procedure until they had voted using all fourteen ballots and taken either one overarching SUS test or fourteen individual SUS tests.

**Measures**

This study utilized the 10-item System Usability Scale created by Brooke (1996) and modified in 2008 by Bangor et al. While the SUS has not undergone a major revision since its inception, the updated 2008 version changes the word “cumbersome” to “awkward” in item 8 of the scale in order to make the instrument easier to understand by a wider demographic (Finstad, 2006). The SUS alternates positive statements with negative statements and gives a likert-scale from 1-5 (1 indicating strongly disagree, 5 indicating strongly agree). The scores were calculated by using the method described by Brooke (1996).

**RESULTS**

The results show that the average SUS score for the task-level assessments were significantly higher than the SUS scores for the test-level assessments (see Figure 2). Results were similar for the ballot and website conditions. Task-level SUS scores on the Honda websites ($M = 65.5$) were significantly higher than the test-level SUS score ($M = 42.8$), $p < 0.0001$. Similar results were observed in the ballot condition, where task-level usability assessments were higher ($M = 59.5$) than test-level assessments ($M = 38.5$), $p < 0.0001$.

**DISCUSSION**

Intuitively, the average score for test-level assessments should be nearly identical to the test-level assessment score. Clearly, this is not the case. In fact the magnitude of the difference is quite surprising, showing a difference of over 20 points for both tested systems. This unexpected result has several potential causes.

First, a primacy or recency effect may be occurring when trying to assess the system as a whole. Ebbinghaus (1902) found that the first and last items in a set of words had the highest percentage of recall. Therefore the first or last task that the participant completed may be the one on which the participants are basing their judgment. In the test-level assessment, if this last task was particularly difficult or easy, it might mean that the participant would overestimate or underestimate the SUS score of the entire system.

In a related vein, it may be that negative experiences are more salient and disproportionality impact the SUS score when only a single measure is taken. For example, a single bad experience, regardless of where it takes place in the test, might be viewed as strong evidence by the user of the lack of usability of the system in general. This phenomenon has been described in the general psychological literature before (Baumeister, Bratslavsky, Finkenauer & Vohs, 2001). All of the positive experiences and successful task completions may not be sufficient to overcome a single negative event. This would then be reflected as a lower score for the product overall. In contrast, an average of a
series of task-level SUS scores would grant the negative task a weight equal to each of the positive experiences, attenuating its negative impact.

Another possibility is that the law of large numbers is coming into play. As the number of times an experiment is performed increase, the average will get closer to the true average (Hoffmann-Jörgensen, 1976). With five data points and fourteen data points, it might be expected that the more parts of a system a participant must judge, the closer that judgment will lie to the true SUS score. Therefore, the task-level assessment may be yielding an actual average for the system that is more accurate than the test-level assessments.

It may also be that the discrepancy in scores may be due to an attempt to correlate the data between users or tasks (Sauro et. al. 2009). If participants complete the task with little error, then there will be very little variation. If the levels of aggregation vary, then participants have more contributing factors in mind when calculating a SUS score.

Finally, it could be that participants are rating something entirely different, depending on when the SUS is administered. The test-level SUS may be measuring qualities of the system that are ignored or under-valued, when compared to the immediacy of the task-level SUS assessment. If this were true, it brings up a larger question: Is usability comprised of the usability of all the separate components of a product (as represented by the task-level assessments) or is usability a more holistic judgment that takes into account more than just the assessment of individual pieces (as reflected in the test-level assessments). This question may be best left to the individual practitioner to decide, based on the goals of the assessment, and is certainly beyond the scope of this paper.

More research needs to be conducted in order to parse apart these different hypotheses. A study that purposely placed high or low usability tasks at the end of the sequence would help determine if the recency hypothesis has merit. Overall usability vs piece-part usability is a more difficult construct to evaluate. Inclusion of a more detailed assessment instrument measuring other product attributes, or more usability dimensions (like the USE Survey, Lund, 2001) might be beneficial. More research should also be conducted to expand the scope of products tested. While there was excellent agreement among the two disparate systems tested here, an expansion to evaluate multiples of other systems would serve to strengthen these findings.

In conclusion, when a researcher administers the SUS appears to have significant consequences on the results that are obtained. Practitioners should be aware of the fact that these measures are not equivalent and understand what they are reporting. Interpretation of other's results should also be done carefully, so an assessment of which method was used can be considered. In the real world, a low comprehensive usability score may impact decisions made when trying to increase the system's usability. Moving forward, it may be interesting to see how task difficulty impacts the comprehensive SUS and to see which one of these SUS administration methods can be deemed more accurate.

REFERENCES


