Quantitative Usability Test Case Study

David Weintraub
March 28th, 2023

Introduction

What Was Tested?

Q2 recently completed a significant redesign of its digital banking landing page, the first page millions of banking customers see after logging into their accounts. In addition to providing an at-a-glance summary of banking and financial activity, the redesigned landing page is a highly interactive platform where users may monitor their spending patterns, review their credit scores, review recent transactions, and so on.

Where Were We At?

After a series of testing and design iterations, the feature was soon to be launched. Stakeholders were eager to report usability metrics, such as a SUS score, to customers (i.e., financial institutions who have or may purchase Q2 as their digital banking provider). In short, the feature was ready for a summative, quantitative usability test.

What Kind of Summative Evaluation?

An effective way to conduct a summative evaluation is to compare metrics between systems. Showing that one thing outperforms another is a much stronger narrative than showing one thing scores well. I initially pushed to compare the redesigned feature to its previous counterpart; however, early on during planning, it was clear that the new feature had substantially changed and offered new functionality. It was not possible to create equivalent tasks to test the two systems. It was decided, in this case, to evaluate the new feature on its own to compare it to industry benchmarks.

Project Timeline

Research Plan

A research plan was created with stakeholders during a kickoff meeting. The document helped me align with stakeholder goals and was an opportunity to set expectations around research methods, timelines, and deliverables.

Create Task Scenarios

The feature was reviewed with stakeholders to create task scenarios, which were designed to give participants a complete experience of the feature.

Task Scenarios for Quantitative Testing Need Extra Care

During quantitative usability testing, it is especially important task scenarios (1) are drafted in a way that participants know clearly when they've completed a task and (2) remain the same across participants.

Pilot

A few participants were piloted before data collection. This gave an opportunity to (1) ensure the testing environment worked and (2) make any final revisions to task scenarios.

Data Collection

Twenty-three participants were tested via remote, moderated sessions.

Why 23 Participants?

Jakob Nielsen (2006) has shown that the inverse effect of sample size on the margin of error has diminishing returns after 20 participants. So, we aimed for 20 participants in our final sample, and we recruited three additional participants in anticipation that some participants would be excluded from the final sample.

Who Is an Outlier?

Outliers should be identified via multiple metrics and observations. In our case, outliers were identified based on (1) task completion rates and (2) time on task. More importantly, however, this operationalization was agreed upon before data analysis, preventing participants from being unfairly excluded (e.g., low SUS score).

Data Analysis

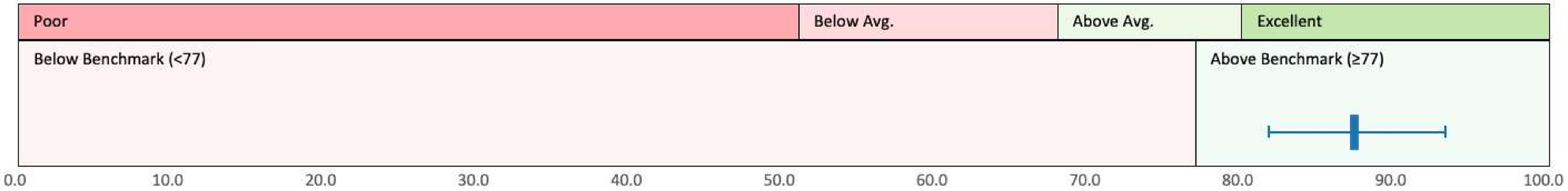
Participants were first screened for outliers; otherwise, data analysis focused on calculating a System Usability Scale (SUS) score and a Summated Usability Metric (SUM).

Report Readout

A polished presentation goes a long way to increase the impact of a research project. A visually compelling slide deck leaves an impression on stakeholders and increases the likelihood that stakeholders share the presentation with other organization members.

Estimated SUS Score Confidence Interval

The average SUS from our sample is 87.4, and there is a 95% likelihood that the actual SUS is somehwere between 81.6 and 93.2.



Don't Gloss Over Confidence Intervals

Stakeholders need education on confidence intervals; otherwise, confidence intervals may overshadowed by average scores. Our measurements are merely a prediction of a true score in our population of interest, and confidence intervals provide a better estimate of whether or not a true SUS score is likely to exceed a benchmark in our population of interest. In the past, I have found myself fumbling to explain confidence intervals in a way that is easy to understand. Therefore, I drafted a script explaining confidence intervals. The planning I put into a script produced a more organized, simpler explanation. Anecdotally, I have noticed an acceptance and discussion of confidence intervals by stakeholders.

The SUS Has Problems Too

The SUS has become an industry standard for good reasons: (1) It produces an intuitive, single score, (2) since the SUS has become an industry standard, a bounty of past research helps researchers better understand SUS scores, and (3) the SUS has established construct validity. I argue, however, that the SUS has significant limitations too.

First, the SUS is entirely subjective. Whether or not this is truly a limitation is debatable. If you're interested in subjective experience, ask about one's subjective experience; however, as you can imagine, measuring subjective experience in this way is not without concern. For example, when provided with subjective rating scales, it is understood that survey respondents are often biased to rate things positively (i.e., acquiescence bias).

Second, the SUS is completed once as a post-test measurement. Participants are expected to complete the SUS after completing many usability tasks. As a result, SUS scores are likely biased more toward participants' recent experiences with a system (*see my article linked below for a further discussion on this topic*).

Alternative measures should combine behavioral and subjective measurements collected throughout testing. A viable measurement includes the Summated Usability Metric (SUM).

Click here to view my article *Task Recency and the System Usability Scale* showing latter tasks disproportionately predict SUS scores during usability testing.

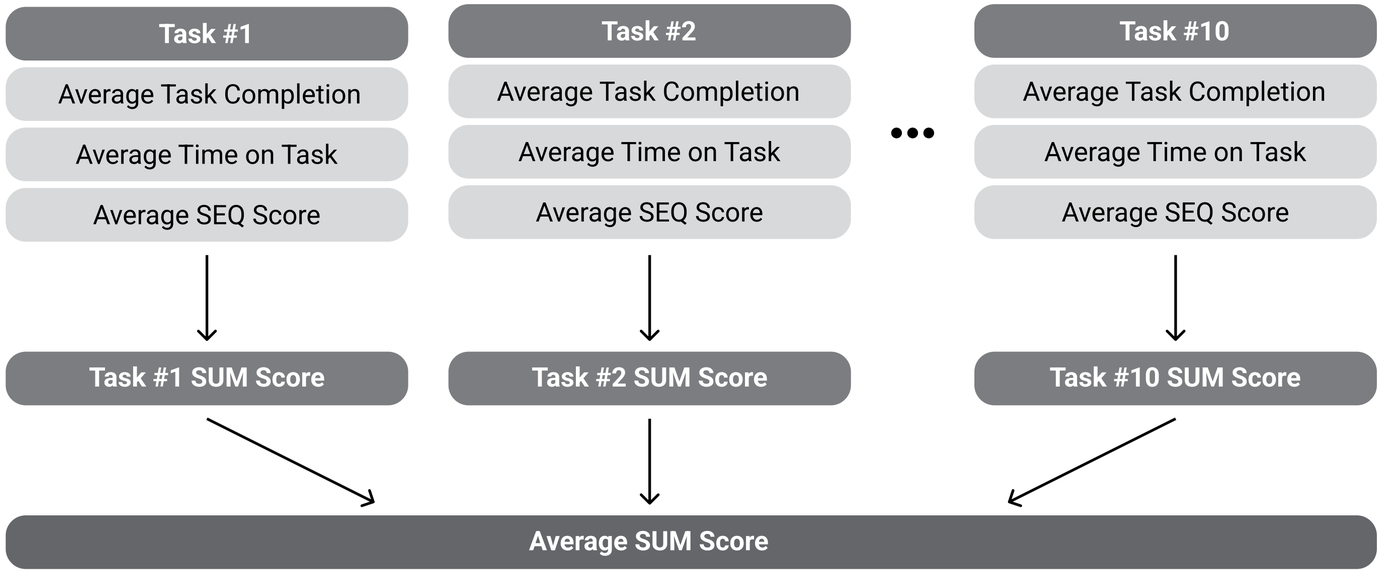
Summated Usability Metric

Calculating Average Time on Task

Since our sample size was less than 25 participants, geometric means were calculated to measure average time on task; otherwise, if our sample were greater than 25, medians would've been calculated to measure average time on task.

How to Measure Task SUM Sores

Since the various task averages reflected measurements on different scales (e.g., task completion and time on task are measured on dichotomous and continuous scales, respectively), it was necessary to standardize each measurement to make them "average-able." Jeff Sauro recommends converting each average measurement into a kind of z-score. In particular, the standard deviations of the task average from a satisfactory score (aka., a specification limit) was calculated for each type of measurement. These scores were than averaged to create a task SUM score.



Advantages of SUM Scores

This was the first project in which SUM scores were presented to a stakeholder within the company. So, during the report readout, I advocated for the SUM. Why should stakeholders care about the SUM? After all, the SUS has become an industry standard.

First, the SUM reflects a combination of behavioral (i.e., task completion, time on task) and subjective measures (i.e., SEQ scores).

Second, the SUM combines measurements collected throughout testing. The SUM places less on participants to have accurate memory retrieval of their testing experience.

Lastly, SUM scores can be used to validate SUS scores. If we conducted future studies of the same feature, increases in SUS scores could be compared to corresponding increases in SUM scores.

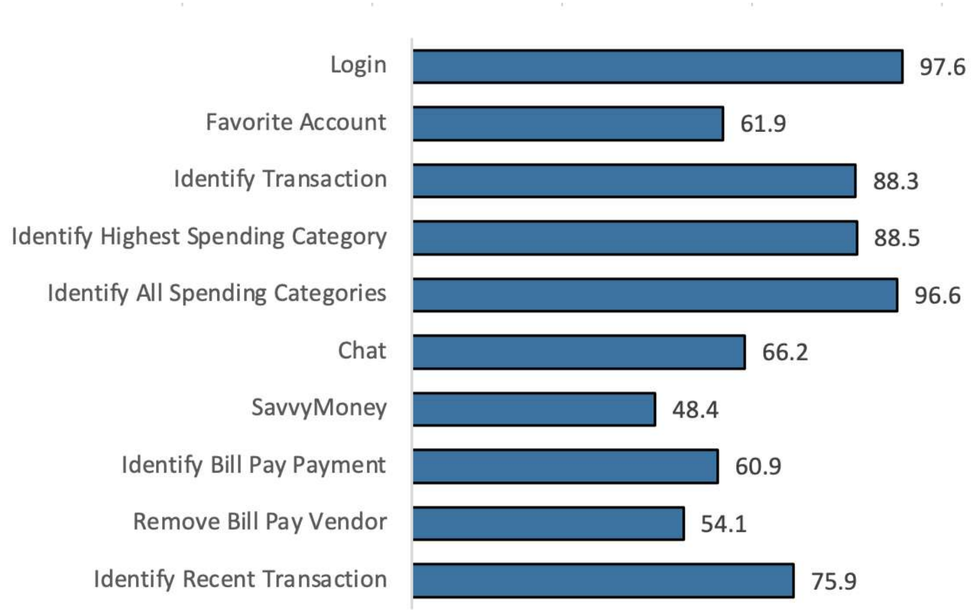
Summated Usability Metric

The average SUM score from our sample was **77.6**, well above a satisfactory score of 50 and within a “Great” range.

SUM != SUS

Since this was the first time SUM scores were presented to the company, I was worried stakeholders would compare SUM scores to SUS scores. So, I explained the differences in interpreting the two scores. Think about the SUS. Based on past research, we know that average scores are around 68. So, a score above 68 would be above average. With SUM, however, any score above 50 is satisfactory. Based on a simple breakdown of the SUM into equivalent ranges, a score of 77.6 is “Great.”

Scale	
Bad	1 to 16
Poor	17 to 33
Sub-Par	34 to 49
Satisfactory	50
Good	51 to 66
Great	67 to 83
Excellent	84 to 100



Impact and Accessibility Testing

Project Impact

This was the first quantitative usability test conducted at Q2, and overall, the project was a major success. The project introduced the SUM to the company and emphasized its value. As a result, SUM scores will be collected for quantitative usability tests moving forward. Anecdotally, I saw stakeholders promote the project's SUM score within the company. The metrics from this project will be used to advertise the new feature during Q2's annual customer conference in May 2023. As well, usability issues uncovered during testing have already led to feature redesigns.

Accessibility Testing

Immediately following this project, I conducted an accessibility evaluation of the redesigned feature. As part of that evaluation, I moderated accessibility testing sessions of the redesigned feature with two individuals with blindness who use screen readers to navigate the web. Like the quantitative usability test discussed here, my accessibility evaluation was the first accessibility test with individuals with blindness conducted within the company. We previously used external vendors to conduct accessibility tests with individuals with blindness. My work on this project has set the foundation for future accessibility tests within Q2.