

MEASURING SATISFACTION WITH STANDARD SURVEY INSTRUMENTS AND SINGLE-BUTTON RESPONSES ON KIOSKS

Laura R. Rabbitt, SAIC,
Jacob A. Hasselgren, SAIC,
Cynthia Cook, SAIC,
Yevgeniy B. Sirotin, Ph.D., SAIC

User satisfaction with a technology is an essential usability metric. Unlike efficiency and effectiveness, which are generally recorded during use, satisfaction is often measured subsequently using questionnaires, such as the modified version of the system usability scale (MSUS). This makes satisfaction the most costly usability measure to acquire in large-scale testing. To mitigate this cost, we compared the performance of a four-button kiosk with a standard SUS instrument for measuring satisfaction. Three hundred and fifty four demographically diverse subjects used the kiosk and completed a SUS questionnaire immediately after using one of two different alternative technologies. Kiosk ratings took only 11.43 ($sd = 7.30$) seconds on average to collect, much faster than 1200 seconds on average for the SUS. Kiosk ratings and MSUS scores were strongly correlated ($r = 0.62$, $p < .005$), showing the same pattern of differences between the tested technologies. However, the index of dispersion for kiosk ratings was 71.74% larger than for MSUS scores. We conclude that satisfaction kiosks are a cost-effective alternative for measuring satisfaction in usability studies with large sample sizes.

INTRODUCTION

The field of usability is a rapidly growing scientific field but is not without its share of controversy over definitions and best practices. When conducting a usability study, there are two major categories: summative and formative (Lewis, 2014). Summative usability studies focus on measurement to determine how well a given product or tool meets a goal (Lewis, 2014). Formative usability studies aim to identify use-errors and design interventions to mitigate use-errors (Lewis, 2014). While the aims of summative and formative usability studies are different, they always include measures of efficiency, effectiveness, and satisfaction. Satisfaction is most commonly assessed with surveys, which are very costly when conducting large-scale human subjects testing. Alternate methods to measure satisfaction that are affordable and fast to administer need to be identified.

Surveys are commonly used to measure satisfaction because they can be tailored to fit the aims of the study or gain insight into underlying traits or cognitive processes that drive participants' perceptions and actions. Additionally, many surveys are standardized with established norm values for different populations. Standardized surveys also offer high reliability, objectivity, and can be replicated and quantified (Nunnally, 1978).

One of the most common standardized instruments to investigate satisfaction is the System Usability Scale (SUS; Brooke, 1996). The SUS is a 10-item survey that is quick to administer and is free to use. Other standardized instruments used for measuring satisfaction include the Usability Metric for User Experience (UMUX; Finstad, 2010). The UMUX also has a shortened version, the UMUX-LITE, which consists of only two questions (Lewis, Utesch, & Maher, 2013). Another recently developed satisfaction metric used in usability studies is the Emotional Metric Outcomes

questionnaire, which assesses the emotional consequences of an interaction (EMO; Lewis & Mayes, 2014). The UMUX and UMUX-LITE have been validated against the SUS and have produced similar findings, indicating that both the long and short versions of the UMUX are valid measures of assessing satisfaction (Borsci, Federici, Bacchi, & Bartolucci, 2015). The EMO has also been cross-validated with the SUS, establishing its ability to measure satisfaction (Lewis & Mayes, 2014). While all these measures have been validated, they are still relatively new compared to the SUS and are not as widely utilized within the published literature.

Standardized surveys offer many benefits and some offer shortened versions. However, when conducting tests with large sample sizes in which surveys are completed in real time, surveys are costly and time intensive to administer for a number of reasons, even when using short versions. First, the survey needs to be prepared for the specifics of the test (i.e., development, modifications, printing, etc.). Second, the survey must be distributed during the test. Given an example sample size of $n=320$, requiring 12 testing sessions, up to four test personnel may be required to distribute and collect the surveys. Third, if the test subjects are to be compensated, the time required to complete the survey must be considered. Fourth, given the larger sample size, the surveys will likely need to be distributed via a paper medium, and will therefore need to be digitized upon completion into a logical database structure to enable efficient analysis. Finally, digitized responses will need to be analyzed. Consequently, the addition of a survey such as the SUS, in a large scale test can increase cost upwards of \$35,000. Given the time/cost tradeoffs associated with large scale tests, alternative methods to assess satisfaction need to be identified.

Other usability studies have utilized satisfaction without questionnaires by employing reaction cards to elicit positive and negative comments from participants (Travis, 2008).

Reaction cards from the “Desirability Toolkit” are used to assess satisfaction and may be used to direct post-test interviews (Benedek & Miner, 2002). An alternate implementation of the reaction cards is to create a checklist with the most commonly selected words and ask participants to select words from the checklist rather than sorting through cards (Travis, 2008). While cards or checklists are effective alternate methods to assess satisfaction, neither method is feasible with large sample sizes.

One alternate measure of satisfaction is through the use of automated processes. However, there has not been any investigation or validation of automated satisfaction methods. In the current study, we aimed to create an automated method of assessing user satisfaction using a kiosk equipped with four buttons. To determine the degree to which the kiosk assessed satisfaction, responses from the kiosk were compared to SUS scores.

Study Objectives

We conducted a study to evaluate the feasibility of obtaining satisfaction with different biometric devices using kiosk responses. To determine if kiosk responses are a viable measure of satisfaction, the relationship between kiosk responses and scores from a standardized instrument, specifically a modified version of the SUS, were compared.

METHODS

Overview

The study we detail here in this paper was a sub-experiment within a larger test that consisted of three semi-independent experiments. The three semi-independent experiments as a whole are referred to as a testing sequence. In the overall testing sequence, travel entry and exit processes equipped with biometric devices were tested along with various forms of feedback and signage. The testing sequence occurred over a two-week period in October 2015. Participants were scheduled to arrive on a specific day for either a morning or afternoon test session that lasted approximately four hours. All participants were consented and were given an anonymized test ID. Here, only satisfaction with two biometric devices will be discussed. The modality and brand of biometric devices are anonymized and are referred to as Biometric Device A and Biometric Device B.

The factors of interest did not overlap between each of the three experiments in the testing sequence with the exception of Biometric Device B. This device was included across multiple experiments to test how experience influenced user behavior so it was explicitly included in the experiment. However, these results will not be discussed in this paper.

In the current study, biometric devices were embedded in a Biometric Transaction Terminal (BTT), to emulate the process of a large group passing through a security checkpoint. Participants were equipped with a paper ticket and a piece of luggage to further emulate this security checkpoint

process. During the test, participants were prompted to form a queue at the BTT and proceed through the BTT one at a time.

A transaction at the BTT consisted of entering the terminal, scanning a paper ticket, followed by completing a biometric transaction, exiting the terminal, rating their experience using the button kiosk, and then proceeding through a hallway to a seating area. Once all participants had completed transactions and were seated, they completed a modified version of the SUS about the BTT specific to the biometric device embedded within the BTT. Participants repeated this process with the BTT equipped with a different biometric device two times. The first time the participants went through the BTT, participants completed a biometric transaction using Biometric Device B. The second time participants went through the BTT, participants completed a biometric transaction with Biometric Device A. The order of biometric devices paired with the BTT remained the same for each treatment group to test an effect in another experiment in the testing sequence. However, the aim of this study is to determine if kiosk and SUS produce similar satisfaction responses and assume there will be order effects on satisfaction.

Responses on the kiosk were collected immediately after participants completed a transaction with the biometric device at the BTT before they walked to the hallway. A modified version of the SUS was administered to participants while they waited in a seating area.

Participants

A total of 354 participants (177 males, 177 females) completed the study from the Washington D.C., Maryland, and Virginia areas. Participants were assigned to one of eight groups that visited a testing facility in the National Capital Region during a morning or an afternoon session over the course of a two-week period. Participants were compensated for their time at the end of the test session.

Modifications of the SUS (MSUS)

To determine participants' satisfaction with each type of biometric device, the SUS was administered after each group had completed a biometric transaction with each. In this study, two statements of the SUS were modified to give greater context to participants. The modified version of the SUS will be referred to as the MSUS for the remainder of this report. The altered wording of these items are in Table 1.

Modifications to the SUS are not unprecedented and have been altered to assist with comprehension for non-native English speakers, and perform in the same manner as the standard SUS instrument (Bangor, Kortum, & Miller, 2008; Finstad, 2006). Other research has demonstrated that only when items were rephrased to extreme positive or negative were there differences in the performance of the SUS (Sauro, 2010). The modifications made to the SUS do not alter the intent of the original statements and were not changed to be phrased to the opposite or extreme valence. Given the prior research and maintaining the intent of each statement, the

MSUS was scored and interpreted in the same manner as the standardized SUS.





Table 1. Modified wording of items 1 and 10 of the SUS.

#	Original Wording	Altered Wording
1	I think that I would like to use this product frequently.	I think that I would like to use this device to verify my identity whenever I travel.
10	I needed to learn a lot of things before I could get going with this product.	I needed many attempts before I figured out how to use this device.

Kiosk

The kiosk used in this test consisted of four buttons, each equipped with a colored emotional face. A framed sign that read “Rate our gate!” was placed behind the buttons and served as instructions to participants when they walked up to the kiosk. The emotional faces on the buttons range from smiling to frowning and correspond to different levels of satisfaction. A depiction of the colored faces on the kiosk buttons and their corresponding satisfaction and numerical values are in Table 2.

Table 2. Kiosk button response images, numerical values, and satisfaction rating.

Button Response				
Numerical Value	1	2	3	4
Satisfaction Rating	Very Happy	Happy	Unhappy	Very Unhappy

The kiosk consisted of four buttons so that participants would not make a neutral response and required them to choose either the positive or negative end of the scale. In some research, the mid-point of a scale is often used as a “don’t know” or “no opinion” option (Baumgartener, & Steenkamp, 2001). To avoid this effect, the kiosks were created with four buttons.

Statistical Analyses

Data was analyzed with R using custom scripts written by the authors (<https://www.r-project.org/>). After data collection, the data was cleaned to remove outlier values, and faulty survey responses (e.g., did not fill out all questions in survey).

Transforming Kiosk Response Values

MSUS responses are standardly quantified into a score ranging from 0-100, with a higher score corresponding to a higher level of satisfaction. The kiosk buttons were assigned

values ranging from 1 (very happy) – 4 (very unhappy), with a higher score corresponding to a higher level of dissatisfaction. Therefore, to compare the two, a transformation of the kiosk button responses was necessary.

Equation 1 below was used to convert kiosk response values.

$$Kiosk\ score = \left(1 - \frac{(x-1)}{3} \right) \times 100 \tag{1}$$

Where *x* is the numerical value of the kiosk response.

Calculating Kiosk Times

The amount of time to submit a response on the kiosk was calculated using timing data from a different experiment. The kiosk in both tests was set up identically, but the other test included a ground truth scan immediately before participants pressed a button on the kiosk. The time between the wristband scan and button response was calculated for each participant. The addition of a wristband scan to establish ground truth would not have contributed significant amounts of time to button response.

The time to administer the MSUS was not recorded during the experiment. However, the test plan allotted 20-minutes during the test to administer the MSUS to participants.

Index of Dispersion and Differences in Dispersion

To calculate the index of dispersion, equation 2 was applied to kiosk values and MSUS scores.

$$D = \frac{\sigma^2}{\mu} \tag{2}$$

To calculate the difference in dispersion, equation 3 was used. The value resulting from this equation was multiplied by 100 to convert this value to a percentage.

$$D_{diff} = \frac{(D_{Kiosk} - D_{MSUS})}{\left(\frac{D_{Kiosk} + D_{MSUS}}{2} \right)} \tag{3}$$

Scoring the MSUS

The MSUS is scored in the same manner as other versions of the SUS. For positive valance items (1, 3, 5, 7, and 9), the scale score has 1 subtracted from it. For negative valance items (2, 4, 6, 8, and 10), the scale score is subtracted from 5. After all items were transformed, the values are summed and multiplied by 2.5 to obtain an overall score ranging from 0 to 100.

RESULTS

Relation of Kiosk and Survey Responses

To determine the relationship between MSUS scores and kiosk responses, a correlation between paired samples was calculated. The correlation between the MSUS scores and kiosk responses were strongly and significantly correlated ($r = 0.62, p < .005$). The positive relation indicates that a

higher response on the kiosk is associated with a higher MSUS score for either technology. This correlation is displayed in Figure 1.

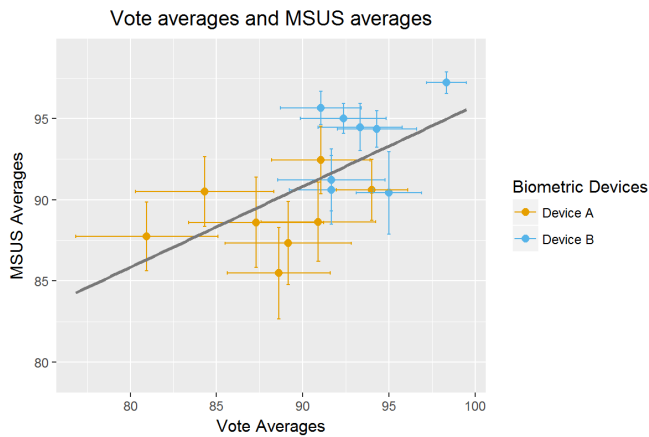


Figure 1. Correlation of Votes Averages and MSUS averages.

Measuring Satisfaction with Kiosk and Survey Responses

The difference between the two biometric methods tested, Biometric Device A and Biometric Device B, was significantly different for both the MSUS scores ($t(7) = 6.41, p < .0005$) and for kiosk responses ($t(7) = 2.75, p < .05$). These results indicate there was a preference for the Biometric Device B over Biometric Device A.

There was no difference between MSUS scores and kiosk responses for Biometric Device B ($t(7) = -0.16, p = .87$). There was no difference between MSUS scores and kiosks responses for Biometric Device A ($t(7) = -0.44, p = .67$). Means and standard deviations of the MSUS and kiosk for each biometric device are displayed in Table 3.

Table 3. MSUS and Kiosk means and standard deviations for each biometric device.

	MSUS		Kiosk		t-test
	M	SD	M	SD	
Device A	88.93	15.59	88.29	22.54	0.16
Device B	93.62	10.41	93.47	15.10	0.44

These findings indicate that the satisfaction scores for each biometric device do not differ between MSUS scores and kiosk responses. These results are displayed in Figure 2.

Average Kiosk Times and Index of Dispersion

The average amount of time to complete satisfaction kiosks took 11.43 seconds. The time to administer the MSUS was not explicitly measured during the experiment. However, prior tests (with similar sample sizes) that used the MSUS guided the amount of time that was allotted during the test to administer the MSUS (DHS S&T & CBP, 2015a; DHS S&T & CBP, 2015b; DHS S&T & CBP, 2016). Given the prior

experiences, 20-minutes (1200 seconds) was allotted during the study to administer the MSUS.

The index of dispersion for both satisfaction measures used equation 1. The index of dispersion for the MSUS was 2.00 while the index of dispersion for the kiosk was 4.24. To determine how much more dispersion there was for the kiosk than the MSUS, equation 2 was used. The difference in index dispersion was 71.74 %, which demonstrates that MSUS scores are far more clustered together closer than kiosk scores.

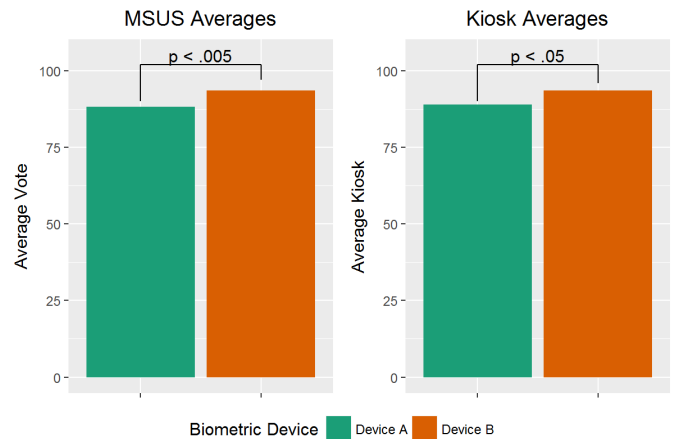


Figure 2. MSUS averages and kiosk averages for Biometric Device A and Biometric Device B.

DISCUSSION

Usability studies focus on a technology’s or process’s effectiveness, efficiency, and satisfaction. When testing large sample sizes, satisfaction measures can be extremely costly to administer. Using shortened versions of surveys or even electronic surveys are still time consuming to administer and score. Satisfaction can be measured alternatively using reaction cards or checklists but these methods are still not feasible with large sample sizes.

The cost of administering surveys is not driven by the cost of materials, rather facility and labor costs. The amount of time to operate a testing facility, and the cost of the individuals conducting the test significantly add to the budget of a large-scale human subjects test. By reducing the time it takes to get satisfaction data greatly lowers the amount of money needed to execute a test.

This study demonstrates that kiosks are an effective method for measuring satisfaction given the relation between kiosk responses and MSUS scores. Both kiosk responses and MSUS scores were able to statistically determine that participants had higher levels of satisfaction with the Biometric Device B than Biometric Device A. There was no difference between MSUS scores and kiosk responses for the same type of biometric device, indicating that participants submitted the same feedback on both instruments. While both the kiosk responses and MSUS scores show the same pattern of responses, indices of dispersion demonstrate that kiosks are only suitable with large sample sizes. While satisfaction with each biometric device may have been influenced by order that

it was used, both instruments detected order effects. Future testing should consider randomizing or counterbalancing the order of treatments and biometric devices to avoid any ordering effects on satisfaction.

The time to administer the MSUS was estimated to take approximately 1200 seconds (20 minutes) to an entire group, while the time to submit responses on the kiosk was estimated to take approximately 11.43 seconds per participant. However, the kiosk was completed concurrently with a task. By measuring satisfaction concurrently, the amount of time to assess satisfaction is reduced by a considerable amount.

The use of kiosks to assess satisfaction are commonly used in marketing studies but have not been implemented into usability studies. To our knowledge, this is the first time that a kiosk to measure satisfaction has been employed and validated in a usability study. Our results demonstrate that this method is a viable satisfaction metric and may be a useful replacement to surveys when large-sample sizes are available. Future directions for the kiosks include introducing dynamic signage to present additional information and feedback when a response is recorded.

Acknowledgments. The research for this paper was fully funded by the Department of Homeland Security Science and Technology Directorate on contract number W911NF-13-D-0006-0003. The views presented here are those of the authors and do not represent those of the Department of Homeland Security or of the U.S. Government.

REFERENCES

- Bangor, A., Kortum, P. T., Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6), 574 – 594.
- Baumgartner, H., & Steenkamp, J. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38, 143 – 153
- Benedek, J., & Miner, T. “Measuring Desirability: New Methods for Evaluating Desirability in a Usability Lab Setting” (Word document) Redmond, WA: Microsoft Corporation, 2002.
- Borsci, S., Federici, S., Bacci, S., Gnaldi, M. & Barolucci, F. (2015). Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International Journal of Human-Computer Interaction*, 31(8), 484 – 495.
- Brooke, J. (1996). SUS: A “quick and dirty” usability scale. *Usability evaluation in industry*, 189(194), 4 – 7.
- Department of Homeland Security Science and Technology Directorate (DHS S&T) & Customs and Border Protection (CBP). (2015a). *Air entry/exit re-engineering (AEER) project Scenario Evaluation Reports Sequence 1*. Washington, D.C.: Author.
- Department of Homeland Security Science and Technology Directorate (DHS S&T) & Customs and Border Protection (CBP). (2015b). *Air entry/exit re-engineering (AEER) project Scenario Evaluation Reports Sequence 2*. Washington, D.C.: Author.
- Department of Homeland Security Science and Technology Directorate (DHS S&T) & Customs and Border Protection (CBP). (2016). *Air entry/exit re-engineering (AEER) project Scenario Evaluation Reports Sequence 3*. Washington, D.C.: Author.
- Finstad, K. (2006). The system usability scale and non-native English speakers. *Journal of Usability Studies*, 1(4), 185 – 188.
- Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, 22, 323 – 327.
- Lewis, J. R. (1993). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1), 57 – 78.
- Lewis, J. R., & Mayes, D. K. (2014). Development and psychometric evaluation of the emotional metric outcomes (EMO) questionnaire. *International Journal of Human-Computer Interaction*, 30(9), 685 – 702.
- Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In M. Kurosu (Ed.), *Human centered design* (pp. 94 – 103). Heidelberg, Germany: Springer-Verlag.
- Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013, April). UMUX-LITE: When there’s no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099 – 2102). ACM.
- Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.
- Sauro, J. (2010). *That’s the worst website ever!: Effects of extreme survey items*. Retrieved from: www.measuringusability.com/blog/extreme-items.php.
- Travis, D. (2008). Measuring satisfaction: Beyond the usability questionnaire. Retrieved from: <http://www.userfocus.co.uk/articles/satisfaction.html>.