A CALL TO REVISIT CLASSIC MEASUREMENTS FOR UX EVALUATION

Stephen Schneider

Microsoft, Stephen.schneider@microsoft.com

Serena Hillman

Microsoft, serena.hilman@microsoft.com

CCS CONCEPTS • Human-centered computing • Human computer interaction • HCI design and evaluation methods

Additional Keywords and Phrases: Summative research, Experience outcomes, KPIs, UX metrics, UX health, OKRs, UX outcomes, user-centered metrics

ACM Reference Format:

Stephen Schneider and Serena Hillman. A Call to Revisit Classic Measurements for UX Evaluation. 2023. CSCW ’23: 26th ACM Conference on Computer-Supported Cooperative Work and Social Computing, October 14-18, 2023, NY. ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

We are user researchers supporting a variety of data and analytics products within Microsoft’s Azure Data organization. One of our prominent ongoing projects involves the comprehensive reevaluation and rejuvenation of user experience (UX) evaluation metrics. In this paper, we present a specific study conducted as part of this broader initiative, which is aimed at aligning and simplifying the System Usability Scale (SUS) for enterprise applications.

The importance of updating and enhancing traditional UX measurement tools such as SUS (System Usability Scale) or UMUX Lite (Usability Metric for User Experience) cannot be overstated. Several critical factors drive the need for this ongoing evolution. Firstly, it is imperative to stay attuned to shifting user preferences in the rapidly evolving digital landscape. Secondly, adapting to the latest technological advancements is crucial to ensure the continued relevance of these metrics. Thirdly, there is a constant drive to improve measurement methods to provide more accurate and actionable insights. Additionally, aligning with stakeholder needs, including the ever-changing landscape of business metrics, is vital for demonstrating the value of UX research. Furthermore, regularly reviewing the reliability and validity of psychometric properties ensures that these metrics maintain their scientific rigor. Lastly, addressing evolving cultural and contextual factors is essential to ensure the applicability of these tools in a global and diverse user base.

One of the cornerstone tools in UX research is the System Usability Scale, which enjoys widespread usage. It comprises of 10 items assessed using an anchored response scale, where 1 represents "strongly disagree" and 5 represents "strongly agree" (as illustrated in Figure 1). An important characteristic of SUS is its proven reliability, even when dealing with relatively small sample sizes.

Understanding and enhancing the usability of enterprise applications is particularly vital in the context of a UX outcomes workshop. It directly impacts user satisfaction, productivity, and ultimately, the success of data and analytics products. Therefore, our efforts to refine and adapt measurement tools like SUS to this specific domain play a pivotal role in achieving meaningful UX outcomes for enterprise applications.

SUS Positive 10-items	**1**	**2**	**3**	**4**	**5**
I think that I would like to use this product frequently.	Strongly disagree	*	*	*	Strongly agree
I found this product unnecessarily complex.	Strongly disagree	*	*	*	Strongly agree
I thought this product was easy to use.	Strongly disagree	*	*	*	Strongly agree
I think that I would need the support of a technical person to be able to use this product.	Strongly disagree	*	*	*	Strongly agree
I found the various functions in this product were well integrated.	Strongly disagree	*	*	*	Strongly agree
I thought there was too much inconsistency in this product.	Strongly disagree	*	*	*	Strongly agree
I would imagine that most people would learn to use this product very quickly.	Strongly disagree	*	*	*	Strongly agree
I found this product very cumbersome to use.	Strongly disagree	*	*	*	Strongly agree
I felt very confident using this product.	Strongly disagree	*	*	*	Strongly agree
I needed to learn a lot of things before I could get going with this product.	Strongly disagree	*	*	*	Strongly agree

2 STUDY GOALS

There are several concerns that warrant attention for SUS and are the focus of this study. Firstly, the inclusion of 10 items in the survey can potentially lead to survey fatigue, especially when conducting benchmark studies that involve task-level questions after each task, such as assessing Satisfaction, Ease of Use, and Appeal. Additionally, there are study-level metrics that are asked at the end of a session, including Satisfaction and SUS (System Usability Scale). In light of this, any reduction in the number of items would be beneficial for reducing survey fatigue.

Another noteworthy aspect of the SUS is its use of an anchored labeled response scale. Anchor response scales, like the SUS's, tend to yield lower data quality and reduced reliability compared to fully labeled response scales. By using a fully labeled response scale, we expect better data quality and a more reliable measure.

However, the most significant concern for us revolves around the alignment of some of the questions with the reality of working with enterprise products. Several of the questions used in the SUS tend to not match the reality faced by enterprise product users. To illustrate this, we provide two examples from the SUS:

- The question "I think I would like to use the system frequently" may not apply to enterprise users who lack the choice of whether to use a particular product or not. Their usage is often mandatory.

- Similarly, the statement "I think I need the support of a technical person to be able to use this product" may not be applicable to IT professionals, as they may not have access to another technical person for assistance.

These concerns highlight the need to adapt and refine the usability assessment tools to better capture the nuances of the enterprise context and ensure the relevance and accuracy of the collected data.

From this, our study goals were to:

- Create a shorter SUS measure.
- Use a fully labeled response scale to improve data quality.
- Create a questionnaire that applies to enterprise and data enterprise products.
- Test the convergent validity of the new questionnaire with satisfaction and SUS to determine whether the new questionnaire is a suitable replacement for SUS.

3 APPROACH

To address our study goals, we adopted a two-fold methodology. First, we conducted an expert review of the SUS with the aim of reducing the number of items and ensuring their applicability to enterprise and data enterprise products. Subsequently, we carried out a retrospective survey to further validate the scale's suitability for these contexts and assess its convergent validity as a potential replacement for the SUS.

During the expert review phase, a panel of nine researchers convened with two primary objectives. The first was to eliminate redundancy within the SUS, and the second was to scrutinize the items to confirm their relevance within the enterprise product domain. As a result of this review, the panel selected the following five items, called the ESUS:

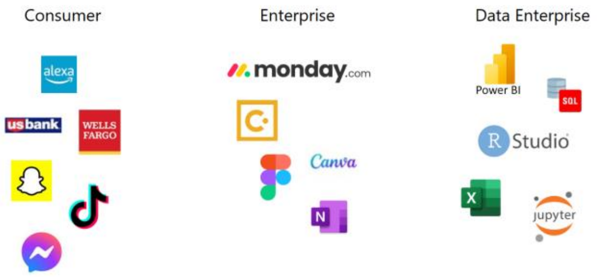
ESUS Items	1	2	3	4	5
How useful is [this product] to you?	Not at all useful	Slightly useful	Somewhat useful	Mostly useful	Very useful
How easy or hard was [this product] to use for you?	Very hard	Hard	Neutral	Easy	Very easy
How confident were you when using [this product]?	Not at all confident	Slightly confident	Somewhat confident	Mostly confident	Very confident
How well do the functions work together or do not work together in [this product]?	Does not work together at all	Does not work well together	Neutral	Works well together	Works very well together
How easy or hard was it to get started with [this product]?	Very hard	Hard	Neutral	Easy	Very Easy

In our assessment of convergent validity, we examined two specific questions:

- Does the ESUS-SAT relationship maintain a similar level of strength compared to the SUS-SAT relationship?

- Is there a strong correlation between ESUS and SUS?

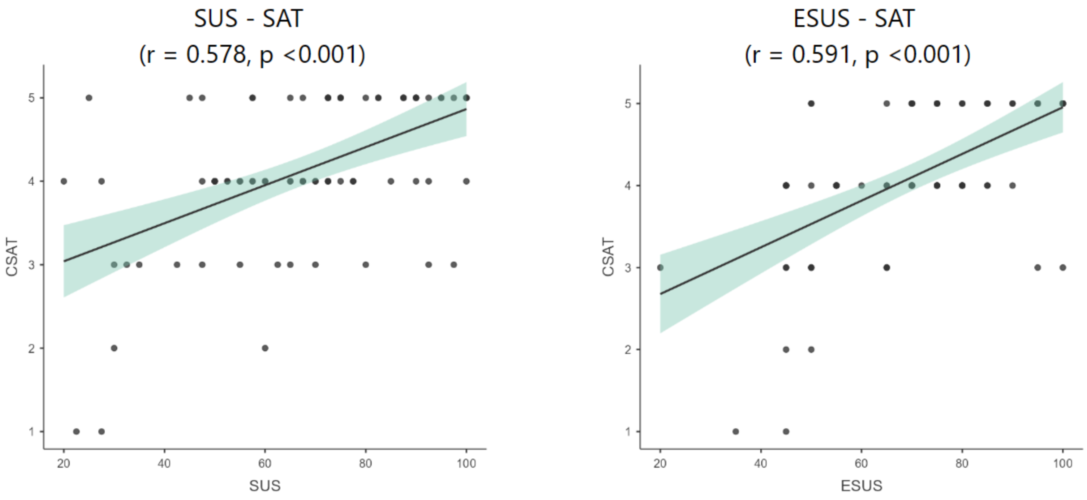
After gathering the data, we categorized the products mentioned by participants into three distinct categories: Consumer, Enterprise, and Enterprise Data (which is a subset specifically focused on data applications within the enterprise context). For the purposes of this study, we did not focus on Consumer scores; instead, we reserved that aspect for exploration in a subsequent paper [2]. Consequently, we excluded Consumer scores from any further analysis. We then examined the scores for Enterprise Data and Enterprise categories. Interestingly, our analysis revealed no significant differences between these two groups, prompting us to combine them for further analysis.



4 STUDY OUTCOMES

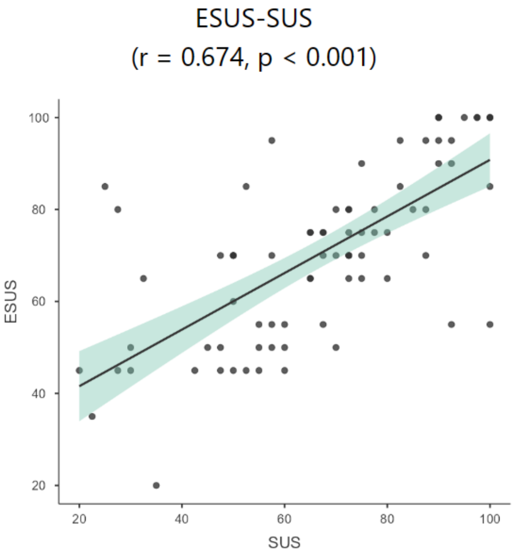
1) Is the strength of the SUS-SAT relationship maintained with ESUS-SAT?

If the ESUS has convergent validity with SUS, the ESUS should have a similarly strong correlation with SAT as the SUS has with SAT. When examining the Pearson correlation coefficients, we find that the SUS-SAT relationship has a coefficient of 0.579, and the ESUS-SAT relationship has a coefficient of 0.591. Both coefficients fall within a similar range, indicating strong relationships. Therefore, we can conclude that the strength of the SUS-SAT relationship remains consistent even when using the new measure, ESUS.



2) Is ESUS strongly correlated with SUS?

A strong correlation between the ESUS and SUS may indicate that they are picking up on the same concept, another key aspect of convergent validity. The correlation coefficient between ESUS and SUS was 0.674. According to Cohen's classification of correlation strength [1], this indicates a strong effect size, particularly when considering human behavioral data, suggesting that the two are picking up on a similar concept.



5 FUTURE WORK

While we plan to continue exploring ESUS by examining its test-retest reliability, we also plan to test ESUS's sensitivity to small sample sizes. Future work should also examine the relationship between the ESUS and other measures of usability, like the UMUX-Lite. Our primary objective is to contribute to the broader field of UX evaluation measurements by revisiting and revitalizing them, ensuring that are metrics keep pace with the ever-changing enterprise product space.

Note: this study is to appear in full in the Journal of User Experience's February 2024 issue [2].

REFERENCES

[1] Jacob Cohen. 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. Routledge

[2] Stephen Schneider, Serena Hillman, Paula Bach and Guoping Ma. 2024. ESUS: Aligning & Simplifying SUS for Enterprise Applications. To appear in the Journal of User Experience. February 2024 Issue.