Journal of User Experience

Vol. 19, Issue 2, February 2024 pp. 60-74

ESUS: Aligning and Simplifying SUS for Enterprise Applications

Stephen Schneider
Researcher
Microsoft
1 Microsoft Way,
Redmond, WA
USA
stschneider@microsoft.com

Serena Hillman
Senior Research Manager
Microsoft
1 Microsoft Way,
Redmond, WA
USA
sehillma@microsoft.com

Paula Bach
Principal Research Director
Microsoft
1 Microsoft Way,
Redmond, WA
USA
pbach@microsoft.com

Guoping Ma
Principal Research Manager
Microsoft
1 Microsoft Way,
Redmond, WA
USA
guopingm@microsoft.com

Abstract

Throughout the last few decades, researchers have developed standard usability questionnaires to evaluate usability and present a single score that represents a product's overall level of ease of use. One of the most notable questionnaires is the System Usability Scale (SUS) (Sauro & Lewis, 2009). However, since the SUS was introduced in 1986, products and services have not only undergone monumental advancements in technology, but Human-Computer Interaction and user experience research practices have matured. These changes are also true in the enterprise space. In this paper, we present preliminary evidence concerning the construct validity of a new usability questionnaire with three advantages for enterprise applications over the original 10-item SUS questionnaire. The Enterprise System Usability Scale (ESUS) offers better measurement of usability for technical products/services, reduced questionnaire items, and alignment with enterprise environments. Results indicate that the ESUS exhibits a similarly strong correlation with satisfaction as the SUS and is strongly correlated with SUS for enterprise and enterprise data products/services.

Keywords

SUS, enterprise applications, usability questionnaires, ease of use, satisfaction, data enterprise applications, ESUS, System Usability Scale

Introduction

Assessing product usability to represent results as a single number is impactful for Human-Computer Interaction (HCI) and UX researchers. HCI researchers have created many standardized usability questionnaires including the System Usability Scale (SUS), Software Usability Measurement Inventory (SUMI), Standardized User Experience Percentile Rank Questionnaire (SUPR-Q®), Single Ease Question (SEQ), Subjective Mental Effort Questionnaire (SMEQ), Usability Metric for User Experience (UMUX), UMUX-lite, Computer System Usability Questionnaire (CSUQ), and the Website Analysis and MeasureMent Inventory (WAMMI). The most common of these usability-specific questionnaires is SUS.

History

The evolution of UX over the last few decades has created a need to capture multiple metrics. UX often spans and explores the relationship between business and established usability metrics like the SUS (Hillman et al., 2022). Usability questionnaires are now often coupled with business metrics that measure business satisfaction like the customer experience (CX) business metrics including Net Promoter Score® (NPS) (Reichheld, 2003), Customer Satisfaction Score (CSAT) (Dixon et al., 2010), and Customer Effort Score (CES). Using metrics to represent user outcomes and business outcomes is complex and often requires companies to capture multiple metrics beyond usability. Yet, capturing multiple metrics in a single study can lead to participants' survey fatigue. Methods such as benchmarking amplify the issue by requiring participants to complete task-level questions in addition to the primary, test-level SUS questions. This challenge illustrates the industry's need to reduce the number of items in the current 10-item SUS evaluation. Moreover, the extreme advancements in technology over the last several decades have changed the language we use to describe technology, how we interact with it, and what usability means.

Expert Panel

As researchers executing benchmarks and summative surveys regularly in practice, we have observed these gaps in SUS firsthand. In the Spring of 2022, a group of nine researchers, who represented teams across approximately 50 applications and services in the Microsoft™ cloud and developer space, participated in expert feedback sessions. We discussed if SUS should continue to be used as a standard usability summative metric within our enterprise space. Based on sharing lived experiences as industry researchers implementing benchmarks and surveys, they raised substantial concern that some items within SUS were too repetitive (that is, items being highly correlated) and not aligned with enterprise products and users (specifically, highly technical enterprise data users). For example, the SUS item, "I think that I would like to use the product frequently," can create confusion for enterprise data professionals because they are not the sole decisionmakers for which software is used by the organization; using specific software is often a requirement of their role. Additionally, the SUS item, "I think I need the support of a technical person to be able to use this product," can also be confusing for highly technical enterprise users. Often, there is no technical support for enterprise users; they are the technical person in question. Based on these discussions and the industry's need to reduce the number of items, we proposed a new 5-item usability questionnaire set.

A 5-Item Usability Questionnaire

The new 5-item usability questionnaire not only reduces the number of items to account for survey fatigue, but it also removes SUS items related to a technical support person and frequency of use; it reduces item redundancy and aligns the scale with common business metrics in a 5-point ordinal scale. The new questionnaire covers the following constructs: ease of use, usefulness, integration with other applications, user confidence in application use, and ease of getting started. We conducted semantic testing on the items to ensure the intent of the questions aligned with user interpretations (see Tables 6 and 7 in the Appendix for a complete list of items and scales).

Research Objectives

In this study, our objective was to explore solutions to the above-noted challenges; specifically, our research questions were: 1. Do different application categories (Enterprise, Consumer, and Enterprise Data) illicit different responses? 2. Would the new 5-item usability questionnaire predict SUS and CSAT? 3. Are there individual differences in scores between the new questionnaire and SUS?

To answer these questions, we surveyed 84 participants to capture SUS scores and scores from the new 5-item usability questionnaire in both unipolar and bipolar formats. We surveyed across the different categories, and we compared the results. For categorization, we defined Enterprise applications as technology that is used within organizations (such as Microsoft Word™), Enterprise Data applications as the technology used to support data professionals in an enterprise setting (such as Power BI™), and Consumer applications (Instagram™) as products or services used by individuals or households for personal use. Our selection of these three categories is based on our intention to compare enterprise and consumer applications and to create a sub-category that targets enterprise data applications. Furthermore, the complexity of the enterprise data space presents an intriguing opportunity for comparison, as it may require specialized tools and expertise to manage and analyze data at scale. Last, our department has a vested interest in researching the enterprise data space, which adds to the significance of this particular comparison. The results led us to a new, SUS-like usability questionnaire for use specifically with enterprise and enterprise data applications. This enterprise-version SUS was dubbed ESUS (see Table 4 for the final ESUS).

Related Work

In this section, we review the task and test-level standardized usability questionnaires, relationships with business-related metrics, and enterprise application usability. Also, we outline the knowledge gaps and how our study addressed these.

Standardized Questionnaires

Test-level and task-level standard usability questionnaires are widely used because they provide more reliability than internally developed alternatives (Sauro & Lewis, 2009; Sauro, 2011). Task-level questions are asked after each task, whereas test-level questions are asked after all tasks are finished. For test-based usability questionnaires, SUS is the industry standard. It was introduced in 1986 to quickly measure a user's subjective perception of the usability of a system (Brooke, 1996). Over the years, it was found reliable even with small sample sizes (Tullis & Stetson, 2004) and has been cited by over 15,000 publications (Brooke, 1996, as cited in Google™ Scholar™), making it the most popular standardized usability questionnaire. To date, changes to SUS have been adopted, the most notable being the introduction of a fully positive item set based on findings from Lewis and Sauro (2017). In their 2011 CHI paper, the authors find no difference in responses between the original SUS items, with both positive and negative items, and a SUS version with only positive items. They did, however, find that the original SUS created mis-scoring by researchers and errors indicative of participants forgetting to reverse their responses. Although SUS provides a reliable standard in an industry setting, 10 items can be lengthy for participants. As mentioned above, the SUS's length is especially challenging when combined with equally important business metrics or a lengthy benchmark session. Additionally, our expert feedback session revealed that SUS has been criticized for having high redundancy, perhaps being too redundant for the measurement of only one construct (Sauro and Lewis, 2011).

UMUX (Finstad, 2010) was developed in 2010 as a shorter alternative to SUS. This questionnaire originally included 4 items, 2 positive and 2 negative items, against a 7-point agree/disagree scale. UMUX was designed to target the ISO 9241 definition of usability covering effectiveness, efficiency, and satisfaction (Finstad, 2010). However, it was later shown that the combination of positive and negative items created a two-factor structure (Lewis et al., 2015). These results were similar to Lewis and Sauro's findings regarding SUS, thus the change to an all-positive scale (2017). UMUX-Lite is modeled after the Technology Adoption Model (TAM)(Davis, 1989), a model that predicts the adoption of new technologies by measuring the perceived ease of use and perceived usefulness of a product with two items. UMUX-Lite has shown promise with high internal reliability (Lewis et al., 2015). Although UMUX and UMUX-Lite

were both developed due to the need for a shorter questionnaire, our objective was to leverage the reliability of SUS to create a SUS-like score better suited for enterprise needs.

Additional questionnaires exist that have more than 10 items. The SUMI (Kirakowski, 1993) has 50 items, WAMMI (Kirakowski & Cierlik, 1998) has 20 items, CSUQ and PSSUQ (Lewis, 1992) both have 16 items. Such questionnaires often measure additional constructs beyond usability, such as SUPR-Q (Sauro, 2015) which measures trust, appearance, and loyalty. It should also be noted that task-level questionnaires such as SEQ and SMEQ (Sauro & Dumas, 2009) bring up discussions around misalignment with a user's actions. For example, it has been reported that approximately 14% of users select the highest possible score for satisfaction despite failing the task (Sauro, 2012). This is less of a concern at the test-level as there is no direct, or one-to-one comparison, between task success and task-level metrics. Task-level questionnaires are an important element to benchmarking and thus notable for us to call out as something we did not focus on in this study.

Business Measurement

Including business metrics is often imperative when conducting usability studies. This is because UX outcomes and business outcomes should complement each other (Hillman et al., 2022). Showing the impact of usability metrics on business metrics is an important story for industry researchers to understand and share (Hillman et al., 2022). NPS (Reichheld, 2003), CES (Dixon et al., 2010), and CSAT (Coelho & Esteves, 2007) are some of the most used business measurements. These metrics summarize the loyalty or satisfaction of a user or customer when interacting with a business. Usability questionnaires have been shown to impact NPS and satisfaction (Bangor et al., 2013; Sauro, 2010). Excluding NPS, which employs an 11-point scale, CES and CSAT have a 5-point ordinal scale. SUS uses a 5-point anchored agreement scale. Our new 5-item questionnaire leverages existing business metric norms with a 5-point ordinal scale. We took this direction because business and usability questions often appear together in benchmarking or survey studies within our organization.

Enterprise Application Usability

It is important to note that enterprise applications have traditionally performed worse than their consumer counterparts for usability. In 2006, Nielsen described this product group as "stuck in the 1990s" and showed that these applications, on average, have a 58% success rate on tasks versus consumer applications at 66% (2006). In 2010, Sauro also explored consumer applications and enterprise applications and found a staggering 2:1 ratio of enterprise-to-consumer usability issues (2017). This gap is thought to exist because enterprise applications are often more complex, specialized, and purchasing/selection often involves a range of different users from different areas of the business (Loranger et al., 2007). Furthermore, enterprise applications sometimes deliberately restrict features based on pay-per-feature models, causing additional complexities around navigation, information architecture, and users' mental models (Loranger et al., 2007). While we acknowledge consumer apps have paywalls as well, there is often more transparency regarding what is included and what you must pay for. As we capture results comparing consumer versus enterprise applications, understanding the basis of how these systems perform, and why, is important to keep in mind.

Method

Study Design

Our research questions follow. RQ1: Do different application categories (Enterprise, Consumer, or Enterprise Data) illicit different responses to the usability scales? RQ2: Is the new 5-item usability questionnaire correlated with the SUS and CSAT? RQ3: Are there individual differences between the three scales?

To address these research questions, we administered a survey using a within-subject study design.

We recognize the learning effect limitation of a within-subject study design, but we selected this approach instead of a between-subject design because we were concerned about being able to

recruit enough participants with enterprise data experience. However, to limit the impact of learning effects, we randomized the order of the survey items for each participant. We then used ANOVA models to compare differences between the application categories to answer RQ1. To answer RQ2, we used correlations to understand how well the questionnaires performed against SUS and CSAT. Finally, we created difference scores between each of the new scales and SUS to answer RQ3 and to examine mean differences and the variance in differences.

Participants

For the survey, we recruited participants from North America, Europe, and Australia from popular third-party recruitment websites (usertesting.com and userinterviews.com). We specifically recruited participants for three distinct application categories: Enterprise, Consumer, and Enterprise Data applications. In the next section, we provide more details about how these participants were screened into or out of the survey.

Procedure

Participants completed a survey that took about 10 minutes. The initial part of the survey requested participants to recall the most recent product, tool, or application they learned to use in the past 6 months. This ensured participants could provide a more accurate response by selecting a tool they had recently learned. If a participant hadn't learned any new product, tool, or application in the last 6 months, they were instructed to close the survey and get in touch with the recruiter. Participants then answered a standard CSAT (Dixon et al., 2010) to evaluate their satisfaction of the application (see Table 5 in the Appendix). Next, participants received the 10-item SUS (see Table 8 in the Appendix), the new 5-item set in bipolar format (see Table 7 in the Appendix), and the new 5-item set in unipolar format (see Table 6 in the Appendix) in a randomized order. Each participant was required to complete all the questionnaires. Again, the focus of this evaluation was what they had identified earlier as the last product, tool, or application they had recently learned.

Since RQ1 sought to examine differences in the usability scales between the three application categories, we conducted an a priori power analysis, which indicated that 102 participants, 34 per application category, would provide sufficient power to detect medium to large differences between the three application categories. However, 18 participants were excluded from the analysis since they named applications in the first section of the survey that the researchers were unaware of or could not find through an internet search. The analytic sample size was 84.

In section 1, participants provided the name of the application they had last learned to use, and their responses were coded into mutually exclusive application categories: Enterprise Data, Enterprise, and Consumer. Two researchers independently completed the coding process for all the responses. After the independent coding, they conducted a joint review of their codes and discussed any discrepancies to come to an agreement and align their codes. Consumer applications were defined as products or services used by individuals or households for personal use. Examples of Consumer applications include the Alexa™ app, personal banking app(s), Snapchat™, smart homes, and Meta™ Messenger. Enterprise applications were defined as technology that is used within organizations. Examples of Enterprise applications we collected included Monday.com™, Concur®, Canva™, Sentinel™ CRM, OneNote™, and Figma™. Our last application category was a sub-set of Enterprise applications, Enterprise Data. These applications are used specifically by data professionals in an enterprise setting around the extracting, transforming, and loading of data, as well as analysis and machine learning activities. Examples of Enterprise Data applications seen in our data set included Power BI, Oracle® SQL Developer, RStudio®. After coding out the different application categories, we had 20 Enterprise, 28 Enterprise Data, and 36 Consumer applications. We then applied the SUS scoring formula to both the new 5-item set raw data and the original 10-item SUS raw data so we could compare the results. The formula was this: For each of the 5 question scores (per participant), subtract 1, add all 5 together (for a total between 0 and 20), and multiply the sum by 5 for a per-participant score ((Σ(Xscore-1))*5).

Results

Results Across Application Categories

We conducted a between-subjects ANOVA comparing the mean SUS, the new 5-item set in both bipolar and unipolar scales, and CSAT scores between Enterprise Data, Enterprise, and Consumer application evaluations to determine if there was a significant difference in scores between application areas. If the ANOVA revealed a significant *F*-test, Tukey's post-hoc test with Bonferroni corrections were used to identify which pair of application evaluations were significantly different.

Table 1. Means (Standard Deviation) of Application Areas

	All areas		**Consumer**	**Enterprise**	**Enterprise Data**
		Cronbach's α			
CSAT	4.12 (0.91)	-	4.25 (0.87)	3.9 (1.17)	4.11 (0.74)
10-item SUS	67.4 (21.0)	0.877	73.6 (22.81)	62.5 (21.00)	63.39 (17.10)
New 5-item SUS Bipolar	71.0 (18.5)	0.775	78.29 (19.33)	66.25 (18.49)	65.18 (14.30)
New 5-item SUS Unipolar	71.9 (19.2)	0.791	79.4(20.4)	65.0(15.8)	66.5(16.4)

CSAT

Although Consumer applications had a higher mean CSAT score (M = 4.25, SD = 0.87) than Enterprise (M = 3.9, SD = 1.17) and Enterprise Data (M = 4.11, SD = 0.74), the mean differences in CSAT scores between the three application categories were not significantly different ($F(2,44) = 0.76$, $p = .490$).

SUS

Consumer applications did have a higher mean SUS score (M = 73.6, SD = 22.81) than Enterprise (M = 62.5, SD = 21) and Enterprise Data (M = 63.39, SD = 17.10). However, the differences in SUS scores between the categories were not significant ($F(2, 46.9) = 2.44$, $p = .098$).

Bipolar

The mean bipolar score for Consumer applications (M = 78.29, SD = 19.33) was also higher than the average Enterprise application score (M = 66.25, SD = 18.49) and Enterprise Data score (M = 65.18, SD = 14.30). However, the ANOVA results indicated a significant difference in mean bipolar scores ($F(2, 46.6) = 5.13$, $p < .05$). Tukey post-hoc tests revealed that the mean Consumer application score was significantly higher than the mean score for Enterprise applications ($t(80) = 2.443$, $p < .05$) and for Enterprise Data applications ($t(80) = 2.94$, $p < .05$). However, the mean scores between Enterprise and Enterprise Data applications were not significantly different ($t(80) = -2.08$, $p = .976$).

Unipolar

Like the bipolar scores, the mean unipolar score for Consumer applications (M = 79.4, SD = 20.4) was higher than the average score for Enterprise applications (M = 65, SD = 15.8) and Enterprise Data (M = 66.5, SD = 16.4). The ANOVA *F*-test revealed a significant difference between the application areas ($F(2, 47.5) = 5.30$, $p < .05$). Post-hoc follow-ups indicated that the mean unipolar score for Consumer applications was significantly higher than the mean for Enterprise ($t(78) = 2.802$, $p < .05$) and Enterprise Data applications ($t(78) = 2.76$ $p < .05$). There was also no significant difference between the mean unipolar scores for Enterprise and Enterprise Data applications ($t(78) = 0.280$, $p = .958$).

Summary

The results show no significant differences in average scores between Enterprise Data and Enterprise applications, suggesting that participants do not rate the two application categories differently on the four metrics analyzed. Based on these results, we merged the Enterprise Data and Enterprise categories for further analysis in this paper.

We found significant differences in average bipolar and unipolar scores between the Enterprise Data and Enterprise versus the Consumer application scores. Participants who reported using a consumer application tended to evaluate the application more favorably than those who used an enterprise application. This supports the historical notion described above that enterprise applications do not perform as well for usability as consumer applications. From this finding, we recommend that participants who evaluated consumer applications should be kept distinct from those who evaluated enterprise applications (Enterprise or Enterprise Data). Following this logic, we surprisingly did not see any significant difference in SUS scores between application groups, which may indicate that SUS is robust across application areas or too noisy to find differences between consumer and enterprise applications.

Prediction of Satisfaction and SUS

To examine the convergent validity of the new 5-item scales, we examined how well the bipolar and unipolar scales were correlated with SUS and CSAT. If the new scales were correlated with CSAT to the same degree as SUS, and if the new scales are strongly correlated with SUS, we can tentatively conclude that the new measures have good convergent validity and may be suitable for use in place of SUS. Because we are interested in comparing the strength of the correlations between the new 5-item scales, we used Williams's test to determine whether the differences between correlation coefficients that share a dependent variable (SUS or CSAT) are significantly different from each other (Williams, 1959). The Williams test, sometimes referred to as Steiger's *Z*-test is more robust, especially at smaller sample sizes than an alternative test, like Hotelling's *t*-test (Steiger, 1980; Dunn & Clark, 1971). Because our earlier ANOVA analysis revealed significant differences between the evaluation of consumer applications and enterprise and enterprise data applications, the two Enterprise categories were combined, and Consumer applications were excluded from the correlation analysis. The exclusion of Consumer applications reduced the analytic sample size to 45.

Of the correlations with CSAT, the new 5-item set with the bipolar scale had the strongest correlation ($r = .633$, $p < .001$) followed by the unipolar scale ($r = .602$, $p < .001$). Although SUS was still correlated with CSAT, the correlation was weakest ($r = .578$, $p < .001$). However, the strength of the correlations between the SAT and responses to the bipolar, unipolar, and SUS scales were not significantly different (unipolar-bipolar and SAT: $t(43) = 0.53$, $p = .6$; SUS-unipolar and CSAT: ($t(43) = 0.26$, $p = .8$); SUS-bipolar and CSAT: ($t(43) = .54$, $p = .59$), which suggests that the SUS, unipolar, and bipolar scales have equally strong correlations with CSAT. While correlations with CSAT are important, as CSAT is a useful metric for measuring the success of a product, these correlations should not be interpreted as a test of which of the three scores is better at measuring usability. This is because CSAT is a measurement of the satisfaction of a product, not necessarily only usability. These findings instead indicate that the relationship between SUS and CSAT is maintained with the new 5-item set in bipolar and unipolar scales.

To understand how well the bipolar and unipolar scales correlate with usability, we next examined the correlations between the new scales with SUS. As previously mentioned, SUS may have interpretability issues for enterprise applications; however, we recognize it is still the standard measurement of usability, and therefore we used it as a comparison. The bipolar scale was strongly associated with SUS ($r = .613$, $p < .001$), but the unipolar scale variation had a slightly stronger correlation with SUS ($r = .691$, $p < .001$). Like CSAT, the correlation coefficients were not significantly different ($t(43) = 1.41$, $p = .16$), which indicates that both the bipolar and unipolar scales have equal convergent validity (that is, how well the new measures correlate with other usability measures).

Table 2. Enterprise and Enterprise Data Correlation Scores

	CSAT	SUS	Bipolar	Unipolar
CSAT	1			
SUS	0.578***	1		
Bipolar	0.633***	0.613***	1	
Unipolar	0.602**	0.691***	0.878***	1
*p < .05, **p < .01, *** p <.001				

New 5-Item Set, Unipolar Versus Bipolar

One concern raised by using the bipolar or unipolar response scales was that the response scales could be confusing for respondents. For example, the confidence and usefulness questions on the bipolar scale, especially with the negative anchor ("Not very useful" or "Very unconfident") may be less readily understood by participants compared to their unipolar response scale counterparts ("Not at all useful" and "Not at all confident"). Therefore, we created and reran the correlational analysis with a fourth usability measure, the bipolar scale with unipolar measures of confidence and usefulness. This new scale, named the "revised hybrid" scale, uses the unipolar variant of the confidence item ("How confident were you when using [this product]?") and the usefulness item ("How useful is [this product] to you?") with the other three bipolar items. The revised hybrid scale was reliable (Cronbach's α = 0.778, M = 70.8, SD = 18.76). Like with the bipolar and unipolar measures, the mean score of the revised hybrid measure was significantly higher ($F(2, 47) = 4.75$, $p < .05$) for Consumer applications (M = 78, SD = 19.3) than Data Enterprise (M = 65, SD = 16.6; $t(78) = 2.81$, $p < .05$) and Enterprise applications (M = 65.8, SD = 17; $t(78) = 2.442$ $p < .05$), but there was no statistical difference between Enterprise and Data Enterprise applications ($t(78)= 0.141$, $p = .98$). So, we again combined Enterprise and Data Enterprise applications and excluded Consumer applications for further analysis.

Although the hybrid CSAT correlation ($r = .591$, $p < .001$) was weaker than the bipolar and unipolar scales CSAT relationships but stronger than the correlation between SUS and CSAT, the correlations between the three scales and CSAT were not significantly different (bipolar-hybrid: $t(43) = 1.05$, $p = .30$; unipolar-hybrid: $t(43) = 0.24$, $p = .81$; SUS-hybrid: ($t(43) = 0.13$, $p = .89$). The revised hybrid scale correlation with SUS ($r = .674$, $p < .001$) was stronger than the bipolar scale relationship with SUS but slightly weaker than the correlation between unipolar and SUS; however, the differences in the size of the correlation coefficients were not statistically significant (bipolar-hybrid: $t(43) = 1.61$, $p = .12$; unipolar-hybrid: $t(43) = 0.42$, $p = .68$). Finally, the revised hybrid scale was strongly correlated with the bipolar ($r = .944$, $p < .001$) and unipolar scales ($r = .932$, $p < .001$). Collectively, the correlation results indicate that the revised hybrid scale has similar convergent validity as the unipolar and bipolar scales. Though we cannot suppose whether the hybrid scale is objectively easier for respondents to understand, the results indicate that the hybrid scale is just as predictive of CSAT and SUS as the unipolar and bipolar scales.

Table 3. Bipolar (Original) and Bipolar Revised (Bipolar with Unipolar Confidence and Usefulness items)

	CSAT	SUS	Revised Hybrid	Bipolar	Unipolar
CSAT	1				
SUS	0.578***	1			
Revised Hybrid	0.591***	0.674 ***	1		
Bipolar	0.633***	0.613***	0.944***	1	
Unipolar	0.602**	0.691***	0.932***	0.878***	1
*p < .05, **p < .01, *** p < .001					

Individual Calculated Scores

In this section, we examined the differences in individual responses to the SUS and the bipolar, unipolar, and hybrid scales. To do this we created three difference scores, in which we subtracted the SUS scores from each of the proposed scales. Recall that each of the scales was standardized to range from 0 to 100. Overall, the individual scores between bipolar, unipolar, and the revised hybrid scales and SUS did not differ much. The average difference between bipolar and SUS scores was only 2.6 points on the 100-point scale, whereas the difference was 3.67 points between the unipolar and SUS scores and 2.39 points between the revised hybrid and SUS scores. The small mean differences in scores suggest that the three scales are comparable to SUS scores.

However, the standard deviation of the difference scores was quite large, ranging from 13.8 for the difference score between unipolar and SUS to 15.4 for the differences between the bipolar and SUS scores. Though we expected participants to provide different answers to the scales due to the scales using different response options, future research should examine whether the response options are driving participants to respond differently.

Conclusion

Discussion

In the Results section, we identified that scores for Enterprise Data and Enterprise applications did not differ across the four metrics analyzed. However, Consumer application scores significantly differed from Enterprise Data and Enterprise application scores for the three new 5-item sets, but not for SUS. This was expected as past research indicates that consumer and enterprise applications differ in usability (Nielsen, 2006; Sauro, 2019). We also found that the unipolar scale was slightly more strongly associated with SUS than the bipolar and revised hybrid scales, but the bipolar scale had a stronger relationship with CSAT than the unipolar and revised hybrid scales did. However, considering the individual questions, the confidence and usefulness of unipolar questions may be more easily understood by participants than by their bipolar and revised hybrid counterparts. However, the revised hybrid scale, with its substitution of the unipolar questions for confidence and usefulness in the bipolar scale, may reduce the burden on participants while maintaining a strong relationship with SUS and the unmodified bipolar and unipolar scale. Because of this, we recommend using the revised hybrid scale for enterprise applications. Taking these findings into account, we propose practitioners use the ESUS (see Table 4) as a new usability standardized questionnaire that focuses on enterprise and enterprise data applications.

Table 4. Proposed ESUS Questionnaire

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

Recommendations

In this paper we described a need for researchers in industry to have a compact standardized usability questionnaire that addresses the unique needs of enterprise applications. We then presented our findings on a study that we conducted to explore a potential solution. Our proposed questionnaire reduces the items in half, removes problematic questions that do not fit with the reality of current enterprise applications, and aligns with existing business metric norms with a 5-point ordinal scale. Both the proposed 5-item bipolar and unipolar sets showed promising results and strong correlations with both the original SUS scores and satisfaction as well as differences between consumer and enterprise applications. With further review, we created a revised hybrid scale, a combination of bipolar and unipolar items, that performed as well as the bipolar and unipolar scales, which we offer as a potential new standard usability questionnaire for enterprise applications.

When comparing the SUS and ESUS individual scores, we found that the average bipolar, unipolar, and revised hybrid scores were very close to the average SUS score for enterprise data and enterprise applications. However, there are large variations in individual scores across the four metrics. More research is needed to explore these relationships.

Overall, we recognize that the introduction of a new standard usability questionnaire requires additional validation. In considering the requirements set forth by Sauro and Lewis (2016) and Sauro (2012) of reliability, validity, and sensitivity, we recommend additional research for reliability and sensitivity as an extension of the work presented in this paper. We look forward to continuing to test and explore the effectiveness of our enterprise-specific version of SUS (ESUS) as well as sharing the questionnaire with the HCI and UXR communities for external review.

Tips for Usability Practitioners

- If you are evaluating the usability of an enterprise application, consider using ESUS versus SUS.
- If your user has a strong technical background or works in a technical role, consider using ESUS versus SUS.
- For a shorter usability questionnaire, consider using ESUS versus SUS.

References

Bangor, A., Joseph, K., Sweeney-Dillon, M., Stettler, G., & Pratt, J. (2013, September). Using the SUS to help demonstrate usability's value to business goals. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *Los Angeles*, *USA*, *57*(1), 202-205. SAGE Publications.

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.

Bangor, A., Kortum, P. T., & Miller, J. T. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P. W. Jordan, B. Thomas, I. L. McClelland, & B. Weerdmeester (Eds.), *Usability evaluation in industry*. CRC Press.

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, *8*(2), 29–40.

Coelho, P. S., & Esteves, S. P. (2007). The choice between a five-point and a ten-point scale in the framework of customer satisfaction measurement. *International Journal of Market Research*, *49*(3), 313–339.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340.

Dixon, M., Freeman, K., & Toman, N. (2010). Stop trying to delight your customers. *Harvard Business Review*, *88*(7/8), 116–122.

Dunn, O. J., & Clark, V. (1971). Comparison of tests of the equality of dependent correlation coefficients. *Journal of the American Statistical Association*, *66*(336), 904–908.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327.

Hillman, S., Jain, S., Jienjitlert, V., & Bach, P. (2022, April). The BLUE framework: Designing user-centered in-product feedback for large scale applications. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, *New Orleans*, *USA*, 1–8.

Kirakowski, J., & Cierlik, B. (1998, October). Measuring the usability of web sites. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *Los Angeles*, *USA*, *42*(4), 424–428. SAGE Publications.

Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology, 24*(3), 210–212.

Lewis, J. R. (1992, October). Psychometric evaluation of the Post-Study System Usability Questionnaire: The PSSUQ. *Proceedings of the Human Factors Society Annual Meeting*, *Los Angeles*, *USA*, *36*(16), 1259–1260. Sage Publications.

Lewis, J. R., & Sauro, J. (2017). Revisiting the factor structure of the System Usability Scale*. Journal of Usability Studies, 12*(4), 183–192.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015a). Investigating the correspondence between UMUX-LITE and SUS scores. In *Design, User Experience, and Usability: Design Discourse*: *4th International Conference*, *DUXU 2015*, *Held as Part of HCI International 2015*, *Los Angeles*, *CA*, *USA*, *August 2–7*, *2015*, *Proceedings*, *Part I*, 204–211. Springer International Publishing.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015b). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31*(8), 496–505.

Loranger, H., Nodder, C., & Nielsen, J. (2007). *B2B website usability for converting users into leads and customers*. Nielsen Norman Group.

Nielsen, J. (2006, May 31). *B2B usability.* Nielsen Norman Group Articles. https://www.nngroup.com/articles/b2b-usability/

Reichheld, F. F. (2004). The one number you need to grow. *Harvard Business Review*, *82*(6), 133–133.

Revelle, W. (2015). *psych: Procedures for Personality and Psychological Research* [Software Version 1.5.8]. Northwestern University, Evanston, Illinois, USA. http://CRAN.R-project.org/package=psych

Sauro, J. (2009, October 9). *Do users fail a task and still rate it as easy?* MeasuringU. https://measuringu.com/failed-sat

Sauro, J. (2010, January 7). *Does better usability increase customer loyalty?* MeasuringU. https://measuringu.com/usability-loyalty/

Sauro, J. (2011, February 3). *Measuring usability with the System Usability Scale (SUS).* MeasuringU. https://measuringu.com/sus/

Sauro, J. (2012, March 27). *8 advantages of standardized usability questionnaires.* MeasuringU. https://measuringu.com/standardized-usability/

Sauro, J. (2015). SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of Usability Studies*, *10*(2), 68–86.

Sauro, J. (2017, October 11). *Measuring usability: From the SUS to the UMUX-Lite.* MeasuringU. https://measuringu.com/umux-lite/

Sauro, J. (2019, September 29). *How common are usability problems?* MeasuringU. https://measuringu.com/problem-frequency/

Sauro, J., & Dumas, J. S. (2009, April). Comparison of three one-question, post-task usability questionnaires. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1599–1608.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. *Proceedings of the SIGCHI conference on human factors in computing systems*, 1609–1618.

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? *Proceedings of the SIGCHI conference on human factors in computing systems*, 2215–2224.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Tullis, T. S., & Stetson, J. N. (2004, June 7-11). A comparison of questionnaires for assessing website usability [Presentation]. *Proceedings of UPA 2004 Conference, Minneapolis, USA*.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(2), 396–399.

Appendix

Item Sets Used in the Survey

Table 5. CSAT Satisfaction Question, Used in the Study Survey

CSAT - Satisfaction	**1**	**2**	**3**	**4**	**5**
How satisfied or unsatisfied are you with [this product]?	Very unsatisfied	Unsatisfied	Neutral	Satisfied	Very satisfied

Table 6. The New 5-Item Set, Unipolar, Used in the Study Survey

New 5-Item set - Unipolar	**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 was [this product] to use for you?	Not at all easy	Slightly easy	Somewhat easy	Mostly 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 in [this product]?	Does not work together at all	Works slightly well together	Works somewhat well together	Works mostly well together	Works very well together
How hard was it to get started with [this product]?	Not at all hard	Slightly hard	Somewhat hard	Mostly hard	Very hard

Table 7. The New 5-Item Set, Bipolar, Used in the Study Survey

New 5-Item set - Bipolar	**1**	**2**	**3**	**4**	**5**
How useful is [this product] to you?	Very not useful	Not useful	Neutral	Useful	Very useful
How easy was [this product] to use for you?	Very hard	Hard	Neutral	Easy	Very easy
How confident were you when using [this product]?	Very unconfident	Unconfident	Neutral	Confident	Very confident

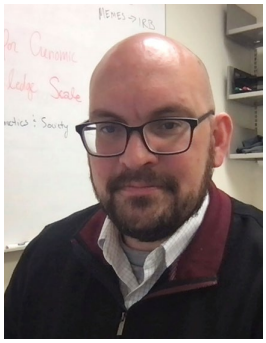
How well do the functions work together in [this product]?	Does not work together at all together	Does not work well together	Neutral	Works well together	Works very well together
How hard was it to get started with [this product]?	Very hard	Hard	Neutral	Easy	Very easy

Table 8. Original 10-Item SUS Used in the Study Survey

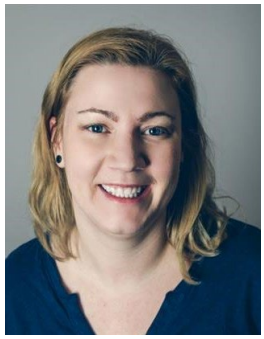
SUS Positive 10-items (anchored)	1	2	3	4	5
I think that I would like to use this product frequently.	Strongly disagree	*	*	*	Strongly agree
I found this product unnecessarily complex.	Strongly disagree	*	*	*	Strongly agree
I thought this product was easy to use.	Strongly disagree	*	*	*	Strongly agree
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

Data Analysis: Data cleaning and analysis was conducted using R. The ANOVAs were conducted using the R base package. Williams test used the psych package in R (Revelle, 2015).

About the Authors



Stephen Schneider, PhD
Dr. Schneider is a researcher supporting Microsoft Purview. Though he loves expanding his methodological tool kit, he finds comfort in using quantitative methods and experimental designs. Stephen earned a PhD in Biopolitics and Political Psychology at the University of Nebraska.



Serena Hillman, PhD
Dr. Hillman is a research manager in Microsoft Azure Data. Her team helps drive the future of data and analytics. She has contributed to the HCI discipline with 35+ publications, presenting at conferences such as Grace Hopper, ACM CHI, ACM CSCW, and ACM MobileHCI. She earned a PhD in HCI at Simon Fraser University.

Paula Bach, PhD
Dr. Bach is a Principal Research Director in Microsoft Azure Data. Her career has spanned over a decade producing and evangelizing customer insights to help teams make more human-centered product decisions. She earned a PhD in HCI at the College of Information Sciences and Technology at Penn State University.

Guoping Ma, PhD
Dr. Ma is a Principal Research Manager in Microsoft Azure Data. She has spent her career driving human insights into both consumer and enterprise products. Her team focuses on Power BI and the greater business intelligence space. She earned a PhD in IST at Indiana University Bloomington.