International Journal of Human-Computer Interaction, vol. 42, no. 20 (2026), pp. 16892-16905
Author version. Published version available at: https://doi.org/10.1080/10447318.2026.2625260

System Usability Scale: A Meta-Analysis of How SUS Relates to Workload, Task Time, and Error Rate

Morten Hertzum, Department of People and Technology, Roskilde University, Roskilde, Denmark

Address: Roskilde University, Universitetsvej 1, 4000 Roskilde, Denmark. Email: mhz@ruc.dk.

Abstract

The System Usability Scale (SUS) is a widely used instrument for measuring perceived usability. This study meta-analyzes 105 studies that reported SUS, workload, task time, and error rate for pairs of systems. Workload was restricted to self-report metrics. The meta-analysis resulted in three main findings. First, SUS and workload were strongly correlated. Second, SUS and performance, as measured by task time and error rate, were related but somewhat independent. Third, task time and error rate correlated with SUS to similar extents. In only 10% of the studies, the system that received the higher SUS score imposed higher workload. In contrast, the system that received the higher SUS score resulted in poorer performance in 24% (task time) and 23% (error rate) of the studies. To conclude, SUS complements performance measures and it largely does so by also being sensitive to perceived workload.

Keywords: Perceived usability, Workload, Performance, Usability evaluation, User experience

1 Introduction

The System Usability Scale (SUS; Brooke, 1996) is a simple and widely used metric for users to self-report how they experience the usability of a system. By summarizing system usability in a single number, the SUS score is a global assessment, which lends itself to benchmarking. It is often administered in usability tests as a quantitative supplement to their qualitative reporting of usability problems (Bangor et al., 2008; Brooke, 2013; Lewis, 2018). To improve our understanding of SUS and, thereby, make it easier to interpret and assign importance to SUS scores, this study seeks to answer the following research question: *To what extent do users give higher SUS scores to systems with which they achieve lower workload, shorter task time, and lower error rate?*

Intuitively, system usability, and thereby SUS, is higher when the workload required to perform a task with a system is low, when the time it takes to complete tasks with the system is short, and when the proportion of incorrectly solved tasks is low. Some studies find this expected association (e.g., Lamontagne et al., 2021). However, others find that SUS is dissociated from workload (e.g., Az-zahra et al., 2019), task time (e.g., Werrlich et al., 2018), or error rate (e.g., Kurosawa et al., 2018). If such dissociations are rare, they can reasonably be explained away as outliers. If they are common, they raise questions about what SUS effectively measures and how it informs usability tests. Clarifying how SUS relates to workload, task time, and error rate is needed to help understand and apply this important usability metric, but it will also – more generally – improve our understanding of whether users' experience of the usability of a system is primarily associated with workload, task time, or error rate.

To investigate how SUS scores relate to workload, task time, and error rate, the present study meta-analyzes published studies that report SUS, perceived workload, task time, and error rate for pairs of

systems. For each pair, the degree to which one of the systems received a higher SUS score from the users is compared with the degree to which this system resulted in lower workload, shorter task time, and lower error rate. This pairwise-comparison approach ensures that the differences in SUS, workload, task time, and error rate contributed by each comparison are from users solving the same tasks in the same use context. Thereby, the present meta-analysis spans diverse users, tasks, and contexts while controlling for their effect on the results.

2 Background

As a self-report instrument, SUS measures perceived usability. Other instruments for measuring perceived usability include the Post-Study System Usability Questionnaire (PSSUQ; Lewis, 1995), the Questionnaire for User Interface Satisfaction (QUIS; Chin et al., 1988), and the Software Usability Measurement Inventory (SUMI; Kirakowski & Corbett, 1993). While these other instruments model usability as a multidimensional construct, SUS assumes a unidimensional structure (Brooke, 1996). Several studies have tested this assumption (e.g., Borsci et al., 2009; Lewis & Sauro, 2009, 2017). Lewis (2018) concluded that SUS appears to have a bidimensional structure (with the positively worded items constituting one factor and the negatively worded items another) but also that this structure is artificial and of little practical interest. Hence, he recommended treating SUS as a unidimensional measure, as originally intended.

The SUS instrument consists of ten items, which alternate between positively and negatively worded statements, see Table 1. Each item is rated on a five-point scale from “Strongly disagree” (1) to “Strongly agree” (5). The SUS score is calculated by subtracting 1 from the odd-item ratings, subtracting the even-item ratings from 5, and multiplying the sum of these adjusted ratings by 2.5, thereby resulting in a score between 0 and 100 (Brooke, 1996). To help interpret SUS scores, Bangor et al. (2008) report that the median SUS score in a database with evaluations of 206 systems was 71. That is, a system had to score 71 or higher to be in the top-50%. For a set of 959 SUS scores, Bangor et al. (2009) also had the users rate the system on a scale where all response options were adjectives. The mean SUS scores corresponding to these adjectives were 13 (“Worst imaginable”), 20 (”Awful”), 36 (“Poor”), 51 (“OK”), 71 (“Good”), 86 (“Excellent”), and 91 (“Best imaginable”).

Table 1. The SUS instrument, originating from Brooke (1996)

#	Item
1	I think that I would like to use this system frequently.
2	I found the system unnecessarily complex.
3	I thought the system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.
5	I found the various functions in this system were well integrated.
6	I thought there was too much inconsistency in this system.
7	I would imagine that most people would learn to use this system very quickly.
8	I found the system very cumbersome to use.
9	I felt very confident using the system.
10	I needed to learn a lot before I could get going with this system.

Several studies have found moderate to strong correlations between SUS and other measures of perceived usability (e.g., Bangor et al., 2009; Borsci et al., 2015; Lewis, 2018). For example, Lewis (2018) reported correlations ranging from .50 to .96. These studies indicate rough agreement about the construct of perceived usability across different instruments for measuring it. In contrast, the correlation between SUS and task success has been reported to range from .22 to .50 (Lewis, 2018), indicating that SUS is a measure of user perceptions and only partially determined by performance.

Drew et al. (2018) corroborated this indication. They found a non-significant correlation of .29 between SUS and task success but a significant correlation of .51 between SUS and perceived success (i.e., the users' rating of "whether or not they felt that they completed each task successfully"). Similarly, Kortum and Peres (2014) reported modest, but reliable, correlations between SUS and task success in both lab and field settings. However, none of these studies compared SUS with performance measures other than task success, that is, none of them compared SUS with task time. In addition, they did not compare SUS with workload.

Workload concerns the balance, or imbalance, between the user's capabilities, the task demands, and the system (Hart & Staveland, 1988). Workload has been found to be somewhat independent from SUS scores (Longo, 2018). As a result of the independence, Longo (2018) found that a model combining SUS and workload predicted task success better than did models based on either SUS or workload; the study did not consider task time. In a study not including SUS, Hertzum (2022) found that workload is also somewhat independent from task time and error rate. SUS, workload, task time, and error rate measure different dimensions of the same overall concept, usability. The ISO 9241-210 (2019) standard defines usability as consisting of the dimensions of effectiveness, efficiency, and satisfaction. SUS is about satisfaction, both workload and task time are about efficiency, and error rate is about effectiveness. In addition, SUS differs from some other satisfaction metrics by typically being measured after users have performed all tasks with a system. Sauro and Lewis (2009) found that such post-test measurements correlate less strongly with performance than satisfaction metrics taken after each task.

Previous studies have mainly investigated how SUS relates to other measures of perceived usability and to task success/error rate. They have not reviewed how SUS relates to all three of workload, task time, and error rate. Workload in particular has been absent in previous SUS studies. Previous studies have also not investigated how often SUS is dissociated from workload or performance and, thereby, contradicts them. The present meta-analysis aims to fill this gap. The included studies were not conducted to investigate SUS but, rather, to compare the usability of a pair of systems. That is, these studies have not previously informed discussions about SUS.

3 Method

This study followed standard advice on how to conduct systematic reviews and meta-analysis (Littell et al., 2008). Note that preference ratings were initially included as an alternative to SUS scores and, thus, feature among the inclusion criteria. However, the study was subsequently restricted to SUS scores to investigate this widely used metric in detail, that is, the preference ratings were excluded, see Figure 1.

3.1 Inclusion Criteria

For a paper to be included in the meta-analysis, it had to fulfil six inclusion criteria. First, papers had to report SUS scores or preference ratings. Second, papers had to report workload. In this study, workload was operationalized as either Task Load Index (TLX; Hart & Staveland, 1988) or Subjective Workload Assessment Technique (SWAT; Reid & Nygren, 1988). Third, papers had to report task time, that is, how long it took users to complete tasks. Fourth, papers had to report error rate. Error rate presupposed that a correct solution existed for each task and, then, involved identifying the tasks that were not solved correctly, the errors. Fifth, papers had to report from studies that empirically compared systems by having users perform the same tasks with the systems. Sixth, papers had to be published in English as peer-reviewed journal articles, conference papers, or book chapters.

3.2 Selection Process

The papers included in this study were selected through a multistep process, see Figure 1. First, ACM Digital Library, Google Scholar, IEEE Xplore, and Scopus were searched for candidate papers in January 2024. The search query covered the first four inclusion criteria and looked, save small syntactic

variations across the databases, like this: ("system usability scale" OR sus OR preference OR satisfaction) AND ("task load index" OR tlx OR "subjective workload assessment technique" OR swat) AND time AND ("error rate" OR "task completion" OR "task success"). Second, the 8183 papers that matched the query were scanned for duplicates and then their title was checked against the inclusion criteria. If they passed this check, then their abstract was checked against the inclusion criteria. Third, the full text of the 1181 papers that passed the title and abstract checks was looked up online. Five of them could not be obtained in this way and were instead requested from their authors, who in three cases kindly supplied a full-text copy. Fourth, the 1179 papers obtained in full text were checked against the inclusion criteria. At this stage, the studies reporting preference ratings rather than SUS scores were excluded. After all the checks, 105 papers remained. Figure 1 gives the reasons for the full-text exclusions.

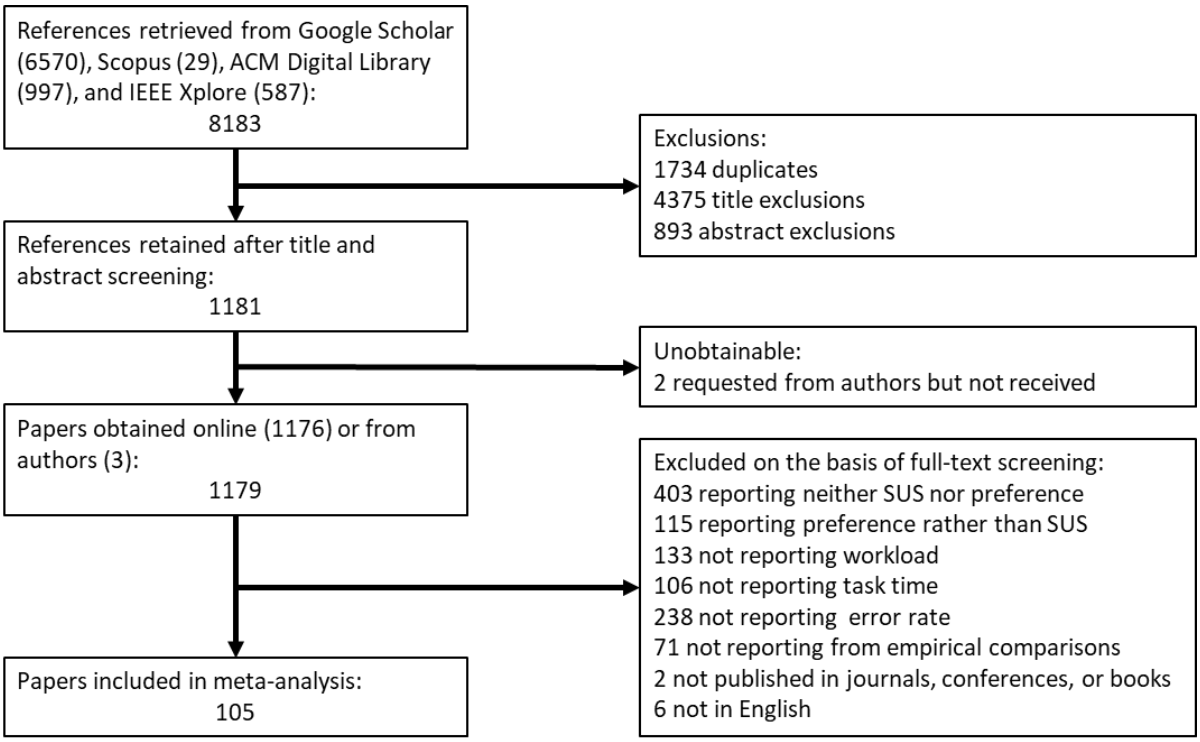


Figure 1. Paper-selection process

3.3 Data Analysis

The analysis of the 105 included papers involved seven steps. Like the paper selection, the data analysis was performed exclusively by the author.

First, each paper was restricted to contributing one pairwise comparison to the meta-analysis. This restriction served to avoid the bias that would otherwise result from including pairwise comparisons that were interdependent because they were from the same paper (Littell et al., 2008). The comparison selected from each paper was the one with the largest difference in SUS between the two compared systems because this situation was considered more important to understand than how workload, task time, and error rate related to SUS for small SUS differences that were, possibly, down to chance. Minor variations in the wording of the SUS items were accepted. As recommended by Finstad (2006), several studies replaced “cumbersome” with “awkward” in Item 8 to make that item

more understandable. Other studies tailored SUS by replacing “system” with the name of the tested system.

Second, data about SUS, workload, task time, and error rate were extracted from the papers for each of the 105 pairwise comparisons. In all comparisons, workload was measured with TLX, not SWAT. The data were in most cases directly available, but sometimes the extraction process involved obtaining values from graphs or computing values from those available. For example, several studies did not report overall workload directly but merely the six TLX subscales (mental demand, physical demand, temporal demand, effort, performance, and frustration), thereby making it possible to compute the overall workload by averaging the subscales. Relatedly, task times (expressed as minutes per word) were in some cases computed from input rates reported in words per minute, and error rates were computed from success rates. The studies predominantly employed raw TLX, that is, they calculated workload as the mean of the TLX subscales. However, a few studies (e.g., Költringer & Grechenig, 2004) also employed the weighting procedure originally proposed by Hart and Staveland (1988). Both these variants of TLX were allowed in this meta-analysis because they have been found to produce comparable results (Byers et al., 1989; Nygren, 1991).

Third, the SUS and TLX data were scores on rating scales. The fixed endpoints of these scales made it harder for a mean SUS or TLX score to come close to the ends of the scale. To compensate for this property of rating scales, the data were rescaled to values in the interval from zero to one and arcsine transformed. Preston and Colman’s (2000) formula was used for the rescaling: (value − lower endpoint) / (max endpoint − lower endpoint).

Fourth, the initial analysis of the data consisted of simply counting in how many studies SUS scores were associated and dissociated with workload, task time, and error rate. Association meant that the difference in SUS score between the two compared systems was in favor of the same system as the difference in workload, task time, or error rate. Dissociation was the reverse; it meant that the system with the higher SUS score was also the system with the higher workload, longer task time, or higher error rate. While this initial analysis accounted for the direction of the difference between the two compared systems, it disregarded the magnitude of this difference.

Fifth, the magnitude of the difference between the two compared systems was quantified by calculating an effect size. Because many papers reported means but not standard deviations, the most widely used effect-size measures were not applicable. Like in Nielsen and Levy (1994), the chosen measure was the logarithm of the ratio of the two means: log(mean 1 / mean 2). That is, the effect size for SUS would be log(80/60) = 0.125 if the mean SUS scores for system 1 and 2 were 80 and 60, respectively. Conversely, the effect size would be log(60/80) = −0.125 if the mean SUS scores for system 1 and 2 were reversed. The sign differs but the absolute effect size is the same irrespective of whether system 1 or system 2 is the more usable. This property of the logarithm function motivated the log-transformation of the ratios. It meant that the outcome of the analysis was unaffected by which system was labelled as the first and the second. The same effect-size measure was used for workload, task time, and error rate. For example, the effect size for workload would be log(35/35) = 0 if the mean workload was 35 for both system 1 and 2.

Sixth, the effect-size measure required non-zero means for both systems. In seven studies, this resulted in an undefined effect size because the mean error rate for one of the systems equaled zero. These error rates were replaced with imputed values by setting the mean to 1/sqrt(users) percent, where sqrt(users) was the square root of the number of users who participated in the study. This formula, also used by Hertzum (2025), ensured an imputed mean below one percent and also incorporated that a mean error rate of zero was harder to attain in a study with more users.

Seventh, information about the involved users, tasks, systems, and domains was extracted from the studies. This information was categorized to characterize the set of included studies, see Table 2.

4 Results

Table 2 gives the profile of the included studies. The users spanned both novices and people with experience, the tasks were with a few exceptions utilitarian, the domains were in most cases not safety critical, and the systems varied across multiple categories. In total, 2570 users participated in the studies.

Table 2. Profile of the included studies, *N* = 105 studies

Category	Count	Percentage
Users		
Novices	54	51
Experienced	38	36
Other	13	12
Tasks		
Utilitarian	103	98
Hedonic	2	2
Domains		
Safety critical	18	17
Not safety critical	87	83
Systems		
Graphical user interfaces	33	31
Augmented reality	25	24
Virtual worlds	27	26
Teleoperation	10	10
Non-visual interfaces	10	10

4.1 Association and Dissociation with SUS

In 60 (57%) of the 105 studies, the users gave a higher SUS score to the system with which they achieved lower workload, shorter task time, and lower error rate, see Table 3. Thus, the data provided support for the intuition that users can in most cases be expected to associate usability with performing well and avoiding elevated demands. SUS was, however, dissociated from one or several of workload, task time, and error rate in the remaining 45 studies. Most notably, the users in three studies (Smith et al., 2020; Xu et al., 2020; Zhao et al., 2020) gave higher SUS scores to the system with which they experienced higher workload, longer task time, as well as higher error rate (Table 3). For example, Smith et al. (2020) compared augmented reality (AR) instructions with paper instructions for an assembly task and found that users experienced lower workload, shorter task time, and lower error rate with paper instructions but gave higher SUS scores to the AR instructions. To explain this finding, Smith et al. (2020) noted that the users "cited the AR system as being more enjoyable to use." The two other studies with threefold dissociations did not comment on them. Instead, Zhao et al. (2020) noted that the SUS scores were high for both systems, and Xu et al. (2020) made their design recommendations on the basis of the results for workload, task time, and error rate.

Table 3. Frequency of the eight possible combinations of association/dissociation with SUS, *N* = 105 studies

Workload	Task time	Error rate	Count	Percentage
Association	Association	Association	60	57
Association	Association	***Dissociation***	14	13
Association	***Dissociation***	Association	15	14
Association	***Dissociation***	***Dissociation***	5	5
Dissociation	Association	Association	4	4
Dissociation	Association	***Dissociation***	2	2
Dissociation	***Dissociation***	Association	2	2
Dissociation	***Dissociation***	***Dissociation***	3	3

Aggregating the information in Table 3 shows association between SUS and workload in 94 studies, between SUS and task time in 80 studies, and between SUS and error rate in 81 studies. The remaining 11 (workload), 25 (task time), and 24 (error rate) studies showed dissociation. That is, SUS scores agreed with workload more often than with task time and error rate. However, the compared systems differed in SUS, workload, task time, and error rate to extents that varied from small and not statistically significant in some studies to large and statistically significant in others. To account for the magnitude of the difference, the following analyses use the logarithm of the ratio between the two systems to measure the effect size.

4.2 SUS versus Workload

Figure 2 plots the logarithm of the ratio of the SUS scores against that for workload. The upper-left and lower-right quadrants of the figure contain the studies where the users gave a higher SUS score to the system that imposed the lower workload. The two other quadrants (lower-left and upper-right) contain the studies that displayed a SUS-workload dissociation. For the studies that displayed a dissociation, the difference in SUS, workload, or both tended to be small. In general, the trend in the data was that a larger difference in SUS corresponded with a larger difference in workload, $r(105) = -.877$, $p < .001$. This strong correlation meant that the variation in workload explained 77% of the variation in SUS. For example, Triantafyllidis et al. (2020) compared a virtual reality head-mounted display (extended with audio and haptic feedback) with a computer monitor (extended with audio but not haptic feedback) for teleoperating a robotic hand. The virtual reality display received a threefold higher SUS score (92.50 vs. 30.00); the difference in workload was similarly large and also in favor of the virtual reality display (26.66 vs. 70.83).

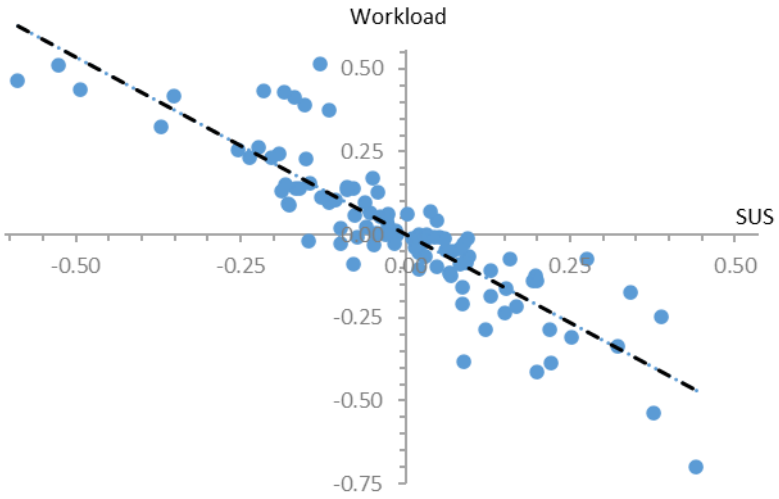


Figure 2. Scatterplot of SUS versus workload, N = 105 studies. Each data point shows the extent to which the users in a study gave a higher SUS score to one of the two compared systems (i.e., the logarithm of the ratio of the SUS scores) plotted against the extent to which workload was experienced as lower with this system (i.e., the logarithm of the ratio of the workload ratings). The dashed regression line gives the trend in the data, $y = -1.065x + 0.001$ (R^2 = 77%).

4.3 SUS versus Task Time

Figure 3 plots SUS against task time. Like for workload, the data displayed the trend that increasingly different SUS scores corresponded with increasingly different task times, $r(105) = -.671$, $p < .001$. This correlation was not as strong as the one between SUS and workload, but the variation in task time still explained 45% of the variation in SUS. Part of the reason for the somewhat weaker correlation was the larger number of dissociations between SUS and task time. In addition, several of these dissociations were in studies with a large difference in task time (see the lower-left and upper-right quadrants in Figure 3). For example, Werrlich et al. (2018) compared an AR system and paper-based instructions for learning manual assembly tasks. While the participants who used the AR system spent twice as long completing the tasks (2747 s vs. 1367 s), they still gave it a higher SUS score than the users of the paper-based instructions gave them (91.83 vs. 80.00). The SUS scores were possibly explained by a significantly lower number of assembly errors with the AR system (1.53 vs. 4.06).

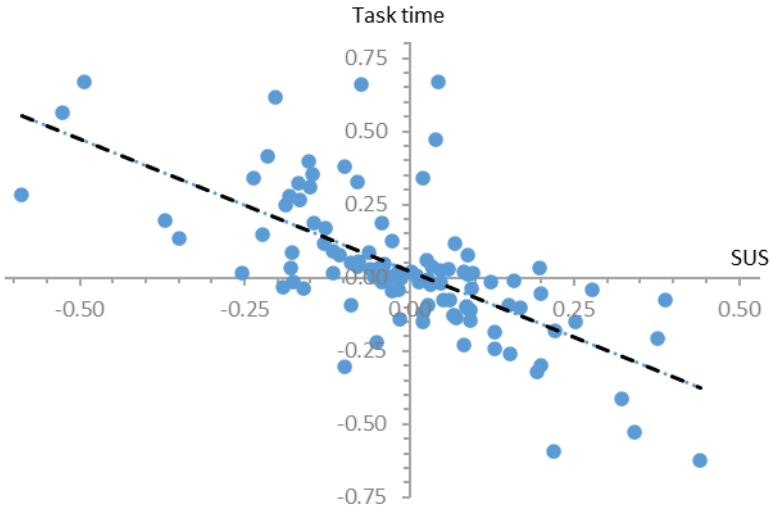


Figure 3. Scatterplot of SUS versus task time, N = 105 studies. Each data point shows the extent to which the users in a study gave a higher SUS score to one of the two compared systems (i.e., the logarithm of the ratio of the SUS scores) plotted against the extent to which task time was shorter with this system (i.e., the logarithm of the ratio of the task times). The dashed regression line gives the trend in the data, $y = -0.903x + 0.022$ (R^2 = 45%).

4.4 SUS versus Error Rate

Figure 4 plots SUS against error rate. The extensive difference in error rate in some of the studies meant that the values for error rate were more dispersed (range: −2.33 to 1.34) than for workload (range: −0.70 to 0.51) and task time (range: −0.63 to 0.67). For example, Schramm et al. (2023) compared eye gazes and hand pointing as techniques for passengers in moving vehicles to select objects outside the vehicle for further interaction. These objects could, for example, be traffic conditions or points of interest. The error rate was 22 times higher with the eye-gaze technique (2.86% vs. 0.13%), yet it received a higher SUS score (85.6 vs. 58.0). Possibly, the users did not assign much importance to the error rates because they were low for both techniques. In spite of some large dissociations between SUS and error rate, the trend in the data resembled, but was weaker than, those for workload and task time, $r(105) = -.584$, $p < .001$. The weaker correlation meant that the variation in error rate merely explained 34% of the variation in SUS.

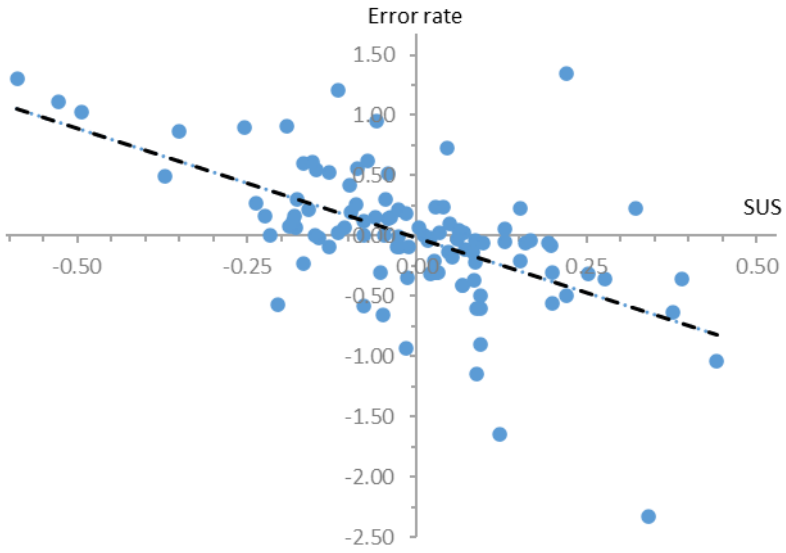


Figure 4. Scatterplot of SUS versus error rate, N = 105 studies. Each data point shows the extent to which the users in a study gave a higher SUS score to one of the two compared systems (i.e., the logarithm of the ratio of the SUS scores) plotted against the extent to which error rate was lower with this system (i.e., the logarithm of the ratio of the error rates). The dashed regression line gives the trend in the data, $y = -1.820x - 0.021$ (R^2 = 34%).

4.5 Partial Correlations

Table 4 (first row) summarizes the correlations among the values depicted in Figures 2-4, that is, the correlations between SUS and each of workload, task time, and error rate. However, the table also shows considerable co-variation among workload, task time, and error rate. For example, the correlation between workload and error rate was .558. This covariation prompted the use of partial correlation analysis to investigate the relationship between SUS and each of workload, task time, and error rate while controlling for the effect of the two others. The first column of the table gives these partial correlations. The correlation between SUS and workload remained strong (r = −.741) when the effect of task time and error rate was partialed away. That is, the meta-analysis shows a genuine relationship between SUS and workload. In contrast, the partial correlations between SUS and task time and between SUS and error rate were weak. That is, when the effect of workload and error rate (respectively, workload and task time) was partialed away, the remaining relationship was still significant but no longer strong. The drop from the correlation coefficients to the partial correlation coefficients revealed that the correlation between SUS and task time (respectively, SUS and error rate) was mainly driven by the co-variation with workload.

Table 4. Correlations (above the diagonal) and partial correlations (below the diagonal) between SUS, workload, task time, and error rate, N = 105 studies

	SUS	Workload	Task time	Error rate
SUS	-	−.877***	−.671***	−.584***
Workload	−.741***	-	.639***	.558***
Task time	−.279**	.135	-	.435***
Error rate	−.213*	.107	.057	-

Note: * p < .05, ** p < .01, *** p < .001 (Pearson correlation/partial Pearson correlation)

4.6 Subset Analysis

Finally, the analysis was redone for subsets of the studies to examine the robustness of the results. The first subset analysis (Table 5) examined whether the results differed across the categories in the profile of the included studies, except that categories with fewer than 15 studies were excluded from the subset analysis. Overall, the results for the subsets resembled those for the full set of 105 studies (Table 4, first row), thereby corroborating the results in Sections 4.1-4.5. However, three differences should be noted: (1) For experienced users, SUS correlated more strongly with task time, suggesting that these users to a larger extent than novice users equated usability with efficiency. (2) For safety-critical domains, SUS correlated more strongly with task time and error rate, suggesting that performance mattered more to perceived usability in these domains than in non-safety-critical domains. (3) For augmented reality, SUS did not correlate significantly with task time and error rate, suggesting that usability was perceived differently than for other systems.

Table 5. Correlations between SUS and each of workload, task time, and error rate for categories of users, domains, and systems

Subset	*N*	Workload	Task time	Error rate
Users				
Novices	54	−.847***	−.592***	−.518***
Experienced	38	−.905***	−.848***	−.612***
Domains				
Safety critical	18	−.889***	−.767***	−.757***
Not safety critical	87	−.884***	−.645***	−.531***
Systems				
Graphical user interfaces	33	−.877***	−.805***	−.650***
Augmented reality	25	−.798***	−.328	−.189
Virtual worlds	27	−.875***	−.720***	−.732***

Note: *** $p < .001$ (Pearson correlation)

The second subset analysis (Table 6) examined whether the results differed for studies that measured SUS, TLX, and error rate in slightly different ways. Four subsets were examined: the studies measuring SUS on the standard 0-100 scale, the studies measuring TLX on the standard 0-100 scale, the studies measuring error rate by the proportion (rather than number) of errors, and the studies that were in all three of these subsets. These subsets were selected because the studies employing nonstandard scales could be suspected of confounding the results. Table 6 shows almost identical correlations for the subsets and the full set of 105 studies, thereby corroborating the robustness of the results.

Table 6. Correlations between SUS and each of workload, task time, and error rate for subsets that employ the standard measuring scales

Subset	*N*	Workload	Task time	Error rate
All	105	−.877***	−.671***	−.584***
0-100 SUS scale	99	−.877***	−.671***	−.586***
0-100 TLX scale	75	−.876***	−.710***	−.609***
Error rate is proportion of errors	52	−.882***	−.671***	−.558***
The three above subsets combined	32	−.882***	−.701***	−.635***

Note: *** $p < .001$ (Pearson correlation)

5 Discussion

This study has meta-analyzed how SUS relates to self-reported workload, task time, and error rate in 105 studies that measure all four constructs for pairs of systems.

5.1 Summary of Findings

The meta-analysis resulted in three main findings. First, *SUS and workload, as measured by TLX, are strongly correlated*. The variation in workload explains 77% of the variation in SUS. Consistent with the strong correlation, SUS and workload are dissociated in only 10% of the 105 studies. In these cases, the system with the higher SUS score imposed the higher workload, but the dissociations mostly involved systems that differed little in SUS, workload, or both.

Second, *SUS and performance, as measured by task time and error rate, are related but somewhat independent*. That is, SUS adds information beyond performance measures. The variation in task time and error rate explains 45% and 34%, respectively, of the variation in SUS. However, much of this explanatory power comes from the co-variation of task time and error rate with workload. Performance and SUS are still significantly correlated when this co-variation is partialed away, but the partial correlation is merely weak.

Third, *task time and error rate correlate with SUS to similar extents*. The variation in task time explains 11 percentage points more of the variation in SUS than the variation in error rate does. Other than that, users appear to assign roughly equal importance to speed and accuracy when they assess the usability of a system. For example, SUS is dissociated from task time and error rate in 24% and 23% of the studies, respectively.

5.2 Implications

SUS complements performance measures such as task time and error rate by adding information about how users perceive the usability of a system. The substantial variation in SUS not explained by the variation in task time or error rate shows that SUS is sensitive to factors beyond performance and, thereby, adds value to usability evaluations. Previous studies have come to the same conclusion by using task success as their measure of performance (Drew et al., 2018; Kortum & Peres, 2014; Lewis, 2018). This meta-analysis extends that conclusion to situations where performance is measured by task time. With one in four studies finding a dissociation between SUS and performance, dissociations are too common to be explained away as outliers. With some frequency, users perceive a system as more usable than another even though they solve tasks faster or make fewer errors with the other system. This finding resembles Nielsen and Levy's (1994) comparison of performance and preference. Their meta-analysis showed that the users in 25% of the studied cases preferred the system with which they performed worse. Simply put, it is not a viable practice to treat SUS scores as indicative of how users will perform, or vice versa.

A key element in what SUS adds to performance is a sensitivity to the workload experienced when using a system. This sensitivity is apparent in several of the items in the SUS instrument (Table 1). While workload is not mentioned explicitly, Items 3, 8, 9, and 10 relate to the effort, frustration, performance, and mental-demand subscales of TLX, respectively. Workload is well recognized in human factors research but has been less standard in usability studies than instruments for measuring satisfaction or perceived usability (Hornbæk, 2006). In addition to SUS, established instruments for measuring usability and satisfaction in usability studies include SUMI (Kirakowski & Corbett, 1993), PSSUQ (Lewis, 1995), and AttrakDiff (Hassenzahl et al., 2003). Their factors (SUMI: efficiency, affect, helpfulness, control, learnability; PSSUQ: system usefulness, information quality, interface quality; AttrakDiff: pragmatic quality, hedonic identification quality, hedonic stimulation quality) may implicitly tap into workload, but none of them mention it explicitly. Similarly, the ISO 9241-210 (2019) standard does not mention workload at all in its widely applied definitions of usability, user experience, and their dimensions. By attending to workload in a merely implicit manner, usability studies forgo the diagnostic accuracy of this concept compared to broader ones, such as efficiency.

Like task time, workload is an efficiency metric, but they capture different components of efficiency. In usability studies, task time is the prevalent metric for measuring efficiency (Hornbæk, 2006). However, this meta-analysis shows that the workload component of efficiency matters to how users perceive system usability. It is an important result of this study that users are more sensitive to workload than task time when assessing the usability of a system. This difference is mainly driven by novice users and non-safety-critical domains. For experienced users and in safety-critical domains, the correlation between SUS and task time is more similar to that between SUS and workload, though still lower. While novice users may partly relate their task time to their learning process rather than to system usability, experienced users relate the perceived usability of a system more directly to their task time. Regarding domains, it is plausible that users are more concerned with performance when safety is at stake. In contrast, SUS did not correlate significantly with task time and error rate for augmented reality. Possibly, augmented reality represents systems that are valued for their novelty but, therefore, also leave users without reference points regarding what constitutes good performance in terms of speed and accuracy. As a result, there may not yet be a basis for significant correlations between SUS and performance for such systems as opposed to, for example, graphical user interfaces.

The strong correlation between SUS and workload could also be interpreted as further evidence that self-report measures tend to co-vary. Previous studies find moderate to strong correlations between SUS and other measures of perceived usability (e.g., Bangor et al., 2009; Borsci et al., 2015; Lewis, 2018); this study adds a strong correlation between SUS and TLX, that is, self-reported workload. It would be interesting to investigate whether SUS also correlates with physiological and other non-self-report workload metrics. Such a study would reveal whether the SUS-TLX correlation is mainly a correlation among perceptions or extends to other kinds of workload metrics. In addition, there may be little reason to measure both SUS and TLX in system comparisons because these two measures are strongly correlated. The six-dimensional structure of TLX makes it a more diagnostic measure than SUS, which for practical purposes is unidimensional (Lewis, 2018). Studies confirm that the six TLX subscales – mental demand, physical demand, temporal demand, effort, performance, and frustration – capture associated but somewhat independent aspects of workload (Hart & Staveland, 1988; Hertzum, 2022).

While usability is widely studied in human-computer interaction, it is also a debated concept. For example, Tractinsky (2018) contends that the concept of usability is too vaguely defined to be theoretically useful because it tends to be operationalized as an umbrella concept. Relatedly, de Winter (2014) laments the poor scientific foundation of the workload concept, arguing that it does not reflect an empirical reality but, at the same time, acknowledging that measuring it serves important operational purposes. The present meta-analysis largely bypasses these conceptual debates by attending to specific operationalizations of usability and workload – the SUS and TLX

instruments. These instruments have well-defined items with prescribed wordings and response scales. Thereby, they differ from the usability and workload concepts, which rather than being definitive are more aptly described as sensitizing concepts (Hertzum, 2018). As sensitizing concepts, they serve researchers and practitioners by giving them a sense of reference and guidance in approaching concrete instances of system use. In the words of Blumer (1954), they "suggest directions along which to look", rather than provide precise specification. The uncertain leap from the SUS and TLX instruments to the usability and workload concepts accentuates that generalization from this meta-analysis to other instruments for measuring usability and workload requires caution and further research.

5.3 Limitations

Three limitations should be remembered in interpreting the results of this study. First, to be included in this meta-analysis, a study had to report SUS, workload, task time, as well as error rate. This requirement limited the number of eligible studies but was necessary to include the interrelations between workload, task time, and error rate in the analysis of their association with SUS. Additional studies report SUS together with a subset of the three other measures. Second, the paper selection and data analysis were performed exclusively by the author. An independent selection and analysis by another person would have strengthened the paper by validating the paper-selection and data-analysis processes. Third, the measures that could be considered for estimating the effect size in the meta-analysis were restricted by the absence of standard deviations in many of the included studies. While the chosen effect-size measure has been used before (e.g., Hertzum, 2025; Nielsen & Levy, 1994), it would have been preferable with a measure that took the distribution of the data into account by incorporating standard deviations.

6 Conclusion

SUS is a widely used instrument for measuring perceived usability. On the basis of 105 included studies, this meta-analysis finds that (1) SUS and workload, as measured by TLX, were strongly correlated, (2) SUS and performance, as measured by task time and error rate, were related but somewhat independent, and (3) task time and error rate correlated with SUS to similar extents. While one in ten studies displayed a dissociation between SUS and workload, one in four studies displayed a dissociation between SUS and task time and between SUS and error rate. These findings document that SUS complements performance measures and that it largely does so by also being sensitive to the workload experienced when performing with a system. Usability practitioners should maintain a clear distinction between performance and perceived usability and, probably, attend more to workload. Researchers should include standard deviations for the reported means of SUS, TLX, and the like to enable the use of recognized effect-size measures in meta-analyses. Furthermore, future research should investigate whether SUS correlates as strongly with non-self-report workload metrics as with TLX.

CRediT authorship contribution statement

Morten Hertzum: Conceptualization, Data curation, Investigation, Methodology, Formal analysis, Visualization, Writing – original draft preparation, Writing – review & editing

Declaration of Conflicting Interests

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Az-zahra, H. M., Fauzi, N., & Kharisma, A. P. (2019). Evaluating e-marketplace mobile application based on people at the center of mobile application development (PACMAD) usability model. In *Proceedings of the SIET2019 Conference on Sustainable Information Engineering and Technology* (pp. 72–77). IEEE. https://doi.org/10.1109/SIET48054.2019.8986067

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

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. https://doi.org/10.1080/10447310802205776

Blumer, H. (1954). What is wrong with social theory? *American Sociological Review*, *19*(1), 3–10. https://doi.org/10.2307/2088165

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, 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. https://doi.org/10.1080/10447318.2015.1064648

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the system usability scale: A test of alternative measurement models. *Cognitive Processing*, *10*(3), 193–197. https://doi.org/10.1007/s10339-009-0268-9

Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–194). Taylor & Francis.

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

Byers, J. C., Bittner, A. C., & Hill, S. G. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? In A. Mital (Ed.), *Advances in Industrial Ergonomics & Safety* (pp. 481–485). Taylor & Francis.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the CHI1988 Conference on Human Factors in Computing Systems* (pp. 213–218). ACM. https://doi.org/10.1145/57167.57203

de Winter, J. C. F. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology and Work*, *16*(3), 289–297. https://doi.org/10.1007/s10111-014-0275-1

Drew, M. R., Falcone, B., & Baccus, W. L. (2018). What does the system usability scale (SUS) measure? Validation using thinking aloud verbalization and behavioral metrics. In *Proceedings of the DUXU2018 Conference on Design, User Experience, and Usability* (pp. 356–366). Springer. https://doi.org/10.1007/978-3-319-91797-9_25

Finstad, K. (2006). The system usability scale and non-native English speakers. *Journal of Usability Studies*, *1*(4), 185–188.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003* (pp. 187–196). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-80058-9_19

Hertzum, M. (2018). Commentary: Usability—A Sensitizing Concept. *Human-Computer Interaction*, *33*(2), 178–181. https://doi.org/10.1080/07370024.2017.1302800

Hertzum, M. (2022). Associations among workload dimensions, performance, and situational characteristics: A meta-analytic review of the task load index. *Behaviour & Information Technology*, *41*(16), 3506–3518. https://doi.org/10.1080/0144929X.2021.2000642

Hertzum, M. (2025). Understanding preference: A meta-analysis of user studies. *International Journal of Human-Computer Studies*, *195*, article 103408. https://doi.org/10.1016/j.ijhcs.2024.103408

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64*(2), 79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002

ISO 9241-210. (2019). *Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems*. International Standard Organization.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, *24*(3), 210–212. https://doi.org/10.1111/j.1467-8535.1993.tb00076.x

Költringer, T., & Grechenig, T. (2004). Comparing the immediate usability of graffiti 2 and virtual keyboard. In *CHI2004 Extended Abstracts on Human Factors in Computing Systems* (pp. 1175–1178). ACM. https://doi.org/10.1145/985921.986017

Kortum, P., & Peres, S. C. (2014). The relationship between system effectiveness and subjective usability scores using the system usability scale. *International Journal of Human-Computer Interaction*, *30*(7), 575–584. https://doi.org/10.1080/10447318.2014.904177

Kurosawa, H., Sakamoto, D., & Ono, T. (2018). MyoTilt: A target selection method for smartwatches using the tilting operation and electromyography. In *Proceedings of the MobileHCI2018 Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 1–11). ACM. https://doi.org/10.1145/3229434.3229457

Lamontagne, C., Sénécal, S., Fredette, M., Labonté-LeMoyne, É., & Léger, P.-M. (2021). The effect of the segmentation of video tutorials on user's training experience and performance. *Computers in Human Behavior Reports*, *3*, article 100071. https://doi.org/10.1016/j.chbr.2021.100071

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78. https://doi.org/10.1080/10447319509526110

Lewis, J. R. (2018). The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, *34*(7), 577–590. https://doi.org/10.1080/10447318.2018.1455307

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In M. Kurosu (Ed.), *Proceedings of the HCI International 2009* (Issue LNCS 5619, pp. 94–103). Springer. https://doi.org/10.1007/978-3-642-02806-9_12

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

Littell, J. A., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford University Press.

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLOS ONE*, *13*(8), e0199661. https://doi.org/10.1371/journal.pone.0199661

Nielsen, J., & Levy, J. (1994). Measuring usability: Preference vs. performance. *Communications of the ACM*, *37*(4), 66–75. https://doi.org/10.1145/175276.175282

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors*, *33*(1), 17–33. https://doi.org/10.1177/001872089103300102

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 185–218). North-Holland. https://doi.org/10.1016/S0166-4115(08)62387-0

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of the CHI2009 Conference on Human Factors in Computing Systems* (pp. 1609–1618). ACM. https://doi.org/10.1145/1518701.1518947

Schramm, R. C., Sasalovici, M., Hildebrand, A., & Schwanecke, U. (2023). Assessing augmented reality selection techniques for passengers in moving vehicles: A real-world user study. In *Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 22–31). ACM. https://doi.org/10.1145/3580585.3607152

Smith, E., Semple, G., Evans, D., McRae, K., & Blackwell, P. (2020). Augmented instructions: Analysis of performance and efficiency of assembly tasks. In *Proceedings of HCI International 2020* (pp. 166–177). Springer. https://doi.org/10.1007/978-3-030-49698-2_12

Tractinsky, N. (2018). The usability construct: A dead end? *Human-Computer Interaction*, *33*(2), 131–177. https://doi.org/10.1080/07370024.2017.1298038

Triantafyllidis, E., Mcgreavy, C., Gu, J., & Li, Z. (2020). Study of multimodal interfaces and the improvements on teleoperation. *IEEE Access*, *8*, 78213–78227. https://doi.org/10.1109/ACCESS.2020.2990080

Werrlich, S., Daniel, A., Ginger, A., Nguyen, P.-A., & Notni, G. (2018). Comparing HMD-based and paper-based training. In *Proceedings of the ISMAR2018 International Symposium on Mixed and Augmented Reality* (pp. 134–142). IEEE. https://doi.org/10.1109/ISMAR.2018.00046

Xu, W., Liang, H.-N., Chen, Y., Li, X., & Yu, K. (2020). Exploring visual techniques for boundary awareness during interaction in augmented reality head-mounted displays. In *Proceedings of the 2020 Conference on Virtual Reality and 3D User Interfaces* (pp. 204–211). IEEE. https://doi.org/10.1109/VR46266.2020.00039

Zhao, L., Liu, Y., Ye, D., Ma, Z., & Song, W. (2020). Implementation and evaluation of touch-based interaction using electrovibration haptic feedback in virtual environments. In *Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces* (pp. 239–247). IEEE. https://doi.org/10.1109/VR46266.2020.00043