ICIC Express Letters
Part B: Applications
Volume **12**, Number **2**, February **2021**

ICIC International ©2021 ISSN 2185-2766
pp. 175–183

SEVERITY OF USABILITY PROBLEMS AND SYSTEM USABILITY SCALE (SUS) SCORES ON AUGMENTED REALITY (AR) USER INTERFACES

Iverton B. Lima, Yewon Jeong, Chaeyeon Lee, Gangseok Suh and Wonil Hwang*

Department of Industrial and Information Systems Engineering
Soongsil University
369 Sangdo-ro, Dongjak-gu, Seoul 06978, Korea
*Corresponding author: wonil@ssu.ac.kr

Received August 2020; accepted October 2020

Abstract. *Severity scales are commonly employed to determine the level of priority of usability problems (UPs) found during a usability inspection, so they can be eventually resolved in the redesign process. Different types of qualitative and quantitative data, besides a list of UPs and severity ratings, can be obtained from usability studies by employing numerous tools. The system usability score (SUS) is one of the most used questionnaires to measure perceived usability of systems and services. Finding relationships between different types of data obtained from a usability study is important because it can help practitioners and developers to have a deeper insight into the system and evaluation procedures. In this study, we examined if the number of UPs reported through usability inspection can predict SUS scores. A total of 24 participants took part in the experiment, where 4 augmented reality (AR) user interfaces (UIs) were inspected and SUS answers were obtained at the end. After the experiment, severity ratings related to the UPs reported were collected from 4 experts. Ultimately, 272 UPs were divided into 3 severity levels: low, moderate, and high. Based on the number of UPs reported by an evaluator related to each severity level, multiple linear regression analysis was performed. Results showed that only the number of high-severity UPs could predict SUS scores. Furthermore, the effects of previous experience with AR UIs and gender (combined with the number of high-severity UPs) on SUS scores were investigated. Results allowed a deeper view into the AR UIs, their issues, and how evaluator characteristics could be related to the usability evaluation.*

Keywords: Severity, Usability problems, SUS, Usability study, AR

1. Introduction.

The adoption of some sort of usability evaluation method, either during prototype design (formative evaluation) or final design (summative evaluation) stages, is an essential part of the development process of user interfaces (UIs) and will result in a list of usability problems (UPs) [1]. Since dealing with all generated problems is generally unattainable, the use of a severity scale is a common practice to classify and prioritize the UPs from the ones that need utmost attention to UPs that have little to no impact on user performance [2]. One of the most important severity scales was developed by Nielsen [2], where the problems were categorized into 'not a problem', 'cosmetic', 'minor', 'major', and 'usability catastrophe' (from 0 to 4). In another study, a rating scale based on 7 different factors was created to determine the severity of a UP [3]. The factors were frequency, difficulty, workflow impact, persistence, frustration, market impact, and fixing effort. In the same work, authors reported that due to simplicity and quick rating process, many practitioners use a simple classification with only 3 categories: 'minor', 'moderate', and 'major' severity. Since the assessment of the severity of UPs is obtained

DOI: 10.24507/icicelb.12.02.175

from an individual's subjective judgment, it is recommended to use the mean of severity judgments of multiple evaluators to increase reliability [2,4].

Usability studies can produce, on top of a list of UPs, substantial qualitative and quantitative data by utilizing different techniques and tools. The analysis of those data allows a greater understanding of UI, users, and the evaluation process (i.e., evaluation method, evaluator effect). Regarding post-test qualitative methods, interviews, focus groups, or open-ended questionnaires are some of the utilized tools. Moreover, ways of collecting post-test or post-task quantitative data are just as important and often employed. System usability scale (SUS) [5], NASA task load index (NASA-TLX) [6], and post-study system usability questionnaire (PSSUQ) [7] are commonly adopted tools for quantitatively assessing respectively perceived usability, perceived workload, and perceived satisfaction. Subjective usability scores are important since they allow developers, usability practitioners, and everyone else involved in the development process to effectively understand each other when debating over the usability attributes of a product or service [8].

The SUS [5], characterized by the author as "quick and dirty", is a freely available and widely used questionnaire to measure perceived usability. While many studies already attested to its reliability and validity, the SUS has been used across numerous kinds of systems and translated to other languages [8-12]. The sensitivity of the SUS was explored in several studies. In the case of usability assessment studies, the SUS questionnaire can be used to effectively compare two or more UIs, two versions of the same system, or even different tasks within a single interface [8]. Furthermore, the same study reported a negative impact of age on the SUS scores, but no significant effect of gender was found. The amount of experience and its effect on SUS scores was studied by many researchers [13-16], where it was reported that more experience results in higher SUS scores. A personality effect was found by Kortum and Oswald [17]; on the other hand, Schmidt et al. [18] stated that SUS was not affected by personality based on their data. In the work of Kortum and Peres [19], after analyzing data from two studies, authors found a strong positive correlation between SUS scores and performance (measured as task success rate), although the relationship was statistically significant for only one of the studies. For a more complete review and analysis regarding the SUS, we refer readers to check the pieces of work done by Lewis [10] and Brooke [20].

The use of mixed methods design (quantitative and qualitative approach) can be considered a common practice in usability studies. For example, Kim et al. [21] employed, among other methods, SUS and interviews to investigate the usability of swallowing training apps. Another example is the study of an e-learning portal for university students, where the authors collected data from usability testing, heuristic evaluation, user experience questionnaire, and an eye-tracking device [22]. However, the adoption of mixed methods design typically means increased cost and increased time when compared to using just one of the approaches (e.g., increased duration of experiments and increased amount of data to be collected and analyzed by practitioners).

During the literature review, although we found many studies in which usability problems were collected and SUS was used to measure perceived usability, we hardly found prior research that looked for a relationship between UPs (either considering type, total number, or severity) and SUS results. Therefore, the objective of this study is to investigate if the number of low-severity, moderate-severity, and high-severity UPs reported by evaluators, through a usability inspection of augmented reality (AR) UIs, can predict SUS scores (single item's score and total score). By looking for a relationship between the number of UPs by severity and SUS, new insights regarding the system could be potentially obtained, which would be very valuable to practitioners and developers, especially in cases where only usability problems are collected due to limited time or budget. Additionally, we also look for the effects of gender and evaluator's previous experience with AR UIs. Such results could be applied, for example, to the selection of specific participants for a usability study. Moreover, we decided to use AR-based applications in this research on the grounds that it is a technology currently present in several areas and used for different purposes (e.g., remote programming of robots [23] and navigation system for liver surgery [24]), with still a lot of room for growth in the next decades. Hence, the study and evaluation of such systems are important to continue stimulating advancements of AR software and hardware.

This paper is structured as follows. In Section 2, details regarding methodology are given, including participants, preparation (materials and apparatus), procedure, and collected data. The results obtained from multiple linear regression analyses are presented in Section 3, where it can be seen not only which items of the SUS could have their score predicted by the number of UPs reported of either low, moderate, or high level of severity, but also the results when certain characteristics of evaluators (gender and previous experience with AR) were explored. Lastly, the paper is concluded in Section 4, where an overview of the study is provided along with final remarks and future research directions.

2. Method.

Following a within-subject design, an empirical study was conducted where participants experienced 4 different AR UIs (in 4 separate sessions) and used a survey to report UPs and answer the SUS questionnaire for each UI.

2.1. Participants.

In total, 24 participants were recruited (12 males and 12 females). The majority were university students in their twenties (mean = 21.7 years old, SD = 2.3). Participants that had no previous experience with AR UIs were a total of 11, while 13 of them had already interacted with some sort of AR UI at least once.

2.2. Preparation.

In the first part of the preparations for the experiment, we utilized Unity and Vuforia engines to design a navigation and a maintenance AR-based application. To reach a total of 4 applications, gaming and learning AR UIs were obtained from the Android Market. All applications were installed and implemented on a tablet. A survey was created to collect the following data: demographics, previous experience with AR UIs, UPs found during the inspection, and SUS answers. From the SUS developed by Brooke [5] along with modification based on the studies of Finstad [25] and Bangor et al. [8], we created the adapted version of the SUS for our study by replacing the word "system" with "AR application". The adapted version was then translated to the Korean

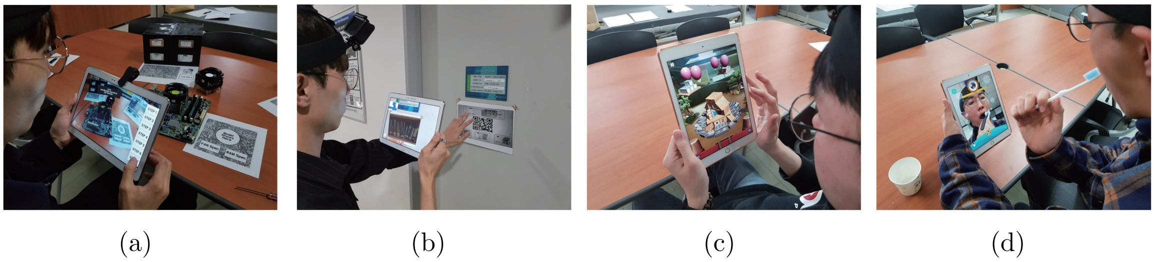
Table 1. Our adapted version of Brooke's SUS questionnaire

[item ID] Adapted SUS statements	
[SUS1] I think that I would like to use this AR application frequently.	[SUS6] I thought there was too much inconsistency in this AR application.
[SUS2] I found the AR application unnecessarily complex.	[SUS7] I would imagine that most people would learn to use this AR application very quickly.
[SUS3] I thought the AR application was easy to use.	[SUS8] I found the AR application very awkward* to use.
[SUS4] I think that I would need the support of a technical person to be able to use this AR application.	[SUS9] I felt very confident using the AR application.
[SUS5] I found the various functions in this AR application were well integrated.	[SUS10] I needed to learn a lot of things before I could get going with this AR application.

Note. *: "awkward" instead of "cumbersome" [8,25] language before it was finally added to the survey. The SUS is constituted of 10 statements, being 5 positive (odd-numbered) and 5 negative (even-numbered) items scored from strongly disagree to strongly agree (5-point scale).

2.3. **Procedures.** Each participant (also referred to as evaluator in this study) randomly experienced one of the AR UIs per session, totaling 96 sessions (24 evaluators × 4 UIs). Sessions occurred only after 24 or more hours after the end of the previous one, serving as a washout period. The first part of the survey, related to demographics and previous experience with AR UIs, was given only once at the beginning of the first session. Each session consists of 5 general procedures: 1) initial instructions regarding the experiment and AR UI were given to the evaluator, 2) time was provided so the evaluator could get familiar with the system, 3) evaluator interacted with the AR UI and performed a task, 4) evaluator inspected the AR UI freely and reported UPs using a survey, and 5) SUS questionnaire was completed by evaluators.

Figure 1 illustrates the 4 different sessions, where the evaluator interacts with only one of the 4 AR UIs. In the maintenance task (a), 2 components of the motherboard were replaced by completing 7 steps. Next, the navigation task (b) was about finding a book inside a laboratory following a path with image targets. The gaming task (c) consists of initially finding an appropriate location to set up the game, and subsequently selecting and completing one of the levels. Finally, the learning task (d) was to follow the directions provided by the app to learn how to perform toothbrushing.



(a) (b) (c) (d)

Figure 1. Evaluator interacting with each one of the 4 AR UIs: (a) maintenance, (b) navigation, (c) gaming, and (d) learning

2.4. **Experimental data.** At the end of the experiment, the following data were collected for statistical analysis: 96 sets of SUS responses and 96 lists of UPs reported by each evaluator. After initial treatment, 272 unique UPs were identified. In a second refinement session, the 272 problems were organized according to their similarities, resulting in 118 different groups. Next, the 118 groups were first separately rated by 4 experts according to their judgment of the severity of the problems within each one. The mean of their severity ratings was calculated, and the UPs were ultimately classified into low, moderate, or high severity. Afterward, the total number of UPs reported by each participant per AR UI was divided into 3 subtotals according to the severity: number of low-severity UPs, number of moderate-severity UPs, and number of high-severity UPs. Additionally, to calculate the overall SUS scores, the equation proposed by Lewis [10] was used: SUSt = 2.5 (20 + SUM (SUS1, SUS3, SUS5, SUS7, SUS9) − SUM (SUS2, SUS4, SUS6, SUS8, SUS10)). As a result, a number between 0 and 100 was computed for each set of SUS responses.

3. **Results.** SUS scores (dependent variables SUSt, which corresponds to the overall score, and SUS1 to SUS10, corresponding to the raw data of one of the SUS items) were predicted through a total of three multiple linear regression analyses. The explanatory variables of the first analysis were the number of low-severity, moderate-severity, and

high-severity UPs reported by an evaluator during usability inspection of each AR UI. In the second analysis, the number of high-severity UPs reported, which was observed as the most relevant predictor of SUS scores, and the evaluator's experience with AR were the explanatory variables. Finally, the last analysis examined the number of high-severity UPs reported and gender as predictor variables.

3.1. **Number of UPs by severity level and SUS scores.** The results for the first multiple linear regression analysis are shown in Table 2. Strong evidence was found regarding the impact of the number of serious problems reported on the SUS scores. Except for SUS1 and SUS5 scores, all estimated regression coefficients (B) related to the number of high-severity UPs reported were found statistically significant (yielding a p-value smaller than .05). Thus, among the 3 levels of severity, only the number of high-severity UPs reported was used as one of the predictor variables in the following analyses, where the effects of AR experience and gender were studied. Lastly, it was observed a significant result regarding the use of the number of low-severity UPs reported to predict SUS4 (need for technical support) score. In summary, it is possible to conclude that if the number of high-severity UPs is known, the scores of almost all SUS items can be predicted, including SUS overall score (SUSt) but with exception of SUS1 and SUS5 scores.

3.2. **Effect of AR experience and number of high-severity UPs.** In the analysis of the effect of AR experience (Table 3), significant coefficient results were found for SUS4 (need for technical support) and SUS10 (amount of learning required) scores ($p = 0.0167$ and $p = 0.0344$, respectively). The only significant coefficient for the interaction was found in the case of SUS6 (perceived inconsistency) score ($p = 0.02472$). When the evaluator has no experience with AR, the equation to estimate the SUS6 score is SUS6 = 1.36312 + 0.13953 $*$ (number of high-severity UPs), as for evaluators that have experienced AR applications, SUS6 = 1.87892 + 0.03216 $*$ (number of high-severity UPs) is the generated multiple linear regression model. A greater value of slope noticed in the first equation indicates that evaluators that have no experience with AR are more sensitive to the number of high-severity UPs on the assessment of perceived inconsistency.

3.3. **Effect of gender and number of high-severity UPs.** As seen in Table 4, no coefficients yielding a p-value smaller than .05 were found when the gender effect was examined. However, when considering coefficient results with a p-value smaller than .1, a suggestively significant impact of gender ($p = 0.0855$) and interaction between gender and the number of high-severity UPs ($p = 0.099$) on SUS8 (perceived awkwardness) score was observed. As a result, the equation generated by multiple linear regression for male evaluators is SUS8 = 2.98231 + 0.01837 $*$ (number of high-severity UPs), while SUS8 = 2.24404 + 0.12785 $*$ (number of high-severity UPs) is the model to estimate the SUS8 score of female evaluators. By looking at the equations, it is possible to see a considerable difference regarding the value of the slope. It can be interpreted as female evaluators are more sensitive to the number of high-severity UPs on the perceived awkwardness ratings.

4. **Discussion and Conclusion.** The SUS is a well-established tool widely used to measure the perceived usability of systems and services. Similarly, the use of severity ratings is a valuable approach to classify and prioritize UPs that will be fixed during the redesign process. In this study, the UPs reported by evaluators during a usability study were classified by experts into 3 levels of severity (low, moderate, and high). Later, the effect of the number of UPs (reported by an evaluator during usability inspection) per severity level on the SUS scores (overall and individual scores) was investigated using multiple linear regression analysis. Significant regression coefficients were found for the impact of the number of high-severity UPs on all SUS scores except for SUS1 and SUS5.

TABLE 2. Estimated regression coefficients (B) and p-values for SUS scores by number of UPs reported for each severity level

	SUSt	SUS1	SUS2	SUS3	SUS4	SUS5	SUS6	SUS7	SUS8	SUS9	SUS10
Low severity UPs	0.23795 (0.88356)	−0.18776 (0.0676) *	0.03133 (0.76569)	−0.00894 (0.93118)	−0.23219 (0.04224) **	−0.03023 (0.742)	0.00196 (0.98106)	0.04435 (0.6657)	0.0749 (0.5245)	−0.04893 (0.64013)	−0.20268 (0.05767) *
Moderate severity UPs	0.19285 (0.72229)	0.003413 (0.92)	−0.03524 (0.3166)	0.019165 (0.5794)	0.01286 (0.73337)	0.024567 (0.424)	0.030654 (0.26777)	0.04808 (0.1627)	0.006608 (0.8663)	0.02662 (0.44658)	0.02983 (0.3991)
High severity UPs	−1.64276 (0.00061) ***	−0.01885 (0.5176)	0.08899 (0.00378) ***	−0.07828 (0.00936) ***	0.09737 (0.00325) ***	0.009042 (0.731)	0.071424 (0.00312) **	−0.07476 (0.0122) **	0.072908 (0.0321) **	−0.081 (0.00787) ***	0.08258 (0.00737) ***

Note. Results are presented as B(p), *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$

TABLE 3. Estimated regression coefficients (B) and p-values for SUS scores by AR experience and number of high-severity UPs

	SUSt	SUS1	SUS2	SUS3	SUS4	SUS5	SUS6	SUS7	SUS8	SUS9	SUS10
AR experience	−11.4128 (0.05891) *	0.23899 (0.54)	0.73806 (0.05949) *	−0.10271 (0.792)	1.0186 (0.0167) **	0.30387 (0.378)	0.5158 (0.09225) *	−0.51806 (0.1792)	0.79196 (0.0718) *	−0.57403 (0.1438)	0.84877 (0.0344) **
High severity UPs	−2.3716 (0.00155) **	0.0004 (0.993)	0.12869 (0.00757) ***	−0.07528 (0.115)	0.16639 (0.00152) ***	0.05311 (0.207)	0.13953 (0.00028) ***	−0.09318 (0.0486) *	0.1335 (0.0134) **	−0.11724 (0.0153) **	0.14836 (0.00272) ***
AR experience × High severity UPs	1.1691 (0.20974)	−0.0244 (0.686)	−0.06743 (0.26397)	−0.00301 (0.96)	−0.10055 (0.12421)	−0.06806 (0.204)	−0.10737 (0.02472) **	0.03224 (0.5883)	−0.09839 (0.1479)	0.06213 (0.3062)	−0.09501 (0.1246)

Note. Results are presented as B(p), *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$

Table 4. Estimated regression coefficients (B) and p-values for SUS scores by gender and number of high-severity UPs

	SUSt	SUS1	SUS2	SUS3	SUS4	SUS5	SUS6	SUS7	SUS8	SUS9	SUS10
Gender	2.27785 (0.7014)	0.143707 (0.704)	0.23759 (0.53449)	0.211054 (0.5767)	0.17347 (0.68255)	−0.00255 (0.994)	−0.11565 (0.7046)	0.015221 (0.968)	−0.73827 (0.0855) *	0.36701 (0.339)	0.26616 (0.5019)
High severity UPs	−1.72194 (0.00901) ***	−0.02041 (0.621)	0.12653 (0.00301) ***	−0.07959 (0.0556)	0.12245 (0.00928) ***	−0.0102 (0.779)	0.06122 (0.0679) *	−0.07245 (0.084) *	0.01837 (0.6924)	−0.06531 (0.12)	0.11224 (0.0106) **
Gender × High severity UPs	0.14068 (0.87828)	0.008503 (0.885)	−0.07943 (0.18106)	0.002987 (0.9592)	−0.03912 (0.5512)	0.040742 (0.431)	0.0247 (0.6008)	0.000503 (0.993)	0.10948 (0.099) *	−0.0283 (0.632)	−0.04755 (0.4382)

Note. Results are presented as B(p), *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$

Little impact was detected by the number of low-severity UPs and no impact was found for the number of moderate-severity UPs, thus not being useful to predict SUS scores. Afterward, regression analysis looked for possible effects caused by the evaluator's prior experience with AR or gender. Evaluators with and without previous experience with AR were found as more sensitive to the effect of the number of high-severity UPs reported on the perceived inconsistency (SUS6) ratings. At last, a statistically weak effect was observed in the analysis of the perceived awkwardness of the system (SUS8) by gender and number of high-severity UPs, indicating that female evaluators are more sensitive to the number of high-severity UPs reported than male ones.

In summary, evidence showed that the number of high severity UPs reported can be used to predict SUS scores. Furthermore, this study was aimed to provide practitioners and developers new ways of using available qualitative and quantitative data to obtain insight into the system. The results related to the analysis of different evaluator characteristics could be used to select specific participants to perform usability inspection and evaluation. Future research on this topic should consider other types of evaluator characteristics, such as age and personality, and also different types of UIs and technologies.

Acknowledgment. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1F1A1054050).

REFERENCES

[1] H. R. Hartson, T. S. Andre and R. C. Williges, Criteria for evaluating usability evaluation methods, *International Journal of Human-Computer Interaction*, vol.13, no.4, pp.373-410, 2001.

[2] J. Nielsen, *Usability Engineering*, Morgan Kaufmann, 1994.

[3] S. Herr, N. Baumgartner and T. Gross, Evaluating severity rating scales for heuristic evaluation, *Proc. of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, San Jose, CA, USA, pp.3069-3075, 2016.

[4] J. Nielsen, Reliability of severity estimates for usability problems found by heuristic evaluation, *Posters and Short Talks of the 1992 SIGCHI Conference on Human Factors in Computing Systems*, New York, USA, pp.129-130, 1992.

[5] J. Brooke, SUS: A 'quick and dirty' usability, in *Usability Evaluation in Industry*, P. Jordan, B. Thomas and B. Weerdmeester (eds.), London, Taylor and Francis, 1996.

[6] S. G. Hart and L. E. Staveland, Development of NASA-TLX (task load index): Results of empirical and theoretical research, in *Human Mental Workload*, P. A. Hancock and N. Meshkati (eds.), North-Holland, Netherlands, Elsevier, 1988.

[7] J. R. Lewis, Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ, *Proc. of the Human Factors Society Annual Meeting*, Santa Monica, CA, USA, vol.36, no.16, pp.1259-1260, 1992.

[8] A. Bangor, P. T. Kortum and J. T. Miller, An empirical evaluation of the system usability scale, *International Journal of Human-Computer Interaction*, vol.24, no.6, pp.574-594, 2008.

[9] M. Gao, P. Kortum and F. L. Oswald, Multi-language toolkit for the system usability scale, *International Journal of Human-Computer Interaction*, pp.1-19, 2020.

[10] J. R. Lewis, The system usability scale: Past, present, and future, *International Journal of Human-Computer Interaction*, vol.34, no.7, pp.577-590, 2018.

[11] S. C. Peres, T. Pham and R. Phillips, Validation of the system usability scale (SUS): SUS in the wild, *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, Santa Monica, CA, USA, vol.57, no.1, pp.192-196, 2013.

[12] T. S. Tullis and J. N. Stetson, A comparison of questionnaires for assessing website usability, *Usability Professional Association Conference*, Minneapolis, MN, USA, vol.1, pp.1-12, 2004.

[13] P. Kortum and M. Johnson, The relationship between levels of user experience with a product and perceived system usability, *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, Houston, TX, USA, vol.57, no.1, pp.197-201, 2013.

[14] P. T. Kortum and A. Bangor, Usability ratings for everyday products measured with the system usability scale, *International Journal of Human-Computer Interaction*, vol.29, no.2, pp.67-76, 2013.

[15] P. Kortum and M. Sorber, Measuring the usability of mobile applications for phones and tablets, *International Journal of Human-Computer Interaction*, vol.31, no.8, pp.518-529, 2015.

[16] U. Lah and J. R. Lewis, How expertise affects a digital-rights-management-sharing application's usability, *IEEE Software*, vol.33, no.3, pp.76-82, 2016.

[17] P. Kortum and F. L. Oswald, The impact of personality on the subjective assessment of usability, *International Journal of Human-Computer Interaction*, vol.34, no.2, pp.177-186, 2018.

[18] T. Schmidt, V. Wittmann and C. Wolff, The influence of participants' personality on quantitative and qualitative metrics in usability testing, *Proc. of Mensch und Computer 2019*, New York, NY, USA, pp.115-126, 2019.

[19] P. Kortum and S. C. Peres, The relationship between system effectiveness and subjective usability scores using the system usability scale, *International Journal of Human-Computer Interaction*, vol.30, no.7, pp.575-584, 2014.

[20] J. Brooke, SUS: A retrospective, *Journal of Usability Studies*, vol.8, no.2, pp.29-40, 2013.

[21] H. Kim, S. H. Lee, N. B. Cho, H. You, T. Choi and J. Kim, User-dependent usability and feasibility of a swallowing training mhealth app for older adults: Mixed methods pilot study, *JMIR mHealth and uHealth*, vol.8, no.7, e19585, 2020.

[22] B. A. Zardari, Z. Hussain, A. A. Arain, W. H. Rizvi and M. S. Vighio, QUEST e-learning portal: Applying heuristic evaluation, usability testing and eye tracking, *Universal Access in the Information Society*, pp.1-13, 2020.

[23] A. E. Solyman, K. M. Ibrahem, M. R. Atia, H. I. Saleh and M. R. Roman, Perceptive augmented reality-based interface for robot task planning and visualization, *International Journal of Innovative Computing, Information and Control*, vol.16, no.5, pp.1769-1785, 2020.

[24] F. Zhang, S. Zhang, K. Zhong, L. Yu and L. N. Sun, Design of navigation system for liver surgery guided by augmented reality, *IEEE Access*, vol.8, pp.126687-126699, 2020.

[25] K. Finstad, The system usability scale and non-native English speakers, *Journal of Usability Studies*, vol.1, no.4, pp.185-188, 2006.