VALIDATION OF SYSTEM USABILITY SCALE FROM EXPERT AND ORDINARY USER PERSPECTIVE

Bayu Kelana[1*], Anggar Riskinanto[2], Mudrikah Nasyiah[3], Muhammad Fajar Fiandhika[4]
Information System Department, Universitas Ary Ginanjar[1, 2, 3, 4]
*Correspondence Email : bayu@esqbs.ac.id

ABSTRACT

This meticulous and comprehensive study aims to provide a thorough overview of the validation results of the ten System Usability Scale (SUS) indicators from the perspectives of experts and ordinary users. The study meticulously compares the validation results of SUS scores in expert and ordinary user groups. The validation results of SUS scores in the expert group are derived from a comprehensive analysis of SUS evaluation results, heuristic evaluation, and interviews with three usability experts. Similarly, the validation results of SUS scores in the ordinary user group are obtained from a rigorous comparison of SUS evaluation results, remote moderated usability testing, and interviews with five ordinary users. While inconclusive, this study's findings shed light on four crucial points. First, there are no universally valid validation results of SUS scores in all SUS indicators from expert and ordinary user perspectives. Second, the ease of use indicator is the only one with a complete and valid evaluation result from the expert perspective. Third, the intuitiveness indicator, while full, yields invalid evaluation results from the standpoint of ordinary users. Fourth, a weak relationship is observed between SUS scores and the results of heuristic evaluation and usability testing. These findings, derived from a robust and meticulous methodology, are expected to significantly contribute to the existing research in application usability, particularly in the context of SUS usage.

KEYWORDS

system usability scale, heuristic evaluation, usability evaluation, usability testing, interview

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International

INTRODUCTION

Testing the software interface is essential to improving user experience from a usability perspective [1]. One such testing method is the System Usability Scale (SUS) evaluation, renowned for its reliability in measuring application usability [2]. Although introduced by Brooke [3] a long time ago, SUS remains one of the widely used methods for measuring usability through questionnaires, especially in healthcare, finance, and social networks [4], [5], [6], [7], [8], [9]. Although the results of quantitative data evaluation like SUS can relatively determine whether a design is good or bad, this type of data finds it challenging to assert absolutely.

Therefore, an assessment like this aims to compare a design with a standard or another design rather than to describe the usefulness of an application [10]. Hence, qualitative evaluation is necessary to complement quantitative evaluation and precisely understand usability issues [10]. Several studies have researched the validation of SUS scores. Drew et al. [11] and Nasyiah et al. [9] conducted research validating SUS scores with usability test results from ordinary users. Besides usability testing, Nasyiah et al. [9] also use interviews to validate the SUS scores. Besides being conducted by ordinary users, experts can also conduct usability evaluations [12], [13]. Validation of SUS scores conducted by experts is still relatively rare. The study of Fiandhika et al. [7] is one of them and uses heuristic evaluation and interview methods to validate the SUS scores. Therefore, validating SUS scores with qualitative usability evaluations conducted by experts and ordinary users can provide new findings for industries using SUS. It is well known that experts have a deeper understanding of how an information system's usability is evaluated compared to that of ordinary users. However, Barth [14] begs to be different, as there is no difference in perspective between an expert and an ordinary user. This study will be important as, from our understanding, this is the first study that investigates the usability evaluation conducted by experts and ordinary users.

System Usability Scale

Brooke [3] developed the SUS questionnaire for the first time, consisting of five statements with positive and five with negative connotations. Riihiaho [13] also supports this questionnaire format. This format avoids potential bias that may lead participants to contemplate each question before determining their answer. By modifying Brooke's method [3], Sauro and Lewis [15] developed a SUS questionnaire consisting of ten statements with positive connotations. Positive connotations aim to facilitate participants' answering of the questionnaire. Additionally, this approach aims to ensure participants' accurate interpretation of statements when researchers need the opportunity to verify the answers received from participants. Because of using Sauro and Lewis's [10] questionnaire style in their studies, the results of Nasyiah et al. [9] and Fiandhika et al. [7] are compared to obtain insight into validating SUS scores with qualitative usability evaluations conducted by experts and ordinary users.

Heuristic Evaluation

Heuristic evaluation is a method in which experts assess a user interface design against a set of guidelines, known as heuristics, to identify and address design problems related to usability [14]. Nielsen developed one set [16], as seen in Table 1. Fiandhika et al. [7] obtained two findings using the heuristic evaluation to validate the SUS score. First, only the ease of use indicator score from all participants could be validated by the heuristic evaluation results, which is a valid score. Second, the score of frequency of use, system integration, and speed of learning factors from all participants could not be validated by heuristic evaluation results.

Table 1. Nielsen's Heuristic

Code	Principles
H1	Visibility of system status
H2	Match between the system and the real world
H3	User control and freedom
H4	Consistency and standards
H5	Error prevention
H6	Recognition rather than recall
H7	Flexibility and efficiency of use
H8	Aesthetic and minimalist design
H9	Help users recognise, diagnose, and recover from errors
H10	Help and documentation

1.1 Remote Moderated Usability Testing

Based on the type of participant involved in the implementation, usability evaluation methods are divided into heuristic evaluation and usability testing. Heuristic evaluation does not include ordinary users, while usability testing involves ordinary users [17]. Based on participant location, usability testing is divided into two, namely in-person testing and remote testing. In-person testing is usability testing conducted in a laboratory room. Remote testing is a valuable approach to reach a wider range of participants and enhance ecological validity through remote testing, where the process utilises tools for sharing video and audio between researchers and users [1], [18].

Based on the researcher's involvement as a facilitator, remote testing is divided into remote moderated testing and remote unmoderated testing. Remote moderated testing allows two-way communication between participants and facilitators, as they are connected online simultaneously during usability testing. Unlike remote moderated testing, remote unmoderated testing does not involve real-time communication between facilitators and participants [19].

Using remote moderated testing to validate the SUS score, Nasyiah et al. [9] obtained three findings. First, there are only two SUS indicators in which all participants could validate the score: ease of use and intuitiveness. Second, the scores from four of five participants are valid, and one of five participants has a disability in the ease of use indicator. Third, in contrast with the ease of use indicator, the scores from all participants are invalid in the intuitiveness indicator. Only the intuitiveness indicator score could be validated completely by all participants. The intuitiveness indicator score of all participants is invalid.

1.2 Interview

An interview is a research method in the field of user experience, where a facilitator asks a participant questions about a topic of interest to study that topic [20]. Interviews are conducted at the end of usability testing to gather verbal feedback from participants regarding their behaviours while using the application [20]. Interviews are conducted to understand participants' experiences and receive critiques and suggestions about the tested application [21], [22]. The interview method has three advantages: 1) Suitable for implementing systems at the outset; 2) Suitable for a small number of participants; and 3) Suitable for obtaining user thoughts [23]. Despite its advantages, the interview method has three disadvantages: 1) Not everyone is comfortable talking to strangers; 2) Not everyone can always remember things in detail and accurately; 3) Participants sometimes think that small things are not important and therefore do not need to be mentioned [20]. Additionally, interviews also require a long time [23]. One type of interview is a semi-structured interview. This interview method focuses on topics of interest to be explored [5]. The semi-structured interview method is commonly found in usability testing research in health [24], [25], [26], [27].

Using interviews to validate the SUS scores, Nasyiah et al. [9] obtained two findings. First, only the ease of use indicator score has a valid measurement, completely from all participants. Second, although not all participants gave a valid score, speed of learning and learning needs indicators obtained valid SUS measurements from four out of five participants. Using the same activity, Fiandhika et al. [7] found that the scores of ease of use, speed of learning, confidence, and learning needs factors are valid

RESEARCH METHOD

To achieve its objectives, this study employs a comparative study approach. It compares the SUS evaluation validation results with each qualitative usability evaluation result from experts and ordinary user groups. This study uses the result of Fiandhika et al.'s [7] study to represent the expert's perspective, while Nasyiah et al.'s [9] study describes the ordinary users' perspective. In Fiandhika et al.'s study, to conduct qualitative usability testing, participants must first try using the application to be evaluated. Then, they must conduct heuristic evaluations and interviews. This is slightly different from Nasyiah et al.'s [9] study, in which the participants tried using the application in a remote-moderated usability testing process before the interview. After collecting the qualitative data, both studies filled out the SUS questionnaire. The illustration of the design of this study can be seen in Figure 1.

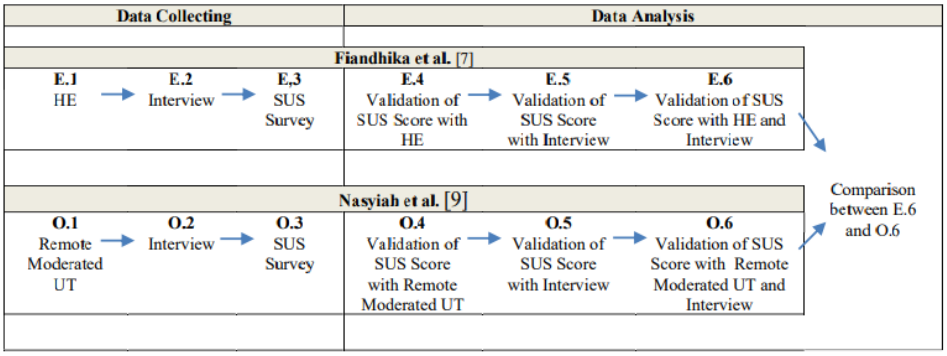

Figure 1. Design of Study

Both studies used different applications as their subjects. The expert group investigated the usability of an electronicbased budget planning application (e-Budgeting) [7], while the ordinary user group evaluated the usability of a social networking application [9]. In collecting data, Fiandhika et al.'s [7] study involved three experts with at least one year of experience as UI/UX specialists who have designed mobile or website applications. According to Nielsen and Molich [27], three to five experts are ideal for conducting heuristic evaluations. Meanwhile, Nasyiah et al.'s [9] study involved five participants from Generation Z who had a job as a worker (two participants) and university students (two participants). The usability testing with five people allows for finding usability issues nearly as effectively as testing with more participants [27]. Therefore, the number of participants involved in the research has met the standard.

1.3System Usability Scale Survey

During this phase, data collection in studies of Fiandhika et al. [7] and Nasyiah et al. [9] involved providing the SUS questionnaire to survey the participants, using a Likert scale ranging from 0 (strongly disagree) to 4 (strongly agree). The SUS questionnaire, derived from Sauro & Lewis [15], employs a positive connotation for all 10 SUS statements, detailed in Table 2.

Table 2. SUS Statement

Code	Statement	Indicator
T1	I think I will use this system often	Frequency of use
T2	I found the system uncomplicated	System simplicity
T3	I think this system is easy to use	Ease of use
T4	I think I need support from a technical person to be able to use this system	Technical support
T5	I found the various functions in this system to be well integrated.	System Integration
T6	I think there are a lot of inconsistencies in this system	Consistency
T7	I imagine that most people will learn to use this system quickly.	Speed of learning
T8	I found this system very complicated to use	Intuitiveness
T9	I feel very confident in using this system	Confidence
T10	I need to learn a lot of things before I can use this system	Learning needs

1.4Heuristic Evaluation

This phase aims to gather usability assessments from an expert viewpoint using the ten heuristic principles established by Nielsen [16]. In Fiandhika et al.'s [7] study, each expert conducted a heuristic evaluation by identifying issues in the ebudgeting application and documenting them in an evaluation form. A thorough examination of the ten business processes within the e-budgeting application was required to align with these principles. Those business processes are 1) Maintaining budget account; 2) Maintaining assumption; 3) Maintaining MPP; 4) Maintaining volume; 5) Maintaining formula; 6) Finalization budget report; 7) Budget entry; 8) Review budget entry; 9) Submit and approval budget; and 10) Checking budget entry all cost centre.

1.5Remote Moderated Usability Testing

Before evaluating using the SUS method, the ordinary user group conducted usability testing using the remote moderated usability testing method. For this purpose, Nasyiah et al. [9] developed a research instrument in the form of test scenarios to guide researchers in instructing ordinary users to perform tasks when trying the social networking application. The test scenarios consist of functions that ordinary users will perform, the duration of completion, and the key steps to answer those tasks. Table 3 provides examples of test scenarios.

Table 3. Test Scenario

Task	Duration	Key Steps
Register the account	120'	1. Search Darisini.com 2. Click "daftar" or "ayo mulai" 3. Continue with Google 4. Choose Google account 5. Fill up personal data on the "buat profile" page 6. Click "simpan"
Log in	20'	1. Logout 2. Click "ayo mulai" 3. Continue with Google 4 4. Log in Success

1.6Interview

The semi-structured interview focuses on exploring engaging topics [5]. During the heuristic evaluation, the experts were asked about their experiences using the e-budgeting application using this method [7]. It is also done to ordinary users for their

experiences using the social network application [9]. The 10 SUS indicators guided the interview. The examples of the interview guide are provided in Table 4.

Table 4. Interview Guidance

Code	Indicator	Guidance
T1	Frequency of use	How often do you expect to use this web application?
T2	System simplicity	In your opinion, what do you think about this web application?
T3	Ease of use	What is your experience when using this web application?

RESULT AND DISCUSSION

The Expert Perspective

Specific results were derived from validating the SUS indicator score through heuristic evaluation. Specifically, the score of T3 among all participants was validated by the heuristic evaluation, confirming its validity. However, the heuristic evaluation results could not corroborate all participants' T1, T5, and T7 scores. The SUS indicator score's validation process with interview results differs from that with heuristic evaluation results. All T3, T7, T9, and T10 scores are deemed valid. Consequently, only the T3 score remains valid after being validated by both heuristic evaluation and interview results.

The Ordinary User Perspective

In remote moderated usability testing, validation of the SUS indicator score confirms only T8's score, making all other participants' scores invalid. This differs from validation via interview results, where only T3's score was valid and collected from all participants. While not all participants provided valid scores, T7 and T10's scores were valid for four out of five participants.

Comparison between the Expert and the Ordinary User Perspectives

Four findings compare the result of SUS score validation between the expert and the ordinary user perspectives. First, no results from expert and user groups across all indicators are entirely valid. Second, T3 is the only indicator that receives comprehensive evaluation findings from the expert group, allowing the SUS score in this indicator to be validated. This may occur because the heuristic evaluation results in the expert group have a strong relationship with the SUS indicator. The validation results of the SUS score in this indicator are valid.

Thirdly, T8 is the only indicator that receives comprehensive evaluation findings from the ordinary user group, allowing the SUS score in this indicator to be validated. This may occur because this ordinary user group's remote moderated usability testing findings strongly correlate with the SUS indicator. The validation results of the SUS score in this indicator are invalid. Fourthly, if the comparison is solely between SUS evaluation results and interviews, then the usability indicator has valid scores from expert and ordinary user groups. This may happen because researchers guide participants more effectively in the evaluation process across the 10 SUS indicators in interviews compared to heuristic evaluation and usability testing.

Discussion

Comparing SUS score validation with qualitative usability evaluations yields three interesting findings compared to previous research on the same subject. Firstly, this study only obtains one validated SUS indicator from each expert and ordinary user group. This may be due to the SUS indicators lacking a strong correlation with the results of heuristic evaluations and remote moderated usability testing. These findings correlate with those of previous research. Drew's study [11] suggests that SUS scores may be more useful as a formative usability testing method to provide a comprehensive overview of user experience. Thus, this comprehensive overview better depicts the application's usability

when evaluated from the perspective of experts and intuitiveness from the standpoint of ordinary user groups. Thirdly, the high number of invalid SUS scores compared to qualitative usability evaluations may also be because this study's SUS questionnaire needed reliability and validity testing. This limitation also occurred in Drew's study [11].

REFERENCES

Muhyiddin. (2020). Covid-19, New Normal and Development Planning in Indonesia. The Indonesian Journal of Development Planning, 4(2), 240–252. https://doi.org/10.36574/jpp.v4i2.118

Ayoib, C.A., & Nosakhare, P.O (2015). Directors culture and environmental disclosure practice of companies in Malaysia. International Journal of Business Technopreneurship, 5(1), 99-114. https://doi.org/10.36574/jpp.v4i2.118

Smith, J. (2021). Advancements in Artificial Intelligence. In A. Johnson (Ed.), Proceedings of the International Conference on Computer Science (pp. 123-135). New York: XYZ Publications.

Arikunto Suharsimi (2016). Research Procedures A Practical Approach. Jakarta: PT Rineka Cipta