System Usability Scale Validation from The Expert Perspective

Muhammad Fajar Fiandhika* and *Bayu* Kelana

STIMIK ESQ, Information System Department, 12560 South Jakarta, DKI Jakarta, Indonesia

Abstract. The System Usability Scale (SUS) is a well-known method for quickly evaluating usability. Although SUS corresponds more strongly with user preference, quantitative data like SUS score is challenging to express in absolute terms. Therefore, the research aims to explore the validity of SUS evaluation results of an e-budgeting system from the expert perspective. With a comparative study approach, the research compared each of the 10 SUS indicators' scores with the result of the qualitative usability evaluation. Data were obtained from three usability experts, who conducted Nielsen's heuristic evaluation, the SUS measurement, and interviews. The research obtained two main findings. Firstly, no entirely valid scores of SUS indicators were found from three usability evaluations of the three usability experts. Secondly, only the score of ease of use shows significant relevance to the heuristic evaluation and interview result. Those findings may provide new insights when evaluating usability.

1 Introduction

Technological advancements are progressing at an unprecedented rate in this digital era compared to previous times. Companies, government entities, and private organizations are facing the imperative to adapt to sustain their operations. To address budget management challenges, e-budgeting has emerged as a pivotal solution, enhancing the efficiency and effectiveness of budget management for businesses and organizations. The e-budgeting system is an information system designed to facilitate the creation of expenditure plans, thereby enhancing productivity, and ensuring the viability of planning systems [1].

However, implementing the e-budgeting system at an Indonesian state-owned enterprise in the insurance sector has encountered several challenges. One of the foremost issues is the difficulty experienced by users who struggle to grasp and utilize the e-budgeting system effectively. User experience encompasses all facets of user interactions with digital systems or products offered by companies [2]. This e-budgeting system is suspected to suffer from usability issues, as reported by budget managers. Non-technical users find it arduous to navigate the application, often requiring assistance from the system development team. This

* Corresponding author: m.fajar.f@students.esqbs.ac.id

strongly suggests that the e-budgeting system at this organization is plagued by user experience problems.

The aforementioned usability challenges have contributed to user resistance, a common issue rooted in a system-centric perspective [3]. Patsiotis et al [4] corroborate this by revealing that resistance often leads to user reluctance or delays in adopting an application or service. To mitigate user resistance and enhance the usability of the e-budgeting system, it is imperative to incorporate usability assessments throughout the implementation process. Testing the software interface is essential to improving user experience from a usability perspective [5]. However, despite its importance, the system development team acknowledges that usability evaluation has not been implemented for the e-budgeting system.

Conducting user experience research on software necessitates the use of appropriate methods to gather usability evaluation data, which can then inform evaluations and recommendations for application enhancements. One such testing method is the System Usability Scale (SUS) evaluation, renowned for its reliability in measuring application usability [6]. In this study, the SUS evaluation employed a questionnaire developed by Lewis & Sauro [7]. However, it's worth noting that SUS does have its limitations, including that suggests SUS is a measure of user perception, and not actual usability [1]. Besides, SUS often overlooks Nielsen's heuristic principles, which are despite being integral to conducting usability assessments [8], [9]. Additionally, the study by Abiwardani et al. [10] indicates that low SUS scores are corroborated by participant interviews, albeit without comprehensive elaboration.

These limitations with SUS align with the research findings of Wahyuningrum et al. [8], suggesting that SUS alone may not provide sufficient insights into usability evaluation outcomes. Therefore, to augment understanding and bolster verification, it is advisable to complement SUS indicators and results with other evaluation methods. Two such methods are Heuristic Evaluation and Interviews. Heuristic Evaluation involves the assessment of usability problems based on heuristic principles with expert input [11]. Meanwhile, according to Melnik [12], interviews are a research technique that yields profound insights into user needs, difficulties encountered, and user aspirations, fostering empathy with users. Through this comprehensive research approach, it is anticipated that a deeper understanding of usability issues and critical patterns can be unearthed, with input from the expert perspective.

2 Materials and methods

The research aims to validate 10 indicators of the SUS evaluation from the usability expert perspectives. Using a comparative study with a non-experimental approach, the analysis was performed in three phases, namely pre-usability evaluation, usability evaluation, and post-usability evaluation. After problems were identified, studied works of literature, and evaluation planning, three usability evaluation methods were accomplished for an e-budgeting application by three usability experts. Each of the experts conducted a heuristic evaluation, then a SUS evaluation, and finally an interview process. All collected qualitative data from heuristic evaluation and interviews were examined by verbatim analysis and theme analysis with a top-down approach to obtain and map its result to every SUS indicator. Meanwhile, all collected quantitative data from the SUS evaluation were calculated to obtain the SUS score. Next step, the score of each SUS indicator was compared to the heuristic evaluation result, and then compared to the interview result. If both are the same, the obtained SUS indicator score is valid. Not valid, if not the same. The SUS indicator score cannot be validated if there is no result from heuristic evaluation. The big picture of the research process can be seen in Figure 1.

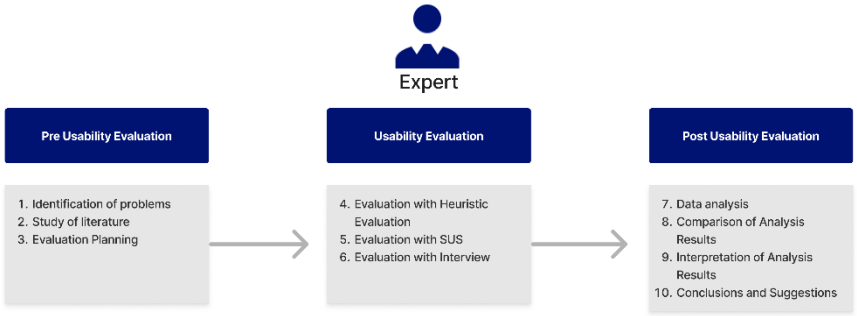

Fig. 1. Usability Evaluation Process

According to Nielsen and Molich [13], three to five experts are ideal for conducting heuristic evaluations. Therefore, the number of participants involved in the research has met the standard. The participants have a minimum of 1 year of working experience as UI/UX specialists and have made mobile or website application designs.

2.1 Heuristic Evaluation

This stage aims to obtain usability evaluation from an expert perspective based on the 10 heuristic principles developed by Nielsen [14]. Heuristic evaluation was performed by every expert by filling in problems in the e-budgeting application to an evaluation form. A total of 10 business processes contained in the e-budgeting application must be explored based on those 10 principles. The form of the heuristic evaluation can be seen in Table 1.

Table 1. Heuristic Evaluation Form

No	Feature	Problem/Usability issue	Principles Violated
1	Maintain Budget Account		
2	Maintain Assumption		
3	Maintain MPP		
4	Maintain Volume		
5	Maintain Formula		
6	Finalization Budget Report		
7	Budget Entry		

No	Feature	Problem/Usability issue	Principles Violated
8	Review Budget Entry		
9	Submit & Approval Budget		
10	Checking Budget Entry All Cost Center		

2.2 SUS Evaluation

At this stage, data was collected by giving the SUS questionnaire to the expert to fill in on a Likert scale from 0 (strongly disagree) to 4 (strongly agree). The SUS questionnaire was adopted from Sauro & Lewis [7], which uses a positive tone for all 10 SUS statements. Those SUS statements form can be seen in Table 2.

Table 2. SUS Statements

Code	SUS 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

Code	SUS Statement	Indicator
T10	I need to learn a lot of things before I can use this system.	Learning needs

2.3 Interview

The semi-structured interview is conducted focusing on topics that are interesting to explore [15]. Using this approach, participants were interviewed about their experience in using the e-budgeting application, when doing the heuristic evaluation before. The interview was performed based on the 10 SUS indicators as guidance. The interview guidance can be seen in Table 3.

Table 3. 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?
T4	Technical support	Do you need help/support from a technical person to use this application to operate this application?
T5	System Integration	Of the various features/tasks that have been carried out, what do you think about the features of this application?
T6	Consistency	Of the various features, appearance, design, and overall that are in the application, what did you find? What are your comments on this application? In your opinion, what is the level of consistency in this application?
T7	Speed of learning	How was your first experience in learning this application?
T8	Intuitiveness	What is your experience in using this application? Does each application test meet expectations?
T9	Confidence	How do you feel when using this application? * Are you afraid of being wrong? Hesitant? And any other feelings?
T10	Learning needs	How do you learn this application? or How was your experience in learning this application?

3 Results and discussion

3.1 SUS Measurement

Based on formula ((*R*1-1)+(5-*R*2)+(*R*3-1)+(5-*R*4)+(*R*5-1)+(5-*R*6)+(*R*7-1)+(5-*R*8)+(*R*9-1)+(5-*R*10)) ∗2.5), the average of SUS score on the e-budgeting is 43.3 (E), which means awful. Expert 3 gave the highest SUS score (52.5) and Expert 2 gave the lowest SUS score (32.5). Both SUS scores mean awful as well. The detailed data and the SUS score from every expert can be seen in Table 4.

Table 4. SUS Score Calculation Results

Participant	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**	**SUS Score**
Expert 1	1	4	0	0	2	0	1	3	1	0	45
Expert 2	1	2	0	2	3	3	1	2	0	0	32.5
Expert 3	0	0	0	0	1	0	0	0	0	0	52.5
Average SUS Score	**0.7**	**2**	**0**	**0,7**	**2**	**1**	**0.7**	**1.7**	**0.3**	**0**	**43.3**

In detail, T3 and T10 obtained the lowest score (0 out of 4). Meanwhile, T2 and T5 obtained the highest score (2 out of 4). Because T10 and T2 have a negative meaning, T10 and T5 have the best result, while T3 and T2 have the worst result. The highest score in the SUS indicator with a negative meaning means the worst result.

To proceed to the comparative analysis stage of the SUS indicator score with the result of heuristic evaluation and interview, it is necessary to convert the score into three categories, namely disagree (X) for 1 and 2, neutral (N) for 3, and agree (Y) for 4 and 5. The result of converting the SUS indicator score can be seen in Table 5.

Table 5. Converted SUS Evaluation Result

Participant	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**	**Number**		
											X	**N**	**Y**
Expert 1	X	Y	X	X	N	X	X	X	X	X	**8**	**1**	**1**
Expert 2	X	N	X	N	Y	Y	X	N	X	X	**5**	**3**	**2**

Expert 3		X	X	X	X	X	X	X	X	X	X	**10**	**0**	**0**
Number	**X**	**3**	**1**	**3**	**2**	**1**	**2**	**3**	**2**	**3**	**3**			
	N	**0**	**1**	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**			
	Y	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**0**			

Expert 3 disagreed with all statements (10 out of 10 statements). Meanwhile, in line with Expert 3, the other experts also disagreed with most of the statements. On the other side, no statement was agreed upon by all experts. In contrast, all participants disagreed with T1, T3, T7, T9, and T10.

3.2 Heuristic Evaluation Analysis

Using a working form as guidance, the heuristic evaluation was done before the participant filled out the SUS questionnaire. The qualitative data from the participants' heuristic evaluation responses were processed and analyzed in some phases. Through the transcription process of the participant's evaluation, the reduction process of the transcript, and the thematic analysis of the reduction result, the heuristic problems in every heuristic principle were obtained. The number of heuristic problems in every heuristic principle can be seen in Table 6. The top three problems were obtained in consistency and standards, visibility of system status, and user control and freedom. Meanwhile, two principles were not obtained heuristic problem, namely error prevention and aesthetic and minimalist design.

Table 6. Result of Heuristic Evaluation

No.	The Nielsen's Heuristic Principles	Number of Principles Violated
1	Visibility of system status.	14
2	Match between the system and the real world.	8
3	User control and freedom.	10
4	Consistency and standards.	20
5	Error prevention.	-
6	Recognition rather than recall.	3
7	Flexibility and efficiency of use.	3

No.	The Nielsen's Heuristic Principles	Number of Principles Violated
8	Aesthetic and minimalist design.	-
9	Help users recognize, diagnose, and recover from errors.	2
10	Help and documentation.	6

Then, the findings in every heuristic were mapped to every SUS indicator. Next, the result of mapping was defined as agreeing or disagreeing with a statement for each. The result was compared with the score for each SUS indicator. If both are the same, the obtained score is valid. Not valid, if not the same. If there is no finding from heuristic evaluation, the score cannot be validated. The validation of the SUS indicator score with the heuristic evaluation result can be seen in Table 7.

Table 7. Comparison Between the SUS Indicator Score and Heuristic Evaluation Result

Indicator	Expert 1			Expert 2			Expert 3			Result		
	S	H	R	S	H	R	S	H	R	TD	TV	V
T1	X	-	TD	X	-	TD	X	-	TD	3	0	0
T2	Y	X	TV	N	-	TD	X	X	V	1	1	1
T3	X	X	V	X	X	V	X	X	V	0	0	3
T4	X	-	TD	N	-	TD	X	X	V	2	0	1
T5	N	-	TD	Y	-	TD	X	-	TD	3	0	0
T6	X	X	V	Y	-	TD	X	X	V	1	0	2
T7	X	-	TD	X	-	TD	X	-	TD	3	0	0
T8	Y	X	TV	N	-	TD	X	X	V	1	1	1
T9	X	-	TD	X	-	TD	X	-	TD	3	0	0
T10	X	X	V	X	-	TD	X	X	V	1	0	2

X = Disagree; Y = Agree; N = Neutral; S = SUS Indicator Score; H = Heuristic evaluation result; R = Result; TV = Not Valid; V = Valid; "-"= No Result; TD = No Validation

Some findings were obtained from the validation of the SUS indicator score with the heuristic evaluation result. Only the score of T3 from all participants could be validated by the heuristic evaluation results, which is a valid score. Meanwhile, the score of T1, T5, and T7 from all participants, could not be validated by heuristic evaluation results.

3.3 Interview Analysis

Using a similar process for analyzing data from heuristic evaluation results, the finding from the analyzed interview results was compared with the score for each SUS indicator. In contrast with the heuristic evaluation result, all participants gave responses to all SUS indicators. Therefore, the score of all SUS indicators could be validated with the interview result.

Table 8. Comparison Between the SUS Indicator Score and Interview Result

Indicator	Expert 1			Expert 2			Expert 3			Result		
	S	I	R	S	I	R	S	I	R	TD	TV	V
T1	X	Y	TV	X	X	V	X	X	V	0	1	2
T2	Y	X	TV	N	X	TV	X	X	V	0	2	1
T3	X	X	V	X	X	V	X	X	V	0	0	3
T4	X	X	V	N	Y	TV	X	X	V	0	1	2
T5	N	Y	TV	Y	X	TV	X	X	V	0	2	1
T6	X	X	V	Y	X	TV	X	X	V	0	1	2
T7	X	X	V	X	X	V	X	X	V	0	0	3
T8	Y	X	TV	N	X	TV	X	X	V	0	2	1
T9	X	X	V	X	X	V	X	X	V	0	0	3
T10	X	X	V	X	X	V	X	X	V	0	0	3

X = Disagree; Y = Agree; N = Neutral; S = SUS Indicator Score; I = Interview result; R = Result; TV = Not Valid; V = Valid; TD = No Validation

The validation of the SUS indicator score with the interview results can be seen in Table 8. In contrast with the validation of the SUS indicator score with heuristic evaluation result, the scores of T3, T7, T9, and T10 are valid. Therefore, only T3 has a valid score when it was validated by the heuristic evaluation results and the interview results.

4 Conclusion

The research assessed the validity of the SUS evaluation results of an e-budgeting system from the expert perspective. The overall usability was rated as poor, with varying expert opinions indicating an "awful" usability level. Detailed analysis of the SUS evaluation result revealed specific tasks with pronounced usability issues, including ease of use and system simplicity, while learning need and system integration showed comparatively better usability. Comparison between SUS scores and heuristic evaluation results indicated discrepancies, with only the score of ease of use aligning with heuristic evaluation findings. During interviews, all participants provided feedback on SUS indicators, confirming the SUS indicator scores, particularly ease of use, speed of learning, confidence, and learning needs,

which have valid scores. Therefore, only the score of ease of use consistently exhibited a valid score from all usability evaluation methods.

These findings in heuristic evaluation analysis support the previous similar research on ordinary user. Participants who encountered difficulties in carrying out their tasks with efficiency and effectiveness frequently justified their performance by citing reasons unrelated to system usability with Nielsen's standard. As a result, they assigned a higher or lower SUS score to the system than an impartial evaluation of their experience would justify [1]. Meanwhile, the other similar research on usability expert did not focus on verification of every score of SUS indicator with interview.

In this study, two limitations were identified. With these limitations, there are suggestions to consider for future research. Firstly, the three usability experts who participated in this study had a maximum of 1 year and 5 months of experience in the field of application usability. Having a larger number of participants with longer work experience as usability experts may yield more diverse findings in future research.

References

1. S. Sutriyani, Sistem e-budgeting Dalam Meningkatkan Akuntabilitas: Perspektif Maqashid Syari'ah [E-budgeting System in Increasing Accountability: Maqashid Syari'ah Perspective], 4, 2, 513 (2023) https://doi.org/10.47467/elmal.v4i2.1742
2. S. Gibbons, UX vs. Service Design, Nielsen Norman Group, 2021, [Online], Available : https://www.nngroup.com/articles/ux-vs-service-design/, [Accessed : 21-Jan-2023]
3. M. Haddara and H. Moen, User resistance in ERP implementations: A literature review, Procedia Comput. Sci., 121, 859–865 (2017)
4. A. G. Patsiotis, T. Hughes, and D. J. Webber, Adopters and non-adopters of internet banking: A segmentation study, Int. J. Bank Mark., 30, 1, 20–42 (2012)
5. K. Moran, Usability Testing 101, Nielsen Norman Group, 2019, [Online]. Available: https://www.nngroup.com/articles/usability-testing-101/, [Accessed: 04-Feb-2023]
6. A. R. Firdaus, N. H. Wardani, and L. Fanani, Evaluasi dan Rekomendasi Perbaikan Usability pada System Autoworksys pada PT. Prima Berkat Gemilang dengan menggunakan Metode Heuristic Evaluation dan System Usability Scale (SUS) [Evaluation and Recommendations for Usability Improvements in the Autoworksys System at PT. Prima Berkat Gemilang using the Heuristic Evaluation Method and System Usability Scale (SUS)], J. Peng. Tek. Inf. dan Ilmu Komp., 3, 6, 5799–5807 (2019)
7. J. R. Lewis and J. Sauro, Item Benchmarks for the System Usability Scale, Usability Stud, 13, 3, 158–167 (2018)
8. T. Wahyuningrum, C. Kartiko, and A. C. Wardhana, Exploring e-Commerce Usability by Heuristic Evaluation as a Compelement of System Usability Scale, Int. Conf. Adv. Data Sci. E-Learning Inf. Sys. ICADEIS 2020, 1–5 (2020) 10.1109/ICADEIS49811.2020.9277343
9. R. Nurlistiani and N. Purwati, Interpretasi Pengujian Usabilitas E-Learning di Masa Pandemi COVID-19 Menggunakan System Usability Scale [Interpretation of E-Learning Usability Testing during the COVID-19 Pandemic Using the System Usability Scale], Sem. Nas. Hasil Pen. dan Pengab. Masy., 1, 0, 164–171 (2021)

10. H. Abiwardani, B. T. Hanggara, and B. S. Prakoso, Evaluasi Usability Aplikasi Usaha Laundry Berbasis Web Menggunakan Metode Usability Testing (Studi Kasus: Aplikasi Smartlink Bos) [Usability Evaluation of Web-Based Laundry Business Applications Using Usability Testing Methods (Case Study: Smartlink Bos Application)], Pengemb. Teknol. Inf. dan Ilmu Komput., 4, 3, 822–829 (2020)

11. I I. N. Arifin, H. Tolle, and R. I. Rokhmawati, Evaluasi dan Perancangan User Interface untuk Meningkatkan User Experience menggunakan Metode Human-Centered Design dan Heuristic Evaluation pada Aplikasi Ezyschool [User Interface Evaluation and Design to Improve User Experience using Human-Centered Design and Heuristic Evaluation Methods in the Ezyschool Application], J. Pengemb. Teknlogi Inf. dan Ilmu Komput., 3, 2, 1725–1732 (2019)

12. N. Melnik, Conducting User Interviews, Usability Testing, and Surveys, UXmatters, 2022, [Online], Available: https://www.uxmatters.com/mt/archives/2022/01/conducting-user-interviews-usability-testing-and-surveys.php, [Accessed: 16-Sep-2023]

13. J. Nielsen and R. Molich, Heuristic Evaluation of User (1990)

14. J. Nielsen, 10 Usability Heuristics for User Interface Design, NNGroup, 2020, [Online], Available : https://www.nngroup.com/articles/ten-usability-heuristics/ [Accessed : 02-Aug-2023]

15. M. G. Davis, M. R. C. Haas, M. Gottlieb, J. B. House, R. D. Huang, and L. R. Hopson, Zooming In Versus Flying Out: Virtual Residency Interviews in the Era of COVID-19, AEM Educ Train, 4, 4, 443–446 (2020) doi: 10.1002/aet2.10486