Perceived Usability Evaluation of IRiS: an Integrated Recommendation Collection System

Jimmy[1] and Kristian Tanuwijaya[2]

[1,2] University of Surabaya, Surabaya, Indonesia
jimmy@staff.ubaya.ac.id

Abstract. This study evaluates the perceived usability of IRiS, which was developed to collect recommendations from senators related to the election of principals in the University of Surabaya (UBAYA). The primary question of this study was "Will IRiS be usable for all senators to use as intended?". The answer to this question is critical, considering that senators were seniors from diverse backgrounds with varying levels of digital literacy. We empirically evaluated the perceived usability using the System Usability Scale (SUS) and asked senators to fill in the SUS questionnaires soon after they used IRiS to submit their recommendations. In general, IRiS was perceived as having "Good" usability and was well-accepted by UBAYA senators. We found the level of perceived usability to be well distributed across gender and age. Nevertheless, we found that participants from STEM faculties perceived IRiS with higher usability scores than participants from non-STEM faculties.

Keywords: User Perception, Usability, Recommendation Collection.

1 Introduction

Usability was formally defined as "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"[1]. ISO 9241-11:2018 stated that the level of usability directly affected the level of effectiveness, efficiency, and satisfaction of users towards the use of a system[1]. Hence, usability was considered a fundamental variable that positively affected system adoption[2], [3].

This paper focused on evaluating the usability of IRiS, an **I**ntegrated **R**ecommendat**i**on collection **S**ystem, as perceived by its users. IRiS was developed to collect recommendations from the university's senators to support the election of principals in the University of Surabaya (UBAYA).

Figure 1 shows the process of principal election in UBAYA. First, each candidate presented their campaign to the senators. Then, each senator was required to submit a set of recommendations for each candidate to the election committee. Each set of recommendations contains an array of close-ended (quantitative) and open-ended (qualitative) questions. Finally, the election committee reported the recommendations

M. Hartono et al. (eds.), *Proceedings of the 4th International Conference on Informatics, Technology and Engineering 2023 (InCITE 2023)*, Atlantis Highlights in Engineering 21,
https://doi.org/10.2991/978-94-6463-288-0_42

to the principal of the targeted role. For example, recommendations to elect a faculty Dean were reported to the Rector of UBAYA.

Figure. 1. The process of principal election in UBAYA

Prior to the adoption of IRiS, the collection of senates' recommendations was performed using pen and paper. The senators provided recommendations in paper-based forms for each candidate, and upon completion, the election committee collected the forms. The committee then manually tallied the quantitative recommendation and reported the overall recommendation. Such a "traditional" process was considered tedious, inefficient, inaccurate, and lack of security to protect the confidentiality of the senators' recommendations.

Nevertheless, the decision to replace the traditional system with a computerized system was not an easy decision to make. The decision was mainly affected by the following considerations:

1. Will the system be secured enough to ensure that only intended parties are allowed to access the recommendations?
2. Will the system be usable for all senators to use as intended?

This paper focused on answering the second question and left the discussion related to the first question for other avenues.

2 System Usability Scale (SUS)

SUS, initially introduced by Brooke in 1996[4], is an instrument to measure the usability of software and/or hardware products [5]. SUS was best known for its simplicity as it used quick and easy-to-answer questionnaires to measure the usability of any system [4], [6]. SUS contained ten statements as shown in Table 1.

The ten statements included five positive statements (statements with odd numbers) and five negative statements (statements with even numbers). Using SUS, users were asked to express their agreement with each of the ten statements on a Likert scale from 1 (strongly disagree) to 5 (strongly agree).

The SUS score ranged from 0 (worst imaginable) to 100 (best imaginable)[7]. Equation 1 formally defines how to compute the SUS Score. First, the 5-level Likert scale (x_i) is mapped into scores between 0 to 4 [4]. For positive statements (S_1, S_3, S_5, S_7, and S_9), the score is the scale level minus 1. For negative statements (S_2, S_4, S_6, S_8, and S_{10}), the score is five minus the scale level. Subsequently, sum the scores from all items and multiply the sum by 2.5 to obtain the SUS score.

$$SUS\ Score = \left(\sum_{i=1,3,5,7,9} x_i - 1 + \sum_{i=2,4,6,8,10} 5 - x_i\right) 2.5 \quad (1)$$

Table 1. The ten statements of SUS [4].

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 of things before I could get going with this system.

Bangor, Kortum, and Miller's study [8] suggested that the average SUS score of their nearly 10-year data is 68. Since then, the 68 was commonly used as a borderline to determine the acceptability of a system's usability. A follow-up study [7] involving 964 participants was performed to determine the adjective rating scale to SUS score from worst imaginable to best imaginable. Figure 2 shows the average SUS score given for each rating as found in Bangor, Kortum, and Miller study [7].

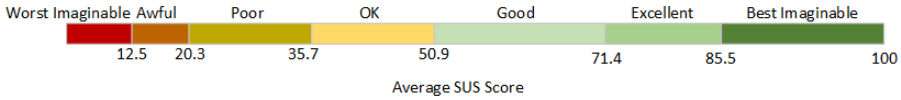

Figure 2. Mapping of average SUS score to adjective ratings

This study used SUS as it provides quick and easy-to-use survey items. The use of uniform survey items allowed comparison between the system being evaluated and other systems. Furthermore, SUS allowed scoring and grading on the usability of the evaluated system.

3 Research Method

To evaluate the user perception of IRiS usability, we prepared a set of questionnaire items based on the SUS items (see Table 2). The questionnaire items were written in Indonesian, following the primary language of IRiS target users. Table 2 lists the questionnaire items in Indonesian (and the English translation). Participating users were asked to tick their agreement level from 1 (Strongly disagree) to 5 (Strongly agree). Prior to filling in the SUS items, we asked participating senators to specify their employee ID for demographic evaluation.

The questionnaire was printed and distributed to senator members who attended each election meeting soon after they submitted their recommendation using IRiS. Each senator was asked to fill out the questionnaire only once. A person can be a

senator at the university level and/or at a faculty level. A senator can submit recommendations for multiple positions. Therefore, a person can use IRiS multiple times. When this occurred, the person was asked to fill in the questionnaire once after their first use of the IRiS. We then performed statistical analysis on the questionnaire results to determine the usability of IRiS. For this, we performed an evaluation on participants demographic data in search of possible biases related to the participant's profile.

Table 2. Questionnaire items (and the English translation).

No.	Pernyataan (Statements)
S_1	Saya merasa ingin sering menggunakan IRIS untuk pengumpulan rekomendasi senat. (I think that I would like to use IRIS to collect senators' recommendations.)
S_2	Saya merasa kompleksitas IRIS terlalu berlebihan. (I think that IRIS unnecessarily complex.)
S_3	Saya merasa IRIS mudah untuk digunakan. (I thought IRIS was easy to use.)
S_4	Saya merasa memerlukan bantuan dari petugas teknis untuk dapat menggunakan IRIS. (I think that I would need the support of a technical person to be able to use IRIS.)
S_5	Saya merasa berbagai fitur IRIS telah terintegrasi dengan baik. (I found the various functions in IRIS were well integrated.)
S_6	Saya pikir terdapat terlalu banyak inkonsistensi di IRIS. (I thought there was too much inconsistency in IRIS.)
S_7	Saya kira kebanyakan orang akan dapat belajar untuk menggunakan IRIS dengan sangat cepat. (I would imagine that most people would learn to use IRIS very quickly.)
S_8	Saya merasa IRIS sangat janggal/aneh/canggung untuk digunakan. (I found IRIS very cumbersome/awkward to use.)
S_9	Saya merasa sangat percaya diri ketika menggunakan IRIS. (I felt very confident when using IRIS.)
S_{10}	Saya perlu mempelajari banyak hal sebelum saya dapat menggunakan IRIS. (I needed to learn a lot of things before I could use IRIS.)

4 Results

We collected 81 questionnaires from 81 senators participating in election meetings across seven faculties in UBAYA between March and April 2023. Seven results were found to be incomplete and were discarded. Therefore, 74 data from the questionnaires were considered. For statistical analysis purposes, we categorized the seven faculties into STEM faculties (i.e., Pharmacy, Biotechnology, and Engineering) and NON-STEM faculties (Creative Industry, Business and Economics, Law, and Polytechnic). Table 3 shows the demographic characteristics of the participants.

We performed further analysis on age and faculty category (i.e., STEM vs. Non-STEM) as these variables affected users' digital literacy and subsequently affected technology adoption (i.e., a person's ability to use digital media or platform)[9], [10]. Figure 3 visualizes age distribution by gender (left) and by faculty category (right). We used an independent t-test on both distributions and found that age distribution by gender and by faculty category was not statistically different.

Table 3. Demographic of participants

Gender (%)	Female	51.35
	Male	48.65
Age	Min	28
	Max	64
	Mean	49.81
Faculty Category (%)	STEM	51.35
	NON-STEM	48.65

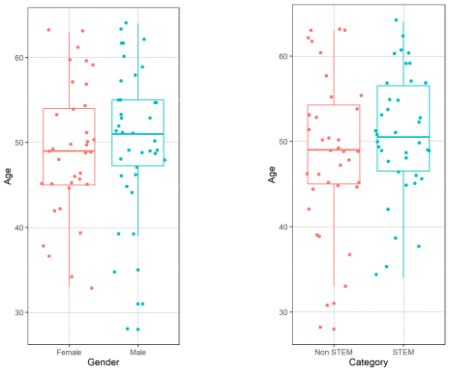

Figure 3. (left) Age distribution by gender, and (right) Age distribution by faculty category

Now that we have a better understanding of the participants, we analyzed IRiS usability as perceived by the participants. On average, IRiS obtained a 70.34 SUS score which was considered an upper "Good" (close to being "Excellent") [7] (see Figure 2). Table 4 lists details of the SUS Score per statement and the sum of scores from all ten statements. We then considered the participants' faculty category and found that participants from STEM faculties considered the usability of IRiS better than those from Non-STEM Faculties. The difference was found to be statistically significant ($p \leq 0.05$). In general, participants from STEM faculties perceived IRiS as having excellent usability. On the other hand, participants from non-STEM faculties perceived IRiS as a system with good usability. Such findings were in line with findings from Das & Bhattacharyya's [9] study, where participants from their STEM group were found to be able to cope with new digital technologies quicker than participants from the non-STEM group.

Table 4. SUS Score per statement and the sum of scores from all participants and groups of participants based on faculty category (STEM vs. Non-STEM). The * symbol signifies significant differences among results from STEM and non-STEM groups.

Data	*n*	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	*Total*
All	74	7.47	5.61	7.84	7.20	7.16	6.52	7.50	7.64	6.93	6.49	70.34
Non-STEM	36	7.64	5.14	7.29	6.60	6.81	6.18	7.29	7.50	6.67	5.49	66.60*
STEM	38	7.30	6.05	8.36	7.76	7.50	6.84	7.70	7.76	7.17	7.43	73.88*

Next, we evaluated the scores obtained from each statement. Figure 4 shows the distribution of scores for each statement. The lowest overall score occurred in S_2 ($\bar{x} = 5.61 \pm 2.48$; $M = 5$), suggesting that over-complexity was perceived as the biggest problem by our participants. This is intriguing as, at the same time, the highest overall score was observed in S_3 ($\bar{x} = 7.84 \pm 1.72$; $M = 7.5$), suggesting that IRiS was perceived as easy to use by most participants. A plausible explanation for these con-

flicting results was that participants used S_2 to measure the complexity of the recommendation instruments. The instruments required senators to provide recommendation for each candidate by answering 34 quantitative and five qualitative (narrative) questions. On average, there were two candidates in a candidature meeting. Hence, at the end of a candidature meeting, each senator answered, on average, 68 quantitative and ten qualitative questions. Hence, while they considered IRiS as easy to use (S_3), they still considered the instruments that they needed to fill in as overly complex and expressed that in S_2.

Subsequently, we evaluated results from individual statements and evaluated their correlations. Table 5 shows that there were no strong correlations between the SUS scores of individual statements. An exception was observed between S_4 and S_{10}, where there was a strong positive correlation between the two variables (r = 0.719). Such observation was expected as S_4 and S_{10} were designed to evaluate the same Learnability factors of a system [11].

Table 5. Correlations between SUS scores from individual statements. Number in bold shows strong correlations between the two statements' scores.

	S_1	**S_2**	**S_3**	**S_4**	**S_5**	**S_6**	**S_7**	**S_8**	**S_9**	**S_{10}**
S_1	1.000									
S_2	0.158	1.000								
S_3	0.408	0.508	1.000							
S_4	0.268	0.424	0.614	1.000						
S_5	0.435	0.457	0.556	0.317	1.000					
S_6	0.378	0.548	0.570	0.573	0.458	1.000				
S_7	0.394	0.388	0.551	0.480	0.536	0.535	1.000			
S_8	0.459	0.476	0.574	0.654	0.313	0.544	0.343	1.000		
S_9	0.415	0.313	0.521	0.323	0.673	0.528	0.580	0.319	1.000	
S_{10}	0.188	0.529	0.529	**0.719**	0.305	0.582	0.395	0.539	0.272	1.000

Furthermore, the jitter points in Figure 4 show another interesting pattern where 0 (zero) scores were observed in even-numbered statements only (with the exception of only one zero point that occurred in S_9). Even numbered statements were statements with negative tones (see Table 3). Further analysis revealed that, in general, scores obtained from the positive statements ($\bar{x} = 7.38 \pm 1.79;\ M = 7.5$) were higher than scores obtained from the negative statements ($\bar{x} = 6.69 \pm 2.31;\ M = 7.5$). The delta was found to be highly statistically significant ($p \leq 0.01$). Such findings raised the question of whether the results were biased based on the polarity of the statements (positive or negative statements).

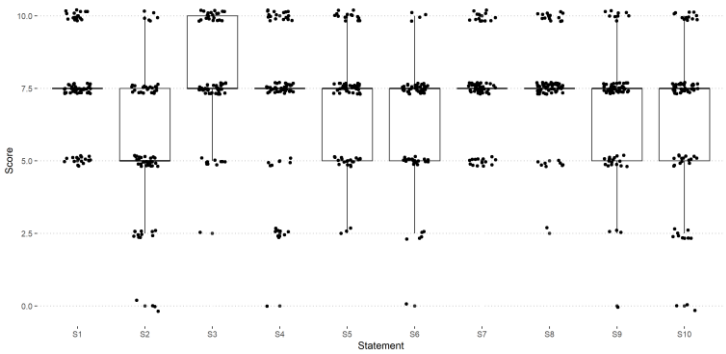

Figure 4. Distribution of scores for each statement

Prior studies presented conflicting conclusions with regard to questionnaire statement polarity. A study by Naji and Ahmad [12] suggested that the statement polarity has no bias towards the validity of Likert scale results. Another group of studies showed that biases on varying statement polarity did exist, yet this bias can be minimized by balancing the use of statements with both polar [13], [14]. On the other end of the spectrum, Alvarez et al. [15] suggested that balancing statements' polarity greatly biased the results. For this, we argued that, in line with the design of SUS, balancing the use of both negative and positive statements was the best compromise to minimize bias from the statement polarity. Nevertheless, we acknowledged the possible polarity bias when analyzing results from individual statements.

We then examined the age and SUS score data to check whether age has a bias on user usability perceptions. Figure 5 shows that there is no correlation between age and user perceptions of IRiS usability (Spearman correlation coefficient R = -0.026). Such observation indicated that IRiS was perceived as having good usability regardless of the user's age. Therefore, we concluded that there was no age bias in the SUS Score.

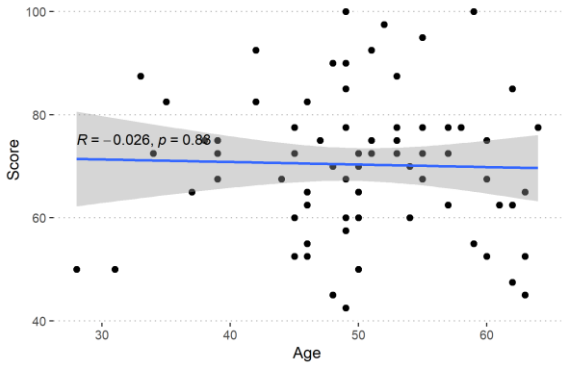

Figure 5. The scatter plot on Age vs. Score distribution shows that there is no correlation between the two variables

6. Conclusions

This study sought to investigate the usability of IRiS as perceived by UBAYA senators. IRiS is an **I**ntegrated **R**ecommendat**i**on collection **S**ystem that was developed to collect senator recommendations in a meeting to elect a principal in UBAYA. To evaluate the usability of IRiS and answer the research question, we used the System Usability Scale (SUS).

We obtained 74 user perceptions on IRiS usability data from 74 senators participating in election meetings across seven faculties in UBAYA. We found no bias related to age distribution or gender distribution across faculty categories (STEM vs. Non-STEM).

Overall, participants perceived the usability of IRiS as "Good". The usability scores were found to be consistent across ages. Nevertheless, we found that participants from STEM faculties perceived IRiS with higher usability scores than participants from non-STEM faculties.

The lowest usability score was identified in S_2 ("I think that IRiS was unnecessarily complex"). Interestingly, the highest score was identified in S_3 ("I thought IRiS was easy to use"). A plausible explanation for these conflicting findings is that participants used S_2 to measure the complexity of the recommendation instrument instead of measuring the complexity of IRiS. The recommendation instrument was indeed cumbersome as, on average, senators were asked to answer 68 quantitative and ten qualitative questions related to the two candidates in an election meeting. Hence, while participants considered IRiS as easy to use (S_3), they still considered the instruments that they needed to fill in as overly complex and expressed it to S_2.

An interesting avenue for the future avenue is to compare and contrast user perception of IRiS usability against users' actual experiences while using IRiS (e.g., time to complete, or learning curve). By doing so, we could evaluate the impact of user perceptions on their experience.

This study benefits those in the domain of system development. This study showed how a digital technology was proven to be usable and was well accepted by senior people (i.e., the average age of our participants was 50 years old) from diverse backgrounds with varying levels of digital literacy.

References

[1] ISO, "ISO 9241-11:2018 Ergonomics of Human-System Interaction, Part 11: Usability: Definitions and concepts," 2018. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en (accessed May 15, 2023).

[2] C. Chang and H. Almaghalsah, "Usability evaluation of e-government websites: A case study from Taiwan," *Int. J. Data Netw. Sci.*, vol. 4, no. 2, pp. 127–138, 2020.

[3] I. Maramba, A. Chatterjee, and C. Newman, "Methods of usability testing in the development of eHealth applications: a scoping review," *Int. J. Med. Inform.*, vol. 126, pp. 95–104, 2019.

[4] J. Brooke, "SUS-A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, 1996.

[5] M. Hyzy *et al.*, "System usability scale benchmarking for digital health apps: meta-analysis," *JMIR mHealth uHealth*, vol. 10, no. 8, p. e37290, 2022.

[6] A. Kaya, R. Ozturk, and C. Altin Gumussoy, "Usability measurement of mobile applications with system usability scale (SUS)," in *Industrial Engineering in the Big Data Era: Selected Papers from the Global Joint Conference on Industrial Engineering and Its Application Areas, GJCIE 2018, June 21--22, 2018, Nevsehir, Turkey*, 2019, pp. 389–400.

[7] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," *J. usability Stud.*, vol. 4, no. 3, pp. 114–123, 2009.

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

[9] A. R. Das and A. Bhattacharyya, "Is STEM a better adaptor than non-STEM groups with online education: an Indian peri-urban experience," *Asian Assoc. Open Univ. J.*, 2023.

[10] A. Tarhini, K. Hone, and X. Liu, "Measuring the moderating effect of gender and age on e-learning acceptance in England: A structural equation modeling approach for an extended technology acceptance model," *J. Educ. Comput. Res.*, vol. 51, no. 2, pp. 163–184, 2014.

[11] J. R. Lewis and J. Sauro, "The factor structure of the system usability scale," in *Human Centered Design: First International Conference, HCD 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009 Proceedings 1*, 2009, pp. 94–103.

[12] M. A. N. Qasem and S. B. A. Gul, "Effect of items direction (positive or negative) on the factorial construction and criterion related validity in Likert scale," *Asian J. Res. Soc. Sci. Humanit.*, vol. 4, no. 4, pp. 114–121, 2014.

[13] J. L. Pimentel and J. L. Pimentel, "Some biases in Likert scaling usage and its correction," *Int. J. Sci. Basic Appl. Res.*, vol. 45, no. 1, pp. 183–191, 2019.

[14] J. L. Pimentel, "A note on the usage of Likert Scaling for research data analysis," *USM R\&D J.*, vol. 18, no. 2, pp. 109–112, 2010.

[15] J. Suárez Álvarez *et al.*, "Using reversed items in Likert scales: A questionable practice," *Psicothema, 30*, 2018.