Usability Evaluation on Educational Chatbot using the System Usability Scale (SUS)

Arief Hidayat*
Computer Science Department,
Wahid Hasyim University
Semarang, Indonesia
rifmillenia@gmail.com

Agung Nugroho
Mechanical Engineering Department,
Wahid Hasyim University
Semarang, Indonesia
agungnugroho3006@gmail.com

Safa'ah Nurfa'izin
Chemical Engineering Department,
Wahid Hasyim University
Semarang, Indonesia
nurfaizin95@gmail.com

***Abstract*—Chatbot is a medium that can be used in education, especially learning, students seem to communicate through chat with educators about learning materials. To assist the learning in computer hardware, AI-based was used to build the chatbot. This research was to asses the effectiveness of myHardware chatbot. The research started with the determination of research object, data used and the scale evaluation in the utility system. The method consisted of scenario determination, data collection, and calculation. Based on the results, the assesment score obtained was 61.03. With a description of marginal acceptability, a grade D scale, with ok rating of adjective and the application had sufficient usability value. Based on the final System Usability Scale (SUS) score, it is possible to conclude that the myHardware chatbot is acceptedable by the students and can function well.**

Keywords—usability, educational, chatbot, system usability scale

I. INTRODUCTION

Chat robots, often known as chatbots, are one of the advanced technologies that may eventually replace human workers. Later humans or users may simply discover the needed information thanks to the work system carried out by chatbots. Chatbots are now being incorporated on websites. In 1966, Joseph Weizenbaum, an MIT professor, created the first chatbot. Chatbots were, of course, built quite simple at the time. The ELIZA chatbot is a software that works with MAC time-sharing. A technology at MIT that allows people and computers to converse in natural language. The decomposition rules triggered by the keywords in the text input are used to examine the input sentences.ELIZA is concerned with the following essential technological issues: (1) keyword detection, (2) identification of minimum context, (3) choice of suitable transformations, (4) creation of answers in the key words absence, and (5) supply of code editor for ELIZA "scripts." [1]. The chatbots on the website often respond to a query submitted by the user. The scope in issue has also been constrained so that it does not extend beyond the stated boundaries. However, there are chatbot programs that do not have a scope restriction, thus when answering a question, it is frequently not in agreement with what the user asked.

The application of a chatbot application in learning is still lackThe use of a chatbot application in education is still insufficient. There is no need for the chatbot application that is utilized as the foundation because it has not been deployed in learning. However, not everyone uses these arguments while developing chatbot apps. Chatbot technology is a type of Natural Language Processing application. NLP is a branch of Artificial Intelligence that explores human-computer communication using natural language. Such computing models will facilitate communication between humans and computers, particularly in terms of learning, giving the impression that students are interacting with lecturers.

Based on previous research conducted by [2] stated the increasing awareness of students in using mobile phones for educational purposes and Patrick bii Who said chatbots can be useful for educational purposes because they are more interactive than traditional e-learning systems? Students can continue to interact with bots by asking questions about specific fields. [3].

Although chatbot systems and conventional web applications are both applications that are built based on the development of web technology. However, in the process of interacting with users, the two are very different. In conventional web applications, users interact with applications through mediums such as buttons, tables, forms, and the like, so in chatbot-based applications, users must interact with agents who focus on their abilities in terms of natural language processing. The myHardware chatbot is built as an AI-based learning chatbot to help learn about computer hardware and its components and functions

One of the fields of science to analyze and evaluate the level of ease of use of software is Usability [4]. Usability is defined as technique in assesing the software testing with the aspects of learnability, efficiency, memorability, errors, and satisfaction. SUS is one of the methods for analyzing or testing usability by involving end users of the process.

SUS has several advantages, including the evaluation stage is easier and understandable by respondents, describes the maximum results even though it involves a small sample, and can distinguish between applications whether they can be used properly or not. SUS has a clear calculation method in conducting evaluations, so it is hoped that the evaluation value obtained has a level of accuracy that can be maintained. Usability is a quality feature that assesses how simple the user interface is to use. A well-designed interface can increase user-system interaction. Furthermore, usability is a factor that has a significant impact on an application's success. Three areas of usability measurement, according to the International Standard Organization, are: (1) Effectiveness is defined as the determination of users in a given environment to achieve a specific goal. (2) Efficient is the user's ability to achieve the goal. (3) Satisfaction is the absence of discomfort and the positive behavior of a product. [5].

II. RELATED WORK

Chatbot apps are utilized for teaching and learning in educational settings. According to research, chatbots may be utilized to convey learning content to students via online platforms as conversational agents capable of supplying users with reliable information. [6], [7]. Educators recognize

2022 Seventh International Conference on Informatics and Computing (ICIC) | 979-8-3503-4571-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIC56845.2022.10006991

the value of using Chatbots in educational settings to provide students with an engaging experience. [8].These bots allow students to ask questions and receive replies. [9], and receive personalized assistance [10]. The advent of learning methodology in education, such as the Chatbot system, has individualized online learning and made learning materials available to students at any time and from any location. Wartman and Combs [11] contend that education should advance in tandem with changes in the professional sector, necessitating the use of Artificial Intelligence (AI) in teaching and learning Chatbots may be utilized during learning to forecast and intensely tailor student learning sessions by modeling individual learning patterns using natural language conversation. [12]. Based on their studies [13]–[15], Chatbots are an excellent technology breakthrough for increasing student learning motivation, cognitive mastery, and accomplishment.

This research is trying to combine engineering pedagogic science itself with information technology. Not just an effort to digitize engineering science using information technology. The development of this learning chatbot application covers all existing aspects, namely engineering pedagogy, information technology, learning, games and character values. All of that will be in research on the development of this learning chatbot application called myHardware chatbot as mentioned in Fig. 1.

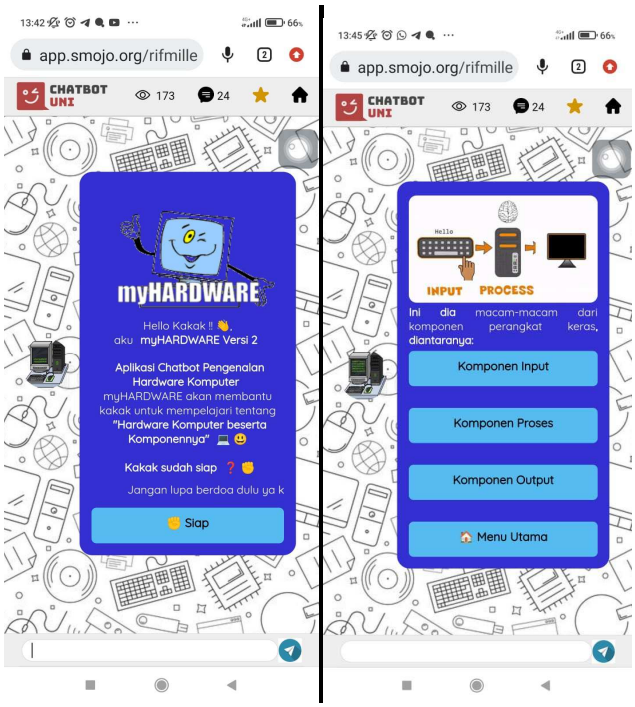

Fig. 1. Chatbot myHardware

The System Usability Scale (SUS), according to past research, is a questionnaire that may be used to quantify the usability of a computer system from the user's subjective point of view [16]. John Brooke has been working on SUS since 1986. SUS has been frequently used to test usability and has various advantages, including: SUS is simple to apply because the outcome is a score of 0-100. SUS is simple to use and does not need complex computations. SUS is free and does not require any additional expenses, and it has been shown to be valid and reliable even with a tiny sample size. [5] [17] [16]. According on the findings of past research, the System Usability Scale (SUS) is a viable and reliable usability assessment tool. [18] [17] [19] [20]. As a result, the researchers utilized SUS to evaluate the usefulness of this myHardware chatbot.

III. Methodology

The research step begins with developing the test scenario, then choosing respondents, testing them, and summarizing the results. The test scenario describes the application to be evaluated as well as the format of a questionnaire. Furthermore, in the process of picking respondents, what is done is to identify the respondents who will analyze the myHardware chatbot application. Respondent testing is a stage in which respondents provide an assessment of the application. SUS is the evaluation instrument employed. The final step is a recapitulation of the findings derived using the SUS computation.

Usability evaluation on the myHardware Chatbot was carried out using the SUS method. SUS has 10 (ten) statements as shown in Table I.

TABLE I. SUS Instruments

No	Statement	Scale
1	I think that I will want to use this app more often	1-5
2	I found that this app, doesn't have to be this complicated	1-5
3	I think the app is easy to use	1-5
4	I think that I will need help from a technical person to be able to use this app	1-5
5	I found the various functions in this app well integrated	1-5
6	I think there are too many inconsistencies in this app	1-5
7	I imagine that most people will find it easy to learn this app very quickly	1-5
8	I find, this app is very complicated to use	1-5
9	I feel very confident to use this app	1-5
10	I need to learn a lot of things before I can start using the app	1-5

The response scale ranges from 1 to 5. 1 indicates strongly disagree, 2 indicates disagree, 3 indicates slightly agree, 4 indicates agree, and 5 indicates definitely agree. Provisions for calculating assessment outcomes using SUS:

1. the answer scale is reduced by 1 on odd statements
2. 5 minus the answer scale on even statements
3. 4 is the most positive response on a scale of 0 to 4
4. the answer scale is added up and multiplied by the number 2.5
5. The average answer is determined from the statement.

The final phase in SUS usability evaluation is to identify the end outcome. Acceptability, grade scale, and adjective rating are the three key characteristics of the grading level. Acceptability is a criteria used to judge whether or not an application is acceptable to users, with degrees ranging from not acceptable to marginal (low and high). A grade scale is a criterion for determining the level of quality with which an application is utilized. The letter grades are A, B, C, D, and E. While the adjective rating is used to judge whether or not an application is beneficial. The adjective evaluations are as follows: worst possible, bad, ok, good, excellent, and best possible.

The outcomes of the application evaluation will be determined by the recapitulation of the score in the form of the average cumulative value of the instrument. The assessment findings are presented in terms of acceptability, grade scale, and adjective rating, rather than the value of each evaluation statement. Conclusions are produced using a grading scale and average terms: A > 80.3, 74 B 80.3, 68 C 74.3, 51 D 68, and F 51.

IV. Results And Discussion

Based on the research method, the results of the research on the usability evaluation of Chatbot myHardware can be explained as follows.

A. Characteristics of Respondents

In this study, respondents were students of the UNWAHAS Informatics Engineering study program. The number of respondents who filled out the questionnaire was 220 students with details of the 2018 batch of 17 students (7.72%), the 2019 class of 40 students (18.18%), the 2020 class of 98 (44.54%), and the 2021 class of 65 students (29.54%). The illustration of respondent variation is presented in Table II.

TABLE II. Respondent Variation

Class Student	Amount	Precentage (%)
2018	17	7.73
2019	40	18.18
2020	98	44.55
2021	65	29.55
Total	**220**	**100.00**

B. Usability Value on Each Instrument

Respondents gave a score for each SUS statement. This value is the usability value. There are ten statements in SUS as a reference for evaluating. Different results will be owned by each respondent's answer on odd numbers. This is due to the different rules in calculating the answer scale. The conditions are:

1. The respondent's answer scale is reduced by 1 (one) for each answer given by the respondent to statements with odd numbers
2. 5 (five) minus the respondent's answer scale, each answer given by the respondent to the statement with an even number.
3. The results of number 1 and number 2 will get an answer scale of 0 to 4, which means 4 is a positive value
4. The average value is sought from the total statement multiplied by 2.5
5. The final result of the usability evaluation is determined by the result in number 4.

TABLE III. Respondent's Answer Recapitulation

Number of Statement	Respondent's Answer Scale					Respondent
	1	2	3	4	5	
Statement 1	6	8	51	109	46	220
Statement 2	4	41	81	66	28	220
Statement 3	1	7	41	114	57	220
Statement 4	24	89	56	40	11	220
Statement 5	2	15	50	123	30	220
Statement 6	9	43	86	62	20	220
Statement 7	1	9	50	110	50	220
Statement 8	38	94	47	33	8	220
Statement 9	1	9	67	111	32	220
Statement 10	17	55	54	61	33	220

Respondents provided an answer scale in the usability evaluation with SUS on the myHardware Chatbot with the recapitulation results as shown in Table III. Furthermore, according to the answer data, SUS number 1 and number 2 are calculated. From the results of the calculation, the data recapitulation of respondents' answers is shown in Table IV.

TABLE IV. Recapitulation of Average Value

Number of Statement	Respondent's Answer Scale					Amount	Respondent	Average Value
	1	2	3	4	5			
Statement 1	0	8	102	327	184	621	220	2.82
Statement 2	16	123	162	66	0	367	220	1.67
Statement 3	0	7	82	342	228	659	220	3.00
Statement 4	96	267	112	40	0	515	220	2.34
Statement 5	0	15	100	369	120	604	220	2.75
Statement 6	36	129	172	62	0	399	220	1.81
Statement 7	0	9	100	330	200	639	220	2.90
Statement 8	152	282	94	33	0	561	220	2.55
Statement 9	0	9	134	333	128	604	220	2.75
Statement 10	68	165	108	61	0	402	220	1.83

The respondent's answer data is the answer recapitulation data, then the calculation is carried out according to the provisions of the SUS. Based on each statement, it can be explained how the respondents' opinions regarding the myHardware Chatbot application are as follows:

a. Statement 1

This statement is to find out the extent to which respondents want to use the application regularly. A total of 6 or 2.73% of respondents gave a score of 1, then 8 or 3.64% gave a score of 2, as many as 51 or 23.18 gave a score of 3, as many as 109 or 49.55% gave a value of 4, and as many as 46 or 20.91% gave a value of 5. In the calculation of the SUS

method , obtained an average value of 2.82 from 220 respondents.

b. Statement 2

In number 2, if the respondent gives a smaller value, the better the result will be. A total of 4 or 1.82% gave a rating of 1, as many as 41 or 18.64% gave a rating of 2, as many as 81 or 36.82 gave a rating of 3, as many as 66 or 30.00% gave a score of 4, and as many as 28 or 12.73% gave a score of 5. In the calculation of the SUS method, obtained an average value of 1.67 from 220 respondents

c. Statement 3

The application is used easily or not is the focus of the 3rd statement. The value given by the respondent is 1 or 0.45% giving a rating of 1, as many as 7 or 3.18% giving a rating of 2, 41 or 18.64 giving a rating of 3, as many as 114 or 51.82% giving a score of 4, and as many as 57 or 25.91% giving a score of 5. calculation using the SUS method, obtained an average value of 3.00 from 220 respondents

d. Statement 4

Whether or not the need for help from other people is stated in questionnaire number 4. A total of 24 or 10.91% gave a rating of 1, as many as 89 or 40.45% gave a rating of 2, as many as 56 or 25.45 gave a rating of 3, as many as 40 or 18.18% gave a score of 4, and 11 or 5.00% gave a value of 5. In the calculation of the SUS method, an average value of 2.34 was obtained from 220 respondents.

e. Statement 5

In number 5 related to the usefulness of the application, it can be seen the distribution of the results of the questionnaire. A total of 2 or 0.91% gave a rating of 1, as many as 15 or 6.82% gave a rating of 2, as many as 50 or 22.73 gave a rating of 3, as many as 123 or 55.91% gave a score of 4, and 30 or 13.64% gave a score of 5. In the calculation using the SUS method, obtained an average value of 2.75 from 220 respondents

f. Statement 6

Regarding application consistency, 9 or 4.09% gave a rating of 1, as many as 43 or 19.55% gave a rating of 2, as many as 86 or 39.09% gave a rating of 3, as many as 62 or 28.18% gave a score of 4, and as many as 20 or 9.09% gave a score of 5. the calculation process using the SUS method, obtained an average value of 1.81 from 220 respondents

g. Statement 7

Regarding the ease of application, whether it is easy to learn is the result of the 7th statement. A total of 1 or 0.45% gave a rating of 1, as many as 9 or 4.09% gave a rating of 2, as many as 50 or 22.73 gave a rating of 3, as many as 110 or 50.00% gave a score of 4, and 50 or 22.73% gave a value of 5. In the calculation process using the SUS method , obtained an average value of 2.90 from 220 respondents

h. Statement 8

Regarding the complexity of using the application, it is presented in number eight. A total of 38 or 17.27% gave a rating of 1, as many as 94 or 42.73% gave a rating of 2, 47 or 21.36 gave a rating of 3, as many as 33 or 15.00% gave a score of 4, and 8 or 3.64% gave a score of 5. In the calculation process with the SUS method, obtained an average score of 2.55 from 220 respondents

i. Sstatement 9

Regarding the students' confidence or optimism in using the application, this number is presented in this number. A total of 1 or 0.45% gave a score of 1, as many as 9 or 4.09% gave a rating of 2, as many as 67 or 30.45 gave a rating of 3, as many as 111 or 50.45% gave a rating of 4, and as many as 32 or 14.55% gave a score of 5. In the calculation process using the method SUS, obtained an average score of 2.75 from 220 respondents.

j. Statement 10

In this number 17 or 7.73% gave a rating of 1, 55 or 25.00% gave a rating of 2, 54 or 24.55 gave a rating of 3, 61 or 27.73% gave a score of 4, and 33 or 15.00% gave a score of 5. In the calculation process with the SUS method, obtained an average value of 1.83 from 220 respondents.

Based on the results of the average scores from statement 1 to statement 10, there are 3 numbers with low average scores. The statement is contained in number 2 (about the complexity of the application), number 6 (about application inconsistency) and number 10 (about the need to learn before using the application). The 3 numbers with good averages are in statement 1 (about the continued use of the application), statement 3 (about ease of use), and statement 7 (about the ease of learning the application).

C. Chatbot Usability Level

The calculation of usability level showed the results of the values listed in Table V. It is known that the evaluation of the myHardware Chatbot gets a value of 61.03. From the value or score, the meaning will be explained in accordance with the provisions of SUS. Application usability has a purpose based on three levels of Acceptability, Grade scale and Adjective rating. Based on Acceptability, the myHardware Chatbot application from the aspect of being accepted by users is categorized as marginal. While the Grade scale assesses the application from the aspect of the quality level. The evaluation results show that the myHardware Chatbot application is in grade D scale. The adjective rating starts the application from the aspect that determines the usability rating, the evaluation results show that the myHardware Chatbot application is categorized as ok.

TABLE V. SUS STATEMENT RECAPITULATION

Number of Statement	Statement Average	Total (Average x 2.5)
Statement 1	2.82	7.06
Statement 2	1.67	4.17
Statement 3	3.00	7.49
Statement 4	2.34	5.85
Statement 5	2.75	6.86
Statement 6	1.81	4.53
Statement 7	2.90	7.2
Statement 8	2.55	6.38
Statement 9	2.75	6.86
Statement 10	1.83	4.57
	Total Value	**61.03**

Although there are some weaknesses (poor scores) from the usability evaluation research of the myHardware Chatbot using the SUS method, when viewed from the final SUS score, it can be concluded that the usefulness of the myHardware Chatbot is still accepted by students. The good scores given by respondents/students include, the application will be more frequent and easy to use, users do not need to need help from others, various functions have been integrated well, ordinary users will easily learn quickly, and feel very confident.

V. Conclusion

The usability evaluation of the myHardware chatbot using the SUS method concluded that the application has a usability value that is quite in accordance with the value given by the respondent for each number. Statement number 1 got an average score of 2.82, number 2 averaged 1.67, number 3 averaged 3.00, number 4 averaged 2.34, number 5 had an average score of 2.75, number 6 averaged 1.81, number 7 average score 2.90, number 8 average score 2.55, number 9 average score 2.75 and number 10 average score 1.83. There are 3 numbers with an average score of low/poor. There are also 3 numbers with a high/good average. MyHardware Chatbot application is categorized as meeting usability standards with a final score of 61.03, with a description of marginal acceptability, grade D scale and adjective rating ok. Although there are several weaknesses (poor scores), based on the final score of SUS, it can be concluded that the myHardware Chatbot is still usefully accepted by students and functions well.

Acknowledgment

The researcher would like to thank LP2M Wahid Hasyim University for providing a grant with contract number 41/LPPM-UWH/PENELITIAN/INTERDISPLINER/DIPA-UWH/2022.

References

[1] J. Weizenbaum, "ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[2] D. Dutta, "Developing an Intelligent Chat-bot Tool to assist high school students for learning general knowledge subjects," *Georg. Inst. Technol.*, p. 13, 2017.

[3] P. Bii, "Chatbot technology: A possible means of unlocking student potential to learn how to learn," *Educ. Res.*, vol. 4, no. 2, pp. 218–221, 2013.

[4] U. Ependi, T. B. Kurniawan, and F. Panjaitan, "System Usability Scale Vs Heuristic Evaluation: a Review," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 10, no. 1, pp. 65–74, 2019.

[5] I. H. N. Aprilia, P. I. Santosa, and R. Ferdiana, "Pengujian Usability Website Menggunakan System Usability Scale Website Usability Testing using System Usability Scale," *J. IPTEK-KOM*, vol. 17, no. 1, pp. 31–38, 2015.

[6] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020.

[7] C. W. Okonkwo and A. Ade-Ibijola, "Python-bot: A chatbot for teaching python programming," *Eng. Lett.*, vol. 29, no. 1, pp. 25–34, 2021.

[8] E. H. K. Wu, C. H. Lin, Y. Y. Ou, C. Z. Liu, W. K. Wang, and C. Y. Chao, "Advantages and constraints of a hybrid model K-12 E-Learning assistant chatbot," *IEEE Access*, vol. 8, pp. 77788–77801, 2020.

[9] G. Hiremath, A. Hajare, P. Bhosale, and R. Nanaware, "Chatbot for Education System," *Int. J. Adv. Res. Ideas Innov. Technol. ISSN*, vol. 4, no. 3, pp. 37–43, 2018.

[10] S. Sinha, S. Basak, Y. Dey, and A. Mondal, "An Educational Chatbot for Answering Queries BT - Emerging Technology in Modelling and Graphics," 2020, pp. 55–60.

[11] S. A. Wartman and C. Donald Combs, "Medical education must move from the information age to the age of artificial intelligence," *Acad. Med.*, vol. 93, no. 8, pp. 1107–1109, 2018.

[12] K. Crockett, A. Latham, and N. Whitton, "On predicting learning styles in conversational intelligent tutoring systems using fuzzy decision trees," *Int. J. Hum. Comput. Stud.*, vol. 97, pp. 98–115, 2017.

[13] M. P. C. Lin and D. Chang, "Enhancing post-secondary writers'writing skills with a chatbot: A mixed-method classroom study," *Educ. Technol. Soc.*, vol. 23, no. 1, pp. 78–92, 2020.

[14] D. F. Murad, E. Fernando, M. Irsan, S. A. Murad, P. M. Akhirianto, and M. H. Wijaya, "Learning support system using chatbot in 'Kejar C Package' homeschooling program," *2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019*, pp. 32–37, 2019.

[15] C. Troussas, A. Krouska, and M. Virvou, "Integrating an adjusted conversational agent into a mobile-assisted language learning application," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 2017-Novem, pp. 1153–1157, 2017.

[16] J. Brooke, "SUS : A Retrospective," *J. Usability Stud.*, vol. 8, no. 2, pp. 29–40, 2013.

[17] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean; adding an adjective rating," *J. usability Stud.*, vol. 4, no. 3, pp. 114–23, 2009.

[18] J. R. Lewis, "The System Usability Scale: Past, Present, and Future," *Int. J. Hum. Comput. Interact.*, vol. 34, no. 7, pp. 577–590, 2018.

[19] J. R. Lewis and J. Sauro, "The factor structure of the system usability scale," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5619 LNCS, pp. 94–103, 2009.

[20] Z. Sharfina and H. B. Santoso, "An Indonesian adaptation of the System Usability Scale (SUS)," *2016 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2016*, pp. 145–148, 2017.