IPTEK The Journal of Technology and Science, 33(1), 2022 (e/pISSN:2088-2033, 0853-4098)

DOI: 10.12962/j20882033.v33i1.12027 | Received 20 Jan, 2022; Revised 19 April, 2022; Accepted 19 May, 2021

ORIGINAL RESEARCH

ADAPTIVE ASSESSMENT AND GUESSING DETECTION IMPLEMENTATION

Akbar Noto Ponco Bimantoro | Umi Laili Yuhana*

Dept. of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Correspondence

*Umi Laili Yuhana, Dept of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. Email: yuhana@if.ts.ac.id

Present Address

Gedung Teknik Informatika, Jl. Teknik Kimia, Surabaya 60111, Indonesia

Abstract

Computerized adaptive testing (CAT) is a context-based adaptive assessment. However, the assessment result may not be valid because the examinee might cheat or guess the answers. Although there are many guessing detection methods, there are not many discussions about their implementation into CAT. Therefore, this paper presents an example of a modification of an existing software so the newly modified software can detect guessed answers and be able to select questions adaptively. The system can detect assuming behavior by recording the examinee's answer time. Also, the designed system can like questions adaptively by connecting Fuzzy logic, which calculates what level the question should select for the next iteration. The system is responded well by elementary and college students. A total of 56.6% felt the system was straightforward to use. The detection methods can detect guessing behavior of about 73%. However, the system's sensitivity is low if the method is forced to classify answers which answered in a long response time / general guessing. Nevertheless, when we limit the data classified within 10s response time (rapid-guessing), the method's sensitivity rises to 68.78%.

KEYWORDS:

Adaptive Assessment, Computerized Adaptive Testing, Guessing Behavior, Question Selection

1 | INTRODUCTION

Lately, technology has been widely used as a medium to support the learning process. The learning process usually consists of studies and evaluations. The evaluation itself is a process to assess knowledge, ability to understand, and attainment of the examinee's skill evaluation[1]. Indonesian elementary schools have three assessment goals: attitude, knowledge, and skills[2]. Knowledge evaluation is divided into several levels: remembering, understanding, applying, analyzing, evaluating, and creating. These levels can be assessed using several tests such as multiple-choice, written tests, oral test, and assignments. This is also applied in junior high schools and high schools, with the same assessment method in the latest curriculum[2–4]. In Indonesia, assessment is usually used to test abilities and new student admissions screening, measure learning outcomes, and as a determinant of graduation. Also, assessment can be used to determine the resource allocation effectiveness in each competency[5], select the student's learning path[6], and adjust the learning content according to the student's abilities[7].

Computerized Adaptive Testing (CAT) is one of the assessment methods that can adjust the question's difficulty based on the examinee's ability interactively[8]. The most uncomplicated CAT research uses a Rule-Based System to select questions based on the examinee's answers[9, 10]. The research that uses Item Response Theory (IRT) modeled the question's difficulty and the probability examinee answer the question correctly mathematically to select the following questions[11–13]. Other studies use fuzzy logic and bloom taxonomy to select questions[14]. While other adaptive assessment methods that use Tree usually select question scenarios from the pre-defined graph (Tree-CAT)[15], Merged Tree-CAT[16] is a computational time improvement version of Tree-CAT. Some studies use four parameters logistic (Item difficulty, item discriminant, probability, and response) on Mamdani fuzzy interference[17]. However, there aren't many CATs that use time as a parameter except Lendyuk et al.[18], who combine RBS with fuzzy rules. The study can select test difficulty based on the true-false percentage and time needed to answer a test level.

Not only question selection, but CAT can also estimate the examinees' ability. However, that ability assessment maybe not is invalid because they are cheating or guessing the answer[19]. There are many methods to detect rapid guessing, such as time threshold. It can be defined using manual inspection of time distribution[20–23], question features such as several characters[24, 25], common k-seconds threshold[26], and mixture model[20]. These detection methods should be able to improve the accuracy of CAT's adaptive assessment since Wise claims rapid-guessed answers should be excluded for scoring[27]. However, there aren't many discussions about its implementation on adaptive assessment (CAT), even though rapid-guessed answers can threaten the validity of the assessment.

Therefore, this study aims to present a software design modification so that the existing system can select questions adaptively based on the examinee's ability and detect rapid guessing. This study proposed a modified CAT that can count the examinee's guessing behavior. We modify ExamITS, known as the TOEFL test platform, by logging several metadata that can be used to detect rapid guessing behavior. This paper discusses previous research about CATs and detection in Section 2. At the same time, the proposed system design and method will be presented in Section 3. Section 4 shows the results and its discussion. ConclusionThe conclusion of this study will be described in Section 5.

2 | PREVIOUS RESEARCHES

From the literature study conducted, we found several CATs architectures. We group the CATs into four groups based on their methods, i.e., Rule-based system, Item response theory, Fuzzy logic, and Tree. The summary of our literature study is described in Tabel 4. From the table, we can see that most research is about software design. Also, most of these methods do not count rapid-guessing detection in their system.

TABLE 1 Previous research about CATs.

Researcher	Method	Focus	Key Points
Louhab et al.[9]	RBS	Software Design	Mobile-based CAT system based on examinee's profile based
Pan and Lin[10]	RBS	Software Design	IEEE-standardized MVC CAT
Vega et al.[11]	IRT	Software Design	Maximum Information IRT on Higher Education
Uto and Ueno[12]	IRT	Improvement	GRM modification of IRT
Cui et al.[13]	IRT	Implementation	Implement IRT on Squirrel AI
Chrysafiadi et al.[14]	Fuzzy	Fuzzy + Bloom	Combining fuzzy and cognitive theories
Lendyuk et al.[18]	Fuzzy	Select	Fuzzy rules for learning path construction
Ridwan et al.[17]	Fuzzy	Fuzzy 4PL Software Design	Ability estimation
Delgado-Gómez et al.[15]	Tree	Tree-CAT	Pre-generated Tree for adaptive tests
Rodríguez-Cuadrado et al.[16]	Tree	Tree-CAT time improvement	Improvement of Tree-CAT

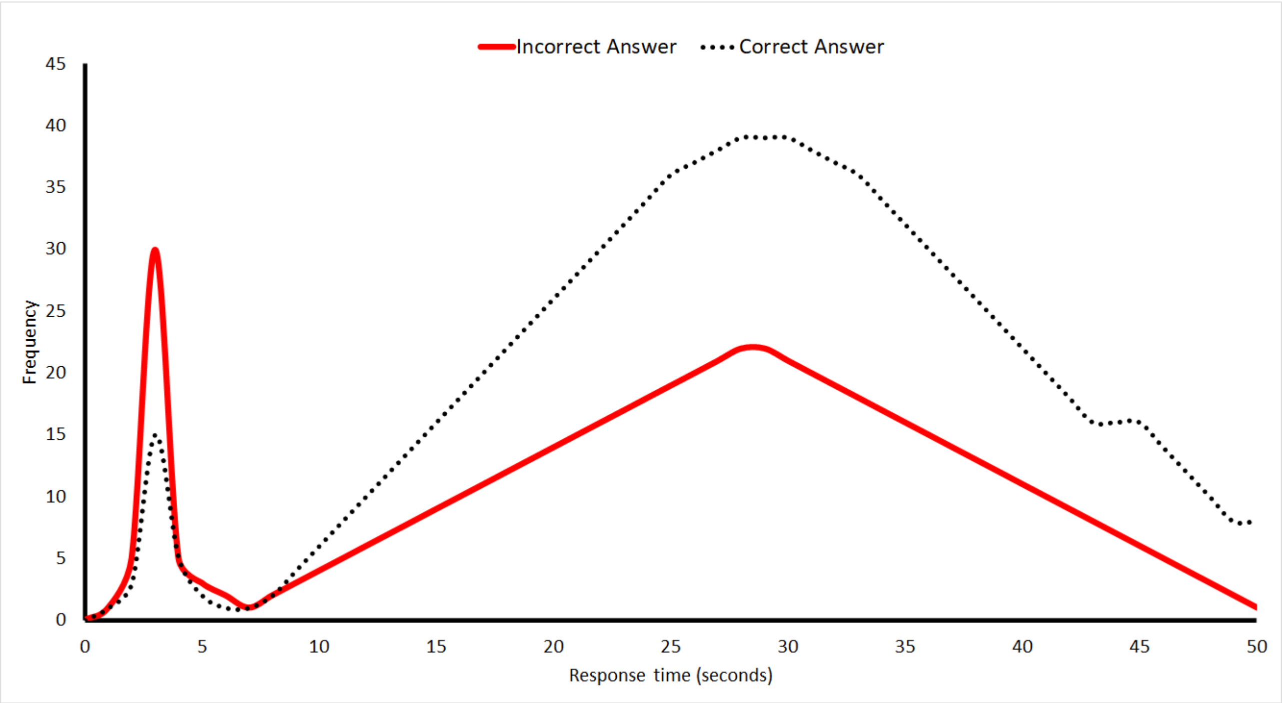


FIGURE 1 RT Distribution of answers.

Fuzzy logic was initially proposed by Zadeh[28]. In its development, Mamdani introduced a new inference model, which is currently known as Mamdani Inference System[29] or min-max fuzzy. This method was used by Ridwan et al. on the CAT 4PL (item difficulty, item discrimination, probability, and examinee's answers) Ridwan et al.[17]. To determine the probability value of the i-th question, equation (1) is used where ϑ is the student's ability to be estimated, c is the guessing factor which was 0.25 constant value, ai is i-th item discrimination, and bi is i-th item difficulty. Item difficulty is calculated by equation (2). While item discriminant is calculated by equation (3), where u is the highest 50% participant (upper bound) and l is the lowest 50% (lower bound).

$$Probability_i(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}} \tag{1}$$

$$b = \frac{numberof\,incorrectanswers}{n} \tag{2}$$

$$a = \left(\frac{number\ of\ incorrect\ answers}{n_u}\right)_u - \left(\frac{number\ of\ incorrect\ answer}{n_l}\right)_s \tag{3}$$

There are several methods to detect rapid guessing behavior. The most commonly used is time threshold, as rapid guessing usually happen when examinee answer the question faster than the time they need to read and solve the question. Rapid guessing detection with response time (RT) threshold was done by Schnipke et al. using RT distribution on 17.415 GRE-CBT participants Schnipke[30]. The RT distribution mapped into a graph shown in Figure 1 is then analyzed. Inspection is carried out by manually selecting a time threshold that separates the two distributions of correct and incorrect answers. From the figure, it can be seen that the frequency of incorrect answers is very high in the first 8 seconds. Also, it can be seen that the frequency of correct answers starts increasing after 8 seconds. Thus, the threshold time in this scenario is 8 seconds. Therefore, the answers in the first distribution (1) are categorized as guessing behavior, and the second distribution is then classified as solution behavior.

In contrast to k-seconds which use a 3-5 seconds threshold regardless of the context of a question Wise[27], the Surface feature uses the question's characteristics such as the question's length, the existence of an image or table, and the subjects being tested.

TABLE 2 Threshold Surface Feature.

Criteria	Threshold
Mathematical Problem / Spatial Reasoning 1	5 seconds
< 200 characters 2	3 seconds
200 – 1000 characters 3	5 seconds
> 1000 characters 4	10 seconds

Thus, the features of a question adjust the time threshold. Each of the proposed thresholds is described in Table 2. Slim et al. use the context of a question and the subject of the test 1,2,3,4[25]. At the same time, Wise and Kong[24] use the amount of character and the existence of an image or table in questions 2,3,4. However, both studies did not explicitly explain the evaluation results of their methods.

Unlike the explained method that uses a time threshold, Lin et al. processed the logit ability of an examinee (e) and item difficulty (b) to detect pseudo guessing. Lin et al. claim that if there is a big difference between examinee logit ability and question's difficulty, then the answer is pseudo-guessed. Lin et al. classify answers as pseudo-guessing if b – e ≤ 2. These detection results were then used to refine the existing Rasch model by deleting all pseudo answers from the dataset[31]. From the experiment result on elementary students, Lin claims that the ability assessment precision of high-ability students increases.

3 | MATERIAL AND METHOD

3.1 | Material

Figure 2 shows the technical design of ExamITS modification into a system that can select questions based on students' ability. The architecture consists of the Laravel-MVC application connected with Flask by REST API and MySQL database. The database stores various data such as questions, question metadata, and user profiles. The REST API and AJAX data from JSON format simplify data exchange between systems. This application is then installed on an elastic cloud to reduce the server's upgrade complexity.

In developing CAT with rapid guessing detection, it is necessary to log metadata or examine activity history. The recorded metadata is adjusted to the type of test used. In the classical exam, the metadata recorded is the examinee's response time (RT). The response time starts when the examinee opens a question and ends when they answer a question. Meanwhile, in adaptive assessment (CAT), additional metadata that needs to be recorded are RT and question's index, such as item difficulty and item discriminant. The recorded metadata is then used both in CAT and guessing detection.

The user interface design in this paper is shown in Figure 3-Figure 7. The interface is made responsive, considering that not all participants have computers or laptops. Most of the elementary students in this study use smartphones. Thus, a web design with a responsive interface that suits any device such as laptops, tablets, and smartphones is a must. Figure 3 presents a dashboard that shows the personal information of the examinee and a list of tests that can be taken. Figure 4 and Figure 5 show the layout of a question for classical test and adaptive assessment. The system shows an alert if the examinee is detected using another application, as shown in Figure 6. While Figure 7 shows an example of the interface's responsiveness.

$$Probability_i(\theta) = (1 - c_i)\frac{e^{1.7a_i(\theta_{i-1}-b_i)}}{1 + e^{1.7a_i(\theta_{i-1}-b_i)}}) \tag{4}$$

$$c = \begin{cases} 1, & RT < threshold \\ 0, & RT \geq threshold \end{cases} \tag{5}$$

$$c = \begin{cases} 1, & (b - e) \geq 2 \\ 0, & RT < 2 \end{cases} \tag{6}$$

The statechart diagram of the system is shown in Figure 8. The figure shows that the CAT will stop selecting questions in the 6-th iteration when there is no estimated ability between it and the previous (i¬-1) iteration or when the exam time is up. As for

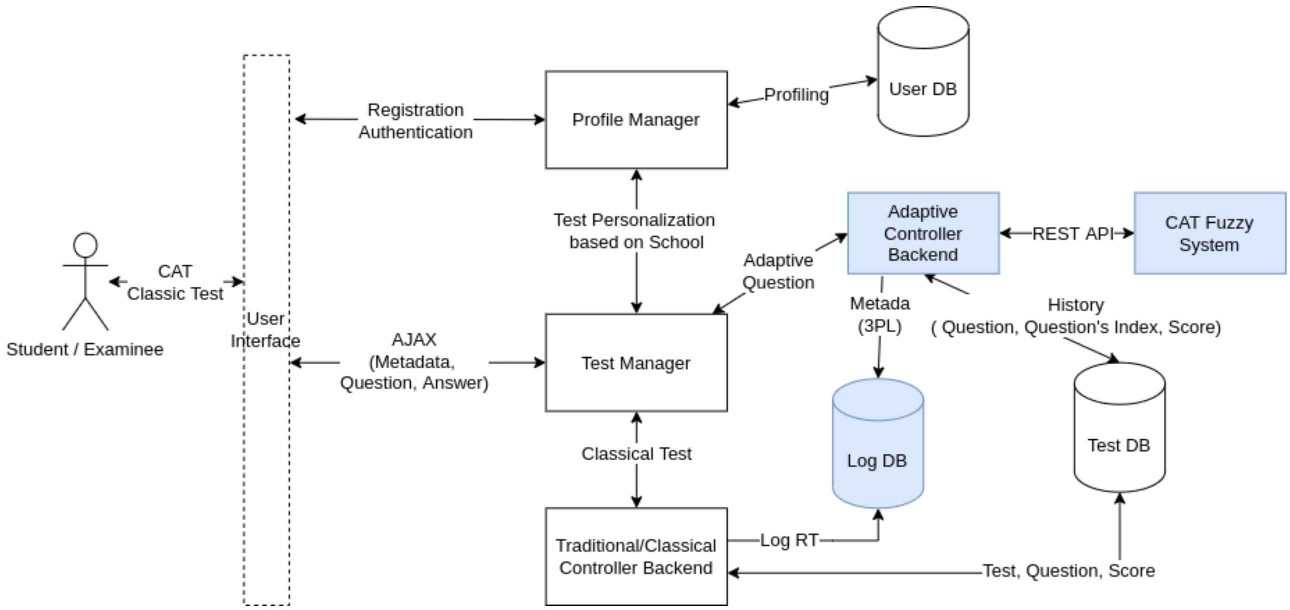


FIGURE 2 The technical design of the proposed system.

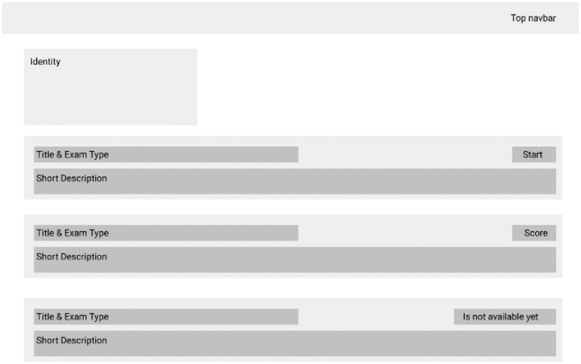


FIGURE 3 The dashboard.

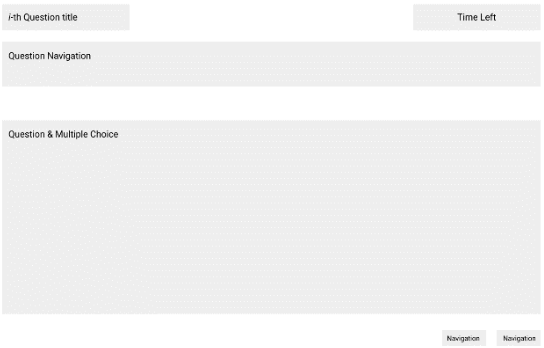


FIGURE 4 The classical exam layout.

the first iteration, the selected questions will be medium-leveled. Also, the time examinee needs to answer the data is recorded when they open the question and send it to the server with their answers. Then, the system needs to classify the answer category to estimate the next question's difficulty.

4 | RESULTS AND DISCUSSION

The system proposed in this study consists of 4 main parts: user interface, personalized exam based on the examinee's school, classical exam, and CAT. The system is tested on elementary students, bachelor students, and magister students with 111 participants. The subject of the tests is mathematical problems, software engineering, and software management. Test difficulty and subject are adjusted with the participant's education level. Also, the test is carried out with several guessing detection methods, namely RT Distribution, Surface Features, K-Seconds, and Ability Logit pseudo guessing. From the data collected, 24.9% of answers are detected as guessing. The accuracy of each method is described in Table 4. From the table, pseudo-guessing detection has the worst results. It occurred because the difference between the question's difficulty and the examinee's ability is too far apart. Most of the answers are detected as guessing behavior. Thus, resulting in a high result of false-negative and low accuracy.

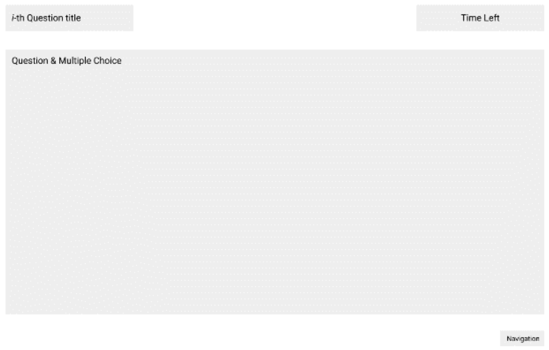
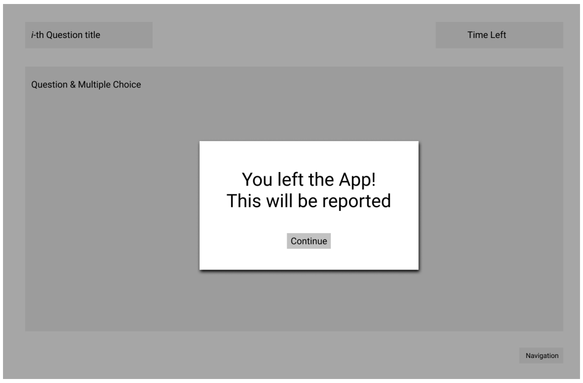
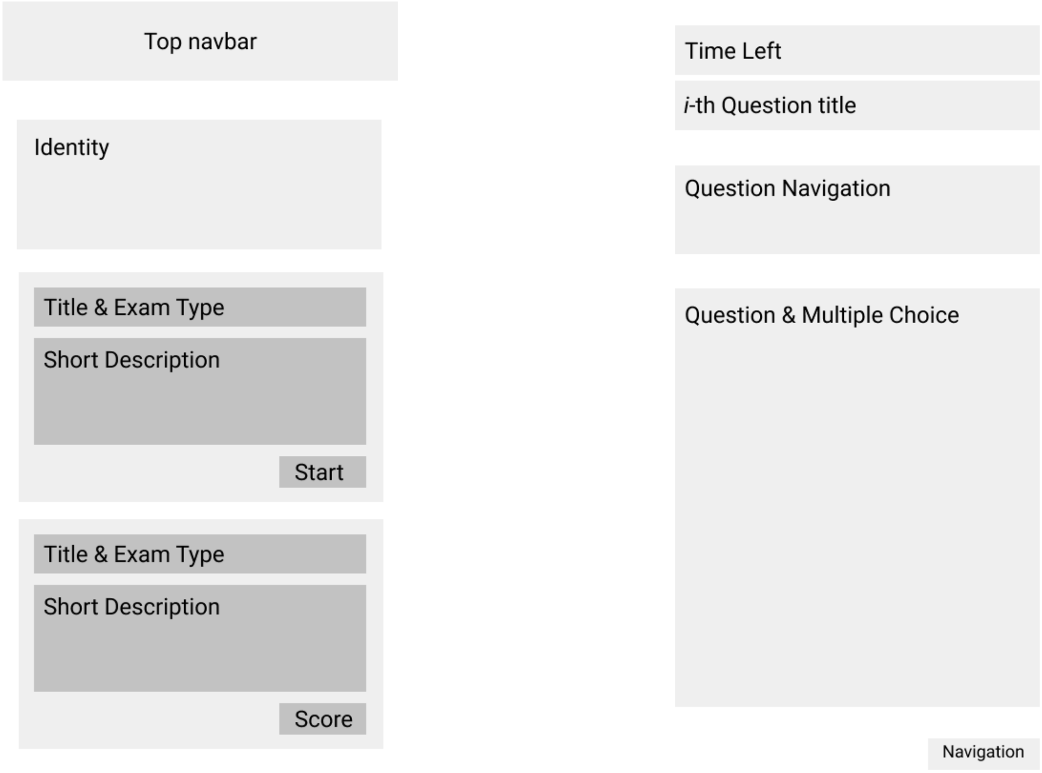


FIGURE 5 The adaptive assessment layout.



You left the App!
This will be reported

FIGURE 6 The possible cheating detection.



(a)

(b)

FIGURE 7 The responsive interface of the dashboard and exam layout.

Meanwhile, Table 4 shows the confusion matrix of time-based methods on data whose response time is below 10s. The purpose of this limitation is to show how the method detects rapid-guessing. The table presents that although the accuracy drops, the sensitivity increases. This means the system sensitivity to detect rapid-guessing is better. Also, we can infer from Table 4 that the high accuracy and low sensitivity results are due to the high value of true negative–solution behavior. This happened because the data classes were imbalanced. Also, most of the answer's response times are above 10 seconds for both guessing and solution behavior. Therefore, the detection methods produce many false-negative values – guessing behavior detected as a solution – since its time threshold maxed at 10 seconds. Nevertheless, we can infer from the table that RT Distribution is the best detection method to detect guessing behavior.

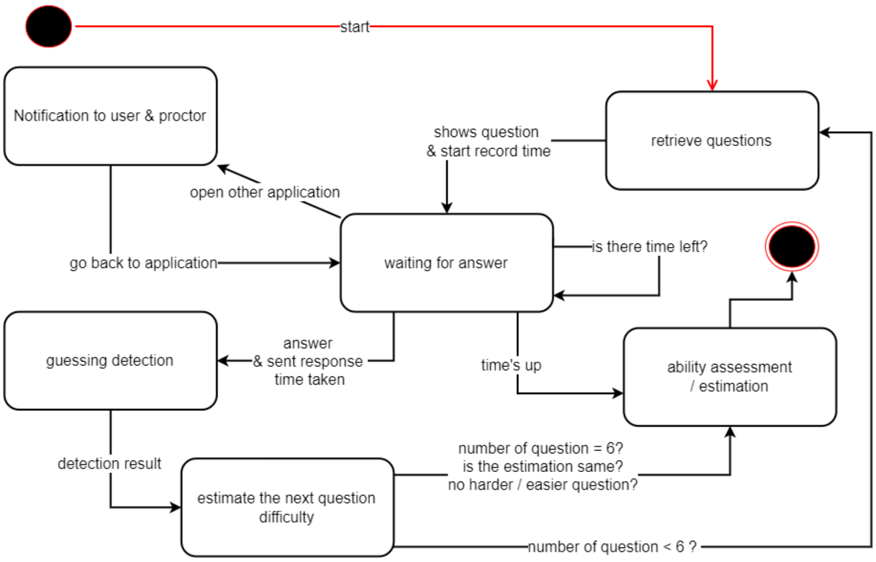


FIGURE 8 The statechart diagram of the Proposed System.

TABLE 3 The accuracy of each detection methods of overall data.

Method	Accuracy	Sensitivity	Specificity	Precision	F1-Score
RT Distribution	73,39%	12,62%	93,13%	37,41%	17,87%
K-Seconds	73,07%	3,29%	95,90%	20,89%	5,69%
Surface features	72,78%	8,00%	93,99%	30,35%	12,66%
Ability Logit	7,70%	0,00%	8,95%	0,00%	0,00%
Modified K-Seconds	73,54%	2,11%	96,91%	18,36%	3,79%

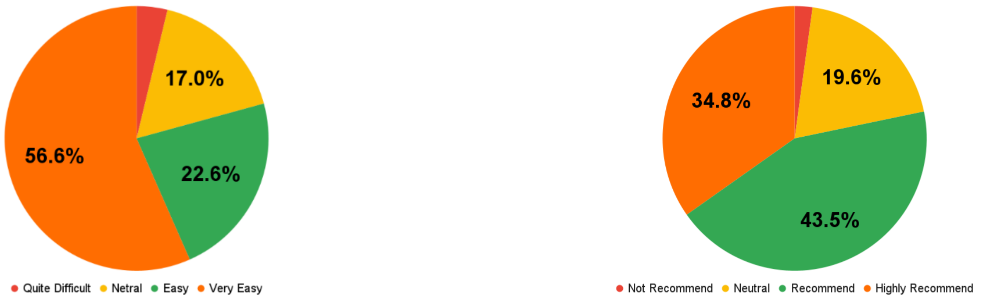
TABLE 4 The accuracy of data which answer's response time below 10 seconds (time-based rapid guessing).

Method	Accuracy	Sensitivity	Specificity	Precision	F1-Score
RT Distribution	45.54%	68.78%	35.85%	30.89%	42.64%
K-Seconds	56.42%	32.29%	66.54%	28.80%	30.44%
Surface features	53.51%	56.50%	52.26%	33.16%	41.79%
Modified K-Seconds	61.46%	25.11%	76.69%	31.11%	27.79%

Meantime, the classification of examinee's answers with a response time below 10 seconds shows a lower accuracy but a higher sensitivity. This higher sensitivity result is due to lower false-negative values. While a higher value of false-positive values causes low accuracy – solution behavior is detected as guessing. This happened for several reasons, such as the question needing several seconds or revisiting (re-answering the previous question). The results of questionnaires regarding the system developed are shown in Figure 9. Based on the examinee's feedback, 56.6% stated that the system was straightforward, and only 3.% felt the system was challenging to use. While from Figure 9b, only 2.2% of participants did not recommend the system to be used as an exam platform in their schools/institutions. Nevertheless, most of the participants agreed and recommended the system be applied. From the questionnaire results, it can be concluded that the system is straightforward. Even the examiners recommend the system to be used in their schools. Furthermore, the cheating detection alert received a good response.

5 | CONCLUSION

In this study, we modified the classical test system so that it can detect guessing behavior and assess the examinee's ability adaptively (CAT). 56.6% of participants stated that the designed system was straightforward to use from the data collected. Only 2.2% did not recommend this system to their schools. Even the examinees highly recommend the use of the system. We also show that the system can not detect general guessing, although it has decent accuracy. Nevertheless, the system shows a higher



(a) Ease of use of the system

(b) Examine feedback on the system compared to their existing platform.

FIGURE 9 The results of the examinee's questionnaire of the developed system.

sensitivity while detecting rapid-guessing, i.e., 68.78%. The guessing detection methods can be implemented to the system by adding response time (RT) logs and question's difficulty. The guessing results are then inserted into CAT's guessing factor. The CAT can easily be added to an existing system by connecting it to the fuzzy system through REST API.

In this study, we modify the system so it can assess and select questions adaptively. Furthermore, we tried to count guessing behavior into CATs. We found that the system can detect rapid-guessing pretty well. However, there are several other examinee behaviors, such as hesitation, where examinee hesitate whether their answer is true or not. Therefore, we'd like to count the examinee's different behavioral patterns during a test into CATs and investigate how they affect CAT's performance and how to implement its detection.

ACKNOWLEDGMENT

This work was supported by the Institut Teknologi Sepuluh Nopember under Grant No. 1149/PKS/ITS/2021.

CREDIT

Akbar Noto Ponco Bimantoro: Data Curation, Writing - Original Draft, Validation, and Investigation. **Umi Laili Yuhana:** Conceptualization, Methodology, Formal Analysis, Writing - Review & Editing, and Supervison.

References

1. Kennedy KJ, Lee JCK. The Changing Role of Schools in Asian Societies. Routledge; 2007. https://www.taylorfrancis.com/books/9781134127306.

2. Kemendikbud. Panduan Penilaian Untuk Sekolah Dasar (SD). Kementrian Pendidikan dan Kebudayaan; 2016.

3. Kemendikbud. Panduan Penilaian oleh Pendidik dan Satuan Pendidikan Atas. Kementrian Pendidikan dan Kebudayaan; 2017.

4. Kemendikbud. Panduan Penilaian Untuk SMA. Kementrian Pendidikan dan Kebudayaan; 2015.

5. Peng SS, Lee JCK. Educational evaluation in East Asia: Emerging issues and challenges. Nova Science Publishers, UK; 2011.

6. Hwang GJ. A conceptual map model for developing intelligent tutoring systems. Computers & Education 2003 4;40:217–235. https://www.sciencedirect.com/science/article/abs/pii/S0360131502001215https://linkinghub.elsevier.com/retrieve/pii/S0360131502001215.

7. Hwang GJ, Sung HY, Chang SC, Huang XC. A fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. Computers and Education: Artificial Intelligence 2020;1:100003. https://doi.org/10.1016/j.caeai.2020.100003https://linkinghub.elsevier.com/retrieve/pii/S2666920X20300035.

8. Tseng WT. Measuring English vocabulary size via computerized adaptive testing. Computers & Education 2016 6;97:69–85. https://linkinghub.elsevier.com/retrieve/pii/S0360131516300501.

9. Louhab FE, Bahnasse A, Talea M. Towards an Adaptive Formative Assessment in Context-Aware Mobile Learning. Procedia Computer Science 2018;135:441–448. https://doi.org/10.1016/j.procs.2018.08.195.

10. Pan CC, Lin CC. Designing and implementing a computerized adaptive testing system with an MVC framework: A case study of the IEEE floating-point standard. Proceedings of 4th IEEE International Conference on Applied System Innovation 2018, ICASI 2018 2018;p. 609–612.

11. Vega YLP, Bolanos JCG, Nieto GMF, Baldiris SM. Application of item response theory (IRT) for the generation of adaptive assessments in an introductory course on object-oriented programming. Proceedings - Frontiers in Education Conference, FIE 2012;p. 0–3.

12. Uto M, Ueno M. Item Response Theory for Peer Assessment. IEEE Transactions on Learning Technologies 2016 4;9:157–170. http://ieeexplore.ieee.org/document/7243342/.

13. Cui W, Xue Z, Shen J, Sun G, Li J. The Item Response Theory Model for an AI-based Adaptive Learning System. In: 2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET) IEEE; 2019. p. 1–6. https://ieeexplore.ieee.org/document/8937383/.

14. Chrysafiadi K, Troussas C, Virvou M. Combination of fuzzy and cognitive theories for adaptive e-assessment. Expert Systems with Applications 2020;161:113614. https://doi.org/10.1016/j.eswa.2020.113614.

15. Delgado-Gómez D, Laria JC, Ruiz-Hernández D. Computerized adaptive test and decision trees: A unifying approach. Expert Systems with Applications 2019;117:358–366.

16. Rodríguez-Cuadrado J, Delgado-Gómez D, Laria JC, Rodríguez-Cuadrado S. Merged Tree-CAT: A fast method for building precise computerized adaptive tests based on decision trees. Expert Systems with Applications 2020 4;143:113066. https://linkinghub.elsevier.com/retrieve/pii/S0957417419307833.

17. Ridwan W, Wiranto I, Dako RDR. Ability estimation in computerized adaptive test using Mamdani Fuzzy Inference System. IOP Conference Series: Materials Science and Engineering 2020 5;850:012004. https://iopscience.iop.org/article/10.1088/1757-899X/850/1/012004.

18. Lendyuk T, Sachenko S, Rippa S, Sapojnyk G. Fuzzy rules for tests complexity changing for individual learning path construction. In: 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 2 IEEE; 2015. p. 945–948. http://ieeexplore.ieee.org/document/7341443/.

19. Wise SL. An investigation of the differential effort received by items on a low-stakes computer-based test. Applied Measurement in Education 2006;19:95–114.

20. Pastor DA, Ong TQ, Strickman SN. Patterns of Solution Behavior across Items in Low-Stakes Assessments. Educational Assessment 2019 7;24:189–212.

21. Demars CE. Changes in Rapid-Guessing Behavior Over a Series of Assessments. Educational Assessment 2007 4;12:23–45.

22. DeMars CE, Wise SL. Can differential rapid-guessing behavior lead to differential item functioning? International Journal of Testing 2010;10:207–229.

23. Setzer JC, Wise SL, van den Heuvel JR, Ling G. An Investigation of Examinee Test-Taking Effort on a Large-Scale Assessment. Applied Measurement in Education 2013 1;26:34–49. http://www.tandfonline.com/doi/abs/10.1080/08957347.2013.

739453.

24. Wise SL, Kong X. Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. Applied Measurement in Education 2005 4;18:163–183. http://www.tandfonline.com/doi/abs/10.1207/s15324818ame1802_2.

25. Silm G, Must O, Täht K. Test-taking effort as a predictor of performance in low-stakes tests. Trames Journal of the Humanities and Social Sciences 2013;17:433. http://www.kirj.ee/?id=23010&tpl=1061&c_tpl=1064.

26. Wise SL, Ma L, Kingsbury GG, Hauser C. An investigation of the relationship between time of testing and test-taking effort. National Council on Measurement in Education 2010;p. 1–18.

27. Wise SL. Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications. Educational Measurement: Issues and Practice 2017 12;36:52–61. https://onlinelibrary.wiley.com/doi/10.1111/emip.12165.

28. Zadeh LA. Fuzzy sets. Information and Control 1965 6;8:338–353. https://linkinghub.elsevier.com/retrieve/pii/S001999586590241X.

29. Mamdani EH. Application of Fuzzy Algorithms for Control of Simple Dynamic Plant. Proceedings of the Institution of Electrical Engineers 1974;121:1585–1588.

30. Schnipke DL. Assessing Speededness in Computer-based tests using item response times. Dissertation 1995;.

31. Lin CK. Effects of Removing Responses With Likely Random Guessing Under Rasch Measurement on a Multiple-Choice Language Proficiency Test. Language Assessment Quarterly 2018 10;15:406–422. https://www.tandfonline.com/doi/full/10.1080/15434303.2018.1534237.

How to cite this article: Bimantoro A.N.P., Yuhana U.L. (2022), Adaptive Assessment and Guessing Detection Implementation, *IPTEK The Journal of Technology and Science*, *33(1):26-35*.