A New Method for Setting Response-Time Threshold to Detect Test Takers' Rapid Guessing Behavior

Deni Hadiana*
Centre for Assessment and Learning
Kementrian Pendidikan dan Kebudayaan
Jakarta, Indonesia
*deni.hadiana@kemdikbud.go.id

Bahrul Hayat
State Islamic University Syarif Hidayatullah
Jakarta, Indonesia
bahrulhayat@uinjkt.ac.id

Burhanuddin Tola
Jakarta State University
Jakarta, Indonesia
burhanuddintola@unj.ac.id

***Abstract*—Anomalous behaviour such as rapid guessing in item response can affect test validity and the utility of decisions made with test scores, making it critical to identify this anomalous behaviour. The author proposes a new nonparametric method for setting a response-time threshold for flagging rapid guessing behaviour of item responses. Using data from the computer base national exam of Junior High School in Indonesia, the author analysis and compares the effectiveness of the new and existing method for detecting rapid guessing behaviour based on response time. Based on ROC curve analysis, the author concludes a new method for setting a response-time threshold is more effective than existing methods.**

Keywords—threshold, response time, rapid guessing, effectiveness

I. INTRODUCTION

Computer-based tests provide information about responses and response time [1]. Both of this information are very useful for identifying the anomalous behaviour of each test taker. The anomaly behaviour as analysed for its relationship with item parameters, responses, response time, and the ability of the test taker will obtain information on the quality of the test results that have an impact on the use of test results. Research to determine anomalous behaviour detection method is very urgent because this method can improve the test taker's ability score estimation to be more accurate.

The method of detecting anomalous behaviour based on response time has been developed by Wise et al [2]. This method uses the determination of the threshold with any criteria. Setting the threshold response time is a crucial step in differentiating rapid guess behaviour and solution behaviour [2]. In this study, setting four methods for the threshold were applied, first, the visualization identification method (VI) developed by Wise [3], and three methods for setting the threshold response time, namely the type of items (TS), 2.5 SD outliers, and 2,75SD outliers that were developing in this study. The four threshold methods will be applied to detect the rapid guess behaviour.

Detecting for the rapid guessing behaviour in this study focuses on the test taker responds to items with a time that is faster than the threshold for each item. The research aims to analyse the most effective method in detecting the behaviour of the test-takers' rapid guessing by setting the threshold response time, namely the VI method, TS method, 2.5SD outliers method, and 2.75SD outliers method.

II. RESPONSE TIME, THRESHOLD, RAPID GUESSING

Response time is the time taken by a test taker to read and answer a test item [4,5]. Response time is calculating from the test taker starts clicking the item to be processed for the first time, answered, until the click to the other item test, including if the test taker chooses to return to processing the same item after completing another item test. The response time is the total time when the test taker starts working on the item test for the first time until the number of attempts, and so on until the test time runs out.

According to Entink [6], studies that correlate response time to item parameters and test-takers parameters are useful. For example, speed analysis in tests, detecting cheating, and differentiating solution behaviour from rapid guessing behaviour, and designing tests like what Bridgeman and Cline have done in 2004, Wise and Kong in 2005, Breithaupt, Chuah, and Zang in 2007, and van der Linden in 2008. The information of response time and responses for each item on computer-based tests has the potential resulting in better anomaly behaviour detection methods [7]. The causes of anomalous responses in testing include behaviours such as cheating, carelessness, low motivation, random responding, creative responses, and guessing [8-13].

Solution behaviour occurs when test-takers respond to items carefully, considering each part of the item test before determining answer choices [14]. Wise and Kong believed that test-takers who made hard efforts in answering the items would show solution behaviour [15]. On the other hand, test-takers who carelessly processed items showed rapid guessing behaviour. Wise and Kong argue test-takers who process an item using a reasonable time to respond to it show solution behaviour by working hard [15]. The more often test-takers show solution behaviour use longer it takes to respond to test items. Rapid guessing behaviour occurs when test-takers respond quickly without processing and carefully considering each item of the test [15].

Appropriately distinguishing between solution behaviour and rapid guess behaviour is very crucial when we are going to correctly recover item parameters and test-takers, thereby ensuring the validity of the conclusions and the validity of the items. Wise and Kong [15] proposed a Response Time Effort (RTE) that begins by calculating the SBij with the following equation.

$$SB_{ij} = \begin{cases} 1 \ if \ RT_{ij} \geq T_i, \\ 0 \quad otherwise \end{cases} \qquad (1)$$

SBij is the solution behaviour of test taker j in item *i*, Ti is the threshold between the time of the rapid-guessing behaviour and the solution behaviour in item i, RTij is the response time of test taker j in item i. Then the RTE is calculated using the following equation.

$$RTE_j = \frac{\sum SB_{ij}}{k} \qquad (2)$$

K denotes the number of items in a test form.

RTE scores range from 0 to 1 and reflect the proportion of items for test-takers who behaved solutions. The RTE score range is getting closer to 1 indicate the test-takers are increasingly solutions behaviour. The challenge in calculating RTE lies in determining the threshold (T) accurately. The threshold response time is a time limit used to differentiate solution behaviour and rapid guess behaviour. Wise et al [2] use the same threshold for all items. Wise and Kong [15] determine threshold response time based on the number of characters, item test that has less than 200 of character are 3 seconds for threshold, items that have between 200 to 1000 of characters are 5 seconds of the threshold time, and items with have greater than 1000 characters are 10 seconds. Wise [2,3] used the same threshold method for all items and method VI on the response time-frequency distribution graph. Kong et al [2] used a two-state mixture model wherein each point two conditions occur rapid guess and solution behaviour. Lee and Jia [7] used the VI method by considering the responses of test-takers, Wise and Ma [16] used the average percentage of response time for each item in determining the time threshold. Guo et al set the threshold response time using the VI method by considering the difficulty level of the item [17]. Demars [18] and Setzer et al [19] concluded that the VI method in setting the threshold is most favourable.

In this study, besides using the VI method for setting threshold response time. The setting for the threshold response time considers the type of item (TS) that has taken into account the characteristics of the test item as the number of words, the presence of image stimuli, tables and illustrations, calculation, the level of difficulty of the items. The method of setting the threshold in this study is also determined based on the data distribution pattern when doing the test so that the outlier response time data calculated based on 2.5 standardized values (2.5 SD) and 2.75 standardized values (2.75SD) by making the median value as a central tendency. The median is using as the central tendency value because of the asymmetrical pattern of response time distribution. The median has been used as a central tendency value by other researchers in analysing response times. In 2000, Tsaousis et al [20] used multiple regressions to determine the relationship between the median response time and differentiation, difficulty level, presence of images, number of words, and item types. According to Whelan [21], many researchers use the median as a central tendency in the analysis of response time. In the data that varies between responses, the median is robust than other methods, let alone a large number of samples [21]. Sahin and Colvin [22] and Michaelides et al [23] used the median in analysing the bimodal distribution pattern of response time. Artner [24] uses the median when conducting a comparison simulation study of a person fit in the Rasch model. Rousselet and Wilcox [25] concluded that the median response time is more appropriate to use than average as a value for central tendency. According to Rousselet and Wilcox [25], bias in the median will not occur if the sampling technique takes into account the diversity of the sample distribution. The median is more appropriate to use than the average when comparing the asymmetric slope distribution. This finding underlies the researchers to use the median as a parameter of central tendency in analysing response times. In the context of determining the threshold time based on the central tendency value, Wise and Ma [16] determined the time threshold considering the data distribution pattern and using the average response time of each item as the central tendency value.

III. Research Methods

The sample in this quantitative research was test-takers of computer base national exam junior high school for the 2017/2018 academic year in natural science subjects from DKI Jakarta with a total of 732 samples. The natural science test consists of 40 multiple choice items with four options. Setting the threshold to determine the solution behaviour and rapid guessing use methods VI, 2,5SD, 2,75SD, and TS. The ROC curve analysis is applied to conclude the effectiveness methods.

IV. Results and Discussion

In method VI, the highest threshold occurs in item 21, which is 28 seconds, and the lowest item threshold is in number 36, which is 5 seconds. When we compare this threshold with the characteristics of large-scale items such as the national exam, most items measure the cognitive level above the ability to remember, which is 90% or 36 items. Meanwhile, the item test that measures the ability to remember is the 10% or four items test. The resulting threshold time and the variation with this method are too low. Method VI is very dependent on accuracy in identifying the frequency distribution graph of the response time, thus causing the subjective methods. Method VI is suitable for items that have a bimodal distribution pattern. While, in this study, not all items test has a bimodal distribution pattern. This phenomenon often occurs in analysing response time, such as reported by Sahin and Colvin [21]. They found one item with a bimodal distribution pattern before the median response time [23]. Although the VI method is most widely used [18,19], the fundamental problem with this method is related to the bimodal distribution that is rarely founding in many items [26].

Also, the VI method does not consider item characteristics such as complexity and cognitive level. Meanwhile, many studies have confirmed that item complexity as the presence of pictures and tables affects response time. That phenomenon has founded in this study. Zenisky and Baldwin [27] concluded that the presence of images, an increase in difficulty, and an increase in the number of words increase response time generally. Hect, Siegle, and Weirich [28] found that the presence of images in items increased the response time of 12 seconds. Counted PG items generally have a longer response time than simple and complex PG [27,28].

The method TS is based on the cognitive level and the complexity of the items. Therefore, to decide the cognitive-process and complexity of the item test, expert judgment is needed by experts who have the knowledge, understanding, and experience of item development. The results of expert judgment concluded that all items have an RVI value ≥ of 0.30. This RVI value means that the items can be said to be valid, meaning that the agreement of experts in determining the cognitive level and types of validity items with the validity range tended to be medium and high [29]. Even items 25 and 21, all experts agree to conclude that these two items are the cognitive level of understanding.

The third and fourth threshold setting methods are done based on the analysis of the data distribution pattern of the total time the test-takers takes to work on the test package. The methods commonly used in analysing response time are central tendency parameters such as mean and dispersion parameters such as standard deviation. The mean difference was then tested by ANOVA using a hypothesis. But this method, according to Wilcox [30] and Ratcliff [31], can reduce the power and fail to detect the real difference in the mean due to the effect of positive slope and response time data outliers. This method is commonly used [22]. To maximize the results based on this method, researchers usually delete outlier data, transform the data, or keep using outlier data but choose parameters that are not too sensitive to outlier data. Thus, it is clear that the mean as a parameter of central tendency cannot reflect response time data that is skew to the right or the left. Many researchers used the median as a parameter of central tendency in the analysis of response time for each item. Because of the data that varies widely between responses, the median is more robust than other methods [22], especially for the large number of respondents [22]. Artner [24] used the median response time for a comparison simulation study of a person fit in the Rasch model and used the median in analysing the ROC curve. Zenisky and Baldwin [27] used the median response time for test development and validation. Smith [27] has analysed the relationship between the median response time and the discriminating power index, difficulty level, and number of words in items. Rousselet and Wilcox [25] concluded that the median response time is more appropriate to use than average as a value for central tendency. According to Rousselet and Wilcox [25], the bootstrap percentile technique can correct the bias of measurement caused by the median. It will not occur if the sampling technique takes into account the diversity of the sample distribution. The median is more appropriate to use than the mean when comparing the slope distribution.

According to Becker et al [32], response time has positively correlated with items that require control over cognitive processing as items on national exams. Becker et al also found that the items are automatically processed immediately and easy to find logically, generally have a faster response time as happened in the number of items of 11, 13, 16, and 18 [32]. Usually, the increased complexity is linear by an increase in response time [32].

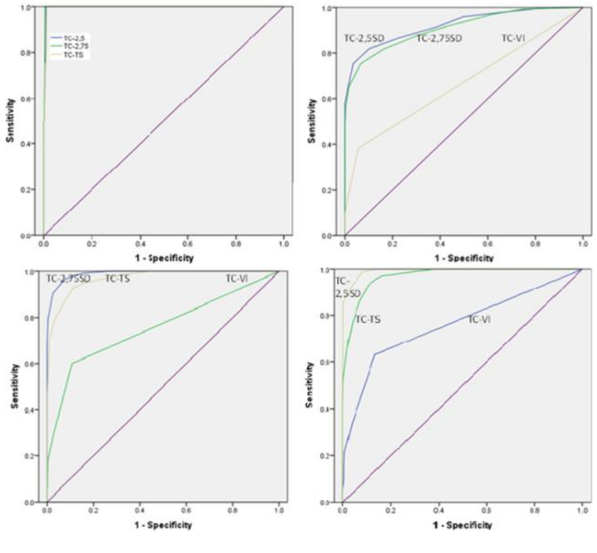

Fig. 1. ROC curve of rapid guessing behavior.

Figure 1 shows the results of the ROC curve plot that illustrates the sensitivity and sensitivity area on the rapid guess criteria greater than 0.25 for the four threshold methods VI, TS, 0.25SD, and 0.75SD.

Based on the calculation of the area under the curve when the method of determining the threshold VI, TS, 2,5SD, and 2.75SD is the variable state, the TS method is the method with the largest area under the curve is 98.25%. The results of the ROC curve plot illustrate the sensitivity and sensitivity area on the rapid guess with a criterion is more than 0.25. For the four methods of setting the threshold, providing information on the sensitivity of TS, 2.75SD, and 2.5 SD methods are always significantly more effective than VI in detecting rapid guess behaviour. Based on the calculation of the area under the curve when the VI, TS, 2.5SD, and 2.75 SD methods are as a variable state, the TS method is the method with the largest area under the curve is 98.25%. Therefore, TS is the most effective method in setting the threshold for detecting rapid guess behaviour than VI.

V. Conclusion and Recommendations

A. *Conclusion*

Based on the results of the ROC curve analysis, the method of determining the threshold developed in this study, namely the TS, 2.5SD, and 2.75SD methods, has a higher effectiveness than the VI method in detecting rapid guessing behaviour. Determination of rapid guess behaviour using the TS threshold method is the most effective in detecting rapid-guessing behaviour than TC-2,5SD, TC-2,75SD, and TC-VI.

B. *Recommendations*

Regarding scientific development, it is necessary to develop a variant VI based on the typology of the graph or the distribution characteristics of each item identified. Regarding the implications of the use of research results, the TS, 2.5SD, and 2.75SD methods have developed in this research can be applied in various computer-based test at the classroom, school, local, and national levels. Especially in the development of a computer-based test application system for gathering ability scores and anomalous behaviour scores of test-takers. In the context of assessment by teachers, schools, and government, these methods can be applied for formative, summative, and intervention functions. In the context of selection, test-taker anomaly information can be considered to determine the conclusion of the selection test result. In the context of learning achievement tests, test-taker anomaly information can become confirmation data on item parameters and ability parameters. Response time information can be used in the development of item banks, test design (test times, randomizing patterns of items in packages), and adaptive tests (setting the length of time per item).

Acknowledgments

Researchers say thank you to the Centre for Assessment and Learning, Agency Research and Development and Book, Ministry of Education, and Culture that has provided support in the form of response time data and responses in this study.

References

[1] Y.H. Lee and H. Chen, “A review of recent response-time analyses in educational testing,” Psychological Test and Assessment Modeling, vol. 53, no. 3, pp. 359-379, 2011.

[2] X.J. Kong, S.L. Wise and D.S. Bhola, “Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior,” Educational and Psychological Measurement, vol. 67, no. 4, pp. 606-619, 2007.

[3] S.L. Wise, “An investigation of the differential effort received by items on a low-stakes, computerbased test,” Applied Measurement in Education, vol. 19, pp. 95-114, 2006.

[4] S. Kupiainen, M.P. Vainikainen, J. Marjanen and J. Hautamaki, “The role of time on task in computer-based assessment of crosscurricular skills,” Journal o f Educational Psychology, vol. 106, pp. 627-638, 2014.

[5] Y.H. Lee and S.J. Haberman, “Investigating test-taking behaviors using timing and process data,” International Journal of Testing, vol. 16, no. 3, pp. 240-267, 2016.

[6] R.H.K. Entink, Statistical models for responses and response times. Thesis, 2009.

[7] Y.H. Lee and Y. Jia, “Using response time to investigate students' test-taking behaviors in a NAEP computer-based study,” Large-scale Assessments in Education, vol. 2, no. 1, pp. 1-24, 2014.

[8] G. Karabatsos, “Comparing the aberrant response detection performance of thirty-six person-fit statistics,” Applied Measurement in Education, vol. 16, no. 4, pp. 277-298, 2003.

[9] R.R. Meijer and K. Sijtsma, “Methodology review: Evaluating person fit,” Applied psychological measurement, vol. 25, no. 2, pp. 107-135, 2001.

[10] B. Thiessen, Relationship between test security policies and test score manipulations. (PhD Thesis). University of Iowa, 2008.

[11] P. Emmen, A person-fit analysis of personality data. (Master Thesis). Department of Social and Organizational Psychology. Vrije Universiteit, 2011.

[12] D.I. Belov, “Detection of test collusion via Kullback–Leibler divergence,” Journal of Educational Measurement, vol. 50, no. 2, pp. 141-163, 2013.

[13] J.N. Tendeiro and R.R. Meijer, “Detection of invalid test scores: The usefulness of simple nonparametric statistics,” Journal of Educational Measurement, vol. 51, no. 3, pp. 239-259, 2014.

[14] C. Wang and G. Xu, “A mixture hierarchical model for response times and response accuracy,” British Journal of Mathematical and Statistical Psychology, vol. 68, no. 3, pp. 456-477, 2015.

[15] S.L. Wise and X. Kong, “Response time effort: A new measure of examinee motivation in computer-based tests,” Applied Measurement in Education, vol. 18, no. 2, pp. 163-183, 2005.

[16] S.L. Wise and L. Ma, “Setting response time thresholds for a CAT item pool: The normative threshold method,” In annual meeting of the National Council on Measurement in Education, Vancouver, Canada. pp. 163-183, 2012.

[17] H. Guo, J.A. Rios, S. Haberman, O.L. Liu, J. Wang and I. Paek, “A new procedure for detection of students’ rapid guessing responses using response time,” Applied Measurement in Education, vol. 29, no. 3, pp. 173-183, 2016.

[18] C.E. Demars, “Changes in rapid-guessing behavior over a series of assessments,” Educational Assessment, vol. 12, no. 1, pp. 23-45, 2007.

[19] J.C. Setzer, S.L. Wise, J.R. van den Heuvel and G. Ling, “An investigation of examinee test-taking effort on a large-scale assessment,” Applied Measurement in Education, vol. 26, no. 1, pp. 34-49, 2013.

[20] I. Tsaousis, G.D. Sideridis and A. Al-Sadaawi, “An IRT–Multiple Indicators Multiple Causes (MIMIC) Approach as a Method of Examining Item Response Latency,” Frontiers in psychology, vol. 9, no. 2177, 2018.

[21] F. Sahin and K.F. Colvin, “Enhancing response time thresholds with response behaviors for detecting disengaged examinees,” Large-scale Assessments in Education, vol .8, pp. 1-24, 2020.

[22] R. Whelan, “Effective analysis of reaction time data,” The Psychological Record, vol. 58, no. 3, pp. 475-482, 2008.

[23] M.P. Michaelides, M. Ivanova and C. Nicolaou, “The relationship between response-time effort and accuracy in PISA science multiple choice items,” International Journal of Testing, vol. 20, no. 3, pp. 187-205, 2020.

[24] R. Artner, “A simulation study of person-fit in the Rasch model,” Psychological Test and Assessment Modeling, vol. 58, no. 3, pp. 531-563, 2016.

[25] G.A. Rousselet and R.R. Wilcox, ”Reaction times and other skewed distributions: problems with the mean and the median,” Meta-Psychology, vol. 4, 2020.

[26] S.L. Wise, “Rapid-guessing behavior: Its identification, interpretation, and implications,” Educational Measurement: Issues and Practice, vol. 36, no. 4, pp. 52-61, 2017.

[27] A.L. Zenisky and P. Baldwin, “Using item response time data in test development and validation: Research with beginning computer users,” Center for educational assessment report no. 593, 2006.

[28] M. Hecht, T. Siegle and S. Weirich, “A model for the estimation of testlet response time to optimize test assembly in paper-and-pencil large-scale assessments,” Journal for educational research online, vol. 9, no. 1, pp. 32-51, 2017.

[29] P.L. Gaol, M. Khumaedi and M. Masrukan, “Pengembangan Instrumen Penilaian Karakter Percaya Diri pada Mata Pelajaran Matematika Sekolah Menengah Pertama,” Journal of Research and Educational Research Evaluation, vol. 6, no. 1, pp. 63-70, 2017.

[30] R.R. Wilcox, “How many discoveries have been lost by ignoring modern statistical methods?” American Psychologist, vol. 53, no. 3, pp. 300, 1998.

[31] R. Ratcliff, “Methods for dealing with reaction time outliers,” Psychological bulletin, vol. 114, no. 3, pp. 510, 1993.

[32] N. Becker, F. Schmitz, A.S. Göritz and F.M. Spinath, “Sometimes more is better, and sometimes less is better: Task complexity moderates the response time accuracy correlation,” Journal of Intelligence, vol. 4, no. 3, pp. 11, 2016.