REiD (Research and Evaluation in Education), 6(2), 2020, 87-97
Available online at: http://journal.uny.ac.id/index.php/reid

Comparison of methods for detecting anomalous behavior on large-scale computer-based exams based on response time and responses

[*1]Deni Hadiana; [2]Bahrul Hayat; [1]Burhanuddin Tola
[1]Graduate School, Universitas Negeri Jakarta
Jl. R. Mangun Muka, Rawamangun, Pulo Gadung, Kota Jakarta Timur, Jakarta 13220, Indonesia
[2]Faculty of Psychology, Universitas Islam Negeri Syarif Hidayatullah Jakarta
Jl. Kertamukti No. 5, Cireundeu, Ciputat Timur, Tangerang Selatan, Banten 15419, Indonesia
[*]Corresponding Author. E-mail: denihadianadua@gmail.com

Submitted: 18 April 2020 | *Revised: 25 June 2020* | *Accepted: 16 July 2020*

Abstract

This study aims to determine the anomalous index (*indeks anomali* or IA) that considers both response time and responses and compares it with response time effort (RTE) or rapid guessing (*tebakan cepat* or TC) on various thresholds. Response time and responses from 732 examinees are in natural science subjects consist of 40 multiple choice items with four answer choices. Response time and responses are analyzed to obtain descriptive statistics related to them, calculate the TC and IA index using two methods of the threshold, the first method (M1) is a visualization of identification, and the second method (M2) is based on the amount of time spent responding to each item related to the complexity of items, as proposed by Nitko. The performance of the IA and TC scores is compared related to validity and reliability. The coefficient alpha of IAM1 score 0.84, the coefficient alpha of IAM2 0.82. Both values of the alpha coefficient have fulfilled the reliability requirements of the index determination. The IA proposed in this study has a high correlation with ERP, which is commonly used to determine the solution behavior's magnitude and rapid guessing. The correlation value of IAM1 with TCM1 0.86, the correlation value of IAM2 with TCM2 0.89, and this high correlation value shows that there is a strong relationship between IA and TC. Determination of threshold time uses three categories of multiple choices item that reveal IA and TC distributions that are close to normal distribution so that it reflects natural empirical conditions.

Keywords: *anomalous index (IA), rapid guessing (TC), threshold, reliability, validity*

How to cite: Hadiana, D., Hayat, B., & Tola, B. (2020). Comparison of methods for detecting anomalous behavior on large-scale computer-based exams based on response time and responses. *REiD (Research and Evaluation in Education), 6*(2), 87-97. doi:https://doi.org/10.21831/reid.v6i2.31260.

Introduction

Cognitive tests, such as computer-based national exams (CBNE), measure the competency of students' knowledge after they complete the learning process for approximately three years for junior and senior high school in certain subjects according to the curriculum. Since 2015, in addition to the paper and pencil-based national exams, CBNE began to be implemented. Even since the implementation of the national exams in 2018, CBNE has become the main mode. Based on the Center for Educational Assessment report on the results of the national exams, the junior high school CBNE examinees continued increasing nationally, in 2015 as much as 0.22%; in 2016 to 3.72%; in 2017 became 32.26%; in 2018 it became 62.97%, even in Jakarta, Indonesia, CBNE was used in 2017 and 2018 with 100% each. CBNE is expected to increase the validity, reliability, and integrity of the exams.

This is an open access article under the CC-BY-SA license.

Computer-based exams, such as CBNE, have many advantages over pencil and paper-based tests. Computerized exams according to Lee and Chen (2011) can provide complex information, because, in addition to providing information about examinee responses, computer-based exams can provide information on the response times that reflect the amount of time that is spent by examinees to respond to each item. Meanwhile, according to Linacre and Rudner, as quoted by Georgiadou et al. (2006), the advantages of computer-based exams are test management flexibility, increased test security, increased motivation in information technology literacy, and time efficiency. A good test security procedure can be a quality control test implementation so the validity of a good test score is guaranteed (Lewis et al., 2014) and can be obtained. Test security is related to performance (validity and reliability) of a test (Cizek & Wollack, 2016, p. 3). Thus, invalid CBNE items and unreliable CBNE tests scores result in information on test results that cannot be used, especially if CBNE scores are used for various strategic interests such as for selection to next education levels, mapping of education quality, and policy interventions to improve the quality of education. CBNE results will be meaningful, appropriate to target, and effective if the scores obtained by CBNE examinees are accurate. This means that the scores obtained by CBNE examinees truly reflect the ability of CBNE examinees. The score obtained in the CBNE is closely related to the response pattern and the response time pattern of the examinees since they can be used to determine anomalous data. Thus, research that is related to response patterns and response time patterns to determine anomalous data is very urgent because the analysis of responses and response times accurately will have a real contribution to improving the quality of the examinee ability estimation (Fox et al., 2007).

The response time that is spent by each examinee during processing and responding to each item and the response can be directly obtained on a computer-based exam. Based on the response time data, we can detect the anomaly response time of the test examinees compared to the response time of other test examinees. Examinees who answer items too quickly compared to other examinees can be indicated as examinees who exhibit anomalous behavior.

Anomalous behavior is likely to occur due to various reasons, among others, the examinee has known information related to the item earlier, rapid guessing, and responded randomly. Anomalous behavior is closely related to testing security, examinee's integrity, item validity, test reliability, fairness, and examinees ability. Thus, Van der Linden (2006), Marianti et al. (2014), Meijer and Sotaridona (2006), Widiatmo and Wright (2015), and Wise and Kong (2005) conclude that if anomalous data were analyzed appropriately, better measurement results for abilities would be obtained, for example, the anomalous data is not included in the estimated ability parameter.

Several methods can be used to determine anomalous data based on the response time. Wise and Kong (2005) believe that examinees with high efforts in responding to each item will show behavioral solutions. On the other hand, the examinees with low effort in responding to each item will show guessing behavior shown by responding to the items rapidly. This rapid guessing can be seen from the short response time where they did not take the time to read the item in full and it is impossible to consider the item carefully. This rapid guessing behavior underlies the determination of the examinees' anomalous behavior, Wise and Kong named it Response Time Effort (RTE). Before determining RTE for dichotomous items, the SBij is calculated first using the equation in Formula (1).

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

SBij is the solution behavior of examinee j on item i, Ti is the threshold between the time of rapid guessing behavior and solution behavior on item i, RTij is the response time of examinee j in item i. Next, the RTE is calculated using the equation in Formula (2), in which k is the number of items in the test, and the range of RTE scores from 0 to 1 reflecting the proportion of items for examinees who have solution behavior.

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

RTE value close to 1 indicates the higher effort in answering all items of the test, on the contrary, the value of RTE getting closer to 0 indicates the lower effort that occurs. In this study the term rapid guessing (TC) will be used as an index that has an inverse relationship with RTE, meaning that the lower of the RTE value or the closer it is to 0, the TC value is higher or closer to 1. From Formula (1) and Formula (2), it is known that RTE calculations do not consider the responses of examinee answers both to the items below the threshold and to the items above the threshold. Examinees who correctly answer selected items with less time than the threshold must be treated differently from examinees who wrong answers selected items with less time than the threshold. Likewise, examinees who correctly answer selected items longer than the threshold must be treated differently from examinees who wrong answers selected items longer than the threshold.

Examinees who answered correctly on selected items with response time above the threshold are normal behavior. Conversely, examinees who answered correctly on selected items with response times below the threshold are anomalous behavior. Based on these two conditions, we propose the normal index (*indeks wajar* or IW) and the anomaly index (*indeks anomali* or IA) that consider the response time and responses. Mathematically, the IW is stated in Formula (3), while the IA is stated in Formula (4).

$$IW_j = \frac{\Sigma\, PW_{ij}}{\Sigma(PW_{ij} + PA_{ij})} \qquad (3)$$

$$IA_j = \frac{\Sigma\, PA_{ij}}{\Sigma(PW_{ij} + PA_{ij})} \qquad (4)$$

IWj is the normal index of examinee j, PWij is reasonable behavior on item i of examinee j which is given a score 1 for the examinee who correctly answers item i (Bi) with response time (WRij) above the threshold (Ti), or PWij = 1 if Bi >Ti. IAj is the anomaly index of examinee j, PAij is an anomalous behavior on item i of examinee j which is given a score 1 for examinees who correctly answer item i with time below or equal to the threshold or PAij = 1 if Bi ≤ Ti. The relationship between IWj and IAj is shown in Formula (5).

$$IW_j + IA_j = 1 \qquad (5)$$

The IW scores getting closer to 1 indicates more normal behavior, while IW scores getting closer to 0 show more anomaly behavior. In contrast, the IA scores getting closer to 1 indicate behavior anomalous increasingly, while the IA scores getting closer to 0 show more normal behavior.

The challenge in determining the RTE, IW, and IA lies in determining the accurate threshold (T). Determination of the threshold must consider the characteristics of items, like the number of words in an item, the presence of stimulus pictures, tables, and illustrations, items with calculation, the difficulty of items. Nitko (Naga, 2013, p. 47) said simple multiple-choice items need 40 to 60 seconds to answer, complex multiple-choice items need 70 to 90 seconds, and multiple-choice items with calculation need 120 to 300 seconds.

Wise and Kong (2005) determined the threshold based on the number of characters. Items with the number of characters less than 200 are a threshold of three seconds, items with characters between 200 to 1000 are five seconds of a threshold and the threshold of items with more than 1000 characters are in ten seconds. Kong et al. (2007) apply several methods to determine the threshold: the common threshold for each item; based on item characteristics such as the number of characters and the presence of pictures; visualization identification of response time-frequency distribution graphs; and estimation using two mixed models.

The RTE that had been developed by Wise and Kong since 2005 did not consider the responses of each item of examinees for determining the index of effort or the index of rapid guessing. Thus, this study was conducted to determine the anomaly index (IA) which considers both response time and responses and comparing with RTE at various thresholds.

Method

Log file data of 732 examinees obtained from the Educational Assessment Center, Research and Development Agency, Ministry of Education and Culture was converted into structured spreadsheet data. The structured data consists of response time in seconds and a dichotomous score of responses from 40 multiple-choice items in Natural Sciences subjects. The data were then screened and processed with the help of Minitab, Microsoft Excel, and SPSS application. The data were then analyzed quantitatively to obtain descriptive statistical information related to the response time and responses, determine the TC index and IA index on the two methods of determining the threshold. The first method (M1) was done by identification visualization (VI), which is looking at the response time for the first time the response time decreases sharply then increases again through visualization of the response time-frequency graph. The second method (M2) was done using the length of time criteria for working on the multiple-choice items proposed by Nitko, namely, simple multiple-choice items using a 40 second, 70 seconds complex multiple-choice, and 120 multiple-choice calculations. For this reason, before the index calculation, those items are grouped into three, namely simple multiple-choice, complex multiple-choice, and multiple-choice calculation. Then, we conducted a comparison of the results of anomalous data analysis with RTE and IA index.

Findings and Discussion

Lindsey (2004, p. 197) explains that the characteristics that must be considered in the selection of the response time distribution: (1) must be positive; (2) short response time is more common than long response time, or in other words, the magnitude of the short response time probability is very large compared to the magnitude of the long response time probability or positive skewed (Van der Linden, 2006). Distributions that match these two characteristics are Lognormal, Weibull, and Gamma distributions (Lindsey, 2004, pp. 203-206). Figure 1 shows the mean and standard deviation and median scatterplots and averages on 40 items and 732 examinees. All response times are positive with a minimum response time of 1 second. This condition is following the response time criteria according to Lindsey (2004, p. 197).

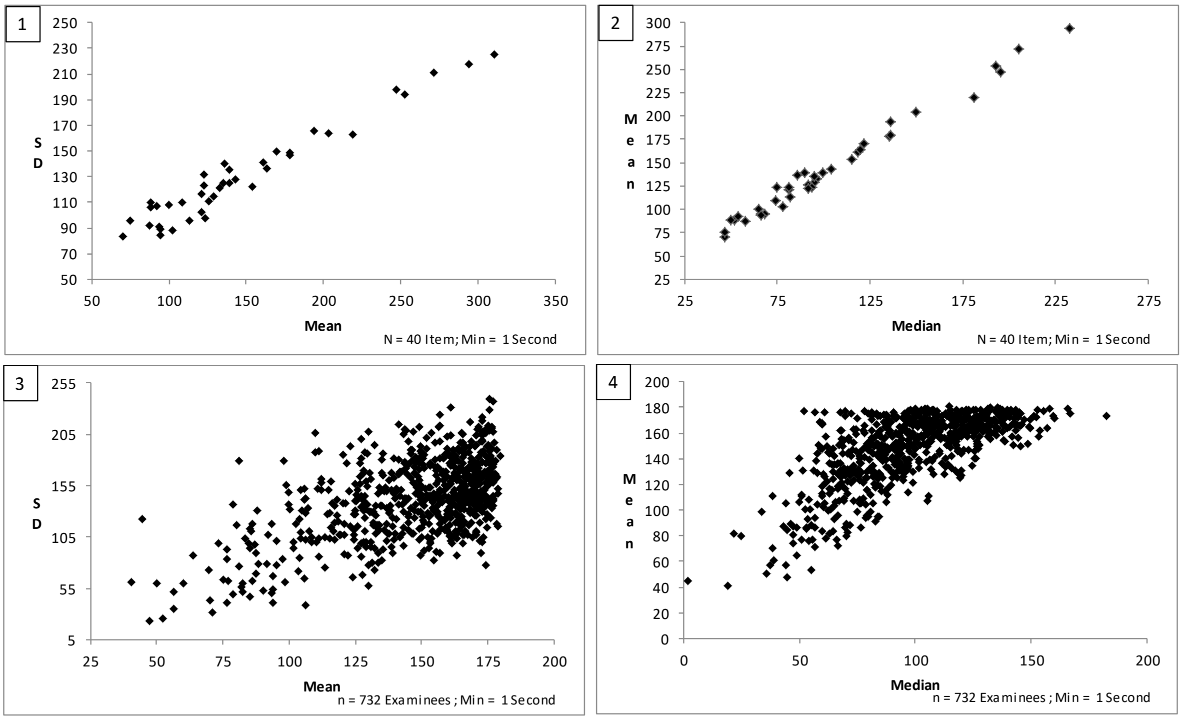


Figure 1. Scatterplots of Mean and Standard Deviation, Mean, and Median of Items and Examinees

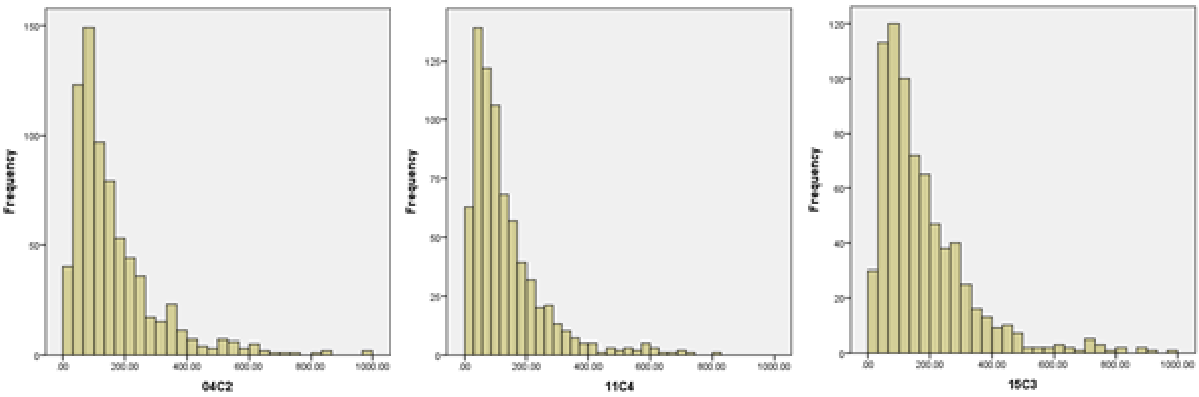


Figure 2. Distribution of Response Time for Items Number 4, 11, and 15

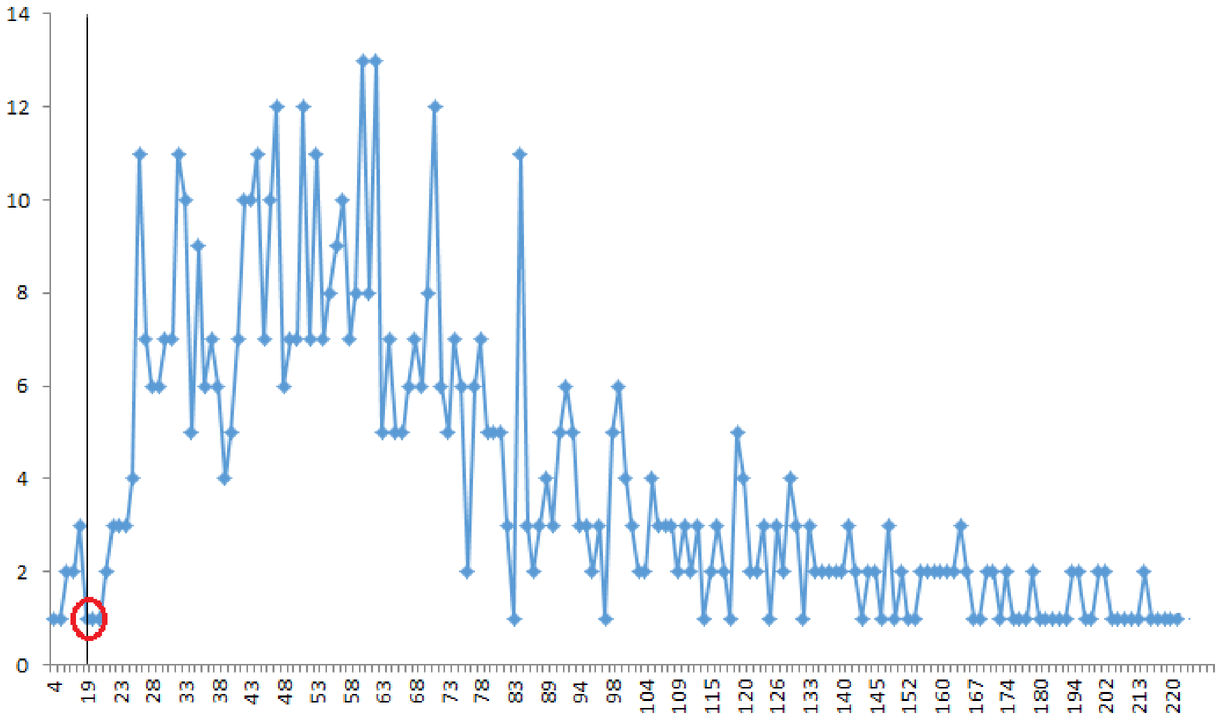


Figure 3. Frequency of Response Time of a Selected Item

In addition, Figure 2 shows the distribution of response time for items number 4, 11, and 15. According to Fox et al. (2007), Van der Linden (2006), Lindsey (2004, p. 197), and Wulansari et al. (2019), the distribution of response time that tends to skew to the right shows that a small portion of the response time-frequency is on the right, meaning that the probabilities of short response time are higher than the probabilities of long response time. This distribution, according to Lindsey in Wulansari (2019, p. 140) is following the characteristics of Lognormal, Weibull, and Gamma distribution.

In this study, the Anderson Darling test was performed to determine the distribution of response time samples. Wulansari (2019, p. 88) states that the distribution with the lowest Anderson Darling value is the most proper distribution for the response time on each item. From the results of comparison of the Anderson Darling on Lognormal, Weibull, and Gamma distributions for 40 items, it can be seen that items number 12, 20, 21, and 23 are appropriate with the characteristics of the Gamma distribution, while 36 other items match the characteristics of the Lognormal distribution. Information on the characteristics of distribution is very crucial when determining the response time threshold through the visualization method of response time-frequency graphs.

Figure 3 provides information about the frequency of response time (seconds) for one of the items used to determine the threshold time using M1. The threshold is set at a sharp drop in response time then rises again, in the figure, the threshold is 19 seconds in the circle. This threshold is a short response time and is insufficient for the examinees to process the items and options.

This method was proposed by Wise (2006). Conceptually, according to Kong et al. (2007), the thre-shold is a short response time for examinees so that they do not have sufficient time to process and determine the correct answer from the items. After visualizing the identification of 40 response time-frequency graphs obtained a threshold in seconds based on the M1 method for items number 1 to 40 respectively: 26; 14; 8; 17; 15; 17; 21; 22; 26; 8; 6; 26; 21; 20; 15; 7; 14; 14; 11; 23; 28; 22; 15; 19; 18; 22; 23; 20; 14; 17; 19; 11; 19; 18; 13; 5; 16; 21; 13; and 13.

Based on the results of an analysis of the characteristics of 40 items including the number of words, the presence of pictures, tables, illustrations, and the cognitive level of items. Items number 2; 3; 5; 6; 7; 14; 25; 26; 29; 30; 31; 32; 33; 36; and 37 are simple multiple-choice type and use a threshold of 40 seconds for each item. Items number 1; 4; 17; 22; 24; 28; 35; 38; 39; and 40 are complex multiple-choice type and use a threshold of 70 seconds for each item. Items number 8; 9; 10; 11; 12; 13; 15; 16; 18; 19; 20; 21; 23; 27; and 34 are multiple choice type of calculation and use a threshold of 120 seconds for each item. The threshold of 40 items is shown in Figure 4.

Determination of threshold using the visualization identification method of the response time graph does not consider the characteristics of the item, such as the complexity of items and items require calculation. This method considers the certainty response time, especially for the response time that decreases sharply for the first time and then increases. As a result, this method produces less stable response time and does not reflect the degree or level of each item, and tends to produce a low threshold so that it impacts on the low percentage of examinees who exhibit anomalous behavior (Hauser & Kingsbury, 2009). If this threshold determination method is applied to detect the level of an anomaly behavior, the examinees who are detected anomaly behavior are very few, especially for the complex items, the items that require calculations, the items that have pictures, and the items that contain many words, as usually found on items in the national testing like in Indonesia.

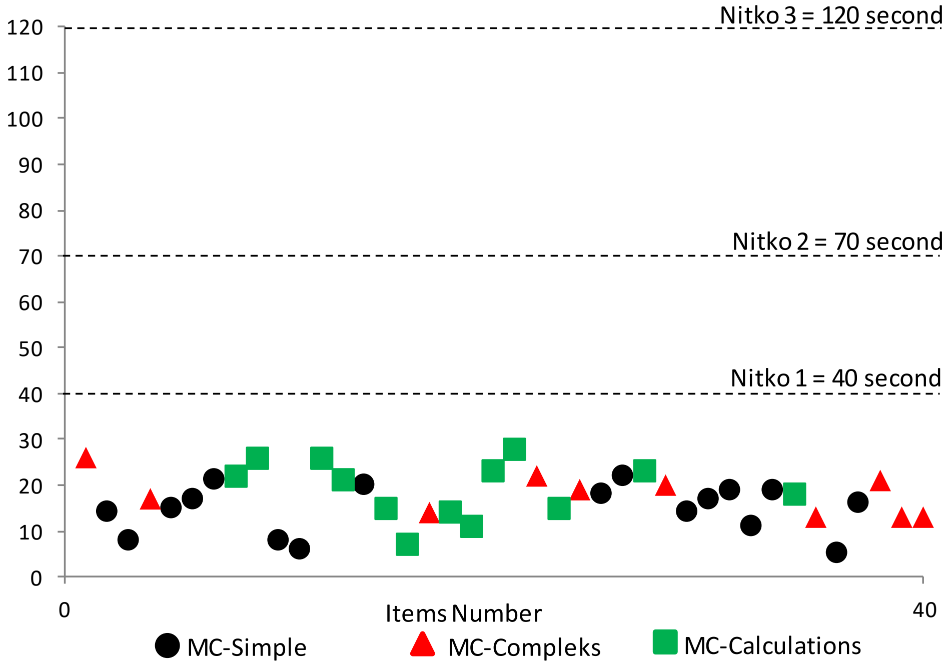
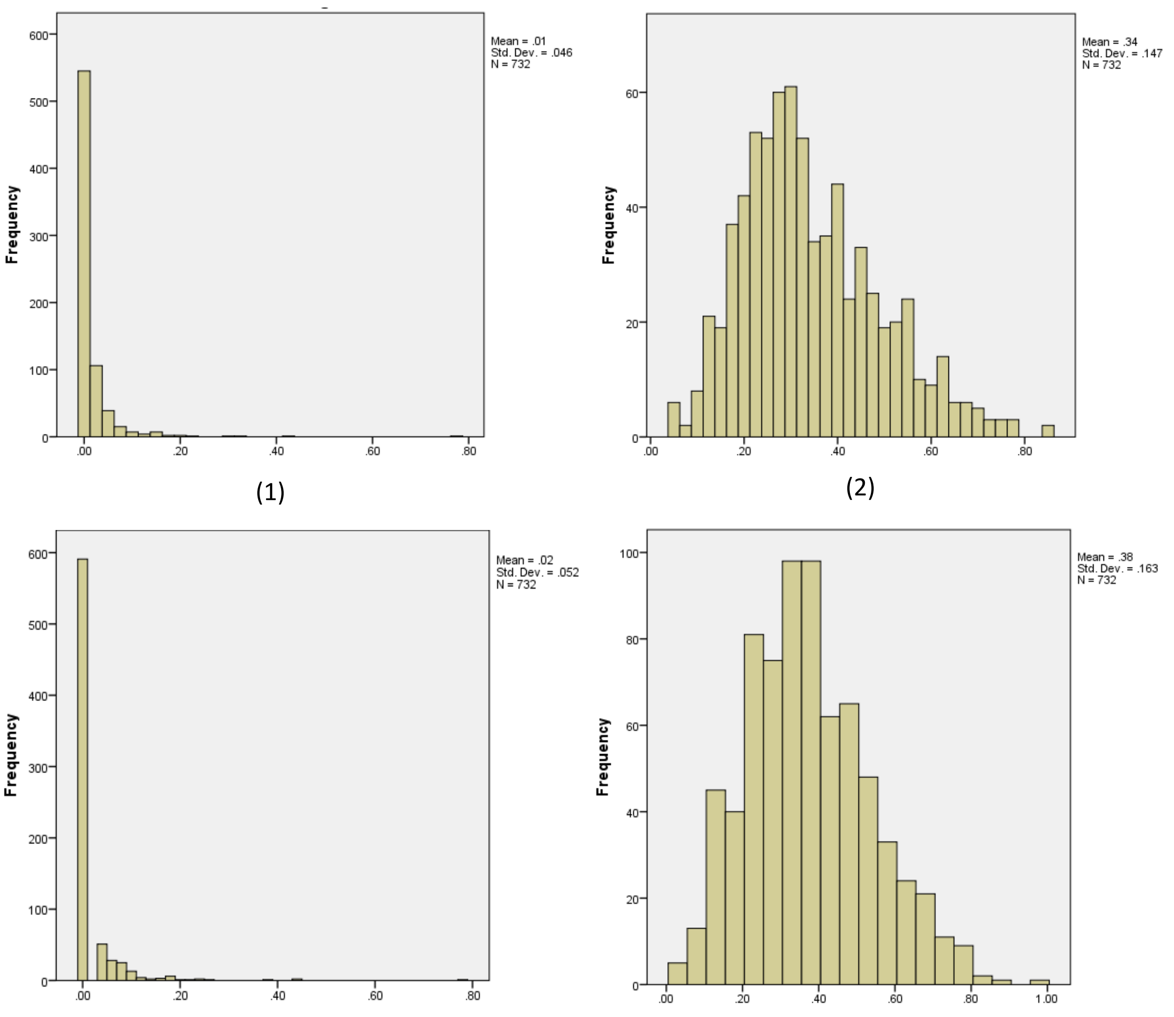


Figure 4. Threshold of 40 Items

Hauser and Kingsbury (2009) states that the use of thresholds that are too short and applied to all items by regardless the characteristics of the item has many limitations, because the characteristics of each item are different than the other ones, each item has a unique psychometric parameter such as difficulty levels, and each item has a unique surface characteristic, like the number of words. Therefore, the determination of the threshold that considers the characteristics of the item or the subgroup of items makes more reasonable as occurred performed in this study.

Wise and Kong (2005) said that the determination of an index must have an adequate degree of reliability. According to Wise and Kong, a minimum alpha coefficient of 0.80 is acceptable for index determination. With 95% CI, the coefficient alpha of TCM1 and TCM2 index is 0.85 and 0.83 respectively. With 95% CI, the coefficient alpha of IAM1 and IAM2 index is 0.84 and 0.82 respectively. This coefficient alpha value reaches the reliability requirements of index determination.

Figure 5 shows the frequency of TCM1, TCM2, IAM1, and IAM1 on the distribution of index scores. In Figure 5, most of the index scores are close to and equal to 0, meaning that most of the examinees are not indicated to rapid guessing behavior, but most of them show solution behavior during responding to the item or show normal behavior. This can be observed in the figure that shows negative skewness distribution. The IAM1 mean is higher than the TCM1 mean because the determination of IA considers the response time and responses of examinees. This finding is following Wise and Kong (2005).



(1) (2)

(3) (4)

Figure 5. Frequency Index of TCM1 (1), TCM2 (2); IAM1 (3), IAM2 (4).

Besides, Figure 5 shows that the frequency of TCM2 and IAM2 index scores is mostly distributed at the middle or moderate index scores and close to the normal distribution. It means that a small number of examinees showed anomaly behavior or natural behavior and most of them behaved normally or moderate degrees both an anomaly and rapid guessing. Thus, this method is following empirical conditions. The IAM2 mean is greater than the TCM2 mean, because the determination of IA takes the response time and responses into account. Therefore, the determination of the M2 threshold reflects or approaches the normal distribution in both TC and IA.

A significant correlation on the TC index developed by Wise and Kong and IA proposed in this study shows that conceptually has a strong relationship between the TC index and the IA index (see Table 1). The relationship is higher when the threshold determination method used is the same. This can be seen in the TCM2 correlation value with IAM2 which is higher than the TCMI correlation value with IAM1. Thus, M2 has a higher relationship than M1.

By using an index range of 0.74 to 1, IAM2 succeeded in detecting 16 examinees, TCM2 succeeded in detecting eight examinees, while IAM1 and IAM2 each had one examinee. IAM2 was the most successful in detecting anomalous examinees because the calculation of the IAM2 threshold considered the characteristics of the items and the responses so that the probability of detecting anomalies was higher as shown in Figure 5.

Table 1. Correlation of Spearmen's Rho TCM1, TCM2, IAM1, IAM2

			TCM1	TCM2	IAM1	IAM2
Spearman's Rho	TCM1	r_{sp}	1.000	.489 **	.859 **	.443 **
		Sig. (2-tailed)	.	.000	.000	.000
	TCM2	r_{sp}	.489 **	1.000	.415 **	.884 **
		Sig. (2-tailed)	.000	.	.000	.000
	IAM1	r_{sp}	.859 **	.415 **	1.000	.419 **
		Sig. (2-tailed)	.000	.000	.	.000
	IAM2	r_{sp}	.443 **	.884 **	.419 **	1.000
		Sig. (2-tailed)	.000	.000	.000	.

**. Correlation is significant at the 0.01 level (2-tailed).

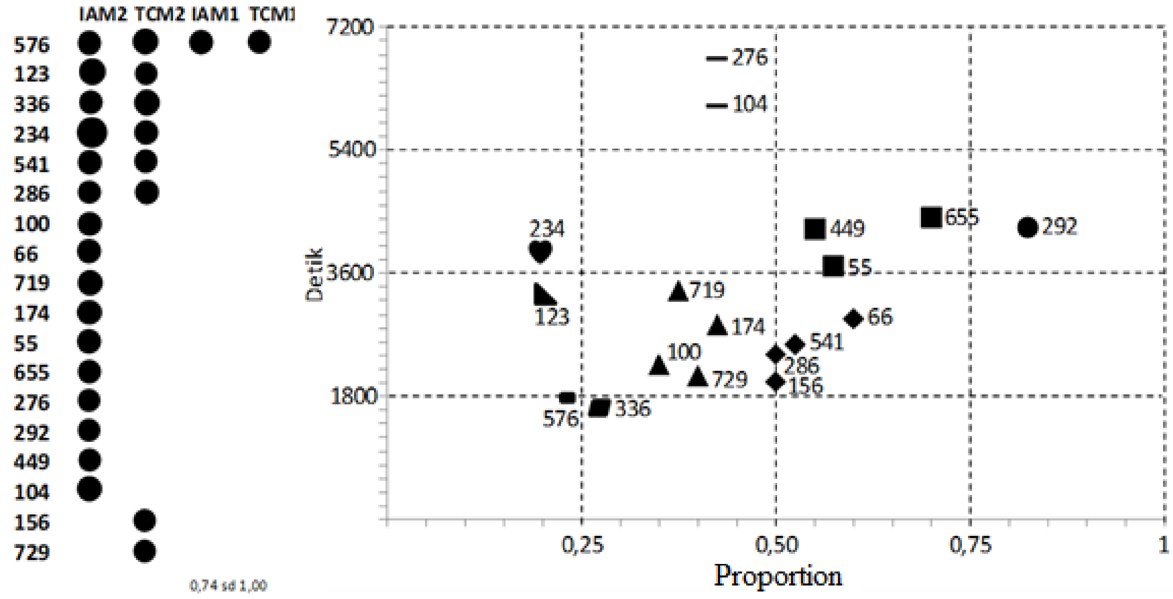


Figure 6. IA and TC Index of Range 0.74 to 1 (left); Total Time Test and Correctness Proportion (right)

From Figure 6, examinee 576 were detected anomaly by IAM2, IAM1 and detected rapid guessing by TCM2, meaning that TCM1 was detected by all indices in this study. Data analysis of the proportion of correct answers and the total test response time provided information that Examinee 576 had the proportion of correct answers low, below 0.25 but spent the total time to answer all items is too short, that is 1800 seconds or 45 seconds per item from the 7200 seconds time allocation. Thus, examinee 576 indicated that the behavior of rapid guessing as examinee 336. Anomalous behavior in examinees 576 and 336 was due to a lack of effort to process and answer items as predicted by Meijer (2003). Examinees 276 and 104 showed a great effort by maximizing the allocation of test time available but the test results were not good enough with the proportion of correct answers below 0.5. Examinee 292 had the highest proportion of 0.75 and was able to complete the test with about 3900 seconds. Then why was the examinee detected anomaly by IAM2? First, this phenomenon may occur due to the quality of the item, for example, there are keywords in the stimulus that lead to options, the construction of item which results in the examinee not having to process every word and symbol in the item but just connecting each of these keywords to answer the item. Second, the options' distractors do not work properly because the distractors are not homogenous and logical. There are various possible types of anomalous behavior based on the results of responses and response time analysis, includeing cheating, creative responses, careless responses, lucky guesses (Meijer, 1996). Therefore, to confirm these various types of anomaly behavior, it is better to analyze item characteristics and psychometric parameters of items such as discrimination index and difficulty level.

Conclusion

A simple method for determining the solution behavior index or rapid guessing behavior (TC) is the ERP that is proposed by Wise and Kong since 2005, this method is still often used mainly for low stakes tests. This ERP method considers the response time only and does not consider the responses of examinees in each item. The IA method that is proposed in this study considers the response time and the responses of each examinee on each item and it is easily implemented. The reliability coefficient alpha of the IAM1 score 0.84, while the coefficient alpha of the IAM1 reliability is 0.82. Both values of the alpha coefficient have fulfilled the reliability requirements of the index determination. IA that is proposed in this study has a high correlation with ERP which is commonly used to determine the magnitude of the solution behavior or rapid guessing behavior. The correlation value of IAM1 with TCM1 is 0.86, the correlation value of IAM2 with TCM2 is 0.89. This high correlation value shows that there is a strong relationship between the IA and ERP (TC). The determination of the threshold must consider the characteristics of the items, such as the presence of pictures and the number of words and psychometric characteristics such as the level of difficulty items. The determination of the threshold uses three groups of multiple-choice items, namely: the simple multiple-choice, complex multiple-choice, and multiple-choice with calculation resulting in IA and TC distributions that are close to normal distribution so that it reflects natural empirical conditions. To conclude the type of anomaly shown by examinees, IA should be confirmed by qualitative and psychometric attributes of the test items and examinees' abilities. To perfect this study, research should be conducted regarding the determination of a more comprehensive threshold by considering the item surface characteristics such as the number of words, the cognitive level of items, the complexity of items, and the psychometric characteristics of items such as difficulty level, discriminating index, and the ability of the examinees.

Acknowledgment

The researchers deliver their gratitude to the Center for Educational Assessment, Research and Development Agency, Ministry of Education and Culture for providing support in the form of data on the response time and responses in this study.

References

Cizek, G. J., & Wollack, J. A. (2016). *Handbook of quantitative methods for detecting cheating on tests* (1st ed.). Routledge. https://doi.org/10.4324/9781315743097

Fox, J.-P., Entink, R. K., & Van der Linden, W. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software, 20*(7). https://doi.org/10.18637/jss.v020.i07

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2006). Evaluation parameters for computer-adaptive testing. *British Journal of Educational Technology*, *37*(2), 261–278. https://doi.org/10.1111/j.1467-8535.2005.00525.x

Hauser, C., & Kingsbury, G. G. (2009, November 4). *Individual score validity in a Modest-Stakes adaptive educational testing setting* [Paper presentation]. The Annual Meeting of the National Council on Measurement in Education, Sandiego, CA. https://www.nwea.org/resources/individual-score-validity-modest-stakes-adaptive-educational-testing-setting/

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from Rapid-Guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. In *Psychological Test and Assessment Modeling* (Vol. 53, Issue 3). http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/06_Lee.pdf

Lewis, C., Lee, Y.-H., & Davier, A. A. Von. (2014). Test security for multistage tests: A quality control perspective. In N. Kingston & A. Clark (Eds.), *Test Fraud (Statistical Detection and Methodology)* (1st ed.). Routledge. https://doi.org/https://doi.org/10.4324/9781315884677

Van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204. https://doi.org/10.3102/10769986031002181

Lindsey, J. K. (2004). *Statistical analysis of stochastic processes in time*. Cambridge University Press. https://doi.org/10.1017/CBO9780511617164

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426–451. https://doi.org/10.3102/1076998614559412

Meijer, R.R., & Sotaridona, L. (2006). *Detection of advance item knowledge using response times in computer adaptive testing* (LSAC research report series No. CT 03-03). Law School Admission Council.

Meijer, R. R. (1996). Person-Fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3–8. https://doi.org/10.1207/s15324818ame0901_2

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, *8*(1), 72–87. https://doi.org/10.1037/1082-989X.8.1.72

Naga, D. S. (2013). *Teori sekor pada pengukuran mental* (2nd ed.). Nagarami Citrayasa.

Widiatmo, H., & Wright, D. B. (2015, April). *Comparing two item response models that incorporate response times* [Paper presentation]. National Council on Measurement in Education Annual Meeting, California, Illionis, USA. https://www.researchgate.net/publication/283711098_Comparing_Two_Item_Response_Models_That_Incorporate_Response_Times

Wise, S. L. (2006). An investigation of the differential effort received by items on a Low-Stakes Computer-Based Test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in Computer-Based Tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wulansari, A. D. (2019). *Model logistik dalam IRT dengan variabel random waktu respon untuk tes terkomputerisasi* [Doctoral Dissertation, Universitas Negeri Yogyakarta]. Eprints UNY. http://eprints.uny.ac.id/id/eprint/66079

Wulansari, A. D., Kumaidi, & Hadi, S. (2019). Two parameter logistic model with Lognormal Response Time for Computer-Based Testing. *International Journal of Emerging Technologies in Learning (IJET)*, *14*(15), 138–158. https://doi.org/10.3991/ijet.v14i15.10580