ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2025; 16(3); 160-178

Comparing Different Methods of Identifying Rapid-Guessing Thresholds

Çağatay COŞKUN * Duygu ANIL **

Abstract

Rapid-guessing behavior demonstrated by individuals in achievement tests is expressed as a factor that threatens validity. Response time data, which can be obtained through computer-based test applications, allows for the distinction between rapid-guess behavior and solving behavior. In the literature, there are various methods for determining the threshold value used to make this distinction. In this study, the response time data, which were binary-coded based on different threshold determination methods, were incorporated into the item response tree model, and the fit of response times determined by which method best suited the model was examined. The data of 3329 individuals who participated in the PISA 2018 implementation in English and in a computer-based environment were used. The response and response time data of these individuals for 18 mathematics items from booklet 2 (M02 – M03) were selected for analysis.

Among the threshold determination methods, the normative threshold method (NT10), visual inspection of response time distributions (VI), graphical review methods based on response accuracy, the cumulative proportion (CUMP) method, median-based, and logarithmic transformation-based methods (logNT10) were selected. The individuals' responses to the items and their response times were analyzed using the "*irtree*" and "*lme4*" packages. Comparative results regarding the model-data fit of the response times obtained by different methods were provided. According to the results, the data obtained with the normative threshold method applied after the logarithmic transformations of the response times showed the best model fit. The logNT10 method can be a good alternative in cases where the RT distribution is non-bimodal and skewed. Furthermore, when compared to other evidence regarding the validity of the threshold values used, it was found to be consistent with the model-data fit results. Based on these results, existing methods and recommendations for determining threshold values for rapid-guess behavior were discussed.

Keywords: response time, IRTree models, identifying thresholds, rapid-guessing behavior

Introduction

It is assumed that tests administered to determine what individuals know and can do accurately reflect their performance. In practice, test takers may deviate from their optimal performance. This may be due to time constraints (test speediness) in high-stakes assessments and lack of motivation in low-stakes assessments (Lee & Chen, 2011). In tests referred to as low-stakes assessments in literature, which do not directly impact individuals' lives, situations can arise where individuals are not sufficiently motivated to reflect their performance or do not show enough effort. In tests like PISA, categorized as low-stakes, opinions on effective participation can be obtained through self-report surveys. However, these responses should be approached with skepticism. For this reason, alternative sources of evidence have been sought. One such piece of evidence is the response times of individuals to items. Computer-based tests provide the opportunity to collect data on individuals' response times. This allows for different evidence regarding whether individuals are exerting the necessary effort to respond to items. The responses of individuals who, for various reasons, do not use sufficient time are referred to as rapid-guess behavior, and this situation poses a threat to validity (Wise, 2017). Furthermore, such responses lead to biased estimates.

* Turkey Ministry of National Education, Ankara, Türkiye, cagataycoskunmeb@gmail.com, ORCID ID: 0000-0002-4166-3447
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara, Türkiye, duygu.anil73@gmail.com, ORCID ID: 0000-0002-1745-4071

To cite this article:
Coşkun,Ç. & Anıl, D. (2025). Comparing different methods of identifying rapid-guessing thresholds. *Journal of Measurement and Evaluation in Education and Psychology, 16*(3), 160-178. https://doi.org/10.21031/epod.1590685
Received: 24.11.2024
Accepted: 22.08.2025

Rapid-guess behavior is defined as responding quickly to an item without understanding its content (Schnipke & Scrams, 2002). Therefore, it is stated that this behavior can be distinguished by looking at the time individuals spend to answer the item. Responses below the time required to read and understand an item are characterized as rapid-guess behavior, while responses above this time are characterized as solving behavior. An important issue is how to determine the threshold value to be used to distinguish this behavior. There are studies in literature that propose different threshold determination methods.

Identifying Threshold

When considering the threshold value as the time necessary for reading and understanding an item, factors such as the number of words in the item, the inclusion of tables and figures, and other varying conditions must be considered. Therefore, determining a unique threshold value for each item seems more logical. Threshold value determination methods are examined in two parts: normative and graph-based methods.

Normative Methods

In early studies on this subject, methods were used where the same threshold value was given to all items (for example, 3 seconds, 5 seconds). In later studies, the normative (NT10) method, proposed by Wise (2006) and known as the normative method, was frequently used. While the method was initially applied by determining a single threshold value for all items in different studies, it later proposed that 10% of the average response time for all items be set as the threshold value. Although this method allows for setting a threshold value for all items, it is influenced by the distribution of response time variables. When examining the distribution of response times for items in the study, it was found that the distributions were skewed to the right and did not meet the assumptions of normal distribution. For this reason, the NT10 method was applied by performing logarithmic transformations on the response time data, and accordingly, the threshold value was determined. This method was referred to as logNT10 in the study and presented in the tables as such. Additionally, due to the skewed distribution, the NT10 method was applied based on median values rather than averages.

Graph-based Methods

Another method used for determining the threshold value is based on the examination of Kernel density plots of response time data. The visual inspection method, proposed by Wise (2006), can be used in cases where the plot shows a bimodal distribution. When the distribution plot is bimodal, the smallest value between the two peaks is determined as the threshold value (Figure 2). However, this method cannot be used when the plot is unimodal. Therefore, the threshold value can only be determined for a limited number of items. The methods mentioned so far are based solely on response times, without considering individuals' answers.

A similar graph-based method was proposed by Lee and Jia (2014), based on the correctness of the items. In this method, it is expected that rapid-guess behavior and solution behavior will diverge in terms of the response accuracy rates of the item. In multiple-choice items (with four options), it is expected that the response accuracy rate of individuals showing rapid-guess behavior will be below .25, while those exhibiting solution behavior will have a rate above this value. Based on this method, when determining the threshold value, the response accuracy rates over time are shown using a line graph. The second at which the response accuracy rate exceeds chance level is determined as the threshold value. However, in this method, the response accuracy rate can exceed the chance level within the first few seconds. In such cases, a threshold value cannot be determined for these items. Additionally, in some studies, the first second in which the response accuracy rate starts to increase continuously can also be determined as the threshold value.

To address the limitations of this method, Guo et al. (2016) proposed the cumulative proportion (CUMP) method. Unlike the previous method, this approach considers the response accuracy rate for a given time value and all the times below it, rather than specific time segments. The second at which this value starts to exceed the chance level is accepted as the threshold value. However, it is not always possible to determine a threshold value for all items using this method, as the response accuracy rate can surpass the chance level from the very first second.

Another issue is how to handle individuals who exhibit rapid guessing behavior once a threshold value has been determined. In some studies, such responses can be completely excluded from the test, or they can be treated as missing data, with estimations made using various methods. When rapid guessing responses are considered as missing data, methods such as multiple imputation or maximum likelihood estimation can be used. Wise & DeMars (2006) effort-moderated IRT model is an example of this. In this study, we used IRTree models, which divide the response process into a series of decision-making processes, to address rapid response behavior.

IRTree Models

One of the model-based approaches is the Item Response Tree (IRTree) Models. In the literature, they are also referred to as Multi-Process Item Response Theory Models. The development of these models was pioneered by the works of Böckenholt (2012) and De Boeck and Partchev (2012). Unlike standard IRT models, Item Response Tree models are based on the idea that the process of responding to an item occurs through a series of decision-making stages. These models analyze the process by identifying nodes within an item, representing this stepwise structure through a tree diagram.

These models are used to model response styles in surveys or Likert-type measurement instruments. They can also be applied to examine the non-negligible missing data that emerges in achievement tests. The tendencies of individuals to leave items blank or their inability to reach the final items due to poor time management can be modeled (Debeer et al., 2017; Huang, 2020). Another issue that arises in achievement tests is that individuals may fail to fully reflect their abilities due to reasons like lack of motivation. It is noted that this situation is more likely to occur in tests that have no significant consequences in individuals' lives, referred to as low-stakes tests. In such cases, separating these types of responses from the intended characteristics being measured by utilizing response times is possible with IRTree models (Leventhal & Pastor, 2024).

In the study, based on response times, individuals exhibiting rapid guessing behavior were coded as 0, incorrect responses as 1, and correct responses as 2, and this dataset was included in the specified linear model. Through IRTree models, the tendencies of individuals to engage in rapid guessing behavior can be separated as a distinct variable, allowing ability estimations to be modeled more accurately.

Where T_i is the threshold value for item i and RT_{ij} is the duration of the response of person j to item I. By comparing the responses of individuals with the help of the determined thresholds, the response time data is converted into binary coded data (Equation 1).

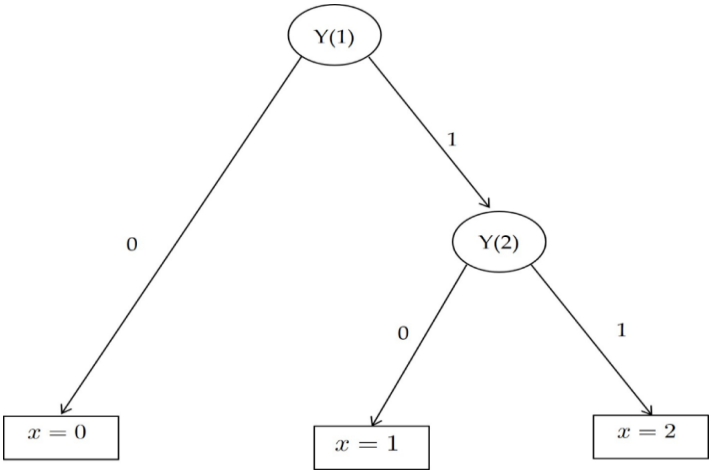
$$SB_{ij} = \begin{cases} 1, \text{if } RT_{ij} \geq T_i, \\ 0, \text{otherwise.} \end{cases} \quad (1)$$

Here, the SB_ij index represents solving behavior when it is 1, and rapid-guessing behavior when it is 0. According to this index, the responses given by individuals to the items have been transformed into three categories. Figure 1 shows the tree diagram proposed by Leventhal and Pastor (2024). Here, sub-item Y(1) represents a variable that indicates the individual's tendency to exhibit solving behavior, while sub-item Y(2) reflects the characteristic intended to be measured by the item. Thus, the responses to an item are divided into two sub-items (pseudo items), and the response process is illustrated through a tree diagram.

IRTree models are flexible and informative. Since these models are generalized linear mixed models (GLMM), their estimates can be obtained using the lme4 R package (Bates et al., 2015). Thus, the ability parameters for both nodes are obtained. In addition, model-fit statistics such as Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) can be obtained.

Figure 1

Linear Response Tree for Three Response Categories



Significance of the Study

Identifying guessing behavior is important for the validity of the measurements. However, evidence must be provided to ensure that the threshold values used to distinguish this behavior based on response times are valid. It is considered crucial to test the proposed different methods on real data, especially data from low-stakes tests like PISA. Another issue is how to handle the responses marked as rapid-guessing behavior, as such responses can affect the psychometric properties of the items and the estimations to be made. This negatively impacts both the test development process and subsequent stages. Low-stakes large-scale assessments such as PISA examine group performance rather than individual performance. Rios & Soland's (2022) findings indicate that different test-taking efforts are effective in intra- and inter-country comparisons. Therefore, identifying and addressing aberrant behaviors such as rapid guessing is important for large-scale assessment projects.

When looking at studies comparing threshold determination methods, the following methods are generally used:

- Guo et al. (2016) CUMP, Visual Inspection, common thresholds (5 sec, 8 sec)
- Wise et al. (2019) NT10, CUMP, Visual Inspection, Visual inspection with response accuracy, Change in information
- Sahin & Colvin (2020) NT10, common thresholds
- Rios & Guo (2020) common threshold (3 sec), NT10, Hybrid (CUMP + MLN)
- Kroehne et al. (2020) CUMP, NT10, NT20, Visual Inspection
- Soland et al. (2021) CUMP, MLN, NT10 and NT30
- Bulut et al. (2023) NT10, NT15, NT20, NT25, Random Search, Genetic Algorithm

In this study, the aim is to comprehensively compare the methods used in existing studies. Therefore, CUMP, Visual Inspection, Visual Inspection with response accuracy, NT10, median10 (Rios & Guo (2020), and logNT10 methods were used. The median10 and logNT10 methods here are different applications of the NT10 method. The logNT10 method proposes an easily applicable method for existing nonparametric methods.

In this study, instead of removing the flagged responses from the test or coding them as incorrect, we applied the item response tree model, which we believe better models the response process. In doing so, evidence for the validity of the determined threshold values can be obtained by comparing the fit of data sets created using different threshold determination techniques. The aim is to evaluate threshold values

at the model-level (like in the study by Bulut et al. (2023)). Here, it is important to utilize IRTree models, which differs from other studies.

Additionally, the consistency between model fit and the validity evidence proposed by Wise & Kong (2005) will be examined. In line with these objectives, the following questions are addressed in this study.

1.How are thresholds established to distinguish rapid-guessing behavior using different methods?

2.Which of these threshold values fits better with the item response tree (IRTree) model?

Method

Sample

The data used in the study belong to the PISA 2018 application and were obtained from the cognitive item data file and cognitive items total time/visit data file sections of https://www.oecd.org/en/data/datasets/pisa-2018-database.html#data. Considering the number of individuals involved, the data for the M02 and M03 mathematics items from booklet 2 were selected. Since response times were needed for the items, only individuals who took the test on a computer were included in the study. Additionally, to avoid variations in word count, only individuals who took the test in English (LANGTEST_COG = 313) were selected. Consequently, analyses were conducted on the responses and response times for 18 items from a total of 3329 individuals. The contents of the items could not be accessed.

Table 1

Item Formats and Clusters

Item Type	M02	M03
Simple multiple-choice	5	3
Complex multiple-choice	3	3
Open-ended	1	3

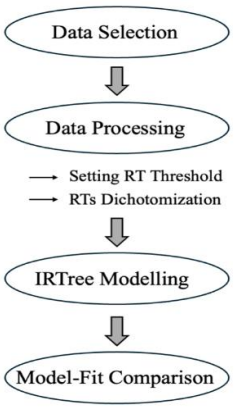
The types of the 18 items analyzed in the study are given in Table 1. In some methods used to determine the threshold value, the chance success of individuals was accepted as .25 in multiple-choice items and .00 in the other two types.

Data Processing

Following the selection of the data for analysis, it must be processed for the IRTree model. The analysis steps of the study are illustrated in Figure 2.

Figure 2

Flow Chart of the Study



Setting RT Thresholds

In this study, six different RT-based threshold determination methods were applied. The methods used were normative threshold (NT10) (Wise & Ma, 2012) and its two different versions (Median10, logNT10), visual inspection (VI) (Wise, 2006), response accuracy visual inspection (VITP or RTRA) (Lee & Jia, 2014), cumulative proportion (CUMP) (Guo et al, 2016). These methods are easy to apply in large-scale computer-based assessments (where RT data is obtained) (Bulut et al., 2023). The selected methods focus on different aspects of the data. Normative methods are determined according to the statistical properties of RT values (mean, median), while the visual inspection method is determined according to the statistical properties of the RT distribution. Other methods consider response accuracy information.

Wise et al. (2004) method, which determines a common threshold value for all items, was not preferred because it is a limited method. The item surface information (Wise & Kong, 2005) method could not be used because the item content could not be accessed. In this method, the threshold value is determined depending on the number of characters in the items. For example, if the number of characters is less than 200, 3 seconds, if it is 1000 characters or more, 10 seconds).

The determination of threshold values was informed by the response time (RT) data with the TT extension of the data set. The descriptive statistics for these variables were calculated. In the NT10 method, 10% of the mean of each RT variable was determined as the threshold value. In the Median10 method, 10% of the median values of the RTs were designated as threshold values. In the logNT10 method, the RT variables were subjected to log-transformation, after which the NT10 method was applied. It was thus possible to approximate the distribution of the RT variables to a normal distribution.

Researchers utilized R codes in graph-based methods. In the visual inspection method, threshold values were marked in bimodal RT variables using density graphs (e.g., Figure 3). In the VI with Response Accuracy method, RT variables were divided into 1-second time segments and converted into categorical variables. The rates at which subjects responded correctly to the item were calculated in these time segments. Line graphs were created for these rates (e.g., Figure 4). In the CUMP method, time segments were determined cumulatively (e.g., 0-1, 0-2, 0-3, etc.). These time segments were then converted into categorical variables, and the correct response rates for each segment were calculated. Consequently, the graphs were prepared as illustrated in Figure 5.

Based on the determined thresholds, the response times of individuals were coded as 1-0, as shown in Equation 1. In this way, six different data sets were created, with the scores (1-0) individuals received from the items being the same. Next, the individuals' responses were converted into three categories by examining the binary time variables. For example, if the time variable for an item was coded as 0 (indicating rapid-guessing behavior), the score for that item was coded as 0, regardless of whether the response was correct or incorrect. If the time variable was coded as 1 (indicating solution behavior), the individual's response to that item was coded as 1 if incorrect, and 2 if correct.

According to the tree diagram used in Figure 1, if the first node Y(1) is 0, the second node does not receive a score and is therefore marked as NA. In this case, a response matrix with three categories, as shown in Table 2, is created. In the table, x=0 indicates observed rapid-guessing behavior, x=1 indicates solution behavior with an incorrect answer, and x=2 represents a correct answer. Y(1) represents the inclination towards solution behavior, while Y(2) represents the measured latent trait.

Table 2

Mapping Matrices T for the Linear Tree

	Y (1)	Y (2)
x = 0	0	NA
x = 1	1	0
x = 2	1	1

The response matrix was arranged as 0, 1, and 2, then converted to a long form for the IRTree model. For this purpose, the *dendrify* function, which is part of the *irtrees* package, was utilized (De Boeck & Partchev, 2012). An exemplification of this conversion is provided in Table 3.

Table 3

Example of a Wide Form

Person	Item Responses			Y (1)			Y (2)		
	1	2	3	1	2	3	1	2	3
1	0	1	2	0	1	1	NA	0	1
2	1	2	0	1	1	0	0	1	NA
3	2	1	0	1	1	0	1	0	NA
…									

As demonstrated in Table 3, Person 1's response in the first item is categorized as rapid guessing behavior. Therefore, Y(1) is 0, and there is no response in the Y(2) sub-node (NA). In general, Y(1) consists of 0s and 1s, indicating rapid guessing and solving behavior. On the other hand, Y(2) denotes the measured trait, and rapid guessing behaviors manifest as NA.

Data Analysis

After the data were prepared for IRTree, model calibration was performed as follows.

$$P(x = 1|\theta_p) = \frac{exp(\theta_{pr} + \beta_{ir})}{1 + exp(\theta_{pr} + \beta_{ir})} \quad (2)$$

Within the context of an IRTree, the internal nodes are modelled as IRT models. In general, the probability that test taker p passes through a specific branch leaving internal node n of item i is modelled as a combination of two factors. Firstly, the test taker's location θ_{pr} is considered, and secondly, the item's location (or threshold) β_{ir} on the latent propensity scale that is relevant for node n (Debeer & De Boeck, 2017).

The probabilities of occurrence of the observed categories 0, 1 and 2 according to the IRTree model are calculated as in the given equations (De Boeck & Partchev, 2012). Here, $-\theta_{p1}$ can be interpreted as the tendency of individuals to show rapid-guessing behaviors. θ_{p2} indicates the ability of the trait to be measured. Likewise, β_{i1} and β_{i2} are the item's rapid guessing threshold and the item's difficulty, respectively. Item's rapid guessing threshold can be interpreted such as item difficulty in sub-node 2.

$$P(x = 0|\theta_p) = P(y(1) = 0|\theta_{p1}) \quad (3)$$
$$P(x = 1|\theta_p) = P(y(1) = 0|\theta_{p1}) * P(y(2) = 0|\theta_{p2}) \quad (4)$$
$$P(x = 2|\theta_p) = P(y(1) = 1|\theta_{p1}) * P(y(2) = 1|\theta_{p2}) \quad (5)$$

The different data sets obtained were included in the same IRTree model (Figure 1). The model was estimated with the *'lme4'* (Bates et al., 2015) and 'irtrees' (De Boeck & Partchev, 2012) packages. The codes of the created model are attached.

In this study, the fit indices of the data to the model, Akaike's information criterion (AIC), Bayesian information criterion (BIC) and likelihood-ratio (LR) statistics were compared. The smaller AIC and BIC values obtained from the models indicate that the data fit the model better. It is stated that the LR

test can be used when there are models of the same size (same number of internal nodes) and data sets of the same size (Jeon & De Boeck, 2016). Since the IRTree model used in the study is the same and the data set with the same number of observations is used, this statistic is used. Model 1 LL_0 has and parameter number p_0, Model 2 has LL_1 and parameter number p_1, degrees of freedom $p_1 - p_0$ and $\chi^2 = -2.(LL_0 - LL_1)$ statistic is compared at $p = .001$ significance level.

Results

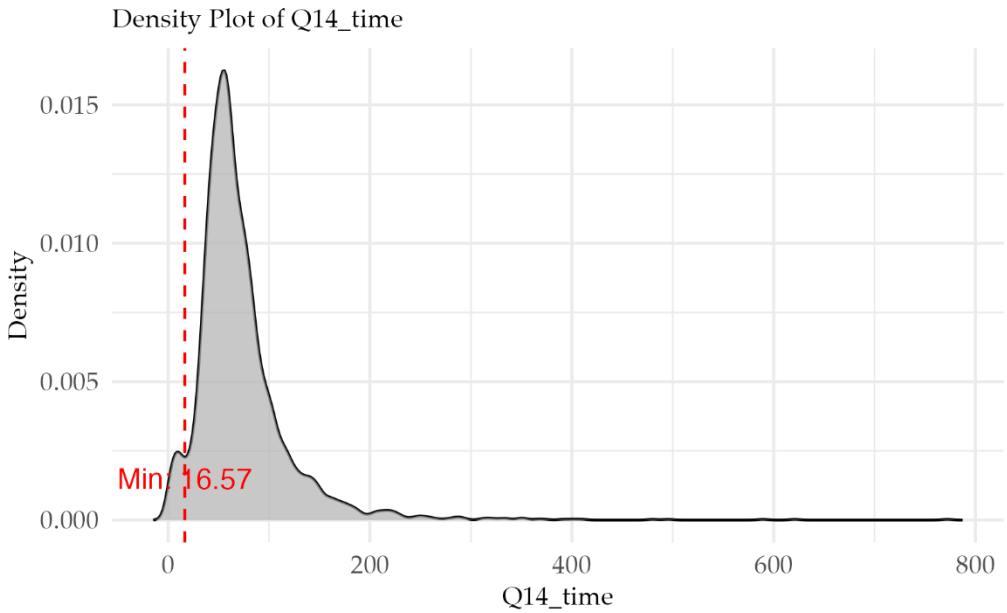
In this section, firstly, the findings regarding the thresholds obtained with different methods and then the findings regarding the fit of the data sets created according to these thresholds to the tree-based item response model are presented.

Results on the Identifying Thresholds

According to the method where the threshold value is determined based on the density graph, it is assumed that there is a difference between the rapid-guessing behavior and the solving behavior in terms of the intensity of the durations. Accordingly, in cases where rapid-guessing behavior is observed, a bimodal structure is identified. When determining the threshold value based on this method, a function is written where the minimum value between the two local peaks corresponds to the threshold value on the x-axis.

Figure *3*

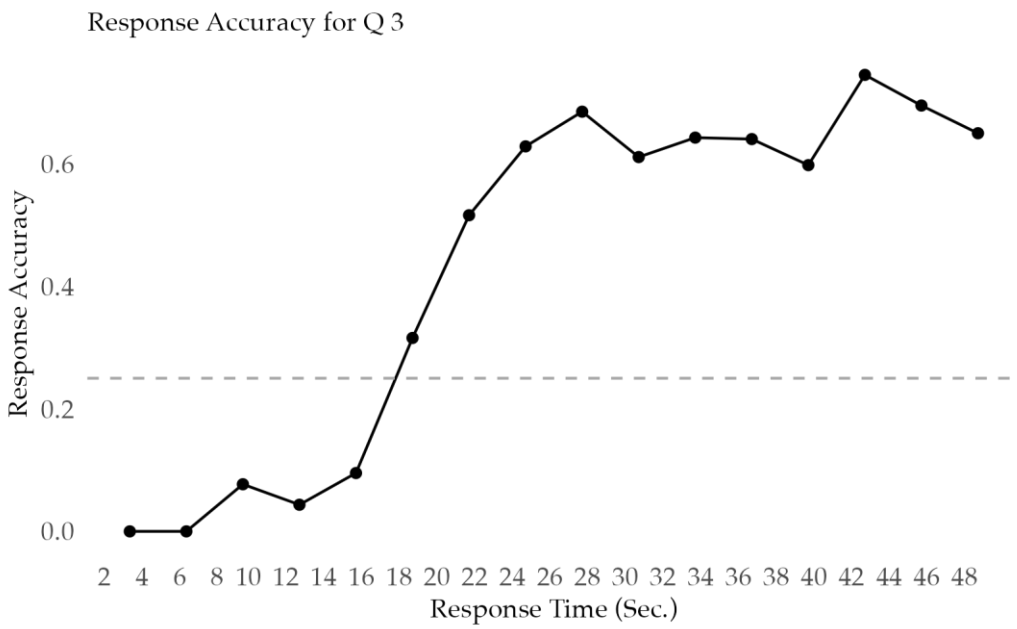
Visual Inspection Method



An example of this is presented in Figure 3. The red dashed line in the figure is thought to represent the distribution of rapid-guessing behavior on the left side and solution behavior on the right side. The determined threshold value in this case is 16.57 seconds. A similar distribution was only obtained in five items, so it was not possible to determine the threshold value for other items using this method.

Figure 4

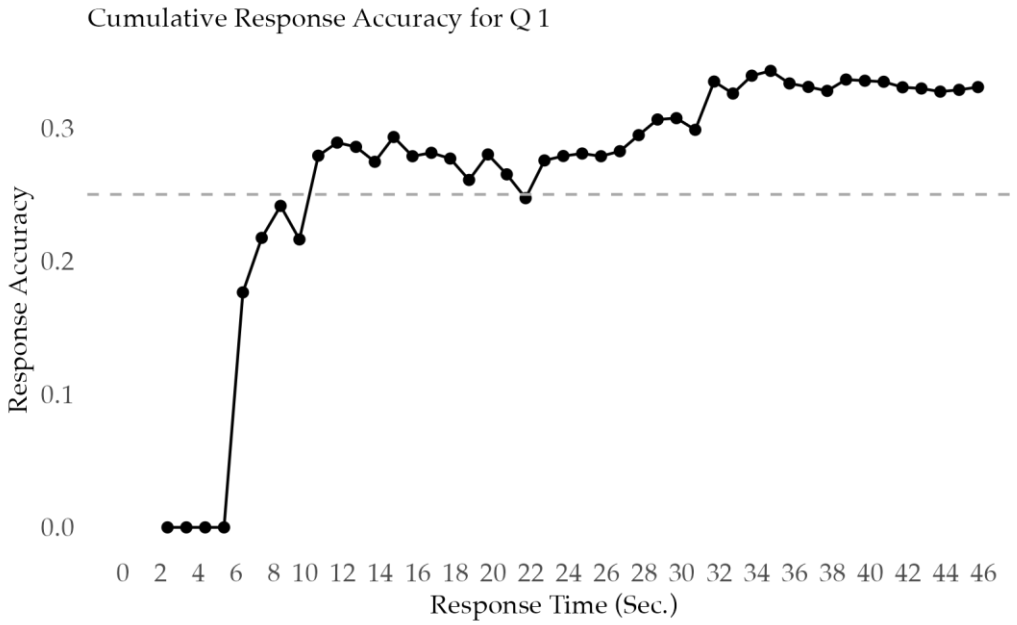
VI with Response Accuracy



In the graph-based method, where only the proportion of correct responses within specific time intervals is considered, the time point at which the correct response rate surpasses chance success is determined as the threshold value. As seen in Figure 4, in Item 3, which is open-ended, the threshold value was set at 6.5 seconds, the point at which the correct response rate exceeded the chance success of .00.

Figure 5

Cumulative Proportion (CUMP) Method



Another graphical method is the threshold determination method based on cumulative response accuracy (CUMP). The accuracy of responses within all response times below a specified time value is calculated and plotted on a graph. As a result, the time point where response accuracy starts to exceed chance level is determined as the threshold value. For example, in Figure 5, since the item is multiple-choice, the chance success is assumed to be .25, and the threshold value is set at 10.5 seconds.

In the NT10 method, which is the most frequently used method for determining the threshold value, 10% of the averages of the response time variables is set as the threshold value. With this method, the threshold value for all items can be determined. In the method called Median10, like the NT10 method, the median values of each item were calculated, and 10 percent of this value was determined as the threshold value. The reason for using this method is the distribution of the response time data. Since the reaction time variables show a right-skewed distribution, it was thought that it would be appropriate to calculate this method based on the median value.

Additionally, based on the distribution of response times, the data transformations were conducted using a logarithmic transformation to approximate a normal distribution, with 10% of the means determined as the threshold value. For this reason, this method has been named logNT10. When examining the distributions of variables related to response time through graphs, it was observed that all of them were right-skewed. Therefore, due to the absence of negative values and the extreme right skewness of the data, it was deemed appropriate to apply the logarithmic transformation (Akbaş & Koğar, 2020). This approach has ensured that the data is closer to a normal distribution.

According to the methods used, item-specific threshold values have been determined. While the density graph-based method was able to establish threshold values for only 5 items, other methods yielded threshold values for all items. According to Table 4, when examining the means of these threshold values, it can be observed that, apart from the logNT10 and density graph-based methods, the other methods provided similar threshold values.

Table 4

Results for the Threshold Identification Methods

Threshold Detection Method	Number of Threshold Identified	Mean	Sd
NT10	18	7,954	1,881
Visual Inspection (VI density)	5	20,606	3,842
VI with Response Accuracy	18	6,528	3,483
Cumulative Proportion (CUMP)	18	6,750	2,353
Median10	18	6,923	1,770
LogNT10	18	2,588	1,064

It is considered that the average of the threshold values determined by the density graph-based method, which is around 20 seconds, is not a suitable value for the threshold. In some studies, using the NT10 method, when the mean exceeds 10 seconds, the threshold is limited to 10 seconds (Wise & Ma, 2012). Therefore, this value is regarded as higher than expected. The average of the threshold values obtained using the logNT10 method corresponds to 2.59 seconds. When examining the averages of the thresholds obtained by other methods, it is observed that lower thresholds have been established.

Validation Thresholds

To verify that the threshold values determined by various methods can accurately distinguish rapid-guess behavior, the method proposed by Wise & Kong (2005) was applied. According to this method, the average of individuals exhibiting rapid-guess behavior must be below chance level. In this study, the multiple-choice items were controlled at .25, while the open-ended items were controlled at .00.

Figure 6

Accuracy Rates for NT10 Threshold

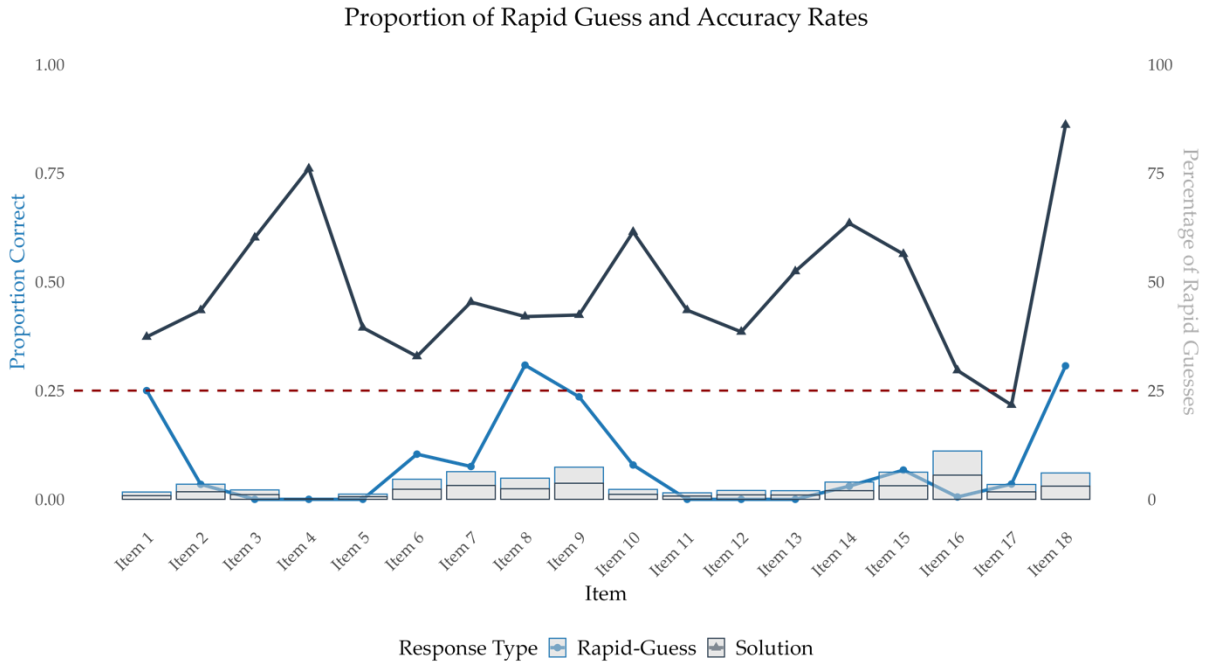


Figure 6 shows the averages of individuals categorized according to the NT10 method. Accordingly, for individuals exhibiting rapid-guess behavior, Items 8 and 18 (multiple-choice) achieved an average result above chance level. Additionally, Item 6 (open-ended) was also answered correctly above chance level. Therefore, it can be stated that this method did not adequately distinguish rapid-guess behavior for these three items.

Figure *7*

Accuracy Rates for LogNT10 Threshold

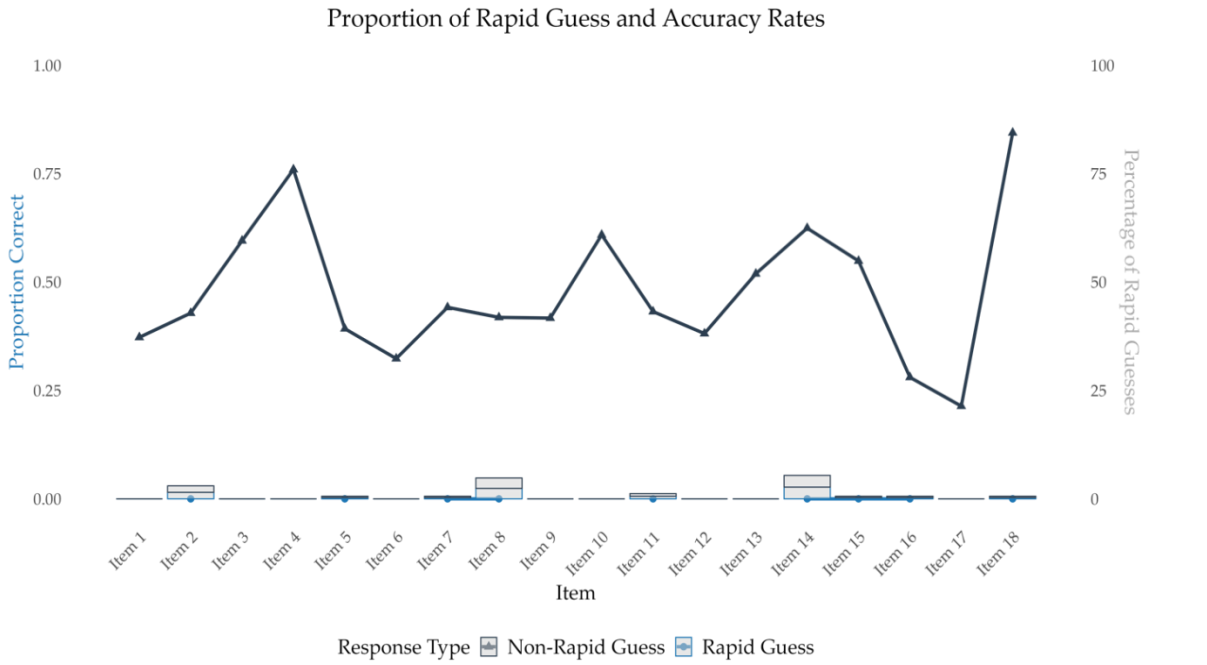
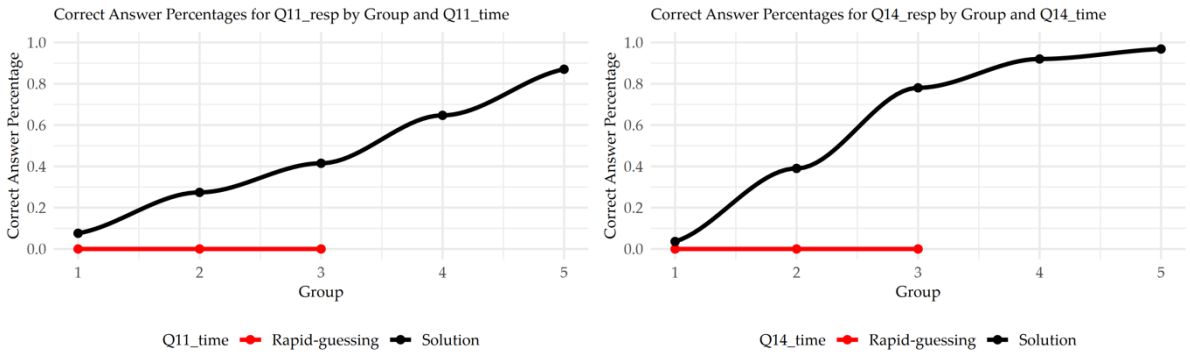


Figure 7 presents the averages of individuals categorized according to the logNT10 method. Although threshold values were determined for all items in this method, rapid-guess behavior was not identified for some items. The averages for the identified items emerged as .00, and the averages for individuals exhibiting solving behavior were observed to be above chance level. Therefore, it can be stated that this method is more effective in distinguishing rapid-guess behavior.

As an additional method for ensuring the validity of the results, the approach employed by Lee and Jia (2014), which involved examining the correlation between the total score and item correct response rates, was utilized. According to this method, a positive correlation is anticipated to be observed between the total scores of individuals exhibiting solving behavior and their performance on the item. Conversely, no correlation is expected to be observed between the total score and item performance of individuals characterized as rapid-guessing behavior. Additionally, it is also expected that individuals in the upper score groups would not display rapid-guessing behavior.

Figure 8

Relationship Between Students' Total Scores and Their Performance on A Selected Item

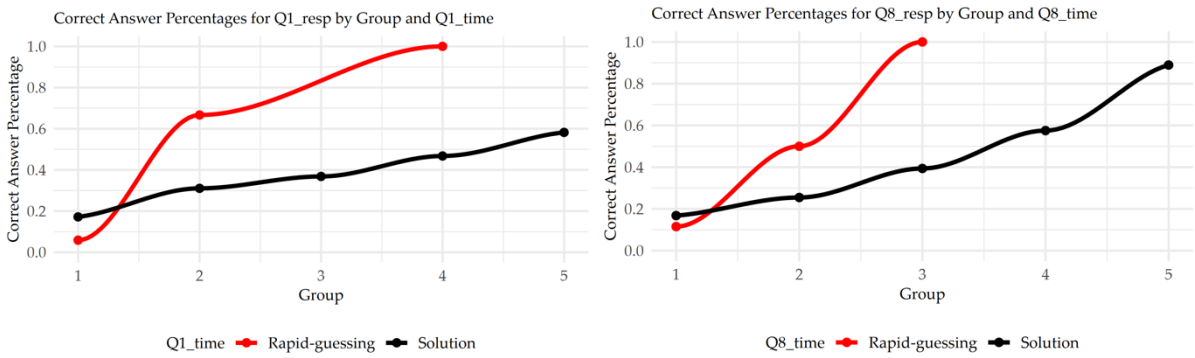


Note: The graph on the right depicts Item 14 categorized by logNT10, while the graph on the left illustrates Item 11 categorized by Median10.

Figure 8 illustrates the items in which the anticipated outcomes were discernible. Figure 8 shows the correct response rates for items 11 and 14, separated by median10 and logNT10 methods. As expected, these items do not show fast guessing behavior in the upper score groups, but solving behavior increases. These items were evaluated as suitable items. The anticipated graphical representations were generated for all items categorized according to the logNT10 method. However, the anticipated graphs could not be generated for some items using methods other than logNT10.

Figure 9

Relationship Between Students' Total Scores and Their Performance on A Selected Item



Note: The graph on the right depicts Item 8 categorized by CUMP method, while the graph on the left illustrates Item 1 categorized by NT10 method.

As illustrated in Figure 9, the proportion of individuals exhibiting rapid-guessing behavior is higher than that of the solving category. Furthermore, it is anticipated that these rates are not associated with the total score within the rapid-guessing category. In this case, *Item 8* (in the CUMP method) and *Item 1* (in the NT10 method) have been assessed as unsuitable items.

Table 5

Proportion of Suitable Items

Threshold Detection Method	Proportion of Suitable Items
NT10	0.67
Visual Inspection (VI density)	0.17
VI with Response Accuracy	0.50
Cumulative Proportion (CUMP)	0.56
Median10	0.67
LogNT10	1.00

When examined according to the method of Lee and Jia (2014), the proportion of suitable items is given in Table 5. According to Table 5, it can be said that the NT10 and Median10 methods can distinguish solving behavior and rapid guessing behavior in 12 items (67%). The logNT10 method is successful in all items. However, at this point, it should be considered that the logNT10 method does not mark rapid guessing behavior in some items (see Figure 7).

Results on the IRTree Model

The model fit statistics for the data sets created using various methods are presented in Table 6. According to the data obtained, the lowest AIC and BIC values were found in the data set generated by applying the normative method after the logarithmic transformation of response times. Conversely, the data created using the NT10 method exhibited the worst fit.

Table 6

Goodness-of-fit Measures for the IRTree Model

Method	AIC	BIC	-LL	Deviance	df
NT10	74756.88	75134.56	-37339.44	74678.88	118629
VI	73557.42	73935.10	-36739.71	73479.42	118632
Accuracy	73870.09	74247.84	-36896.04	73792.09	118855
CUMP	73889.43	74267.20	-36905.72	73811.43	118898
Median10	74022.33	74400.07	-36972.16	73944.33	118836
LogNT10	69559.35	69937.41	-34740.68	69481.35	119776

The LR test $\chi^2(1144;\ .001) = 3998.06)$ concluded that the data set obtained with the logNT10 method provided a significantly better fit than the data set created based on the density plot.

Discussion

Setting a Threshold

The threshold determination methods used in the study can be categorized under two headings. The first includes methods commonly used in this field, referred to as normative threshold methods, while the second consists of methods based on graphical analyses.

Normative Methods

Although normative methods were initially used to determine a common threshold value with the same value for all items, they have gradually evolved into item-specific threshold determinations. According to this method, 10% of the average response time of an item is determined as the threshold value. This method, known as NT10, is frequently used due to its ease of application and its capacity to generate threshold values for all items. However, upon examining the distributions of response time data, it was found that they do not conform to a normal distribution, as nearly all of them exhibit right-skewed distributions, making it inappropriate for these data to be represented by the mean. Therefore, versions that approach normal distribution through logarithmic transformation (logNT10) and those that use the median value (median10) were employed. It is stated that the most suitable measure of central tendency for the distribution of time data, such as response times, is the median (Rousselet & Wilcox, 2020). Various studies in the literature utilize the median value (Sahin & Colvin, 2020; Michaelides et al., 2020). Hadiana et al. (2021) propose a threshold determination method that utilizes standard deviations while employing the median value. In their study, which compares it with the visual inspection method, it is stated that the proposed method yields better results.

When comparing these three normative methods, the best results were obtained from the logarithmic transformation. The method based on the median provided better model fits than the classical NT10 method, which is based on the mean. Overall, when examining normative methods, ease of application and the ability to generate threshold values for each item are significant advantages. However, when using these methods, it is advisable to check the distributions of response time variables and perform logarithmic transformations to approach normality.

Graph-based Methods

The other threshold value determination methods used in the study can be categorized based on graphical analysis. The first of these methods is the density graph-based (visual inspection) method. In this method, it is assumed that the density graphs of response time variables exhibit a bimodal distribution, representing rapid-guess behavior and solving behavior. The value between these two modes is determined as the threshold value. However, this method has several limitations. This hypothesized distribution may not always be encountered. In the conducted study, only 5 out of 18 items exhibited this distribution. To address these limitations, Rios & Guo (2020) propose a new method. By performing a logarithmic transformation of the distribution graphs, they determine the lowest point between the two modes as the threshold value. In this respect, the method called MLN (mixture log-normal) can be considered a different version of the visual inspection method, using logarithmic transformation. It is stated that the threshold values obtained as a result of this transformation yield better outcomes.

The other two methods based on graphical analysis are those that determine threshold values by examining the response accuracy of individuals at specific time intervals. These graphical analysis methods have several disadvantages. First, they tend to contain relatively subjective judgments. The VI with response accuracy (or RTRA) method may contain sparse observations in the determined time segments. Therefore, the response accuracy graph may fluctuate within an RT interval. In this case, the threshold value determination may be based on subjective judgments (Guo et al, 2016).

In some studies, the value at which correct responses begin above chance success is accepted, while in others, the value at which correct response rates exceed zero is considered. Similarly, the point at which the correct response rate starts to continuously increase may be accepted as the threshold value. Another aspect that can be interpreted differently in application is the width of the determined time segments. In some studies, time intervals are created by rounding response times to the nearest whole number, while in others, intervals of 3 or 5 seconds are used. This can lead to different applications in determining the threshold value. Another disadvantage of these methods is that they do not always allow for the determination of threshold values. In the resulting graphs, the correct response rates may completely fall above or below chance success, making it impossible to establish a threshold value in such cases. This situation is attributed to item difficulty. In such cases, it has been suggested that the threshold value be determined as the point where the graph begins to show a trend towards item difficulty (Guo et al., 2016). Another reason for this issue is the correct answer options of the items. It has been shown that individuals tend to choose option B, which is in the middle, more frequently in four-option items (Wise,

2017; Wise & Kuhfeld, 2020). In cases where the correct answer is this option, individuals exhibiting guessing behavior may achieve correct response rates above chance success. Therefore, this method may also fail to determine threshold values for all items. While threshold values can be determined for all items in this study, sufficient evidence has not been provided in terms of model fit or validity.

In general, the methods based on graphical analysis yielded similar results in terms of model fit. However, while the density graph yielded higher threshold values, the two methods based on correct response rates produced lower and similar values. When normative methods and methods based on graphical analysis are compared in general, graphical methods fit the model better than NT10 and median10 methods.

One of the findings of the study is that lower threshold values yield better results in terms of model fit. The threshold values obtained using the logNT10 method produced lower values compared to other methods. A review of studies comparing threshold determination methods indicates that no consensus has been reached on this issue. While some studies suggest that the threshold should be limited, others indicate that more liberally determined threshold values yield better results. For example, Setzer et al. (2013) propose that the threshold in the NT10 method should not exceed 10 seconds. In their study, which employs the threshold determination method based on distribution graphs, they note that values as high as 29 seconds for certain items are excessive for characterizing rapid-guess behavior. Therefore, they used a maximum threshold duration of 10 seconds. Considering that in this study, the determined threshold values are around 20 seconds, it can be stated that these values are not appropriate. This method can be applied at different values such as NT15, NT20, and NT30. Soland et al. (2021) compared different threshold determination methods and noted that the NT30 method (i.e., 30% of the mean) is a different version of the NT10 method, along with the MLN, NT10, and CUMP methods, indicating that these methods produce different threshold values. Soland et al. (2021) and Rios (2022) suggest using methods that allow for more flexible threshold determination. However, according to the findings of this study, the logNT10 method produced lower threshold values, and the data set obtained with this method showed the best fit for the model. On the other hand, additional evidence may be required to say that this time represents the minimum time required to read, understand the content of an item, and respond. Because it cannot be compared with information such as the content of the item and the number of words. Wise (2017) states that response time thresholds are typically much shorter than the time a test taker would require to read and comprehend an item.

In addition, the cognitive level measured by the items can also be considered when determining the threshold value. The items selected from the PISA application range from Level 1 to Level 5. These levels can also be considered when interpreting the threshold values.

The determination of the threshold value may vary depending on the measurement needs. According to Wise (2019), if the aim is to interpret individuals' test scores, more conservative threshold values may be determined. If the aim is item calibration, more liberal threshold values may be used.

When examining threshold determination methods in terms of producing similar results, it has been found that the response accuracy, cumulative (CUMP), and median10 methods generated threshold values with similar averages. In their study, Kong et al. (2007) compared methods based on three seconds per item, word count, and distribution graphs, and the results indicated that different methods produced similar outcomes.

In summary, it cannot be said that there is an ideal single method for determining the threshold values to distinguish rapid-guess behavior. Wise (2017) suggests using multiple methods for threshold determination. It would be more appropriate to examine individuals not only based on response times but also on response accuracy and item discriminability. Additionally, this study has shown that analyzing the distributions of response time variables and applying logarithmic transformations for right-skewed distributions will yield better results.

Results on Model Fit

Once individuals exhibiting rapid-guess behavior are identified, how to handle these responses is another matter. Since such behaviors pose a validity threat, it is important to make predictions without including these responses. In some studies, efforts have been made to remove these responses from the data set (Bulut et al., 2023; Guo et al., 2016). It has been concluded that this intervention improves model fit. In their study comparing threshold determination methods, Bulut et al. (2023) evaluated the

effectiveness of the NT10, CUMP, and MLN methods, as well as the random search and genetic algorithm methods proposed by researchers. According to the results, removing responses that involve rapid-guess behavior yields better model fit, and the proposed methods are noted to be more advantageous than the existing methods.

Tree-based IRT models allow analyses to be conducted without removing these responses. In the study, a tree diagram was used to differentiate between rapid-guess behavior and solving behavior. Thus, the fit values of this model were accepted as evidence for the validity of the threshold determination method. Additionally, methods were compared as proposed by Wise & Kong (2005). As a result of this comparison, it was observed that the threshold values determined using the logNT10 method produced more valid outcomes. Therefore, model fits were presented along with additional validity evidence.

Limitations and Further Research

Lack of knowledge about the content of the items used in the study can be considered a significant limitation. This is because understanding the content of the items is essential for determining threshold values to distinguish rapid-guess behavior. By definition, the threshold value can be influenced by the word count of the item and whether it contains tables and figures. Therefore, it may not be sufficient for threshold determination studies to rely solely on statistical methods. It is recommended that these studies be integrated into the question preparation process. Consulting field experts could be beneficial in determining the time needed for an individual at a certain skill level to read and comprehend the item. Since the content of the items is unknown, the item surface features method used by Kong et al. (2007) could not be applied. In studies where the items are accessible, this method could also be included for comparison.

Although logNT10 showed the best fit when evaluated at the model level, a limited number of threshold values were flagged in this data set. Therefore, it may be recommended to perform the study on different data sets and with different item types. In addition, the study can be repeated with simulation data that can be produced under different conditions.

The two-state mixture model (Schnipke & Scrams, 1997), which is based on Bayesian methods, was not included in the study. This is because this method assumes that the response time variables have a bimodal distribution. Since very few of the available variables meet this assumption, this method was not utilized in the study. Sahin & Colvin (2020) found that only one of the seven items in their study using PISA 2012 data exhibited a bimodal distribution. In this study, however, the ratio is 5 out of 18. Therefore, it can be stated that bimodal response time distributions are not frequently observed in the PISA data.

Wise (2019) proposes an information-based method for threshold determination that relies on item-total correlation. According to this method, it is assumed that in cases where rapid-guess behavior is observed, this correlation will be lower than that of solving behavior. A threshold value is accepted when this correlation exceeds .20. Similar studies could be expanded to include this method. Generally, information-based methods can be associated with item difficulty and item discriminability. Accordingly, the effectiveness of threshold determination methods is influenced by the psychometric properties of the items. When examining the behaviors of individuals taking the test, the characteristics of the items can also be analyzed, and the method to be used can be determined based on these findings. In summary, when selecting the threshold determination method to identify rapid-guess behavior, it is important to consider the psychometric properties of the items and the distributions of response time data. Additionally, different decisions can be made regarding the method based on the content of the item and its position within the test.

Alfian et al. (2023) used a different version of the common k-second method in their study. According to this method, different common threshold values were assigned based on the difficulty indices of the items. Accordingly, a threshold of 3 seconds was determined for difficult items, while a threshold of 2 seconds was set for medium and easy items. This method was compared with other common threshold determination methods based on word count, and the NT10 method was also used. It was noted that the proposed method produced better results. Based on this, studies could be conducted to compare methods based on item difficulty indices with those based on graphical inspection.

The validity of the determined threshold is important not only in terms of item characteristics but also regarding the position within the test. Goldhammer et al. (2017) state that items at the end of the test

may have larger threshold values. This is because it should be considered that items located toward the end of the test may take longer to read and understand due to test fatigue. Future studies could expand by including other factors, such as the position of the items in the test and log records showing the individual's interaction with the items.

Declarations

This paper was adapted from the first author's doctoral dissertation, and part of it was presented at the 9th International Congress on Measurement and Evaluation in Education and Psychology.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: This study utilized publicly available data. Therefore, ethical approval was not required. The authors confirm that all ethical guidelines were followed.

References

Akbaş, U., & Koğar, H. (2020). *Nicel araştırmalarda kayıp veriler ve uç değerler*. Ankara: Pegem Akademi Yayıncılık.

Alfian, M., Yuhana, U. L., Pardede, E., & Bimantoro, A. N. P. (2023). Correction of threshold determination in rapid-guessing behaviour detection. *Information, 14*(7), 422. https://doi.org/10.3390/info14070422

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods, 17*(4), 665–678. https://doi.org/10.1037/a0028111

Bulut, O., Gorgun, G., Wongvorachan, T., & Tan, B. (2023). Rapid guessing in low-stakes assessments: finding the optimal response time threshold with random search and genetic algorithm. *Algorithms, 16*(2), 89. https://doi.org/10.3390/a16020089

De Boeck, P., & Partchev, I. (2012). Irtrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, 48*(1), 1–28. https://doi.org/10.18637/jss.v048.c01

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using irtrees. *Journal of Educational Measurement, 54*(3), 333–363. https://doi.org/10.1111/jedm.12147

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education, 5*(1). https://doi.org/10.1186/s40536-017-0051-9

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173–183. https://doi.org/10.1080/08957347.2016.1171766

Hadiana, D., Hayat, B., & Tola, B. (2021). A new method for setting response-time threshold to detect test takers' rapid guessing behavior. *Advances in Social Science, Education and Humanities Research, 545*, 46–50. https://doi.org/10.2991/assehr.k.210423.062

Huang, H. Y. (2020). A mixture irtree model for performance decline and nonignorable missing data. *Educational and Psychological Measurement, 80*(6), 1168–1195. https://doi.org/10.1177/0013164420914711

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*(3), 1070–1085. https://doi.org/10.3758/s13428-015-0631-y

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

Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling, 62*(2), 147–177. https://doi.org/10.25656/01:23630

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359–379.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education, 2*(1). https://doi.org/10.1186/s40536-014-0008-1

Leventhal, B. C., & Pastor, D. (2024). An illustration of an irtree model for disengagement. *Educational and Psychological Measurement, 84*(4), 810–834. https://doi.org/10.1177/00131644231185533

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing, 20*(3), 187–205. https://doi.org/10.1080/15305058.2019.1706529

Rios, J. A. (2022). Assessing the accuracy of parameter estimates in the presence of rapid guessing misclassifications. *Educational and Psychological Measurement, 82*(1), 122–150. https://doi.org/10.1177/00131644211003640

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education, 33*(4), 263–279. https://doi.org/10.1080/08957347.2020.1789141

Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing, 22*(2), 154–184. https://doi.org/10.1080/15305058.2022.2036161

Rousselet, G. A., & Wilcox, R. R. (2020). Reaction times and other skewed distributions. *Meta-psychology, 4*. https://doi.org/10.15626/mp.2019.1630

Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-scale Assessments in Education, 8*(1). https://doi.org/10.1186/s40536-020-00082-1

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410612250-20

Setzer, J. C., Wise, S. L., Van Den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-scale Assessments in Education, 9*(1). https://doi.org/10.1186/s40536-021-00100-w

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. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education, 32*(4), 325–336. https://doi.org/10.1080/08957347.2019.1660350

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*, 19–38. https://doi.org/10.1111/jedm.2006.43.issue-1

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

Wise, S. L., & Ma, L. (2012, April 13–17). 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. https://www.nwea.org/resources/setting-response-time-thresholds-cat-item-pool-normative-threshold-method/

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego

Wise, S., & Kuhfeld, M. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. Margolis, & R. Feinberg (Eds.), *Integrating Timing Considerations to Improve Testing Practices*, Routledge.

Appendix

Appendix 1. IRTree codes

```
library("irtrees")
library("lme4")

data1 <- as.matrix(data1)

# define mapping matrix and prepare data for analysis
datamap <- cbind(c(0, 1, 1), c(NA, 0, 1))
dataraspT <- dendrify(data1, datamap)
# long form

summary(dataraspT)

# The multidimensional model for linear response trees
(model1 <- glmer(value ~ 0 + item:node + (0 + node | person),
                 family = binomial, data = dataraspT))
ranef(model1)
```