Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: www.tandfonline.com/journals/hame20

Gauging Misclassification in Rapid Guessing Identification in a Fast-Paced Vocabulary Test

Santeri Holopainen, Jari Metsämuuronen, Mikko-Jussi Laakso & Janne Kujala

To cite this article: Santeri Holopainen, Jari Metsämuuronen, Mikko-Jussi Laakso & Janne Kujala (2025) Gauging Misclassification in Rapid Guessing Identification in a Fast-Paced Vocabulary Test, Applied Measurement in Education, 38:1, 25-42, DOI: 10.1080/08957347.2025.2533124

To link to this article: https://doi.org/10.1080/08957347.2025.2533124

APPLIED MEASUREMENT IN EDUCATION
2025, VOL. 38, NO. 1, 25–42
https://doi.org/10.1080/08957347.2025.2533124

Routledge
Taylor & Francis Group

Gauging Misclassification in Rapid Guessing Identification in a Fast-Paced Vocabulary Test

Santeri Holopainen [a], Jari Metsämuuronen [a], Mikko-Jussi Laakso [a], and Janne Kujala [b]

[a]Turku Research Institute for Learning Analytics, University of Turku; [b]Department of Mathematics and Statistics, University of Turku

ABSTRACT

In low-stakes testing, rapid-guessing behavior (RG) presents a significant challenge to the validity of test scores. This study investigates the misclassification produced by nine different response time (RT) threshold methods in identifying RG, using large-scale assessment data from a fast-paced vocabulary test and introducing choice reaction time (CRT) as a ground truth variable. Although the methods varied mostly in their ability to estimate thresholds at all and not in the misclassification rates nor in their nature, the results show significant misclassification rates across methods, ranging from .080 to .096 (Finnish speakers) and from .087 to .164 (non-Finnish speakers). All methods were more conservative than liberal, with false negatives outnumbering false positives. The findings emphasize the problem of the binary mind-set in RG identification, and suggest that there is a need for approaches that identify RG at the participant-by-item level in order to improve the accuracy of RG identification.

1. Introduction

When administering achievement tests that measure the test takers' knowledge, skills, and abilities (KSAs), the test giver assumes that the test takers show full effort and that their responses to the test items reflect their true KSAs. However, especially when a test is low-stakes from the test takers' perspective, i.e., there are few or no consequences for them, some test takers may be disengaged, showing low effort, and answering in ways that do not reflect their true KSAs (Wise & Kong, 2005). Therefore, test-taker disengagement is considered to be the the main construct-irrelevant factor that can have a negative impact on test score validity in low-stakes testing settings (Wise, 2020). Because of the emergence of digital assessments and the ability to measure the test takers' response times (RTs), disengagement research has focused on differentiating between *rapid-guessing behavior* (RG), a response process in which a test taker gives an answer to an item so quickly that he or she may not have fully considered it, and *solution behavior* (SB), a response process in which a test taker attempts to actively solve the problem presented in an item (Wise, 2017). Furthermore, Wise (2017) proposed that there are three potential conditions for rapid guessing, of which the first is a strategic choice from motivated test takers running out of time in high-stakes tests, the second represents responses from unmotivated test takers in low-stakes tests, and the third represents responses from test takers who recognize quickly that they do not have the required KSAs to answer the presented item and then choose to respond with a rapid guess. Presumably, the response time under the third condition is longer than the response time under the second condition.

CONTACT Santeri Holopainen sjholo@utu.fi Turku Research Institute for Learning Analytics, Faculty of Science, University of Turku, Turku 20014, Finland

Supplemental data for this article can be accessed online at https://doi.org/10.1080/08957347.2025.2533124

In applied research and operational settings, the most common approach to differentiate between RG and SB has been to find appropriate *item-specific time thresholds* using response time (RT) and/or response accuracy (RA; or response correctness) information in item response data (Rios & Deng, 2021; Ulitzsch et al., 2023). In these so-called *response time threshold methods*, a response to an item is classified as a *rapid guess* (RG) if the respondent's RT to the item is less than or equal to the identified time threshold, and as an engaged response (SB) if the test taker's RT to the item is greater than the time threshold (Wise & Kong, 2005). Another type of approach to differentiate rapid guesses from engaged responses in RG research has involved item-response theory (IRT) models that directly incorporate the engagement state (e.g., RG or SB) into the model as a latent class variable that is allowed to vary across and within test takers (e.g., Pokropek, 2016; Wang & Xu, 2015). However, in the current study, our focus was exclusively in the RT threshold methods, because the rationale of the study was to fill a gap in the RG literature and to examine misclassification in RG identification produced by the different response time threshold methods in real assessment data. Therefore, the article is structured as follows. First, we briefly present the RT threshold methods proposed in previous RG research, after which we discuss misclassification in RG identification. Next, we discuss the rationale of the study and present the research questions and the methodology used to address the research questions. Finally, we present the findings and discuss the limitations and implications of the study.

1.1. Response Time Threshold Methods

All of the RT threshold methods proposed in RG research share the assumption that there is a region in the item-specific RT distributions where rapid guesses and engaged responses overlap, and where one can find an appropriate time threshold for differentiating between these responses. The methods can be categorized into three typologies based on the type of information they use from the item response data to identify the thresholds (Rios & Deng, 2021). The first typology consists of the *Common K-time Threshold* (CKT; Wise et al., 2004) and the *Inspection of Item Characteristics* (IIC; Wise & Kong, 2005) methods, neither of which considers the test takers' responses or RTs to the test items, nor any other type of information in the data. Wise et al. (2004) proposed a common 3-second threshold to be used for all items (CKT method) in their computerized adaptive testing (CAT) settings where test items are selected from pools that may contain hundreds or thousands of items. Wise and Kong (2005) based their time thresholds on item characteristics, specifically the length of the item as measured by the number of characters and the amount of additional reading required (IIC method).

The second typology consists of the *Visual Inspection* (VI; Schnipke, 1995) method, the *Normative Threshold* (NT; Wise & Ma, 2012) method with 10% (NT10; proposed by Wise & Ma, 2012) and 30% (NT30; proposed by Wise & Kuhfeld, 2021) thresholds, and the *mixture modeling-based approaches* (MMAs), such as the *Mixture Lognormal* (MLN) method of Rios and Guo (2020), as they use only the RT information to find the appropriate thresholds. Schnipke (1995) conceptualized that the overall RT distribution is bimodal, with a small mode of fast responses representing rapid guesses and a larger mode of slower responses representing engaged responses, and proposed the VI method to set the threshold at the time point where the two distributions intersect. Schnipke and Scrams (1997) continued the work of Schnipke (1995) by fitting a standard two-state mixture model to the RT data, which allowed them to estimate the proportion of rapid guesses as well as the location and scale parameters of the two underlying RT distributions. Later, Rios and Guo (2020) developed the MLN method, which fits a mixed lognormal distribution to the RT data and defines the threshold as the time point between the two normal distributions where the density function of the overall distribution reaches its minimum value. Another approach to differentiate between rapid guesses and engaged responses in CAT settings was introduced by Wise and Ma (2012), who argued that a common 3-second threshold was unacceptable for some items in their testing system. They presented the NT10 method, which sets the time threshold for an item as 10% of the item's mean RT, up to

a maximum threshold. Later, Wise and Kuhfeld (2021) proposed the NT30 method because they found that it performed better than NT10 in estimating retest scores and reducing bias in test scores.

Finally, the methods that use both RT and RA information (third typology) are the *Visual Inspection with Conditional P+ Information* (VITP; Lee & Jia, 2014), *Cumulative Proportion* (CUMP; Guo et al., 2016), *Change in Information* (ChInf; Wise, 2019), and *Change in Information and Accuracy* (ChIA; Wise, 2019) methods, and the *Random Search* (RS; Bulut et al., 2023) and *Genetic Algorithm* (GA; Bulut et al., 2023) approaches. The main assumption underlying VITP and CUMP is that rapid guesses are random, which means that the accuracy of rapid guesses should be (approximately) equal to the chance level. When using VITP, one examines the RT frequency distribution and the plot of the conditional accuracy levels evaluated at each possible RT value simultaneously, and the goal is to find a time point at short RTs where the conditional accuracy level stops fluctuating around the chance level (Lee & Jia, 2014), whereas CUMP uses a mathematical formula to find the maximum RT value at which the cumulative accuracy level is less than or equal to the chance level (Guo et al., 2016). Wise (2019) motivated the development of ChInf and ChIA by the fact that rapid guesses do not reflect the true ability of test takers, meaning that these responses are uninformative in the sense of test score validity. In ChInf, the idea is to find the first time segment after which the item-total correlations start to be consistently above a predefined low value, which is assumed to represent uninformative responses, and then to set the time threshold to the lower limit of that time segment. The ChIA method extends the ChInf method by finding a second time threshold, which is the lower bound of the first time segment at which the conditional accuracy level begins to consistently trend toward the value representing engaged responses. In ChIA, the final threshold is the unweighted average of the two time thresholds. Finally, Bulut et al. (2023) assumed that the presence of RG leads to a weaker IRT model fit to the data and proposed the RS and GA approaches, which search for the time threshold that maximizes the adjusted deviance difference between a baseline IRT model fit to the full data and an adjusted model fit to the data where rapid guesses have been removed based on a predefined threshold.

1.2. Misclassification in RG Identification

In RG identification, balancing the number of false positives (incorrect RG classifications) and false negatives (incorrect SB classifications) is a complicated task because the overlapping nature of the RT distributions of rapid guesses and engaged responses inevitably leads to misclassifications, and one cannot be sure that an RT below a certain time threshold truly represents a rapid guess. Furthermore, many RT threshold methods tend to be conservative because test givers tend to avoid misclassifying engaged responses as rapid guesses, i.e., false positives, resulting in a large number of misclassified engaged responses that should have been rapid guesses, i.e., false negatives (Rios & Deng, 2021; Wise, 2017). This can be problematic when scoring individual test takers, as researchers have found that misclassifying rapid guesses as engaged responses can lead to biased estimates of item difficulty and ability parameters (Rios, 2022a, 2022b; Rios & Deng, 2024) and biased score estimates (Wise & Kuhfeld, 2021).

Researchers have also raised concerns that identified engaged responses may reflect less than full engagement (Demars, 2007; Lindner et al., 2019; Wise & Smith, 2011). Specifically, Wise and Kuhfeld (2021) conceptualized disengagement as manifesting in three forms, namely rapid guessing, semi-rapid guessing, and partial engagement, of which rapid guessing is largely detected by existing RT threshold methods even with conservative thresholds, semi-rapid guesses are given relatively quickly by test takers, but is not detected by conservative thresholds, which is related to the tendency of test givers to avoid false positives, and partial engagement is a response behavior in which a test taker does not respond quickly, but gives less than full effort to the items. Presumably, semi-rapid guessing corresponds to one of the three conditions for rapid-guessing (Wise, 2017), where the rapid guesser quickly glances the item content and then makes a rapid guess. Overall, it appears that the binary

mind-set of setting time thresholds to identify RG is problematic in itself, as disengagement is a more complex phenomenon than the simple binary classifications of responses as RG or SB can capture.

To bring another perspective to this discussion, there is a fundamental problem with thinking about response behaviors, such as RG and SB, from the perspective of the test items and setting *item-specific thresholds*, when, in fact, specific response behaviors are more often manifestations of people's, i.e., test takers,' reactions to external stimuli, i.e., test items. This response-stimulus type of thinking corresponds to the speed-accuracy trade-off, which, as van der Linden (2007) noted, is a within-person phenomenon. In addition, the probability of disengaged responses has been shown to be related to both participant and item characteristics (e.g., Wise et al., 2009). Therefore, it would be more meaningful to examine rapid guessing at the participant-by-item level, i.e., time thresholds would be calculated on the basis of both the item and the participant. For example, assuming that ability and speed are related, time thresholds might decrease for high ability individuals and increase for low ability individuals. Notably, some of the aforementioned IRT models for response engagement account for this in part by allowing the latent class variable for the engagement state to vary between and within individuals. However, they are limited to the IRT framework and cannot be used in other situations that do not involve IRT modeling (Nagy & Ulitzsch, 2022).

While it is known that misclassification is inevitable in RG identification, there appear to be no empirical studies on the rate and/or type of misclassification produced by the different RT threshold methods. In particular, while there have been studies of RG misclassification, their focus has not been on misclassification per se, but rather on simulation studies manipulating the rate and type of RG misclassification to examine the robustness of different RT threshold scoring procedures[1] to RG misclassification (e.g., Rios, 2022a, 2022b; Rios & Deng, 2024). Thus, there appears to be a gap in the RG literature in examining the rate and type of RG misclassification that RT threshold methods produce in real assessment data. This is not surprising, however, because when we only have the information about the participants' RTs and RAs on the achievement test, we cannot be sure of the true engagement state underlying the identified rapid guesses and engaged responses, making it impossible to know which responses were misclassified. Thus, to assess RG misclassification, we would need an external ground truth variable that would tell us the true engagement state of the responses.

1.3. Choice Reaction Time as a Ground Truth Variable in RG Identification

In psychology, reaction time is a widely studied variable. Specifically, there have been two common approaches to measuring reaction time, namely simple reaction time and choice reaction time, the latter of which requires the subject to select an appropriate option from a set of stimuli. Because reaction time items have low cognitive demands and require very little time to solve (Deary et al., 2011), we hypothesize that an engaged test taker's choice reaction time would be indicative of the absolute minimum time it takes him or her to respond to items in a more complex achievement test. This implies that the individual's RTs on the achievement test that are less than his or her choice reaction time would be due to RG. Therefore, we propose choice reaction time, or more specifically, the mean log choice reaction time, as a ground truth variable in RG classification.

2. Research Questions

We found that there is a gap in the RG literature in examining the rate and type of misclassification in RG identification in real assessment data. Using choice reaction time as a ground truth variable in RG classification, we aim to fill this gap in the literature by addressing the following research questions (RQs):

[1]See, e.g., Rios and Deng (2024) for an introduction of the RT threshold scoring procedures.

- Research Question 1 (RQ1): What is the rate and nature of the misclassification in RG identification produced by the RT threshold methods?
- Research Question 2 (RQ2): Are there differences between the RT threshold methods in the misclassification rate?
- Research Question 3 (RQ3): Are there differences between the test items in the misclassification rate?

By the nature of RG misclassification, we mean the balance between the number of false positives and false negatives. That is, do the methods produce conservative thresholds, as has been found in previous research? By answering the RQs presented, we hope to provide further evidence for the presence of misclassification and indications of its nature when identifying RG using any of the RT threshold methods.

3. Method

3.1. Measures

3.1.1. Developmental Lexize

Developmental Lexize (d-Lexize) is based on a test that was originally designed to assess the Finnish vocabulary knowledge of non-native speakers (Salmela et al., 2021). More precisely, it measures vocabulary breadth, i.e., the number of words that the test takers know or can recognize. In d-Lexize, participants are asked to indicate whether a visually presented string of letters is a real word by answering "yes" or a pseudoword by answering "no." In total, there were 102 items, of which 68 were real words and 34 were pseudowords (Bertram et al., 2025). Notably, this version of d-Lexize was the first version of the test. To date, more data have been collected with improved test versions where poorly behaving items have been removed in terms of their psychometric properties, such as item response theory-based item discrimination. The development of d-Lexize is reported in a preprint article by Bertram et al. (2025). As the study by Bertram et al. (2025) has not yet been published, we did not have the permission to reveal the test content and its psychometric properties in more detail.

The test items were presented to each participant in random order. Participants could not move on to the next item until they had given an answer to the current item or a time limit of four seconds had been reached. In addition, they were not allowed to review their answers at any time during the test. Although the test had a time limit of six minutes, we treated it as an unspeeded test because the order of the items was randomized and 90% of the participants were able to complete 90% of the items within the time limit.

3.1.2. Choice Reaction Time Task

In the *Choice Reaction Time Task* (CRTT), participants are presented with two colored squares of equal size 16 times. In each of the 16 items, the squares are either the same color, i.e., black-black or red-red, or different colors, i.e., black-red or red-black, so there are only four variations of different CRTT items. In each item, participants are asked to indicate as quickly as possible whether the squares presented are the same or different. CRTT is a two-choice reaction time task, in contrast to, for example, the commonly used Deary-Liewald task, which is a four-choice reaction time task (Deary et al., 2011).

3.2. Participants

The participants were 6,664 students (3,347 boys, 50.2%) from several different schools in Finland, who participated in the assessments in the spring of 2022. Of the participants, 2,482 (37.2%) were 3rd graders, 2,557 (38.4%) were 4th graders and 1,625 (24.4%) were 7th graders. In addition, the

sample was diverse in terms of the participants' language backgrounds, as 1,062 (15.9%) participants' home language was something other than Finnish and 688 (10.3%) participants' home language was a combination of Finnish and some other language. Gender and home language were self-reported by the students, and grade level was reported by the students' teachers. Some of the participants were removed before the main data analyses. See the Data Analysis section for more details.

All necessary ethical support and approvals are in place for the research. Students participated anonymously and voluntarily during regular school hours, and their parents were informed about the assessments. The Finnish law and the guidelines of the Finnish National Board on Research Integrity TENK were followed carefully throughout the study.

3.3. Procedure

The assessments were administered in ViLLE (Laakso et al., 2018), a collaborative digital learning platform developed by the Turku Research Institute for Learning Analytics (TRILA), University of Turku. The teachers of the participating students administered the assessments during regular school hours. To participate in the assessments, students logged into their own ViLLE accounts via a web browser on their computers or tablets. The ViLLE system collects all user interactions and timings (in milliseconds) for further analysis. CRTT and d-Lexize were given in the same day within one session among some other tasks related to reading and mathematics. CRTT was the first task in the whole assessments and d-Lexize was the second. The tasks were introduced with instructions, and d-Lexize was also introduced with two five-item practice tasks. Finally, as d-Lexize is used exclusively as a diagnostic tool for teachers, and it is part of annual assessments of mathematics and reading skills, from which the results are used primarily for research and evaluating the students' skills in mathematics and reading and not in any selection processes nor for grading the students, we treated d-Lexize as a low-stakes test.

3.4. Data Analysis

3.4.1. Initial Data Cleaning

We cleaned the initial item response data from the d-Lexize test with an initial sample size of 6,664 participants with a total of 663,676 item responses. First, in cases where a technical time limit of 4 seconds was reached, the responses were treated as "empty answers," effectively removing them from the data. There were 45,392 (6.8%) of the 663,676 item-level observations with an empty answer. This affected the item responses of a total of 5,523 participants, with the number of empty answers ranging from 0 to 67. On average, there were 6.8 empty answers per participant, and three-quarters of the participants (4,998) had a maximum of nine empty answers. Second, due to technical errors, a very small fraction of RTs (146, <0.1%) were less than or equal to 0. We did not know why this had happened, so we treated these observations as missing, effectively removing them from the data. This affected the item-level observations of 99 participants in total, with the highest number of RTs that were less than or equal to 0 being 4. This data cleaning procedure left us with 6,664 participants with a total of 618,138 item responses.

Regarding the CRTT data, we noticed a warm-up effect, as the first two items had considerably higher median RTs than the rest of the items (1,820 ms and 1,018 ms, respectively, whereas the others varied between 818 ms and 884 ms), so we only considered the data from the participants who managed to answer all the remaining 14 items. In addition, since the probability of correctly guessing at least 10 of the remaining 14 items is less than 5%, we removed participants with less than 10 correct answers, leaving 5,626 individuals (2,838 boys, 50.4%) for the main data analyses. Of the remaining individuals in the sample, 4,790 (85.1%) had Finnish or a combination of Finnish and some other language as their home language (Finnish-speaking group), and the rest (836; 14.9%) did not speak Finnish at home (non-Finnish-speaking group).

3.4.2. Response Time Thresholds

The RT threshold methods that we selected for this study were CKT, NT10, NT30, VITP, CUMP, ChInf, ChIA, an automated version of the visual inspection method, namely the *Automated Visual Inspection* (AutoVI) method, where one estimates the kernel densities of the observed RTs and then locates the local minimum between the first two local maxima, and a mixture modeling-based approach, namely the *Mixture Response Time Quantile* (MRTQ) method, where one fits a two-state mixture model to the log RT data and finds the RT quantile corresponding to the estimated mixture weight of the component with the smaller mean parameter value, i.e., the RG component. Since d-Lexize involves language, we suspected that non-native speakers needed more time to respond to the items and hence the methods would produce larger thresholds for these individuals. For example, Goldhammer et al. (2017) showed that disengagement increased when the test language was not the participant's native language. Therefore, thresholds were estimated separately for Finnish and non-Finnish speakers. The detailed description of applying each of the selected methods to the d-Lexize data is presented in the Appendix.

Of the presented methods, we did not use the IIC method or the RS and GA approaches for several reasons. First, we felt that the IIC method was inapplicable in our setting due to the similarity of the d-Lexize items and their overall simplicity, as the number of characters varied between 4 and 9 and there were no ancillary materials. Second, based on the study by Bulut et al. (2023), their RS and GA approaches seemed to produce thresholds very similar to the NT30 thresholds, as the RS- and GA-based thresholds varied between 29% and 30% of the average item RTs (with the exception of one item with the RS algorithm). Third, Bulut et al. (2023) found that the RS and GA approaches were computationally very intensive, as indicated by the average computation times of 11.5 and 86.6 minutes per item, respectively, making these approaches impractical in operational settings. Therefore, combining the previous points regarding the RS and GA approaches, we considered them inapplicable in our setting.

3.4.3. Main Analysis

The selected RT threshold methods were used to determine a boundary between rapid guesses and engaged responses for each d-Lexize item by applying them to the cleaned data separately for Finnish and non-Finnish speakers using the R 4.3.0 software (R Core Team, 2023). To address the RQs, we examined how well the thresholds proposed by the different RT threshold methods, i.e., the *item-specific thresholds*, identified the same rapid guesses and engaged responses that were identified by the *participant-specific* thresholds based on the participants' choice reaction times (CRTs; see Deary et al., 2011). That is, we used choice reaction time (CRT) as the ground truth variable in the differentiation between rapid guesses and engaged responses. We calculated the participants' CRT estimates on the log scale as mean log RTs in CRTT. See Table 1 for an illustration of the RG classifications when using CRT as the ground truth variable.

Because the log CRT estimates were only approximations of the participants' true log CRTs, there was uncertainty in using choice reaction time as the ground truth variable in the RG/SB classifications. To estimate this uncertainty, we utilized the nonparametric bootstrap method (Efron & Tibshirani, 1986) to calculate standard errors (SEs) of the log CRT estimates as well as the proportions of true

Table 1. Illustration of the classification table for differentiating between rapid-guessing behavior (RG) and solution behavior (SB) with the response time threshold methods using choice reaction time (CRT) as the ground truth variable.

Observed log response time to a d-Lexize item in relation to the participant's log CRT estimate	Ground truth	RG classification	
		$Log\ RT_{lex} > C$ SB	$Log\ RT_{lex} \leq C$ RG
$Log\ RT_{lex} > Log\ CRT_{est}$	SB	TN	FP
$Log\ RT_{lex} \leq Log\ CRT_{est}$	RG	FN	TP

$Log\ RT_{lex}$ = log response time to a d-Lexize item; $Log\ CRT_{est}$ = log CRT estimate; C = estimated RG threshold; TN = true negative; FN = false negative; FP = false positive; TP = true positive.

positives, true negatives, false positives, and false negatives. More specifically, we considered those d-Lexize responses with log RTs that were less than or equal to the log CRT estimates to be rapid guesses and that the uncertainty here came only from the uncertainty in the estimates. However, we considered those responses with log RTs that were greater than the log CRT estimates to be engaged responses, and since the log RT of an engaged response is a sum of the true log CRT and the additional log time required to solve the problem in the corresponding d-Lexize item, the uncertainty in this scenario came partly from the uncertainty in the log CRT estimates and partly from this additional time required, meaning that a response with a log RT somewhere between the estimated log CRT and the estimated log CRT plus the additional time could still have been a rapid guess. For the Finnish speakers, the mean and standard deviation of the CRT estimates in log scale (standard errors are in parentheses) were 6.737 (0.001) and 0.345 (0.002), respectively, and transformed back to the original scale, they were 896 ms (1) and 331 ms (2), respectively. For the non-Finnish speakers, the corresponding values were 6.821 (0.003) and 0.349 (0.004) in log scale and 975 ms (3) and 348 ms (5) in the original scale.

4. Results

4.1. Initial Analysis of the Identified Time Thresholds

For the Finnish speakers, there was a clear bimodal pattern in the RT distributions across all items, resulting in 102 thresholds identified for the AutoVI, NT10, NT30, and MRTQ methods. However, for the non-Finnish speakers, the observed RT distribution was unimodal for 12 items based on their kernel density estimates, and hence AutoVI failed to estimate thresholds for them. Interestingly, by fitting a two-state mixture model to these items' data, we were still able to detect bimodality, as MRTQ produced suitable thresholds for these items with an average of 678 ms and standard deviation of 114 ms.

For both Finnish and non-Finnish speakers, there were items for which the accuracy level and/or item-total correlation did not behave as expected as a function of RT, and therefore VITP, CUMP, ChInf, and ChIA were unable to estimate thresholds for 8, 28, 22, and 23 items for the Finnish-speaking group, respectively, and 11, 29, 40, and 50 items for the non-Finnish-speaking group, respectively. For example, the CUMP method could not estimate 28 (29) thresholds for the Finnish speakers (non-Finnish speakers), because the cumulative accuracy levels of the corresponding items remained above or below the chance level for the entire search space. Furthermore, both ChInf and ChIA could not estimate thresholds for 22 (40) items, because these items' conditional item-total correlations remained above or below .2 or varied around .2 for the entire search space. However, this is not entirely a deficit in the RT threshold methods, as typically those types of items for which item-total correlation is below .2 would be excluded as they are low-quality items that cannot discriminate between low- and high-achievement test takers. That is, in most practical settings with well-calibrated tests, this would probably not be an issue. Notably, ChIA also failed to estimate thresholds for one (Finnish) and 10 (non-Finnish) items because there was no clear time point after which the conditional proportion of correct responses started to show a trend toward the overall proportion of correct responses.

Table 2 shows the threshold statistics for the 50 and 31 common items for which all methods could estimate a threshold for the Finnish and non-Finnish speakers, respectively. Based on the mean threshold values, AutoVI, NT10, NT30, ChInf and ChIA produced smaller thresholds than MRTQ, VITP, and CUMP in both language groups. In general, the RT-only methods had less variation in the thresholds between the items than the RT-and-RA methods. An exception is the AutoVI method for the non-Finnish-speaking group, for which the method produced variation similar to CUMP and ChInf. Interestingly, in terms of mean threshold, the MRTQ method was similar to the VITP and CUMP methods, whereas in terms of the variation between the thresholds, MRTQ was more in line with the other RT-only methods.

Table 2. Statistics of the estimated response time (RT) thresholds and proportions of the rapid-guessing behavior (RG) classifications using choice reaction time as the ground truth variable.

		Threshold statistics		Proportions of the RG classifications			
Method	RG %	*M*	*SD*	TN	FN	FP	TP
Home language = Finnish (50 items; 4,790 participants)							
Ground truth	**11.7**	**896(1)**	**331(2)**	**.8835(8)**	-	-	**.1165(8)**
CKT	3.3	408	0	.8833(8)	.0838(8)	.00013(4)	.03268(4)
AutoVI	3.1	372	40	.8834(8)	.0853(8)	.00011(4)	.03123(4)
NT10	2.1	158	17	.8835(8)	.0957(8)	.00001(2)	.02086(2)
NT30	3.5	474	51	.8832(8)	.0815(8)	.00026(5)	.03502(5)
MRTQ	3.8	536	65	.8826(8)	.0790(8)	.00084(5)	.03749(5)
VITP	4.4	552	120	.8791(8)	.0773(8)	.00441(7)	.03925(7)
CUMP	5.0	592	148	.8815(8)	.0750(8)	.00842(9)	.04156(9)
ChInf	3.4	412	205	.8815(8)	.0840(8)	.00200(5)	.03247(5)
ChIA	3.8	483	135	.8818(8)	.0807(8)	.00172(5)	.03586(5)
Home language = other than Finnish (31 items; 836 participants)							
Ground truth	**22.7**	**975(3)**	**348(5)**	**.773(2)**	-	-	**.227(2)**
CKT	12.5	503	0	.772(2)	.103(2)	.0002(2)	.1247(2)
AutoVI	13.1	500	231	.764(2)	.105(2)	.0087(2)	.1219(2)
NT10	6.4	159	10	.773(2)	.164(2)	.00000(3)	.06378(1)
NT30	12.1	477	31	.772(2)	.106(2)	.0002(1)	.1211(1)
MRTQ	15.8	709	86	.764(2)	.078(2)	.0090(4)	.1492(4)
VITP	14.8	635	105	.766(2)	.086(2)	.0064(3)	.1413(3)
CUMP	16.4	670	209	.755(2)	.081(2)	.0175(4)	.1462(4)
ChInf	12.4	494	222	.768(2)	.108(2)	.0045(2)	.1191(2)
ChIA	13.4	555	117	.769(2)	.096(2)	.0032(2)	.1311(2)

Values in parentheses are the bootstrap standard errors of the last digits of the corresponding statistic. RG % = percentage of identified rapid guesses; *M* = mean; *SD* = standard deviation; TN = true negative; FN = false negative; FP = false positive; TP = true positive.

Excluding NT10, for the Finnish-speaking group, the percentage of identified rapid guesses varied between 3.1% and 5.0%, which is considerably lower than the average RG percentage of 10% found in low-stakes assessments (Rios et al., 2022). For the non-Finnish-speaking group, however, the RG percentage varied between 12.1% and 16.4%. Regarding the NT10 method, even though the mean NT10 threshold values were almost identical between the language groups (158 ms versus 159 ms), the RG percentage for the non-Finnish speakers was 6.4% which was over three times the RG percentage of 2.1% of the Finnish speakers. Relatively speaking, this suggests that for non-Finnish speakers the condition for rapid guessing was more often the second condition of the three conditions proposed by Wise (2017) than for the Finnish speakers.

Finally, we noticed that the test takers tended to rapid guess more at the end of the test compared to the beginning. For example, excluding the first response from all test takers (to account for warm-up effect), the average RG percentage during the first 20 items identified with the MRTQ method was 1.8% in the Finnish-speaking group and 9.5% in the non-Finnish-speaking group, whereas during the last 20 items the average RG percentage was 5.8% in the Finnish-speaking group and 21.7% in the non-Finnish-speaking group. In addition, for the Finnish and non-Finnish speakers, the average increase in RG percentage between two consecutive items was 0.06 and 0.23 percentage points, respectively. This type of pattern was consistent across the methods, indicating that fatigue had a notable role in disengagement.

4.2. Misclassification in RG Identification

4.2.1. Motivating the Use of Choice Reaction Time as the Ground Truth Variable

Before addressing the RQs, we first motivate the examination of the RG misclassification using choice reaction time as the ground truth variable with the help of an illustration in Figure 1, which shows the scatterplot of the mean log RTs in d-Lexize and the log CRT estimates grouped by home language. The ascending lines with slopes of 1 and intercepts of 0 are the

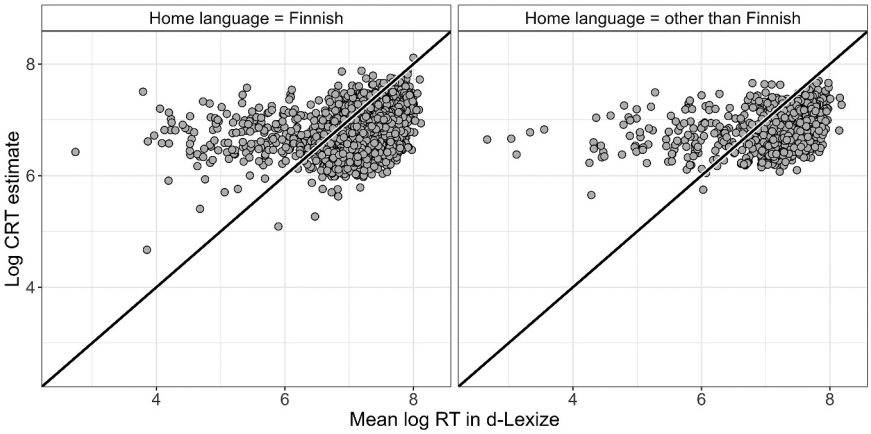

Figure 1. Relationship between the mean log response time (RT) in d-Lexize and the log choice reaction time (CRT) estimate for the Finnish and non-Finnish speakers.

boundaries between the participants who responded faster on average to the d-Lexize items than to the CRTT items (left side of the line) and the participants who responded slower on average to the d-Lexize items than to the CRTT items (right side of the line). Most of the data are in the data clouds on the right side of the boundaries (for Finnish speakers 91.7%, and for non-Finnish speakers 77.4%), which is consistent with the hypothesis that the CRTs of engaged participants will be smaller than their RTs to items in an achievement test that requires additional time to solve, such as the d-Lexize test here. The data clouds, which are very close to the boundaries, indicate that, on average, the d-Lexize items required little additional time to solve. For the Finnish speakers, the data points lying on the left side of the boundaries indicate that some participants were most likely rapid guessing throughout most of the d-Lexize test, as their mean accuracy level of .63 was lower compared to the mean accuracy level of .74 of the rest of the participants on the right side of the boundary. For the non-Finnish speakers, the corresponding accuracy levels were .52 and .56, respectively, which indicates that even the engaged non-Finnish speakers had problems with identifying words and pseudowords, which, in turn, is most probably one of the reasons why the response accuracy-based methods failed to estimate thresholds for more items for the non-Finnish speakers than for the Finnish speakers.

4.2.2. Rate and Nature of RG Misclassification

To address RQ1, we calculated proportions of true negatives (TNs), false negatives (FNs), false positives (FPs), and true positives (TPs), across the common 50 and 31 items for which all the methods could estimate a threshold for Finnish and non-Finnish speakers, respectively. Table 2 contains the proportions of TNs, FNs, FPs, and TPs, and the corresponding bootstrap standard errors (SEs) of the last digits. For both Finnish and non-Finnish speakers, the misclassification rate (proportion of FPs + proportion of FNs) was significant across the methods as it varied between .080 and .096 (Finnish) and .087 and .164 (non-Finnish). Regarding the nature of RG misclassification, as the proportion of FNs varied between .075 and .096 (Finnish) and .078 and .164 (non-Finnish) with only marginal uncertainty (standard error was 0.0008 for all methods in the Finnish-speaking group and 0.002 for all methods in the non-Finnish-speaking group) across methods, while the proportion of FPs varied between 0 and .008 (Finnish) and 0 and .018 (non-Finnish) with essentially no uncertainty (standard error was below 0.001 for all methods in both language groups), the results were consistent with the discussion of Wise (2017) and Rios and Deng (2021) that the RT threshold methods tend to be conservative, resulting in fewer FPs compared to FNs.

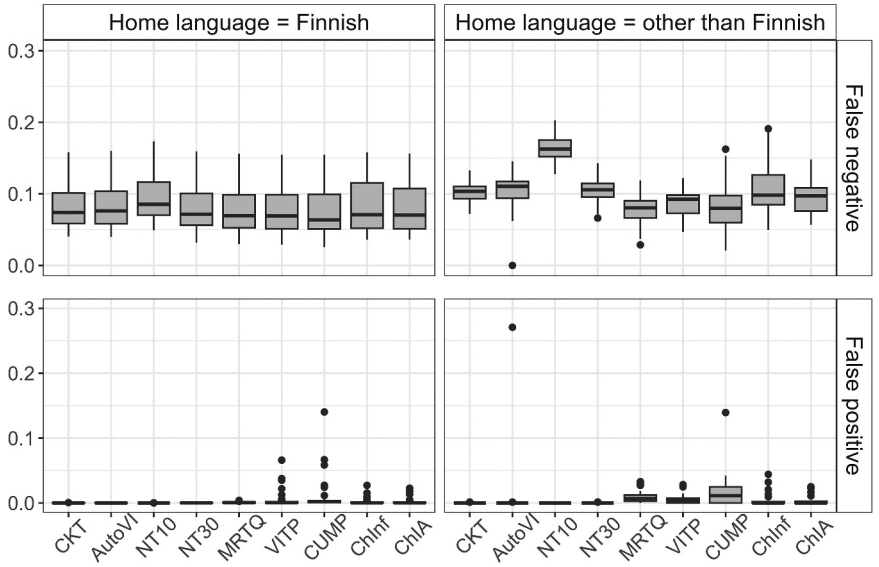

Figure 2. Distributions of the false negative and false positive proportions of the 50 (31) items for which all response time threshold methods could estimate a threshold in the Finnish-speaking group (non-Finnish-speaking group).

4.2.3. Differences Between the Methods and the Items

To address RQ2 and RQ3, we examined the distributions of the proportions of FNs and FPs across the 50 and 31 common items for Finnish and non-Finnish speakers, respectively, which are shown in Figure 2 as box plots. For the Finnish speakers, there were only marginal differences between the methods, while in the non-Finnish speakers' group, the differences between the methods can be seen both in the locations of the distributions and in the scales; while NT10 was the most conservative as it produced the highest proportions of FNs on average, CUMP and ChInf produced the greatest variation in these proportions between the items. In terms of the proportions of FPs, the differences were harder to see, as the proportions for many items were close to zero for all methods. However, for both language groups, VITP, CUMP, ChInf, and ChIA were noticeably different from the others in that they produced a few outliers, as indicated by the data points at the top of the box plots, although the differences between the outliers and the rest of the items were small with a few exceptions.

Finally, we checked the proportions of FNs and FPs of all 102 items for the CKT, AutoVI, NT10, NT30, and MRTQ methods in the Finnish-speaking group and for the CKT, NT10, NT30, and MRTQ methods in the non-Finnish-speaking group, and the results were identical to those shown in Figure 2. In addition, we checked the proportions of all those items for which VITP, CUMP, ChInf, and ChIA could estimate a threshold (94, 74, 80, and 79 items for the Finnish-speaking group, respectively, and 91, 73, 62, and 52 items for the non-Finnish-speaking group, respectively); for the proportions of FNs for all of these methods and for the proportions of FPs for VITP, ChInf, and ChIA, the results were similar to those in Figure 2, but for the proportions of FPs for CUMP, more outliers (four in both language groups) emerged with very high proportions of FPs (> .25). All in all, it appeared that the biggest differences between the methods were not in the misclassification rates nor in the nature of the misclassifications, but in their abilities to estimate appropriate thresholds at all for items with different types of distributions of RTs (AutoVI, NT10, NT30, and MRTQ), changes in the conditional or cumulative accuracy rates as functions of RT (VITP, CUMP, and ChIA), or changes in the item-total correlations as functions of RT (ChInf and ChIA).

For two items (item A and item B), Figure 3 shows the scatterplots of the log RTs to the items and the log CRT estimates, of which the left panel shows the AutoVI threshold and the right panel shows the CUMP threshold as vertical lines. For item A, CUMP produced a relatively high proportion of FPs

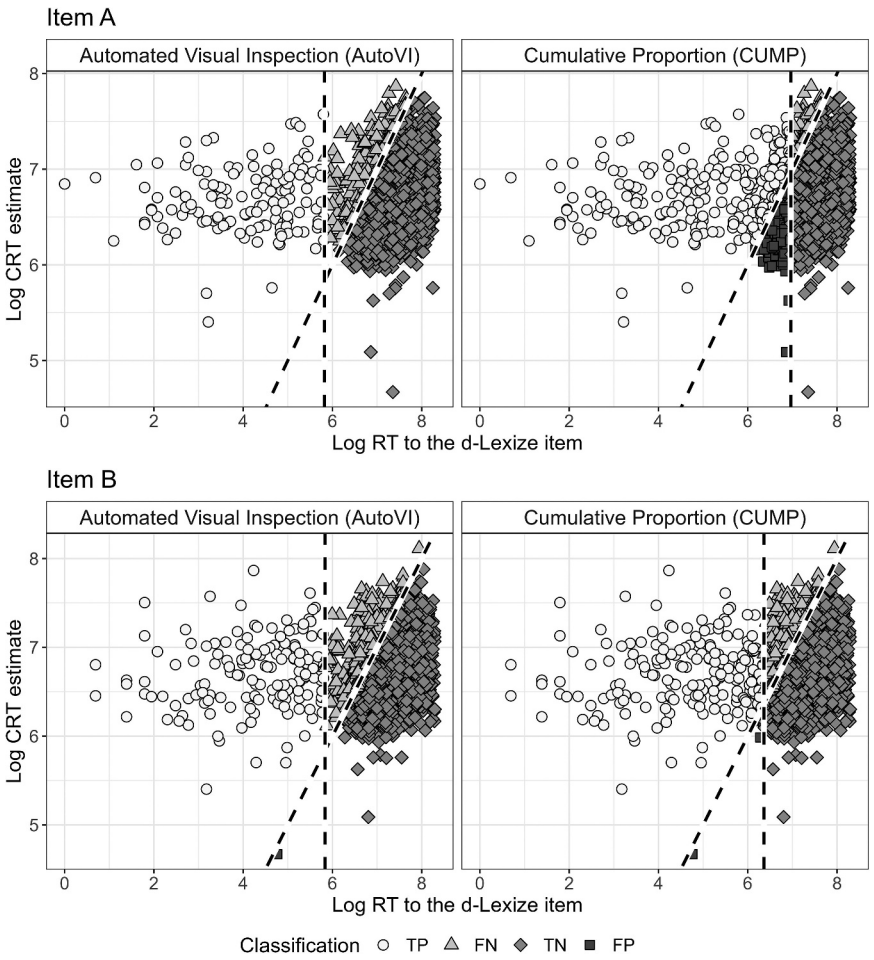

Figure 3. Relationship between the mean log response time in d-Lexize and the log choice reaction time (CRT) estimate in the Finnish-speaking group for an item with a high proportion of false positives (.140; item A) and for an item with a low proportion of false positives (.001; item B) based on the CUMP method. Note. TP = true positive; FN = false negative; TN = true negative; FP = false positive. The vertical dashed lines indicate the time thresholds produced by the corresponding method (AutoVI or CUMP).

(.140), whereas for item B, CUMP produced only a low proportion of FPs (.001). That is, Figure 3 illustrates how the more (i.e., AutoVI) and less (i.e., CUMP) conservative methods produced different types of misclassifications. Since AutoVI is the more conservative of the two methods, its threshold was quite low for both items (338 ms, 5.82 in log scale [item A] and 342 ms, 5.83 in log scale [item B]), which resulted in zero FPs for item A and only one FP for item B and some FNs for both items (proportions of FNs were .07 [item A] and .11 [item B]). In contrast, the CUMP threshold for the item A was quite large (1057 ms, 6.96 in log scale), resulting in fewer FNs (proportion of FNs was .03) at the cost of a significant proportion of FPs (.14).

5. Discussion

In low-stakes testing settings, where there are essentially no consequences for the test takers, disengagement is the main construct-irrelevant factor that threatens the validity of observed test scores (Wise, 2020). Disengagement can manifest as rapid-guessing behavior (RG), a response behavior in which a test taker answers too quickly to show full effort, as opposed to solution behavior (SB) in

which the test taker is engaged (Wise, 2017). The ability to accurately identify RG is important because it allows one to adjust for potential biases in scoring. Therefore, researchers have proposed various approaches to identify RG from data, namely response time (RT) threshold methods (e.g., Guo et al., 2016; Lee & Jia, 2014; Schnipke, 1995; Wise, 2019; Wise & Kong, 2005), which have been widely used in applied research and operational settings (see Rios & Deng, 2021; Ulitzsch et al., 2023). However, since the goal of these methods is to define an appropriate time threshold that can serve as a binary cutoff to differentiate between RG and SB for each test item, the overlapping nature of the RT distributions of these two response behaviors inevitably leads to misclassifications. To date, the rate and type of misclassification in RG identification produced by different RT threshold methods in real assessment data has not been empirically investigated. The current study was designed to fill this gap in the literature.

We applied most of the existing RT threshold methods to large-scale assessment data collected from 3rd, 4th, and 7th grade students using a fast-paced computerized test of Finnish vocabulary knowledge (d-Lexize; Bertram et al., 2025). We examined the rate and type of RG misclassification produced by the different RT threshold methods, using choice reaction time (CRT) as the ground truth variable. The results showed that the misclassification rate was substantial for all methods in the two language groups (Finnish speakers and non-Finnish speakers), varying between .080 and .096 (Finnish) and .087 and .164 (non-Finnish) across the methods. Overall, all the methods appeared to be more conservative than liberal, as the proportions of false negatives (FNs) were all between .075 and .096 in the Finnish-speaking group and between .081 and .164 in the non-Finnish-speaking group, but the proportions of false positives (FPs) were all close to zero in both language groups.

In general, the mixture model-based approach, MRTQ, appeared to perform the best, as it was able to produce thresholds for all 102 items, and for the common 50 (Finnish) and 31 (non-Finnish) items for which all methods could estimate a threshold, it produced among the lowest misclassification rates in both language groups (.080 [Finnish] and .087 [non-Finnish]) with only small variation in the proportions of FNs and FPs across items. Notably, MRTQ performed well even in the non-Finnish-speaking group where the observed RT distribution based on the kernel density estimates of RT appeared to be unimodal for some items. On the one hand, while NT10 was also able to produce thresholds for all items, it produced the highest misclassification rates (.096 [Finnish] and .164 [non-Finnish]) and performed the worst from this perspective. On the other hand, ChInf and CUMP were only able to produce thresholds for 80 and 74 items in the Finnish-speaking group, respectively, and 62 and 73 items in the non-Finnish-speaking group, respectively, and across the common items, ChInf (CUMP) produced the largest variation in the FN (FP) proportions in both language groups, indicating that these methods were not as robust as the others, and they performed the worst from this perspective. These results highlight how difficult it can be for the RT threshold methods to avoid misclassifications, especially FNs, i.e., true rapid guesses that are misidentified as engaged responses, as our findings are consistent with the findings of Wise (2017) and Rios and Deng (2021) that many of the RT threshold methods tend to be conservative, resulting in a higher number of FNs than FPs.

5.1. Limitations and Future Research

Some limitations of this study need to be addressed. First, we were limited to only one type of a test measuring a certain type of psychological construct with only two-choice items that require little time to solve, which reduces the generalizability of the results. The d-Lexize is a highly atypical test for educational settings; although the d-Lexize items are partly comparable to traditional true-false items with yes-no responses in teacher-made tests (e.g., Mehrens & Lehmann, 1991), the outcome of the test reflects knowledge only at a lower mental process level (e.g., recognition), and hence it differs from those used in other large-scale assessments, such as the Programme for International Student Assessment (PISA) where the item domain varies substantially more (reading, mathematics, science) and the four main item types are simple multiple choice, complex multiple choice, computer-scored open response, and human-coded open response instead of true-false (OECD, 2024). Most of the

previous RG research has looked at these types of more traditional multiple-choice items with four or more response options that require the test takers to use higher mental processes without time constraints (e.g., Demars, 2007; Wise, 2019; Wise & Kong, 2005; Wise & Ma, 2012). For example, it has been found that the rate of rapid guessing can be domain-specific (e.g., Kröhne et al., 2022) and that rapid guessing is related to item characteristics (e.g., Wise & Kong, 2005). Therefore, future research could examine the rate and type of RG misclassification in different types of tests with different types of items, especially with more varying time demands than in our study.

Second, it may be that the baseline provided by CRT was too low in the sense that it only detected those rapid guesses that were given by unmotivated test takers that wanted to complete the assessments quickly, while the potential rapid guesses given by those test takers that recognized quickly that they do not know the answer and then chose to rapid guess remained undetected. However, since d-Lexize is a fast-paced test that requires even the engaged test takers to respond quickly and thus produces RT distributions where the modes of rapid guesses and engaged responses are relatively close to each other, it was highly likely that the number of these types of rapid guesses was minimal, or if in reality there was a substantial number of these types of rapid guesses, they highly likely had similar RTs than the fastest engaged responses, and hence it can be argued that none of the current RT threshold methods could detect these types of rapid guesses reliably in the d-Lexize data. That is, for our purposes, CRT could be meaningfully used as a ground truth for RG.

The discussion above brings forward the question whether CRT is a valid ground truth for rapid guessing at all. There are two answers to this question, depending on the type of rapid guessing we are working with; that is, for the total thoughtless type of rapid guessing, CRT is valid, whereas for those types of rapid guesses given by test takers who recognize that they do not know the correct answer and then choose to rapid guess, it probably is not. However, if the goal of RG identification is to remove uninformative responses from the data prior to the main data analysis, one could argue that the former type of rapid guessing is nevertheless more important to identify than the latter, because the former type of rapid guesses are uninformative regarding the test takers' knowledge, skills, or abilities (KSAs), but the latter type of rapid guesses actually are informative in the sense that they tell us that the test takers did not have the required KSAs to respond to the items (see, e.g., Wise, 2017 for more discussion on this matter). In contrast, if we are not interested in why the rapid guesses occurred, then it is not relevant to differentiate between the types of rapid guessing and hence CRT would probably not be a valid ground truth. In this type of situation, CRT could nevertheless be used such that thresholds below the estimated choice reaction times would be too small to detect even the fastest rapid guesses.

While our approach yielded results consistent with the discussion in previous research that the RT threshold methods tend to be conservative (Rios & Deng, 2021; Wise, 2017), future research could examine RG misclassification more broadly in different settings to gain deeper insights into the types of misclassification to which the different methods are prone. In addition, we discussed that it would be more meaningful to examine rapid guessing at the participant-by-item level and calculate the time thresholds based on both the item and the participant. While our approach of using CRT as the ground truth variable in RG identification considered the participant aspect of this perspective, it did not consider the item aspect. To this end, a possibly more comprehensive approach would be to specify item-specific models that map the relationship between the log RTs to the items in the achievement test of interest and the estimated choice reaction times of the participants, and then to define the time thresholds in two dimensions.

5.2. Implications

This study has important implications for measurement practitioners. First, one needs to be mindful of the possibility of RG misclassification when using any of the RT threshold methods, as this may introduce bias into the scoring of individual test takers. In practical settings, if the observed RT distribution is bimodal, we would recommend using a mixture

modeling-based approach, such as the one utilized in this study, and if the observed distribution is not bimodal, these approaches might still work as it is entirely possible that the bimodality remains hidden through visual inspection of the data or kernel density estimation, but fitting a two-state mixture model to the data could reveal this structure. However, if the mixture modeling-based approaches fail, we recommend using the CUMP method. This type of combined approach is similar to the hybrid rule used by Rios and Guo (2020) who first employed the CUMP method, and if that failed, they employed their mixed lognormal distribution-based method.

Ultimately, however, it is up to the practitioner to consider what the goal of identifying RG is and how it relates to the goal of the study: if the goal is to minimize the number of FPs in order to avoid misclassifying true engaged responses as rapid guesses, one should probably use conservative thresholds, such as NT10, NT30, or AutoVI, whereas if the goal is to minimize the number of FNs in order to avoid misclassifying true rapid guesses as engaged responses, one should instead use more liberal thresholds, such as CUMP. Finally, examining the relationship between RTs on the achievement test of interest and individuals' choice reaction times should provide valuable insight into this issue. As we discussed above, regardless of the goal of RG identification, one could use choice reaction time as a tool to validate the time thresholds identified, such that thresholds below the estimated choice reaction times would be too small to identify even the totally thoughtless rapid guesses.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The present study is part of the EDUCA Flagship funded by the Research Council of Finland [#358924, #358947] and the EDUCA-Doc Doctoral Education pilot funded by the Ministry of Education and Culture [Doctoral school pilot #VN/3137/2024-OKM-4].

ORCID

Santeri Holopainen http://orcid.org/0000-0002-6777-6247
Jari Metsämuuronen http://orcid.org/0000-0001-6027-0799
Mikko-Jussi Laakso http://orcid.org/0000-0001-9163-2676
Janne Kujala http://orcid.org/0009-0009-3787-6712

Data Availability Statement

The authors do not have permission to share the data.

References

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). Mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, *32*(6), 1–29. https://doi.org/10.18637/jss.v032.i06

Bertram, R., Rautaoja, T., Holopainen, S., Häikiö, T., Enges, P., Hyönä, J., Lehtonen, M., Pugh, K. R., Rueckl, J. G., Salmela, R., Siegelman, N., & Räsänen, P. (2025). *Assessing vocabulary skills of school children aged 9 to 15 in Finland: Tracking the gender and home language gap* [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-6448049/v1

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

Deary, I. J., Liewald, D., & Nissan, J. (2011). A free, easy-to-use, computer-based simple and four-choice reaction time programme: The Deary-Liewald reaction time task. *Behavior Research Methods*, *43*(1), 258–268. https://doi.org/10.3758/s13428-010-0024-1

Demars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*(1), 23–45. https://doi.org/10.1080/10627190709336946

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*(1). https://doi.org/10.1214/ss/1177013815

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), 18. 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

Kröhne, U., Deribo, T., & Goldhammer, F. (2022). 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

Laakso, M.-J., Kaila, E., & Rajala, T. (2018). ViLLE - collaborative education tool: Designing and utilizing an exercise-based learning environment. *Education and Information Technologies*, *23*(4), 1655–1676. https://doi.org/10.1007/s10639-017-9659-1

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), 8. https://doi.org/10.1186/s40536-014-0008-1

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology*, *10*, 1533. https://doi.org/10.3389/fpsyg.2019.01533

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (Wadsworth Publishing) (4th ed.).

Metsämuuronen, J. (2020). Somers' D as an alternative for the item-test and item-rest correlation coefficients in the educational measurement settings. *International Journal of Educational Methodology*, *6*(1), 207–221. https://doi.org/10.12973/ijem.6.1.207

Metsämuuronen, J. (2022). The effect of various simultaneous sources of mechanical error in the estimators of correlation causing deflation in reliability: Seeking the best options of correlation for deflation-corrected reliability. *Behaviormetrika*, *49*(1), 91–130. https://doi.org/10.1007/s41237-022-00158-y

Nagy, G., & Ulitzsch, E. (2022). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement*, *82*(5), 845–879. https://doi.org/10.1177/00131644211045351

OECD. (2024). *PISA, 2022 technical report*. PISA, OECD Publishing. https://doi.org/10.1787/01820d6d-en

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics*, *41*(3), 300–325. https://doi.org/10.3102/1076998616636618

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/,

Rios, J. A. (2022a). 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. (2022b). An examination of individual ability estimation and classification accuracy under rapid guessing misidentifications. *Applied Measurement in Education*, *35*(4), 300–312. https://doi.org/10.1080/08957347.2022.2155653

Rios, J. A., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-Scale Assessments in Education*, *9*(1), 18. https://doi.org/10.1186/s40536-021-00110-8

Rios, J. A., & Deng, J. (2024). A comparison of response time threshold scoring procedures in mitigating bias from rapid guessing behavior. *Educational and Psychological Measurement*, *84*(2), 387–420. https://doi.org/10.1177/00131644231168398

Rios, J. A., Deng, J., & Ihlenfeldt, S. D. (2022). To what degree does rapid guessing distort aggregated test scores? A meta-analytic investigation. *Educational Assessment*, *27*(4), 356–373. https://doi.org/10.1080/10627197.2022.2110465

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

Salmela, R., Lehtonen, M., Garusi, S., & Bertram, R. (2021). Lexize: A test to quickly assess vocabulary knowledge in Finnish. *Scandinavian Journal of Psychology*, *62*(6), 806–819. https://doi.org/10.1111/sjop.12768

Schnipke, D. L. (1995, April 19–21). *Assessing speededness in computer-based tests using item response times* [Paper presentation]. National Council on Measurement in Education (NCME) Annual Meeting 1995, San Francisco, CA. Retrieved from the ERIC database (ED400276). Retrieved October 30, 2024, from https://eric.ed.gov/?id=ED383742

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

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, *43*(1), 97–99. https://doi.org/10.1111/j.2517-6161.1981.tb01155.x
Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, *27*(6), 799–811. https://doi.org/10.2307/2090408
Ulitzsch, E., Domingue, B. W., Kapoor, R., Kanopka, K., & Rios, J. A. (2023). A probabilistic filtering approach to non-effortful responding. *Educational Measurement Issues & Practice*, *42*(3), 50–64. https://doi.org/10.1111/emip.12567
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z
Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054
Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement Issues & 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. (2020). Six insights regarding test-taking disengagement. *Educational Research & Evaluation*, *26*(5–6), 328–338. https://doi.org/10.1080/13803611.2021.1963942
Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program* [Paper presentation]. National Council on Measurement in Education (NCME) Annual Meeting 2004, San Diego, CA. Retrieved October 30, 2024, from https://www.researchgate.net/publication/264877650_An_Investigation_of_Motivation_Filtering_in_a_Statewide_Achievement_Testing_Program
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., & Kuhfeld, M. R. (2021). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*, *58*(1), 130–149. https://doi.org/10.1111/jedm.12275
Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method* [Paper presentation]. National Council on Measurement in Education (NCME) Annual Meeting 2012, Vancouver, Canada. Retrieved October 30, 2024, from https://www.researchgate.net/publication/265407579_Setting_Response_Time_Thresholds_for_a_CAT_Item_Pool_The_Normative_Threshold_Method
Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650
Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139–153). American Psychological Association. https://doi.org/10.1037/12330-009

Appendix

Detailed Description of Applying the RT Threshold Methods to the d-Lexize Data

Identification of the common threshold for all items (CKT) and the item-specific AutoVI and MRTQ thresholds involved assessing the number of modes of the overall RT distribution (CKT) or the item-specific RT distributions (AutoVI and MRTQ). In each of these methods, we estimated the Gaussian kernel densities of the RTs using Silverman's rule of thumb bandwidth (Silverman, 1981) and assessed the number of modes in the RT distributions by finding the local maxima of the densities within the RT domain. For CKT and AutoVI, the next step was to define the time thresholds as the locations of the antimodes between the two modes, which represent the conceptualization of rapid guesses and engaged responses with distinct RT distributions. Notably, our automated version of the VI method differed from the original version in which RT histograms were examined item by item and then the time points with the lowest frequencies between the two modes were found (Schnipke, 1995). Instead, AutoVI was closer to the MLN method proposed by Rios and Guo (2020), who fitted mixture lognormal distributions to the log-transformed item-specific RT data and defined the time thresholds as the time points between the two normal distributions where the density functions of the overall distributions reached their minimum values. The difference was that we did not transform the RTs to the logarithmic scale in order for the AutoVI method to be as similar to the original VI method as possible. For the MRTQ method, however, we did log-transform the RT data; if the RT distribution was found to be bimodal, we used the log-transformations of the RT mode locations as starting values for the component means when fitting two-state mixture models to the log-transformed item-specific RT data, whereas if the RT distribution was found to be unimodal, we let the algorithm to randomly generate starting values for the component means. Using this method, we defined the item-specific thresholds as the RT quantiles indicated by the estimated mixture weights of the components with the smaller mean

parameter values corresponding to rapid guessing. We used the R Package mixtools to fit the mixture models (Benaglia et al., 2009).

Using the NT method, we defined the time thresholds as the 10% (NT10) and 30% (NT30) of the item-specific mean RTs. Applying the NT10 method to the d-Lexize data was natural because it was originally recommended by its inventors, Wise and Ma (2012). We also chose to use the NT30 method because this method was more recently recommended by Wise and Kuhfeld (2021). In addition, the RS and GA approaches proposed by Bulut et al. (2023), which were not used in this study, seemed to produce thresholds very similar to the NT30 method. Therefore, we felt that using NT30 in addition to NT10 would add value to this study.

We implemented the VITP and CUMP methods in a largely similar manner to their original implementations in Lee and Jia (2014) and Guo et al. (2016), respectively. However, some differences and nuances should be noted. First, both methods require the definition of an accuracy level corresponding to random guessing, i.e., the chance level. In d-Lexize, all items are two-choice questions (word or pseudoword), and intuitively the chance level would be .5 for all items. However, based on initial data analysis, we noticed that when the data was grouped by the participants' home language (Finnish or non-Finnish) and items' lexicality (word or pseudoword), the chance level estimated from all responses that were answered in less than 200 ms was .622 for words and .379 for pseudowords in the Finnish-speaking group and .534 for words and .443 for pseudowords in the non-Finnish-speaking group. This phenomenon, where the observed chance level does not match with the theoretical chance level, may be due to cognitive biases that favor one option over the other. In our case, where the position of the response options ("yes" or "no") remained the same for all items and the position of the correct answer depended only on the lexicality of the presented item (which was random for each participant-by-item interaction as the items were drawn from the item pool randomly for each individual), it appeared that the rapid guessers in both language groups favored the "yes" option, and hence their accuracy in word items was higher than .5 and in pseudoword items lower than .5. Interestingly, this phenomenon was much stronger for the Finnish speakers than for the non-Finnish speakers. In other words, it appeared that the rapid-guessing non-Finnish speakers were less systematic and more random in their response option selection than the rapid-guessing Finnish speakers. In the end, we chose to use the estimated chance levels group-wise for VITP and CUMP.

Second, regarding the VITP method, in the original version, RT was measured in seconds, and Lee and Jia (2014) were able to compute item-specific conditional accuracy levels at all possible RT values, whereas we had to divide RT into 100 ms wide segments to have enough data for each conditional accuracy level. This was due to the precision of our RT measurements, which were in milliseconds. Third, regarding the CUMP method, we set the lower end of the search space for the thresholds to 200 ms to account for sparse data problems at very low RTs. Finally, regarding both VITP and CUMP, in their original versions in Lee and Jia (2014) and Guo et al. (2016), the respective authors did not consider the possibility that a test item could be so hard that the trend of conditional or cumulative accuracy level as a function of RT could be negative, meaning that the overall accuracy level is lower than the chance level. We noticed that d-Lexize had several hard items that behaved this way, so we took this into account by changing the logic in the VITP and CUMP algorithms, and by changing the formula used to calculate the CUMP thresholds.

Finally, with respect to the ChInf and ChIA methods, our approach was very similar to the original approach of Wise (2019). There were two main differences between our approach and the original approach. First, Wise (2019) used the traditional Pearson correlation when calculating the conditional item-total correlations within time segments, whereas we used Somers' D (Somers, 1962) because it has been shown to provide better estimates of an item's discrimination power and to cause less deflation in reliability, especially when there is a mismatch between the scales of the item score and the total test score and when the item's difficulty level is extreme (Metsämuuronen, 2020, 2022). In our setting, the item scores were binary and the total test score for each student was the proportion of correct responses, and some of the item difficulty levels were extreme. Therefore, we considered D to be a superior alternative to Pearson correlation. Of the three directions of D, the direction DgX, in which the score (X) explains the behavior on item (g), or traditionally "X dependent," was used as proposed by Metsämuuronen (2020; Metsämuuronen, 2022; see the discussion of the directions in 2020). Second, the original approach of Wise (2019) did not consider the possibility that, for some items, the trend of the conditional item-total correlations as a function of RT could be negative, meaning that the overall correlation is lower than the initial correlation at low RTs. We accounted for this possibility in the ChInf and ChIA algorithms by adding a step stating that no threshold could be set for the item in question if the total correlation was less than the predefined low correlation value of .2. In addition, similarly to VITP and CUMP, in order to work, ChIA requires that there is a noticeable change in the accuracy level as a function of RT. That is, if there was no time segment after which the conditional proportion of correct responses trended consistently toward the overall proportion of correct responses, no threshold could be set with this method.