International Journal of Assessment Tools in Education
2026, Vol. 13, No. 1, 123–144
https://doi.org/10.21449/ijate.1721066
journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

Disengagement matters: A response-time informed approach to scoring low-stakes assessments

Okan Bulut[1*], **Joyce Xinle Liu**[2], **Hatice Cigdem Bulut**[3]

[1]University of Alberta, Faculty of Education, Centre for Research in Applied Measurement and Evaluation
[2]University of Alberta, Faculty of Education, Measurement, Evaluation, and Data Science
[3]Northern Alberta Institute of Technology, Education Insights, Data and Research

ARTICLE HISTORY

Received: June 16, 2025

Accepted: Nov. 13, 2025

KEYWORDS

Low-stakes testing,

Disengagement,

Rapid guessing,

Idling,

Item response modeling.

Abstract: Low-stakes assessments in K-12 education play a crucial role in monitoring student progress, yet their validity is often compromised by disengaged test-taking behaviors, such as rapid guessing and idling. This study introduces a novel application of the Nested Logit Model (NLM) to integrate test-taking engagement, as indicated by response times (RTs), with item responses to improve ability estimation. Using data from a low-stakes reading assessment in the United States (n = 27,556 students in grades 5 to 8), we compared six scoring approaches, including traditional dichotomous scoring, effort-moderated scoring, and nominal scoring that categorized responses based on RT-informed engagement. Our results demonstrated that nominal scoring approaches, particularly those distinguishing rapid guesses and idle responses, yielded superior model fit, increased measurement precision, and provided nuanced insights into examinee behaviors compared to dichotomous scoring methods. Latent class analysis further identified three distinct engagement profiles—effortful responders, rapid guessers, and idle responders—highlighting the need to address both rapid and idle behaviors in modeling. This study emphasizes the value of leveraging RT data to enhance the accuracy of low-stakes assessments while preserving response information. Findings also suggest the NLM framework as a practical and accessible tool for researchers and practitioners seeking to address disengaged behaviors and ensure the reliability of low-stakes assessments.

INTRODUCTION

In K-12 settings, low-stakes assessments are often used to monitor students' learning progress throughout a school year. Teachers use the scores from these assessments to identify areas where students need more support, and education administrators may use them for broader analysis or accountability purposes (Filiasov & Sweetman, 2023; Finn, 2015). For students, low-stakes assessment may not be considered important because there are no personal benefits or consequences tied to their performance (Wise, 2006). As a result, some students may not invest sufficient effort in responding to items, which could compromise the validity and reliability of test scores (Eklöf, 2006; Wise, 2015). Non-effortful or disengaged test-taking behavior is a widely recognized concern in the measurement literature, particularly in low-

*Corresponding author at: Okan BULUT bulut@ualberta.ca Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5 Canada

e-ISSN: 2148-7456

stakes testing contexts where low motivation is more prevalent (Finn, 2015). Disengaged behavior lowers performance and introduces construct-irrelevant variance to test scores, which could bias item statistics and distort validity and reliability measures (Finn, 2015; Rios & Deng, 2022; Wise & DeMars, 2009). Thus, there have been increasing calls for analytical methods that account for disengaged test-taking to ensure score validity (Finn, 2015).

With the advent of computerized testing and the ability to record process data during a test event, response time (RT) has emerged as a popular source of data for measuring test-taking engagement. RT, which is usually captured at the item-level, refers to the total time duration that the student spends on an item (e.g., from viewing the item to submitting a response). Extensive research has been done over the past two decades by Wise and colleagues (e.g., Wise, 2015; Wise & Gao, 2017; Wise & Kong, 2005; Wise & Ma, 2012) to make a validity argument for the use of RT-based metrics to identify disengaged responses. The general theory behind using RT as an indicator of engagement is that students who are disengaged or distracted tend to have RTs that are substantially shorter (i.e., rapid guessing; Wise, 2017) or longer (i.e., idling; Gorgun & Bulut, 2021) than students with sufficient test-taking effort. Unusually rapid responses suggest guessing or random selections, while idling may lead to other types of issues, such as failing to complete all the items within the allotted time.

Researchers have proposed various methodologies to identify and handle disengaged responses using RTs to mitigate their impact in the scoring process. Many of the previous studies have focused on identifying rapid guessing behavior and suggested filtering out rapidly guessed responses (i.e., recoding such responses as missing) or examinees with a large number of rapidly guessed responses (e.g., Rios *et al*., 2014; Wise, 2015; Wise *et al*., 2006). In the current study, we argue that this approach may be limited in several ways. First, removing a portion of response data through either examinee- or response-level filtering may lead to information loss and potentially jeopardize the validity of the assessment. Second, recent research suggests that in addition to rapid guessing, idling also represents another type of undesirable test-taking behavior as it often leads to not-reached items on the test (Gorgun & Bulut, 2023; Yildirim-Erbasli & Bulut, 2020). Hence, when evaluating disengagement with RTs, both types of behaviors (i.e., rapid guessing and idling) should be considered. Third, filtering could introduce bias into mean ability estimates if there is an association between students' ability levels and disengaged behaviors. Therefore, before deciding how to handle disengaged responses, it is important to first examine the relationships between RTs and students' abilities (Kuhfeld & Soland, 2020).

In this study, we introduce a novel application of the Nested Logit Model (NLM; Suh & Bolt, 2010), integrating both item responses and test-taking engagement derived from RTs within a unified framework. Traditionally applied to analyze nominal response options in multiple-choice items, we adapt the NLM model to categorize different test-taking behaviors—namely, rapid guesses, idle responses, and effortful responses—and incorporate these categories into the estimation of item parameters and examinee ability levels. To demonstrate the effectiveness of this approach, we utilized real data from a low-stakes reading assessment, comparing various scenarios in which item responses were modeled with and without accounting for RT-based test-taking behaviors. The results of this study provide insights for researchers and practitioners on how to harness the added value of RTs in modeling item responses, rather than discarding them from the data, thereby enhancing the accuracy and interpretability of assessment results.

Literature Review

Test-taking engagement in low-stakes assessments

Low-stakes assessments are often characterized as tests where the outcomes hold greater importance for test administrators than for test-takers (Wise & Kong, 2005). Since these tests do not directly impact students' grades or advancement, their motivation to put forth consistent effort on the test items is generally lower compared to high-stakes assessments (Finn, 2015; Wise & DeMars, 2005). Previous research showed that disengaged test-taking behavior in low-stakes assessments can occur at any point during the test, unlike in high-stakes assessments,

where it typically concentrates towards the end (Wise & Kong, 2005). Students may also alternate between disengaged and engaged response behaviors, depending on factors such as perceived item difficulty (Csányi & Molnár, 2024; Pastor *et al*., 2019; Wise & Kong, 2005).

In certain low-stakes testing contexts, test-taking engagement is of significant concern, especially if important educational or accountability decisions are made based on students' performance (Rios *et al*., 2014). Examples include statewide assessments that contribute towards accountability measures and policy decisions, universal screening and progress monitoring tools that help teachers identify at-risk students in core subjects such as reading and mathematics, and low-stakes tests that are administered to students for the purpose of piloting items intended for high-stakes exams (Wise & DeMars, 2006). In large-scale assessments such as the National Assessment on Educational Progress (NAEP) and the Programme for International Student Assessment (PISA), students do not receive feedback on their performance after the test, so the results are primarily used by administrators to track changes in achievement or to compare performance across countries or states. When disengaged test-taking occurs to a non-trivial extent in these assessments, students' test performance may not accurately reflect their true capabilities, hence raising concerns about the validity of the interpretations of their test scores (Rios *et al*., 2014; Wise, 2015). It is, therefore, important to always assess the extent of disengaged behavior in low-stakes assessments and develop strategies to mitigate its impact.

Disengaged test-taking behavior can take many forms. In the literature, it is usually characterized by rapid guessing, idling, omitting items, or providing random responses that do not reflect genuine effort or thought (Finn, 2015; Gorgun & Bulut, 2023; Soland *et al*., 2021; Wise, 2017). Researchers have proposed various methods to measure test-taking engagement, which broadly falls into two categories: self-report measures and behavioral measures (Eklöf, 2010; Finn, 2015). Self-report questionnaires are generally administered either at the end of a test (e.g., Butler & Adams, 2007; Swerdzewski *et al*., 2011) or at regular intervals throughout the test (e.g., Penk & Richter, 2017), asking examinees to report their level of engagement or effort. Common criticisms of self-report measures include their vulnerability to responder bias and their limited ability to capture variations in engagement throughout the test (Rios *et al*., 2014; Wise & Kong, 2005). As computer-based testing became more widely adopted, behavioral measures, particularly RTs, emerged as a valuable source of data for assessing disengaged behavior, due to their ability to be captured objectively and unobtrusively during digital assessments, and the granular insight they can provide at the item level rather than just at the student or test level (Wise & Ma, 2012).

RT as an indicator of test-taking engagement

RTs capture the length of time an examinee spends on each item on a test, which lend insight into their level of engagement during the test-taking process. Schnipke (1995) used RTs to distinguish between two types of response strategies on a high-stakes strictly-timed assessment—*solution behavior*, where students engage with an item to try to arrive at the right answer, and *rapid guessing behavior*, where students respond to an item in an unrealistically short amount of time. Following this line of research, a multitude of studies have extensively explored rapid-guessing behavior using RTs in low-stakes assessments, often identifying it as the main indicator of disengaged test-taking behavior (e.g., Lee & Jia, 2014; Leventhal & Pastor, 2024; Michaelides & Ivanova, 2022; Wise & Kong, 2005).

Recent studies have highlighted the importance of considering instances where some examinees spend an unusually long time on items, which may signal disengagement and can lead to rapid guessing in later sections of the test due to poor time management (Gorgun & Bulut, 2021; Hauenstein *et al*., 2024; Yildirim-Erbasli & Bulut, 2020). This phenomenon, known as *idling*, has been observed in several low-stakes assessments, including the Problem Solving and Inquiry tasks in the Trends in International Mathematics and Science Study (TIMSS) (e.g., Liu

et al., 2024), PISA (e.g., Pools & Monseur, 2021), and various progress-monitoring assessments in the United States (e.g., Gorgun & Bulut, 2021, 2023). The effect of idling on test performance is particularly pronounced under timed conditions, as examinees are expected to manage their time efficiently to avoid not reaching the final items or resorting to rapid guessing toward the end (Gorgun & Bulut, 2023; Tijmstra & Bolsinova, 2018; Zheng *et al*., 2023). Idling is also a significant concern in automaticity-based assessments, where speed and ability are jointly measured. Examples include reading assessments that require simultaneous word decoding, comprehension, and inference (e.g., Samuels & Flor, 1997), or math assessments that demand rapid and accurate recall of mathematical facts (e.g., Stickney *et al*., 2012).

Identifying disengaged responses using RTs typically involves setting suitable time thresholds to distinguish between effortful responses and other types of responses (i.e., rapid guesses and idle responses). To date, multiple methods have been proposed in the literature for establishing thresholds for rapid guessing, including: (1) using surface features such as item characteristics (Wise & Kong, 2005), (2) applying a common *k*-second threshold across all items (Wise *et al*., 2004), (3) visually inspecting RT distributions, either manually or automatically (Rios & Guo, 2020; Schnipke, 1995), (4) using a percentage of the average RT value (Normative Threshold [NT] method; Wise & Ma, 2012), (5) model-based methods that integrate response accuracy information with RTs (e.g., Cumulative Proportion [CUMP] method; Guo *et al*., 2016), and (6) data-driven methods to iteratively search for the optimal RT threshold within a particular range (Bulut *et al*., 2023). Of the above, the NT method (Wise & Ma, 2012) has been a popular choice among researchers and practitioners because it is relatively simple to implement and is currently being used in some operational low-stakes assessments (Soland *et al*., 2021). Recently, researchers have also extended this method to identify thresholds for idling, by applying a percentage of the average RT at the upper end of the RT distribution (Gorgun & Bulut, 2021, 2023).

Another line of research employed exploratory, person-centered techniques, such as latent class analysis and mixture modeling, to classify students based on their test-taking engagement. For instance, Lindner *et al*. (2019) analyzed response times from a low-stakes, computer-based science test and identified five distinct rapid-guessing patterns: one class showing no rapid guessing and four additional classes differing in when disengagement began (early, intermediate, late onset, or consistently low effort). Zheng *et al*. (2023) used mixture modeling with both RT and item scores in the digital version of TIMSS (eTIMSS) 2019 and detected multiple test-taking behavior profiles, including steady pacing, disengaged responding, speeded behavior near the end of the test, and an efficient group that worked quickly but accurately. In PISA 2022, Erdem Kara (2025) applied latent profile analysis to RT, number of actions, and self-reported effort, and identified three prominent engagement profiles: disengaged (10%), moderately engaged (83%), and highly engaged (7%). Most recently, Yıldırım Hoş and Saraç (2025) incorporated RT directly into a mixture item response theory (IRT) framework using PISA mathematics data and identified two latent classes: an engaged ability class where students allocated time strategically and a disengaged non-ability class characterized by rapid responses and guessing.

Utility of RT in modeling and handling disengaged behavior

While many studies have focused on identifying disengaged responses, determining how to handle them is also an important step in mitigating their impact on assessment outcomes. Wise and Kuhfeld (2020a) noted that disengaged responses provide little psychometric information and should be excluded from scoring to maximize measurement accuracy. They argued that such responses represent a momentary cessation of measurement and should be excluded from the scoring process if maximizing measurement accuracy is a priority. Researchers recommended motivation filtering methods, such as response-level or examinee-level filtering,

to remove such responses, assuming they are unrelated to true ability (Wise, 2015; Wise *et al*., 2006; Wise & DeMars, 2005). Although this assumption has been supported in some studies, others have challenged it (e.g., Deribo *et al*., 2021; Kuhfeld & Soland, 2020; Lindner *et al*., 2019), arguing that filtering responses can result in data loss and bias in ability estimates.

Researchers have also developed various methodologies to integrate RT information directly into IRT models to account for test-taking engagement. These models often define engagement as a dichotomous indicator of solution behavior (i.e., 1=effortful response and 0=rapid guessing) using RTs, allowing it to vary across items within individuals (i.e., examinees may show different levels of engagement for different items). In the effort-moderated IRT (EM-IRT) model by Wise and DeMars (2006), effortful responses are modeled using a traditional IRT framework where the probability of a correct response increases with ability, while rapid guesses are represented by a flat response function, indicating the probability of a correct response at a chance level. Other researchers have expanded on this approach, using different IRT models, broadening the scope of disengagement, or allowing for more flexible estimation of thresholds to account for individual differences (e.g., Nagy & Ulitzsch, 2022; Ulitzsch *et al*., 2020; Wang & Xu, 2015; Wang *et al*., 2024). For example, Leventhal and Pastor (2024) introduced an IRTree method (De Boeck & Partchev, 2012) that simultaneously models ability and disengagement (focused on rapid guessing) using a two-stage process. The first stage differentiates between engaged and disengaged responses, while the second stage distinguishes between correct and incorrect responses, allowing for a dependency between disengagement and ability.

In this study, we propose an alternative modeling approach that incorporates different response behaviors as ancillary information in the parameter estimation process. We first define a scoring scheme with nominal response categories based on two criteria: response accuracy (correct or incorrect) and test-taking behaviors (effortful responding, rapid guessing, or idling) determined by RT values. We then identify the most desirable outcome and use it as the key in the NLM (Suh & Bolt, 2010) to assign dichotomous scores during model estimation. For instance, if effortful and correct responses are identified as the key, they are scored as "1," while all other outcomes (e.g., incorrect rapid guesses, incorrect idle responses) are scored as "0." Unlike previous studies that treat disengagement as a binary variable (engaged or disengaged), our approach differentiates between various types of test-taking behaviors and incorporates them directly into the parameter estimation. Further details on our proposed approach are provided in the following section.

NLM for modeling test-taking engagement

The NLM approach was first introduced by Ben-Akiva (1973) in the context of travel demand modeling, and subsequently formalized and applied in different ways to model choice behavior (McFadden, 1981; Ben-Akiva & Lerman, 1985). In NLMs, the preferences for different choices available to a decision-maker are represented by a latent utility function, which consists of both observable factors (e.g., the attributes of the choices) and unobservable factors captured by error terms. Available choices are divided into mutually exclusive subsets called nests, and the error terms associated with the utility of each choice are allowed to be correlated within these nests (Aboutaleb *et al*., 2020). This allows the model to account for similarities between choices within the same nest, providing a more accurate representation of the decision-making behavior, particularly in cases where the decision is made through a sequential or hierarchical process (Bolt *et al*., 2012).

Suh and Bolt (2010) adapted the NLM approach to analyze response data from multiple-choice items where examinees make a choice among discrete response options. In their model, discrete response options are separated into two nests, one with the correct response only, and the other with all the distractors (incorrect responses). Using these nests, the NLM operates on a higher level that distinguishes between a correct and an incorrect response, modeled using logistic IRT

models such as the 2-parameter (2PL) or 3-parameter logistic (3PL) model; and a lower level (conditional upon an incorrect response) that distinguishes between distractors, modeled using Bock's Nominal Response Model (NRM; Bock, 1972). Under the 2PL-NLM model (Suh & Bolt, 2010), the probability of selecting an incorrect response category v ($v = 1,2,\dots,M$) for item *i* can be written as:

$$P(X_i = 0, U_{iv} = 1|\theta) = \left[1 - \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}\right]\left[\frac{e^{(\zeta_{iv} + \lambda_{iv}\theta)}}{\sum_{k=1}^{M} e^{(\zeta_{ik} + \lambda_{ik}\theta)}}\right], \quad (1)$$

where X_i is the examinee's response to item *i* ($X_i = 1$ if correct, and 0 otherwise), U_{iv} denotes the response for incorrect response option v, θ represents the examinee's ability level, a_i and b_i are the discrimination and difficulty parameters for item *i*, and ζs and λs are the intercept and slope parameters for the incorrect response options. The first bracketed term in Equation 1 describes the probability of an incorrect response using the 2PL model, while the second term computes the probability of selecting an incorrect category v based on Bock's (1972) NRM.

The 2PL-NLM model in Equation 1 can be reorganized to take test-taking engagement into account through nominal indicators of engagement operationalized through normative RT thresholds. Assume that Y_i refers to the positive outcome for item *i* (e.g., a correct and effortful response) and N_{iv} is the nominal indicator of negative outcome v ($v = 1,2,\dots,D$) for item *i* (e.g., 1: correct or incorrect response with rapid guessing, 2: incorrect idle response, and 3: effortful but incorrect response). With these new components, the 2PL-NLM can be written as follows:

$$P(Y_i = 0, N_{iv} = 1|\theta) = \left[1 - \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}\right]\left[\frac{e^{(\zeta_{iv} + \lambda_{iv}\theta)}}{\sum_{k=1}^{D} e^{(\zeta_{ik} + \lambda_{ik}\theta)}}\right], \quad (2)$$

where $Y_i = 1$ if the response is a positive outcome and 0 otherwise, and $N_{iv} = 1$ indicates the presence of a negative outcome (0 otherwise). In Equation 2, the first bracketed term computes the probability of a user-defined positive outcome (e.g., correct and effortful responses) while the second term focuses on the probability of a negative outcome, which can involve exhibiting a disengaged response behavior.

Unlike the original 2PL-NLM where the probabilities of incorrect responses are evaluated independently of the correct response probability, the proposed 2PL-NLM evaluates the probabilities of correct or incorrect responses depending on the type of test-taking behavior exhibited for each item. This implies that the operationalization and interpretation of ability (θ) changes depending on how the positive outcome is defined. For example, if the positive outcome is defined as correct and effortful responses, θ estimated through the 2PL-NLM approach in Equation 2 then represents an effort-prioritized ability that excludes disengaged responses (i.e., rapid guesses and idle responses).

Current Study

In this study, we explored the application of the 2PL-NLM to estimate item parameters and ability levels while incorporating test-taking engagement as reflected by RTs. We began by categorizing dichotomous responses from a low-stakes reading assessment into distinct response types, distinguishing between behaviors such as incorrect but effortful responses, incorrect responses with rapid guessing, incorrect responses with idling, and correct but effortful responses, each under distinct scoring schemes. These response categories were then grouped into two main nests: a positive outcome group (e.g., correct and effortful responses) and a negative outcome group (e.g., incorrect responses with rapid guessing or idling). Utilizing the 2PL-NLM, we estimated item parameters and ability levels by accounting for both response accuracy and test-taking engagement. The main research questions guiding the current study

are: 1) Compared with traditional and effort-moderated IRT models, does the NLM framework integrating RT-based engagement data yield more accurate information about students? 2) Does the performance of the NLM vary across various scoring scenarios? We hypothesize that if students' test-taking engagement levels vary significantly, accounting for this variation in the model estimation should enhance the accuracy of the model parameters.

METHOD

Sample and Instrument

To conduct this empirical study, we used data from a low-stakes computer-based reading assessment administered to 27,556 students (5th graders: 29.41%, 6th graders: 28.39%, 7th graders: 21.57%, and 8th graders: 20.63%) in the United States. The assessment was designed as a K-12 screening tool to evaluate students' reading ability and monitor their progress throughout the school year (fall, winter, and spring). It comprises four subtests, each with 30 short multiple-choice questions, totaling 120 items. All items have four response options. These items assess a range of reading skills, such as word recognition and synonym matching. The assessment system recorded both students' responses and their RTs for each item. Students were given a maximum of 10 seconds to complete each item. If a student failed to provide a response within 10 seconds or skipped the item without choosing a response option, the system treated it as incorrect, which occurred in less than 0.001% of all item responses.

Engagement-Based Scoring

We used RTs to identify three types of response behaviors: rapid guessing, idling, and effortful responding. To define rapid guessing, we adopted the NT20 method (Wise & Ma, 2012), which sets a time threshold at 20% of the average RT for each item. Even though NT10 was originally recommended by Wise and Ma (2012), more recent research suggested that it may be too conservative (Wise & Kuhfeld, 2020b). To minimize the impact of outliers in the RT distribution, we used the median RT instead of the mean to represent the average. Following the approach introduced by Gorgun and Bulut (2021) and recently used by Yildirim-Erbasli and Gorgun (2025), we identified idling for each item by setting a time threshold at 180% of the median RT. For example, suppose the median RT for a particular item is 3 seconds, then responses under 0.6 seconds were classified as rapid guesses, while those taking over 5.4 seconds were classified as idle responses. Any responses between the 20% and 180% thresholds were identified as effortful responses.

Dichotomous scoring

We utilized six scoring approaches to evaluate items with and without incorporating RTs (see Table 1). The first three cases maintained the dichotomous nature of the response dataset. Case 1 involved the original response dataset, where incorrect responses were scored as 0 and correct responses as 1. For this case, the 2PL model was used to estimate the item and ability parameters based on dichotomous responses alone.

Case 2 followed established recommendations from the literature to exclude disengaged responses from scoring (e.g., DeMars, 2007; Wise, 2017; Wise & Kuhfeld, 2020a). Here, responses identified as rapid guesses or idle were recoded as missing and subsequently excluded from parameter estimation. We used the 2PL EM-IRT model for parameter estimation in this case.

Case 3 was similar to Case 2 in terms of retaining effortful and correct responses as 1. However, all other responses in Case 3 were recoded as 0. The NT20 method was used to identify rapid guessing, idling, and effortful responding. This information, along with response accuracy (i.e., 1=correct and 0=incorrect), was used to recode the responses. The hypothesis underlying this case is that only effortful and correct responses should indicate the true ability level, while all other response types—whether due to incorrect effortful responses, rapid guessing, idling, or

even correct responses achieved without full engagement—do not provide valid evidence of ability. This implies that accuracy combined with engagement is necessary to reflect genuine ability. For this case, the 2PL model was used for parameter estimation.

Table 1. *Models and scoring schemes used for the six cases.*

Case	Model	Scoring Scheme	
		Positive Outcome	Negative Outcomes
1	2PL	1: (ER, Id, RG) & C	0: (ER, Id, RG) & I
2	EM-2PL	1: ER & C	0: ER & I
3	2PL	1: ER & C	0: (ER, Id, RG) & I 0: (Id, RG) & C
4	2PL-NLM	4: (ER, Id) & C	1: RG & (C, I) 2: Id & I 3: ER & I
5	2PL-NLM	4: ER & C	1: (RG, Id) & I 2: (RG, Id) & C 3: ER & I
6	2PL-NLM	4: ER & C	1: RG & (C, I) 2: Id & (C, I) 3: ER & I

Note. ER: Effortful responding; Id: Idling; RG: Rapid guessing; C: Correct; I: Incorrect. EM-2PL refers to the effort-moderated 2PL IRT model where disengaged responses were discarded by recoding them as NA in the response dataset. For Cases 4, 5, and 6, the category of 4 was scored as 1, and the other categories were scored as 0.

Nominal scoring

For the remaining cases (i.e., Cases 4 to 6), we transformed dichotomous responses into nominal categories by incorporating RTs using the three categories identified through the NT20 method: rapid guessing, idling, and effortful responding. Each case generated a dataset with four response categories (1 through 4), with category 4 designated as the positive outcome. The 2PL-NLM was applied to estimate parameters across these cases.

In Case 4, rapid guesses were coded as 1, irrespective of accuracy, while idle and incorrect responses were coded as 2. Effortful but incorrect responses were coded as 3, and correct responses that were effortful or idle were coded as 4. This case posits that idle but correct responses might reflect a level of engagement similar to effortful responses, even though they were made more slowly than rapid guesses. By coding idle and correct responses together with effortful responses as the positive outcome, the hypothesis is that both response types contribute to a shared ability. This case aims to test whether idle yet correct responses should be classified as evidence of ability similar to effortful responses.

In Case 5, incorrect responses characterized by rapid guessing or idling were coded as 1, while correct responses with the same disengaged behaviors were coded as 2. Effortful but incorrect responses were coded as 3, and effortful and correct responses as 4. In this case, the hypothesis is that both idle and rapid responses—whether correct or incorrect—may represent different levels of engagement that reflect a shared latent ability. By distinguishing between correct and incorrect idle/rapid responses while treating both behaviors as negative outcomes in scoring, this case seeks to determine if engagement, as shown through response type (idle or rapid), is a consistent indicator of underlying ability regardless of response accuracy.

In Case 6, rapid guesses were coded as 1, and slow responses as 2, regardless of accuracy. Effortful but incorrect responses were coded as 3, and effortful and correct responses as 4. This case hypothesizes that rapid and idle responses may represent a type of low test-taking engagement that impacts ability estimates, regardless of response accuracy. Here, the focus is on assessing whether RT alone, independent of accuracy, serves as an indicator of test-taking

engagement. By recoding rapid and idle responses uniformly and analyzing their contribution to ability, the hypothesis is that these response behaviors might dilute the ability estimate, regardless of whether the response was correct. All of the IRT models for the six cases described above were estimated using the mirt package (Chalmers, 2012) in R (R Core Team, 2023).

Evaluation Criteria

The first evaluation criterion used in the current study was the model-data fit for the IRT models based on the six cases described above. To assess the model-data fit, we used several fit indices, including the Root-Mean-Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Following established guidelines (Hu & Bentler, 1999), CFI and TLI values above 0.95 and RMSEA values below 0.08 indicated good fit. Additionally, lower AIC and BIC values reflected a better relative fit, helping to identify the model that best aligned with the observed data.

The second evaluation criterion involved analyzing the test information function (TIF) and conditional standard error of measurement (cSEM) for each of the six cases. TIF was examined to determine the ability levels at which the test provided the most information, while cSEM indicated the precision or reliability of the test at those levels. Notably, as TIF values increased for a given ability level, cSEM values correspondingly decreased, signifying enhanced measurement precision.

The third evaluation criterion was the similarity and consistency of the ability estimates from the six cases. We computed Pearson correlations among the ability levels from the cases. In addition, we calculated the intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979), which shows the consistency of the ability estimates when the same students were evaluated under different scoring conditions. The ICCs helped us determine if the six cases yielded consistently similar estimates of ability.

Lastly, we performed latent class analysis (LCA) to identify student subgroups within the response dataset based on patterns of test-taking behavior (e.g., 1=rapid guessing, 2=idling, or 3=effortful responses based on the NT20 method). This analysis was used to explore if distinct profiles of test-takers exhibited similar response patterns across cases and if these profiles were associated with differences in ability estimates. After identifying latent classes, we examined how each scoring case affected ability estimates within each class. This analysis aimed to generate further insights into whether cases were robust across different response behavior profiles. LCA was performed using the poLCA package (Linzer & Lweis, 2011) in R (R Core Team, 2023).

RESULTS

Model Fit

Table 2 shows the model fit results for the six cases. In examining model fit across the three dichotomous scoring cases (Cases 1 to 3), distinct patterns emerged, underscoring the impact of disengaged responses on parameter estimation and overall model performance. Case 1, which utilized a straightforward dichotomous scoring system (0 for incorrect, 1 for correct), yielded the poorest model-data fit. This result aligns with expectations, as the inclusion of rapid guesses and idle responses likely distorted the underlying ability estimates, introducing noise that impaired the 2PL model's accuracy. This finding highlights the sensitivity of dichotomous models to disengaged responses, which do not reflect true ability and degrade data quality.

Case 2 addressed these issues by treating disengaged responses (those identified as rapid guesses or idle) as missing, excluding them from the parameter estimation. This adjustment improved model-data fit, as the EM-2PL model no longer contended with potentially misleading data. By excluding disengaged responses, the model leveraged only effortful

responses, enhancing its capacity to produce more accurate estimates of true ability. This case offers compelling evidence that disengaged responses significantly impair model performance and that data quality can be improved by filtering out such responses.

Table 2. *Model fit results for the six cases.*

Case	Model	CFI	TLI	RMSEA
1	2PL	0.914	0.912	0.048
2	EM-2PL	0.998	0.997	0.010
3	2PL	0.930	0.929	0.046
4	2PL-NLM	0.950	0.948	0.035
5	2PL-NLM	0.948	0.945	0.030
6	2PL-NLM	0.959	0.957	0.028

Note. The first three cases use dichotomous responses, while the last three cases use nominal engagement categories in addition to dichotomous responses. The model fit indices for Case 2 (EM-2PL) computed after excluding NAs resulted from rapid guessing and idling.

Case 3 introduced a more nuanced scoring approach, retaining effortful and correct responses as 1 but coding all other responses—including rapid guesses, idling, and incorrect effortful responses—as 0. While Case 3 improved model-data fit compared to Case 1, it was less effective than Case 2. This outcome suggests that penalizing disengaged responses by recoding them as incorrect retains some of their noise, ultimately impacting model performance to a lesser extent than excluding them entirely.

Across the three cases employing nominal scoring (Cases 4 to 6), the use of the 2PL-NLM produced consistently better model-data fit compared to the 2PL models where all responses, regardless of engagement, were treated equivalently (as in Cases 1 and 3). This finding highlights the value of incorporating RT-based engagement categories, which capture varying levels of test-taking engagement and offer a more nuanced view of examinee behavior.

The model-data fit in Case 4 suggested that idle, correct responses could serve as partial indicators of ability, aligning with the hypothesis that correct idle responses might reflect a form of engagement more akin to effortful responses than to rapid guesses. However, Case 4 did not achieve the strongest model fit among the nominal scoring cases, indicating that while this approach improved fit, it might still capture some noise from idle responses.

By distinguishing accuracy within the categories of idle and rapid responses, Case 5 examined whether disengaged behaviors of different accuracy levels reflected a shared latent ability. Although the model-data fit was superior to the 2PL model of Cases 1 and 3, it did not outperform Case 6. This finding suggests that the engagement type (rapid or idle) may reflect underlying ability to an extent, but that further refinement is needed to capture engagement as an indicator of ability fully.

Case 6, which yielded the best model-data fit among the three nominal cases assumed that both rapid and idle responses may represent distinct categories of test-taking engagement, regardless of accuracy. The model fit in Case 6, meeting Hu and Bentler's (1999) guidelines for good model-data fit, supports the notion that RT alone can be a reliable proxy for operationalizing test-taking engagement. This result suggests that responses made with low engagement—whether rapid or idle—may indeed dilute the ability estimation. Therefore, distinguishing disengaged responses from effortful ones, independent of correctness, may provide a clearer measure of underlying ability by minimizing the impact of disengaged responses.

In summary, the comparison between dichotomous (Cases 1 to 3) and nominal (Cases 4 to 6) scoring approaches provides valuable insights into the impact of test-taking engagement on IRT model estimation. Generally, the nominal scoring cases demonstrated superior model-data fit compared to the dichotomous cases, with Case 6 achieving the best overall fit. Notably, while

Case 2 in the dichotomous group showed competitive fit by excluding disengaged responses, the nuanced categorization of engagement in the nominal cases offered a more refined model that better accounted for variations in examinee behavior.

Comparison of TIF and cSEM

Figures 1 and 2 present a comparative analysis of TIF and cSEM across the six cases to further clarify how engagement-informed scoring influenced the precision of ability estimation across different levels of the ability spectrum. The TIFs for Cases 1, 2, and 3 showed relatively low information across the ability range compared to the nominal cases in the bottom panel in Figure 1. Case 2 indicated the highest TIF among the dichotomous cases, indicating that excluding disengaged responses enhanced the model's ability to capture ability accurately. This supports the idea that removing disengaged responses could improve data quality. Cases 1 and 3 provided relatively less information, with Case 3 showing the lowest peak. This finding suggests that including disengaged responses (Case 1) or penalizing them by recoding them as incorrect (Case 3) limits the accuracy of ability estimation.

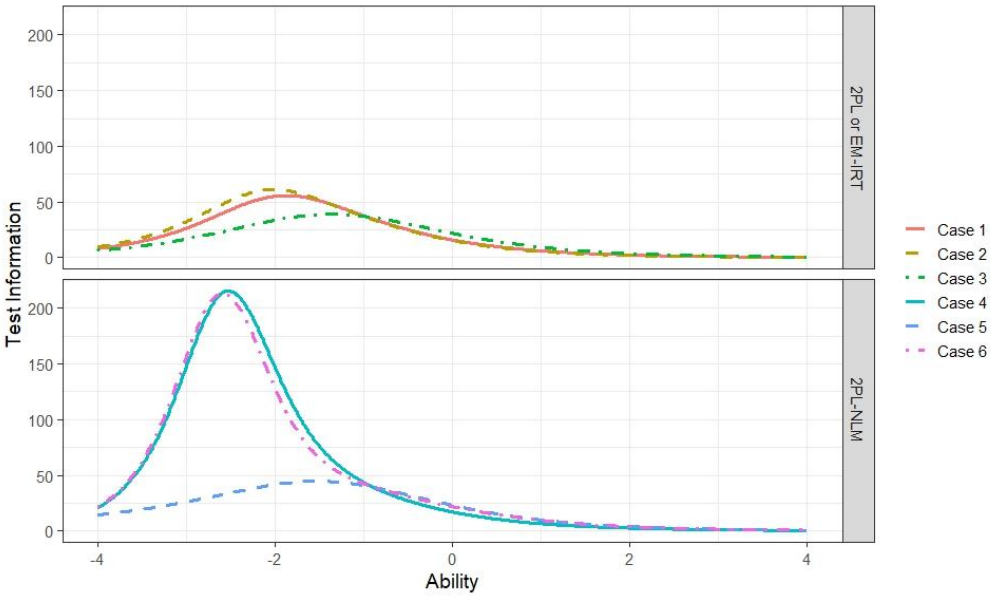

Figure 1. *Test information functions (TIFs) of the six cases.*

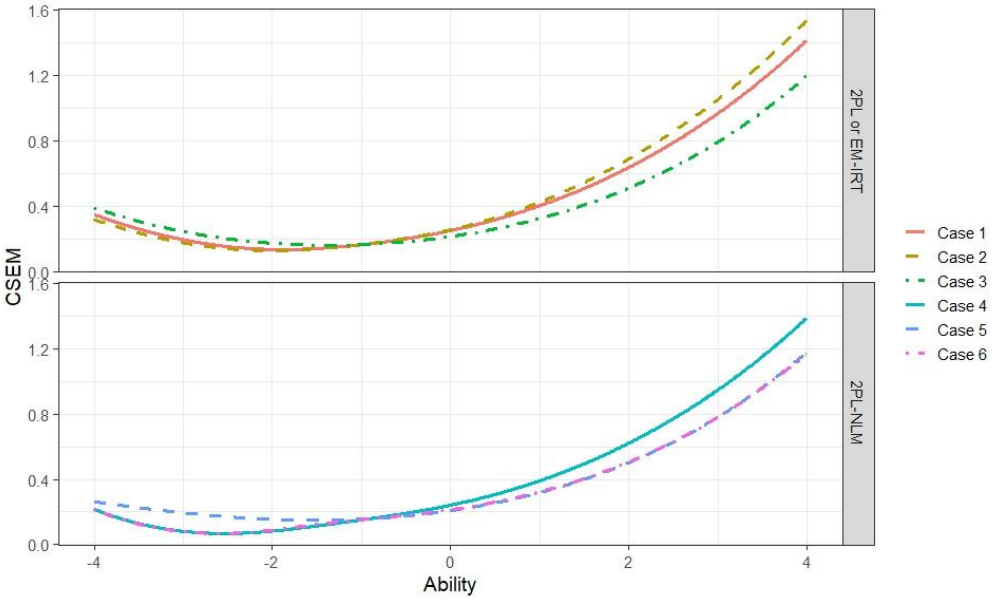

Figure 2. *Conditional standard error of measurement (cSEM) of the six cases.*

Two of the nominal cases (Cases 4 and 6) demonstrated a substantial increase in the test information, particularly in the ability range between θ=-4 and θ=-1. This finding indicates that incorporating engagement through nominal scoring enhanced the model's precision in estimating lower ability levels. Both Cases 4 and 6 reached their highest TIF peaks around θ=-2.5, suggesting that these scoring approaches—where rapid and idle responses were consistently treated as indicators of low engagement regardless of accuracy—provided the most accurate reflection of lower abilities. This finding supports the hypothesis that differentiating types of disengagement (e.g., rapid guessing vs. idling) is a key factor in improving measurement precision. While Case 5 also accounted for test-taking engagement, it yielded notably lower levels of test information compared to Cases 4 and 6, though it still performed comparably to the dichotomous cases.

The cSEM for Cases 1, 2, and 3 was relatively high across the ability range, particularly for θ>0, indicating less precision in estimating higher levels of ability (see Figure 2). This aligns with the TIF results, where these cases provided lower information. Cases 2 and 3 yielded the lowest cSEM for lower and higher ability levels, respectively. By contrast, Case 1, which included all responses without adjustment, produced higher cSEM, reinforcing its poorer performance in capturing ability accurately. The nominal scoring cases (Cases 4 to 6) exhibited substantially lower cSEM, especially around the low to middle range of ability. This trend corresponds with the high TIF values seen in the nominal cases, where these cases captured the ability more effectively. Case 6 showed the lowest cSEM, reflecting the highest precision across the ability range. This complements its superior TIF, indicating that categorizing responses based on engagement level, independent of accuracy, provides the most precise ability estimates. Cases 4 and 5 also show reduced cSEM compared to the dichotomous cases, but with slightly higher values than Case 6.

Similarity and Consistency of Ability Estimates

The ability estimates derived from the six cases were strongly correlated ($r \geq .89$; see Table 3), indicating that the overall rankings of students remained largely consistent across models. This result was anticipated because the 2PL, EM-IRT, and 2PL-NLM models processed correct responses similarly. Additionally, the low incidence of disengaged responses in the dataset (rapid guessing < 2% and idling < 5%) meant that retaining or excluding these responses had minimal impact on ability estimates across the six cases. Furthermore, the high ICC coefficient of .95 ($F(27527, 137635)=107$, $p < .001$) reflects a strong consistency in ability estimates across cases, supporting their consistency (Koo & Li, 2016). The high ICC suggests that the ability scores produced by each case were highly stable and that disengaged response behaviors did not substantially influence the measurement of ability.

Table 3. *Correlations among the ability estimates from the six cases.*

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Case 1 (2PL)	1					
Case 2 (EM-IRT)	.970	1				
Case 3 (2PL)	.940	.896	1			
Case 4 (2PL-NLM)	.983	.950	.926	1		
Case 5 (2PL-NLM)	.949	.895	.997	.938	1	
Case 6 (2PL-NLM)	.939	.896	.979	.957	.982	1

Note. The first three cases use dichotomous responses, while the last three cases use nominal engagement categories in addition to dichotomous responses.

Ability Estimates by Latent Engagement Classes

LCA with two- to five-class solutions was conducted using nominal indicators of test-taking engagement (1 = rapid guessing, 2 = idling, 3 = effortful responding) derived from RT values.

Model fit was evaluated using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the likelihood ratio statistic (G^2), with lower values indicating a better fit. Also, entropy was calculated to estimate how well students were divided into classes based on their posterior class membership probabilities as an indicator of classification quality. Entropy improves when classes are more distinct and individuals are assigned clearly.

Table 4 displays the LCA results for models with two to five classes. Although AIC and G^2 continued to decrease as additional classes were added, indicating incremental improvement in model fit, the reduction in these indices diminished sharply after the 3-class model. Specifically, the improvement from 3 to 4 classes was marginal relative to the increase in model complexity (G^2 decreased by 35,980), and the change from 4 to 5 classes was even smaller (only 15,370). In contrast, the improvement from 2 to 3 classes was substantially larger (G^2 decreased by 86,668), suggesting that the major structure in the data is already explained by three classes. Additionally, the 3-class solution offers the strongest separation as it produced the highest entropy value. Despite having lower BIC values, the 4-class and 5-class solutions split existing classes and produced very small classes that were difficult to distinguish from one another, further supporting the choice of a more parsimonious 3-class solution.

Table 4. *LCA results based on engagement indicators derived from response times.*

Class	*n* of Parameters	AIC	BIC	G^2	Entropy
2	481	2262305	2266260	1724910	0.913
3	722	2176119	2182056	1638242	0.957
4	963	2160621	2168540	1602262	0.920
5	1204	2155733	2155633	1586892	0.931

Note. G^2 = (Likelihood ratio/deviance statistic).

Figure 3 demonstrates the proportions of each test-taking engagement indicator within each latent class in the three-class solution. In each class, effortful responding had the highest overall percentage. However, Class 1 included a substantial proportion of rapid guesses, while Class 2 had a high proportion of idle responses. Based on these patterns, the three latent classes were labeled as follows: Class 1 = Rapid guessers, Class 2 = Idle responders, and Class 3 = Effortful responders.

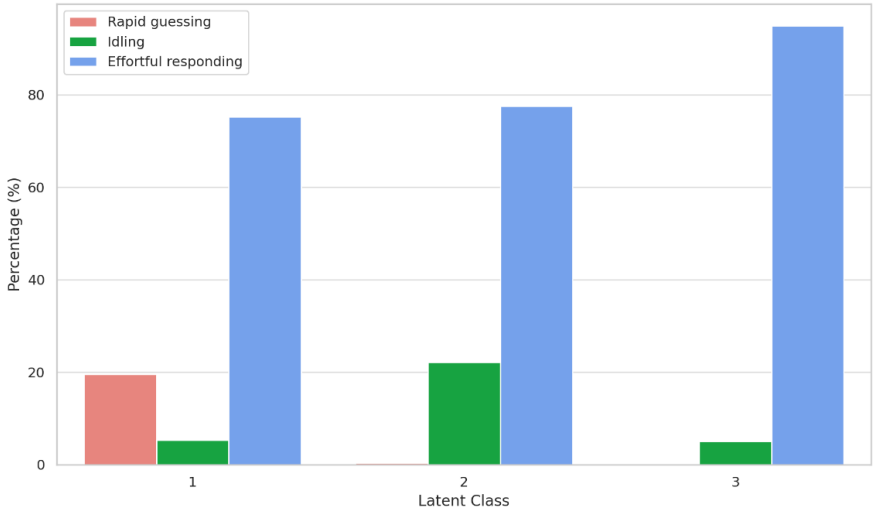

Figure 3. *Percentages of test-taking engagement indicators by latent engagement classes.*

Figure 4 shows the distribution of the ability estimates from the six cases by these latent classes. Across Cases 1 to 3, the distributions of ability estimates seemed somewhat similar for each engagement class. Effortful responders generally showed higher ability estimates compared to idle and rapid responders, which seemed consistent across all three dichotomous cases. Rapid

responders displayed a relatively low ability distribution in each case, with the peak of their distribution shifted further left than the other groups, Also, there was a small variation in the shapes of distributions for each class across Cases 1 to 3, suggesting that the dichotomous scoring approaches may have limited sensitivity to differentiate ability estimates across engagement types.

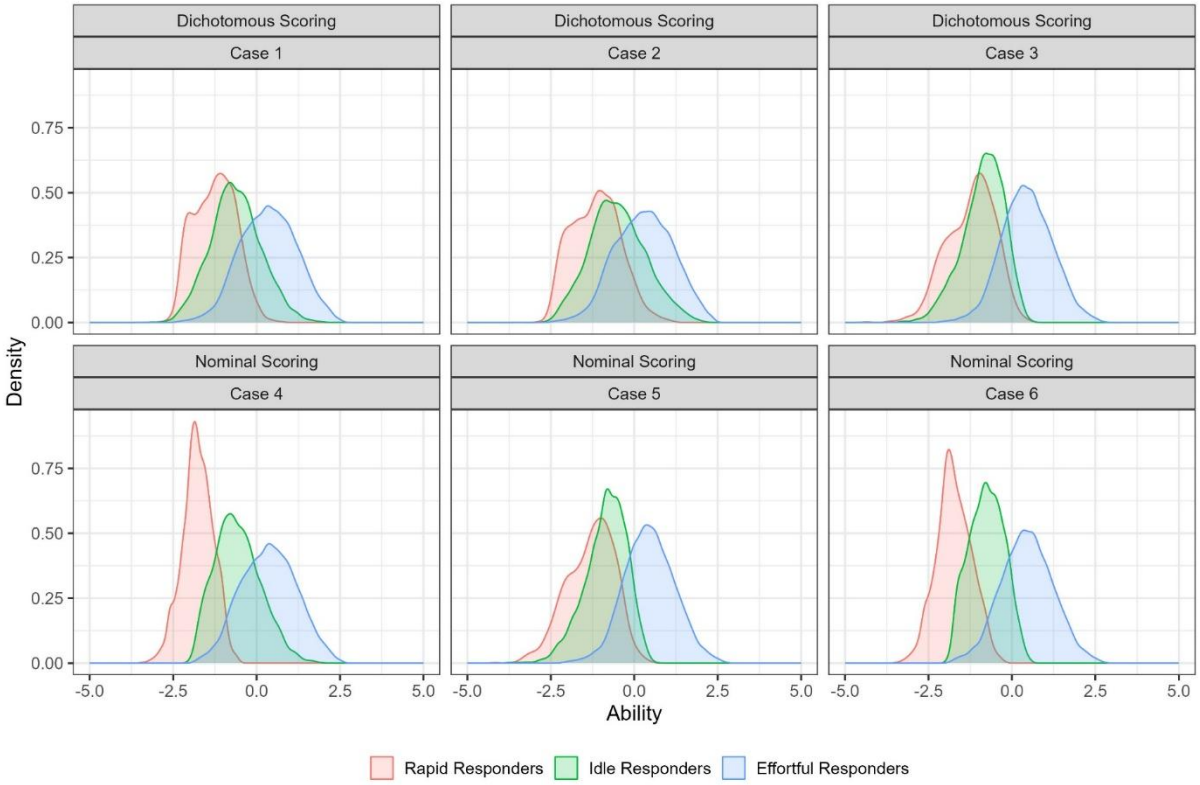

Figure 4. *Density plots of ability estimates from the six cases by engagement classes.*

Cases 4 to 6 exhibited more noticeable variation in the distribution shapes across three engagement classes compared to the dichotomous scoring cases, indicating that nominal scoring could capture nuances in engagement behavior more effectively. For each nominal scoring case, effortful responders maintained a higher ability distribution, peaking further to the right. However, the nominal scoring appears to accentuate differences between effortful responders and other groups, especially in Case 6. Idle and rapid responders showed lower ability estimates under nominal scoring, with distributions for rapid responders sharply peaking on the left, especially in Cases 4 and 6. This finding highlights that nominal scoring, by distinguishing disengaged behaviors more explicitly, might be better at penalizing low-engagement responses, resulting in lower ability estimates for these classes.

The consistency in higher ability estimates for effortful responders across all cases supports the reliability of this class as representative of genuinely engaged students. The sharper differentiation in ability estimates between engagement classes in the nominal cases suggests that nominal scoring may provide a more accurate reflection of ability for students with varying engagement levels. Overall, the results suggest that nominal scoring might be particularly useful in low-stakes assessments where engagement levels vary widely, as it allows for a more nuanced interpretation of ability based on test-taking behavior.

DISCUSSION and CONCLUSION

The lack of test-taking engagement is a common phenomenon in low-stakes assessment settings. Because of its potential to distort ability estimates and undermine the validity and reliability of assessment results, it is important to develop methods to effectively identify and

handle disengaged responses. In this study, we explored a novel application of the 2PL-NLM (Suh & Bolt, 2010) to estimate item parameters and ability levels by integrating RT-based engagement categories with item responses. Our analysis involved comparing six distinct cases that differed in how student responses were coded for use in either the 2PL model, the EM-2PL model, or the 2PL-NLM. RTs were first used to classify responses as disengaged (rapid guesses, idle responses), or engaged (i.e., effortful responses) using the NT method (Wise & Ma, 2012). The first three of the six cases employed dichotomous scoring, either retaining all disengaged responses in the dataset (Case 1), setting them to missing (Case 2), or setting them to 0 (Case 3). The remaining three cases employed nominal scoring with four response categories to further differentiate between different response behaviors. In particular, Cases 4 and 6 distinguished between rapid guesses and idle responses in the response categories for negative outcomes. However, Case 4 also included idle but correct responses as part of the positive outcome. In contrast, Case 5 distinguished between correct and incorrect disengaged responses but did not separate rapid guesses from idle responses.

Among the six cases, Case 1, which did not take the RT-based engagement information into account, yielded the poorest model-data fit and showed relatively lower test information and higher cSEM across the ability range compared to the other cases. This finding supports the claim that there is value in incorporating RT information into measurement models (De Boeck & Jeon, 2019; Sideridis & Alahmadi, 2022), and that responses exhibiting disengaged behavior should be treated separately from effortful responses (e.g., Deribo *et al*., 2021; Wise & DeMars, 2006).

Case 2 incorporated RTs by removing disengaged responses from the scoring process, akin to response-level filtering as seen in the literature (e.g., Rios *et al*., 2017). Notably, this approach yielded the best model-data fit among all the cases, offering evidence that disengaged responses contribute to noise in the data and that response-level filtering works well to improve data quality. In Rios *et al*.'s (2017) study, the authors also acknowledged that this type of filtering can effectively improve data quality, recommending it over examinee-level filtering in most cases due to its less restrictive underlying assumptions. In terms of test information and cSEM however, there were less pronounced differences compared to those observed for Case 1. We argue that even though filtering can help to reduce noise in the data due to disengaged responses, it may not be as effective at improving the quality of measurement. Moreover, recent literature recommended that filtering, either at the examinee or response levels, should be undertaken with caution because it could introduce bias in parameter estimates and aggregate scores (Deribo *et al*., 2021; Kuhfeld & Soland, 2020; Michaelides *et al*., 2024). Prior to applying filtering, the relationship between students' ability and other factors, such as propensity (or tendency) to exhibit disengaged behavior, should be investigated (Deribo *et al*., 2021).

In Case 3, disengaged responses were treated as incorrect, and our results indicated that this approach was not an effective way of handling these responses as there appeared to be no clear benefit to either model fit or measurement precision in terms of improving test information. In the literature, some studies have suggested that treating all disengaged responses (specifically rapid guesses) as incorrect may be appropriate under unspeeded or high-stakes test conditions and could even improve ability estimation (Deribo *et al*., 2023; Wright, 2016). However, this approach is problematic in low-stakes assessment contexts, as it may be unclear whether a rapidly guessed item could have been answered correctly with full engagement (Deribo *et al*., 2023). The results from Cases 1 to 3 highlight the need for more nuanced methods to handle disengaged responses.

In Cases 4 to 6, we demonstrated the use of the 2PL-NLM to model nominal response categories that incorporated RT-based test-taking engagement. Our findings indicated that these cases generally achieved better model-data fit compared to the dichotomous cases that had the same

treatment for all responses regardless of engagement. Notably, the ability to account for rapid guessing and idle responding as distinct disengagement behaviors offers promising benefits for improving measurement precision at lower ability levels, as seen by the substantial increase in test information in Cases 4 and 6.

Our results from LCA provide additional evidence supporting the existence of three latent engagement classes. Several recent studies have expanded the conceptualization of disengaged behaviors, as indicated by RTs, beyond rapid guessing to include idling (e.g., Gorgun & Bulut, 2021, 2023; Lundgren & Eklöf, 2023; Yildirim-Erbasli & Bulut, 2021). However, most of these studies have treated rapid guesses and idle responses as a single category of disengagement during analysis. Gorgun and Bulut (2021), in their investigation of methods for handling not-reached items, proposed a polytomous scoring approach that assigns partial credit based on both the accuracy and speed of responses. In one scenario they tested, rapid or idle correct responses were assigned distinct scores; however, this approach did not outperform cases in which disengaged responses were treated uniformly. In contrast, our study differentiated response types by using nominal categories to classify disengaged behaviors, rather than assigning scores based on response accuracy or engagement. This approach avoids the implicit assumption that rapid guessing or idle responding is inherently superior. The NLM framework facilitates a more nuanced analysis of disengaged behaviors. Our findings demonstrate the potential of the NLM approach as a valuable tool for leveraging item RTs to extract additional information and enhance the accuracy of low-stakes assessments.

Implications

The results have direct implications for educators, assessment developers, and policy stakeholders who rely on low-stakes assessments for screening, placement, and progress monitoring. First, applying the 2PL-NLM with nominal response categories that distinguish between rapid guesses and idling responding (as in Cases 4 and 6) offers a more effective approach for enhancing ability estimation in low-stakes assessments, especially at lower ability levels. One of the main goals of these assessments is to track progress and identify areas where students need more support (Filiasov & Sweetman, 2023). Considering that low-ability students may be more likely to exhibit disengaged behavior (Gorgun & Bulut, 2023; Wise *et al*., 2009), obtaining a more accurate reflection of ability in the lower theta range would be especially useful to guide interventions in practice.

Second, removing rapid guesses and idle responses (i.e., filtering), or treating them as incorrect, do not seem to be effective solutions for treating disengaged responses. While these strategies help to reduce noise in the data, our results did not indicate any clear benefits to measurement precision across the ability spectrum. Our study fills a gap in the literature as previous studies that examined filtering typically focused on its impact on mean or aggregate scores rather than on individual ability estimates (e.g., Michaelides *et al*., 2024; Rios *et al*., 2017; Wise *et al*., 2006). The practical implication is that practitioners should be cautious about relying on these conventional strategies, as they may oversimplify the disengagement problem and not necessarily lead to increased measurement precision or accuracy.

Third, considering both rapid guessing and idle responding together within the NLM framework can add further value to the reliability and validity of low-stakes assessments compared to approaches that focus solely on rapid guessing. Our findings support the presence of three latent engagement classes: effortful responding, rapid guessing, and idling. Systematically accounting for different types of disengaged behaviors in the measurement model ensures that test scores more accurately reflect the intended constructs, and that there is greater consistency in ability estimates across individuals with varying engagement patterns. When deciding on a model to handle disengaged responses in a low-stakes assessment, we recommend, based on this study's findings, for practitioners to choose a model that (1) does not assume independence between disengaged behavior and ability, (2) uses all available response data, and (3) accounts for and

distinguishes between different types of disengaged test-taking behavior beyond just rapid guessing.

Fourth, the findings from this study also carry meaningful implications for a wide range of assessment applications, particularly adaptive testing systems and large-scale assessment programs that routinely rely on RT data but often underutilize it in scoring. In computerized adaptive tests (CATs), where item selection depends on precise interim ability estimates, disengaged responding can destabilize the item selection algorithm and misdirect the measurement trajectory. Integrating RT-informed nominal scoring within a CAT framework can mitigate these disruptions by more accurately distinguishing low-ability responses from low-engagement responses, thereby reducing the risk of inappropriate item difficulty reductions or inflated measurement error.

Fifth, although this study used data from a U.S. low-stakes reading assessment with short, timed items, the broader framework is applicable across domains and geographic settings. Research in mathematics, science, and problem-solving assessments has already shown that rapid guessing and idling occur across content areas and cultures, particularly in low-stakes or mixed-stakes conditions. The NLM-based scoring approach does not rely on reading-specific processes; it requires only that RTs can meaningfully distinguish between rapid, effortful, and idle responses. As such, the method is well-suited to international large-scale assessments, such as PISA and TIMSS, where disengagement can vary systematically across countries due to motivational norms, familiarity with computer-based testing, or test language demands.

Lastly, in considering the utility of the modeling approach proposed in this study, an important question is whether it can be practically implemented in real-world assessments. Although many IRT-based models of varying complexity have been developed to model test-taking engagement along with item responses, Nagy and Ulitzsch (2022) noted that many of these may be difficult to implement in practice due to their limited availability in widely used statistical software. An advantage of the 2PL-NLM is that it is freely accessible through the mirt package (Chalmers, 2012) in R (R Core Team, 2023), thereby alleviating some of these challenges for practitioners.

Limitations and Future Research

This study has several limitations. One limitation is that the dataset used to create the six cases consisted entirely of short items, each intended to be completed within ten seconds. Given this short time limit, the RT distributions may not be representative of those from other types of assessments, such as those with a test-level time limit or those where individual items require a longer time to complete. The short time thresholds for identifying rapid guessing and slow responding, often just a few seconds or less, could also make disengaged behavior harder to detect. Hence, the results of this study should be replicated with other assessment data.

Future research could explore in more detail how differentiating between rapid guessing and idling behaviors impacts ability estimates in different contexts. For example, researchers could investigate differentiating responses in other assessments that require more cognitive load, such as mathematics assessments. As explained earlier, this approach would yield more diverse data in terms of the proportions of rapid guesses and idle responses for each item. It would also be useful to examine the effectiveness of the NLM for cases not considered in this study, such as removing either rapid responses or idle responses separately. Alternatively, different thresholds for categorizing responses based on RTs could be used. While we used four response categories in the 2PL-NLM, future research could use more categories to investigate respondents' RT-based response behaviors in more detail with different scenarios. Such studies could provide more comprehensive insights into the potential of NLMs to enhance ability measurement and inform educational decision-making.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: University of Alberta, Pro00154093.

Data Availability Statement

The dataset used in this study was obtained from a proprietary reading assessment. Therefore, it cannot be distributed by the researchers in any form.

Contribution of Authors

All authors: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, Validation, and Writing-original draft.

Orcid

Okan Bulut https://orcid.org/0000-0001-5853-1267
Joyce Xinle Liu https://orcid.org/0009-0008-2278-8966
Hatice Cigdem Bulut https://orcid.org/0000-0003-2585-3686

REFERENCES

Aboutaleb, Y.M., Ben-Akiva, M., & Jaillet, P. (2020). Learning structure in nested logit models. *ArXiv*. https://doi.org/10.48550/arXiv.2008.08048

Ben-Akiva, M.E. (1973). *Structure of passenger travel demand models* [Doctoral thesis, Massachusetts Institute of Technology]. DSpace@MIT. https://dspace.mit.edu/handle/1721.1/14790

Ben-Akiva, M.E., & Lerman, S.R. (1985). *Discrete choice analysis: Theory and application to travel demand*. Cambridge, MA: MIT Press.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bolt, D.M., Wollack, J.A., & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika*, *77*(2), 339–357. https://doi.org/10.1007/S11336-012-9257-5

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). https://doi.org/10.3390/a16020089

Butler, J., & Adams, R.J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement*, *8*(3), 279–304.

Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Csányi, R., & Molnár, G. (2024). Item- and person-level factors in test-taking disengagement: Multilevel modeling in a low-stakes context. *International Journal of Educational Research Open*, *7*. https://doi.org/10.1016/j.ijedro.2024.100373

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00102

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

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

Deribo, T., Kroehne, U., & Goldhammer, F. (2021). Model-based treatment of rapid guessing. *Journal of Educational Measurement*, *58*(2), 281–303. https://doi.org/10.1111/jedm.12290

Deribo, T., Goldhammer, F., & Kroehne, U. (2023). Changes in the speed-ability relation through different treatments of rapid guessing. *Educational and Psychological Measurement*, *83*(3), 473–494. https://doi.org/10.1177/00131644221109490

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, *66*(4), 643-656. https://doi.org/10.1177/0013164405278574

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345-356. https://doi.org/10.1080/0969594X.2010.516569

Erdem Kara, B. (2025). Exploring Test-Taking Disengagement in the Context of PISA 2022: Evidence from Process Data. *International Electronic Journal of Elementary Education*, *17*(2), 305-315. https://doi.org/10.26822/iejee.2025.380

Filiasov, S., & Sweetman, A. (2023). Low-stakes standardized tests in British Columbia, Canada: System accountability and/or individual feedback? *Education Economics*, *31*(2), 145–165. https://doi.org/10.1080/09645292.2022.2091113

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, *2015*(2), 1–17.

Gorgun, G., & Bulut, O. (2021). A polytomous scoring approach to handle not-reached items in low-stakes assessments. *Educational and Psychological Measurement*, *81*(5), 847-871. https://doi.org/10.1177/0013164421991211

Gorgun, G., & Bulut, O. (2023). Incorporating test-taking engagement into the item selection algorithm in low-stakes computerized adaptive tests. *Large-scale Assessments in Education*, *11*(27). https://doi.org/10.1186/s40536-023-00177-5

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

Hauenstein, C.E., Embretson, S.E., & Kim, E. (2024). Psychometric modeling to identify examinees' strategy differences during testing. *Journal of Intelligence*, *12*(4). https://doi.org/10.3390/jintelligence12040040

Hu, L.-T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Koo, T.K., & Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kuhfeld, M., & Soland, J. (2020). Using assessment metadata to quantify the impact of test disengagement on estimates of educational effectiveness. *Journal of Research on Educational Effectiveness*, *13*(1), 147–175. https://doi.org/10.1080/19345747.2019.1636437

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*(8). 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

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*. https://doi.org/10.3389/fpsyg.2019.01533

Linzer, D.A., & Lewis, J.B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*(10), 1–29. https://doi.org/10.18637/jss.v042.i10

Liu, J.X., Bulut, O., & Johnson, M.D. (2024). Examining position effects on students' ability and test-taking speed in the TIMSS 2019 problem-solving and inquiry tasks: A structural equation modeling approach. *Psychology International*, *6*(2), 492–508. https://doi.org/10.3390/psycholint6020030

Lundgren, E., & Eklöf, H. (2023). Questionnaire-taking motivation: Using response times to assess motivation to optimize on the PISA 2018 student questionnaire. *International Journal of Testing*, *23*(4), 231–256. https://doi.org/10.1080/15305058.2023.2214647

McFadden, D. (1981). Econometric models of probabilistic choice. In C. F. Manski, & D. McFadden (Eds.), *Structural analysis of discrete data with econometric application* (pp. 198–272). Cambridge: MIT Press

Michaelides, M.P., & Ivanova, M. (2022). Response time as an indicator of test-taking effort in PISA: country and item-type differences. *Psychological Test and Assessment Modeling*, *64*(3), 304–338.

Michaelides, M.P., Ivanova, M.G., & Avraam, D. (2024). The impact of filtering out rapid-guessing examinees on PISA 2015 country rankings. *Psychological Test and Assessment Modeling*, *66*, 50–62. https://doi.org/10.2440/001-0012

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

Pastor, D.A., Ong, T.Q., & Strickman, S.N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment*, *24*(3), 189-212. https://doi.org/10.1080/10627197.2019.1615373

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, 29*(1)*, 55–79. https://doi.org/10.1007/s11092-016-9248-7

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-scale Assessments in Education*, *9*. https://doi.org/10.1186/s40536-021-00104-6

R Core Team. (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rios, J.A., & Deng, J. (2022). Quantifying the distorting effect of rapid guessing on estimates of coefficient alpha. *Applied Psychological Measurement*, *46*(1), 40-52. https://doi.org/10.1177/01466216211051719

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., Guo, H., Mao, L., & Liu, O.L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*, 74–104. https://doi.org/10.1080/15305058.2016.1231193

Rios, J.A., Liu, O.L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *2014*(161), 69–82. https://doi.org/10.1002/ir.20068

Samuels, S.J., & Flor, R.F. (1997). The importance of automaticity for developing expertise in reading. *Reading & Writing Quarterly*, *13*(2), 107–121. https://doi.org/10.1080/1057356970130202

Schnipke, D.L. (1995, April). *Assessing speededness in computer-based tests using item response times.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://psycnet.apa.org/doi/10.1037/0033-2909.86.2.420

Sideridis, G., & Alahmadi, M.T.S. (2022). The role of response times on the measurement of mental ability. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.892317

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*(8). https://doi.org/10.1186/s40536-021-00100-w

Stickney, E.M., Sharp, L.B., & Kenyon, A.S. (2012). Technology-enhanced assessment of math fact automaticity: Patterns of performance for low- and typically achieving students. *Assessment for Effective Intervention*, *37*(2), 84–94. https://doi.org/10.1177/1534508411430321

Suh, Y., & Bolt, D.M. (2010). Nested logit models for multiple choice item response data. *Psychometrika*, *75*(3), 454–473. https://doi.org/10.1007/s11336-010-9163-7

Swerdzewski, P.J., Harmes, J.C., & Finney, S.J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Tijmstra, J., & Bolsinova, M. (2018). On the importance of the speed-ability trade-off when dealing with not reached items. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00964

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 83–112. https://doi.org/10.1111/bmsp.12188

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

Wang, B., Huggins-Manley, C., Kuang, H., & Xiong, J. (2024). Enhancing effort-moderated item response theory models by evaluating a two-step estimation method and multidimensional variations on the model. *Educational and Psychological Measurement*, 1-23. https://doi.org/10.1177/00131644241280727

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. (2015). Effort Analysis: Individual Score Validation of Achievement Test Data. *Applied Measurement in Education*, *28*(3), 237–252. https://doi.org/10.1080/08957347.2015.1042155

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

Wise, S.L., & DeMars, C.E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1-17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S.L., & DeMars, C.E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S.L., & DeMars, C.E. (2009). A clarification of the effects of rapid guessing on coefficient α: A note on Attali's "Reliability of speeded number-right multiple-choice tests." *Applied Psychological Measurement*, *33*(6), 488–490. https://doi.org/10.1177/0146621607304655

Wise, S.L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343-354. https://doi.org/10.1080/08957347.2017.1353992

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, Vancouver, Canada.

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. (2020a). A cessation of measurement: Identifying test taker disengagement using response time. In M.J. Margolis, & R.A. Feinberg (Eds.), *Integrating Timing Considerations to Improve Testing Practices* (pp. 150–164). Routledge.

Wise, S.L., & Kuhfeld, M.R. (2020b). 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 presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

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, V.L., Wise, S.L., & Bhola, D.S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*(1), 65–83.

Wright, D.B. (2016). Treating all rapid responses as errors (TARRE) improves estimates of ability (slightly). *Psychological Test and Assessment Modeling*, *58*(1), 15–30.

Yildirim-Erbasli, S.N., & Bulut, O. (2020). The impact of students' test-taking effort on growth estimates in low-stakes educational assessments. *Educational Research and Evaluation*, *26*(7–8), 368–386. https://doi.org/10.1080/13803611.2021.1977152

Yildirim-Erbasli, S.N., & Gorgun, G. (2025). Disentangling the relationship between ability and test-taking effort: To what extent the ability levels can be predicted from response behavior? *Technology, Knowledge, and Learning*, *30*(3), 1475–1497. https://doi.org/10.1007/s10758-024-09810-w

Yıldırım Hoş, H., & Uysal Saraç, M. (2025). How does incorporating the response times into mixture modelling influence the identification of latent classes for mathematics literacy framework in PISA 2022? *Education and Science*, 50(Supplement 1), 129–146. https://doi.org/10.15390/EB.2025.14125

Zheng, X., Sahin, F., Erberber, E., & Fonseca, F. (2023). Identification and cross-country comparison of students' test-taking behaviors in selected eTIMSS 2019 countries. *Large-scale Assessments in Education*, *11*. https://doi.org/10.1186/s40536-023-00179-3