Gorgun and Bulut
Large-scale Assessments in Education (2023) 11:27
https://doi.org/10.1186/s40536-023-00177-5

Large-scale Assessments in Education

METHODOLOGY Open Access

Incorporating test-taking engagement into the item selection algorithm in low-stakes computerized adaptive tests

Guher Gorgun[1*] and Okan Bulut[2]

*Correspondence:
gorgun@ualberta.ca

[1] Measurement, Evaluation, and Data Science, Faculty of Education, University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5, Canada
[2] Centre for Research in Applied Measurement and Evaluation, Faculty of Education, University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5, Canada

Abstract

In low-stakes assessment settings, students' performance is not only influenced by students' ability level but also their test-taking engagement. In computerized adaptive tests (CATs), disengaged responses (e.g., rapid guesses) that fail to reflect students' true ability levels may lead to the selection of less informative items and thereby contaminate item selection and ability estimation procedures. To date, researchers have developed various approaches to detect and remove disengaged responses after test administration is completed to alleviate the negative impact of low test-taking engagement on test scores. This study proposes an alternative item selection method based on Maximum Fisher Information (MFI) that considers test-taking engagement as a secondary latent trait to select the most optimal items based on both ability and engagement. The results of post-hoc simulation studies indicated that the proposed method could optimize item selection and improve the accuracy of final ability estimates, especially for low-ability students. Overall, the proposed method showed great promise for tailoring CATs based on test-taking engagement. Practitioners are encouraged to consider incorporating engagement into the item selection algorithm to enhance the validity of inferences made from low-stakes CATs.

Keywords: CAT, Low-stakes, Test-taking engagement, Response time, Maximum fisher information

The tacit assumption when administrating an assessment is that examinees will invest maximal effort while attempting the items (AERA et al., 2014; Rios & Soland, 2021). The effort here refers to the individuals' investment of mental exertion to achieve a task (Inzlicht et al., 2018). For example, an examinee who knows how to decode a particular word may fail to do so during a test due to insufficient effort or disengagement (Frederick, 2005). The effort exerted by examinees during a test is usually defined as *test-taking engagement*. Therefore, the response behavior characterized as effortful is referred to as solution or engaged response behavior (e.g., Pastor et al., 2019; Schnipke & Scrams, 2002), whereas the response behavior characterized as non-effortful is known as disengaged response behavior (e.g., Soland & Kuhfeld, 2019). Disengaged responses typically do not reflect students' true ability levels (Swerdzewski et al., 2011; Wise, 2006; Wise &

Kong, 2005), and thus they may contaminate statistical estimates, psychometric properties, classifications, individual scores, test inferences, and conclusions if left untreated (Lindner et al., 2019; Rios & Soland, 2021; Wise et al., 2009).

The absence of personal consequences and intrinsic value for examinees distinguishes an assessment setting as low-stakes testing. Although teachers, schools, or district authorities may use these assessments to adjust teaching practices, identify at-risk students, or compare the relative performance of teachers and schools, students generally do not observe this process (Finn, 2015). Limited personal consequences for students (Penk & Schipolowski, 2015) may violate the assumption that the assessment reflects students' ability levels, and the lack of effort thereof may lead to the underestimation of students' ability levels (Rios et al., 2017). Therefore, researchers and practitioners face a challenge regarding whether the observed low performance is due to low engagement or low ability when statewide (e.g., National Assessment on Educational Progress) (Braun et al., 2011; Swerdzewski et al., 2011) or large-scale international assessments (e.g., Programme for International Student Assessment (PISA), International Reading Literacy Study, and the Trends in International Mathematics and Science Study) (Eklöf, 2007; Liu et al., 2014) are used for accountability purposes, teacher or program evaluations, funding decisions, or country performance comparisons (Eklöf, 2006, 2010; Finn, 2015; Thelk et al., 2009).

Researchers have proposed several *reactive* remedies to mitigate the negative influence of test-taking disengagement. Motivation filtering approaches based on response time thresholds (e.g., Wang & Xu, 2015; Wise & DeMars, 2005; Wise & Kong, 2005) or self-report measures which are used to rate student motivation or effort while taking the assessment (e.g., Eklöf, 2006; Sundre, 1999; Sundre & Moore, 2002; Sundre & Wise, 2003) suggest removing disengaged responses or examinees. Wise and colleagues also proposed effort monitoring with proctor notification to improve test-taking engagement (Wise et al., 2019). Accordingly, the test proctor is notified after an examinee demonstrates a disengaged response, and the proctor urges the examinee to show effortful response behavior. Nevertheless, these reactive remedies deal with test-taking disengagement after it occurs and may be limited in sustaining the optimal testing conditions. For example, if an examinee displays disengaged response behavior for most of the items, it may not be possible to estimate the examinee's ability levels. Therefore, a *proactive* approach is necessary to minimize the occurrence of disengaged responses during an assessment administration.

Understanding the onset of disengaged response behavior may help researchers and practitioners inhibit disengaged responses in a low-stakes assessment. Thus, researchers have long been interested in explaining the examinee or test characteristics that may trigger disengagement in low-stakes settings. While some argued that examinee disengagement in non-adaptive fixed tests is due to the potential mismatch between the difficulty of the items on the test and the examinee's ability level (Tonidandel et al., 2002), others argued that it is rather examinees' intrinsic motivation that determines whether an examinee invests effort or not (Wise & Smith, 2011). Expectancy-Value Theory (Wigfield & Eccles, 2000) may account for these two different explanations of examinee disengagement in low-stakes settings. That is, some examinees may not expend effort if they perceive a task as unattainable (e.g., the item is perceived as too difficult; *expectancy*).

Likewise, some may not attempt the task due to the lack of personal consequences (e.g., the item is considered to have no direct impact on the examinee; *value*).

On a par with the expectancy aspect of the Expectancy-Value Theory, several researchers claimed increased levels of student engagement in adaptive tests because items are optimally selected according to the examinee's interim ability level (e.g., Linacre, 2000; Mead & Drasgow, 1993; Weiss & Betz, 1973). Hence, compared with non-adaptive tests, computerized adaptive tests (CATs) may provide examinees with a better testing experience by selecting and administering the items that match their ability levels more closely. Yet, studies found mixed empirical evidence regarding the positive association between CATs and test-taking engagement. For example, Martin and Lazendic (2018) and Ross and colleagues (2018) found a positive association between adaptive tests and test-taking engagement. On the other hand, others found no significant positive impact of adaptive tests on test-taking engagement (e.g., Bergstrom et al., 1992; Häusler & Sommer, 2008; Ling et al., 2017; Lunz & Bergstrom, 1994), indicating that using an adaptive test may not be sufficient to ensure a high level of test-taking engagement and like fixed-item tests, CATs are also susceptible to examinee disengagement.

The current study proposes a *proactive* remedy to tackle test-taking engagement during an assessment administration in low-stakes adaptive tests. Specifically, we aim to extend the use of response time distributions to the item selection procedure in CATs (Wise, 2014, 2020; Wise & Kingsbury, 2016). First, we design a post-hoc simulation study based on real data to evaluate whether incorporating the test-taking engagement into the item selection process could improve the performance of a CAT system. Our analyses show the promise of integrating engagement into item selection to enhance the CAT's accuracy. Next, we propose a novel item selection algorithm where optimal items are selected adaptively based not only on item information at a given ability level but also on test-taking engagement. After describing how the proposed item selection algorithm works, we evaluated its performance using a simulation study based on real data from a timed computer-based (non-adaptive) large-scale assessment.

Literature review

Disengagement in low-stakes assessments

Compared with high-stakes assessments, low-stakes assessments are more prone to test-taking disengagement (Finn, 2015; Wise, 2006; Wise & DeMars, 2005). In high-stakes settings, examinees may get disengaged towards the end of the test due to running out of time or test fatigue (Schinke & Scrams, 1997, 2002). However, in low-stakes settings, disengaged responses may occur anywhere during test administration because examinees may switch back and forth between disengaged and engaged response behaviors during the test (Lindner et al., 2019; Wise & Kong, 2005). Some researchers argued that examinees begin the test with an engaged response behavior, but once they enter the disengaged response mode, they are likely to stayed disengaged until the end of the assessment (Bolt et al., 2002; Jin & Wang, 2014). In contrast, Cao and Stokes (2008) argued that examinees are more likely to demonstrate disengaged response behaviors (e.g., rapid guessing) after they encounter difficult items on the test.

Previous research primarily associated disengaged or non-effortful responding with rapid guessing (e.g., Rose et al., 2010; Ulitzsch et al., 2020; Wise, 2019). As examinees

perceive that they need to spend a long time to answer the item correctly, the time cost becomes higher, and the expectation of answering the item correctly does not compensate for the cost of effort (i.e., speed-accuracy trade-off), especially in timed tests (De Boeck & Joen, 2019). Although longer response times are often associated with higher accuracy, response accuracy is likely to decrease after a certain length of response time (i.e., dual-preprocessing theory; De Boeck & Jeon, 2019). As an examinee foresees that an item requires a longer time to find the correct answer, the effort becomes more costly, and the examinee may eventually cease to invest enough effort to answer the item. These types of response behaviors are referred to as *slow errors* and can be due to attentional lapses, lack of automation, and uncertainty (Novikov et al., 2017). Under ideal testing conditions (e.g., a sufficient duration is granted to examinees to complete the items), examinees would be expected to show effortful response behavior throughout the test by adjusting their speed to complete the test in time and minimizing the occurrence of not-reached items (Gorgun & Bulut, 2021; Tijmstra & Bolsinova, 2018).

Disengaged test-taking behavior can be instantiated as either *rapid guessing* (i.e., the use of an unrealistically short amount of response time) or *idle responding* (i.e., the use of an unrealistically long amount of response time when attempting the items on the test) (Gorgun & Bulut, 2021; Wise, 2017; Yildirim-Erbasli & Bulut, 2021). Rapid guessers may attempt to randomly select a response option without reading and understanding the item content, whereas idle responders may not work on the item (i.e., invest effort; Lindner et al., 2019) such as through engaging in daydreaming leading to allotted time to expire.. Although the presence of rapid guessing is problematic for all types of assessments, idle responding becomes a major concern in automaticity-based assessments where speed and ability are operationalized together, such as reading assessments that require students to decode words, comprehend the information, and make inferences at the same time (e.g., Samuels & Flor, 1997) or math assessments that require students to recall math facts rapidly and accurately (e.g., Stickney et al., 2012). Further, it could be argued that idle responding could be a concern for timed assessments where examinees may have to deal with a speed-ability trade-off by adjusting their speed to complete all the items within the allotted time and avoid having not-reached items as much as possible (Tijmstra & Bolsinova, 2018; Ulitzsch et al., 2020).

Response times recorded for each item during computer-based tests, including CATs, have been widely used as a proxy (see Rios & Deng, 2021) to classify examinees' test-taking behavior as effortful (i.e., engaged) or non-effortful (i.e., disengaged). Several threshold-based and model-based methods have been proposed for identifying and handling disengaged responses in low-stakes assessments. For example, Wise and Ma (2012) proposed a Normative Threshold (NT) method that uses a particular percentage (i.e., 10%, 15%, or 20%) of the average response time for each item as a threshold to identify students with rapid guessing behavior. There are also model-based approaches that aim to quantify test-taking engagement based on ability and speed (e.g., Guo et al., 2020; Pohl et al., 2014; Pokropek, 2016; Ulitzsch, 2020, 2021; Wang & Xu, 2015). These methods are typically used for flagging and removing disengaged responses from the data after test administration is completed. With dynamic assessment tools such as CATs, it might be possible to identify, and overturn disengaged response behaviors in real-time for low-stakes assessments.

Computerized adaptive tests

CATs have numerous advantages over paper-and-pencil and computer-based tests with fixed items, such as optimal item selection based on each examinee's interim ability level (Eggen, 2012; Lord, 1980; Veldkamp, 2003), substantial reductions in test length and test duration (e.g., Bulut & Kan, 2012; Choe et al., 2018, Weiss, 1982; Weiss & Kingsbury, 1984), increased measurement precision (Davey, 2011; Weiss, 2004), and the availability of response time information for a large pool of items (van der Linden, 2008; Wise & Kingsbury, 2016). In CATs, optimal items for each examinee can be adaptively selected using a variety of methods, such as maximum Fisher information (MFI; Birnbaum, 1968), Kullback–Leibler information (Chang & Ying, 1996), and maximum likelihood weighted information (Veerkamp & Berger, 1997). Among these methods, MFI is the most common method for selecting items adaptively in operational CATs (Thompson & Weiss, 2011). With the MFI method, the item with the highest Fisher information at the interim ability level is iteratively selected until a test termination criterion is satisfied (Lord, 1980; Weiss, 1982). The MFI method implicitly assumes that each examinee puts their maximum effort into answering the items until the end of the test. Hence, as more items are administered, the interim ability estimate is expected to become more precise (i.e., closer to the true ability level) and thereby better guide the item selection procedure.

Although the maximal effort assumption can be reasonable for high-stakes CATs (e.g., licensure or certification exams), it may not necessarily hold for low-stakes CATs (e.g., universal screening and progress monitoring measures in K-12). The presence of significant test-taking disengagement would be detrimental to the item selection process in CATs. For example, if an examinee responds to the item through rapid guessing, their probability of answering the item correctly would be lower than the expected probability based on the true ability level. Therefore, the CAT is likely to underestimate the examinee's interim ability level and select an easier item from the item bank (Betz & Weiss, 1976; Wise, 2020; Wise & DeMars, 2005; Wise et al., 2014). Therefore, the item selection algorithms need to be modified to minimize the negative impact of test-taking disengagement on item selection in CATs.

Integration of response times into item selection

Integrating response times into the item selection algorithm can be promising for optimizing the item selection procedure during a CAT administration. Several researchers proposed new algorithms for item selection by inversely weighting the MFI algorithm with the expected response time (e.g., Choe et al., 2018; Fan et al., 2012). With these algorithms, the optimal item selection is based on the maximum information per unit of expected response time. That is, the measurement efficiency depends on the number of items and the total time required to complete the CAT. Similar approaches involving response times in adaptive tests have been proposed to identify aberrant response behaviors such as cheating during a CAT (e.g., van der Linden & Guo, 2008). These approaches often rely on the hierarchical modeling framework introduced by van der Linden (2007) to estimate the examinee's expected speed and ability jointly; however, they do not necessarily consider the response time as an indicator of test-taking engagement.

Wise and Kingsbury (2016) and Wise (2020) proposed an effort-guided CAT approach based on two modifications. In the first modification, the CAT would ignore the examinee's disengaged responses when calculating the final ability estimate after the test is terminated. In the second modification, if the examinee's response time for a given item is below a normative threshold, the response is flagged for rapid guessing and thus not used for updating the interim ability estimate. Instead, the MFI algorithm selects the next item based on the previous estimate of the interim ability. The first modification can be a remedy for situations where the examinees had only a few disengaged responses that did not occur in succession during the CAT administration. Otherwise, depending on the maximum number of items to be administered, the item selection algorithm may not accommodate the negative impact of disengaged responses on the estimation of interim ability. The second modification can be an effective solution for situations where the examinee re-engages with the test after giving a few disengaged responses. Both modifications follow a reactive approach by responding to disengagement after it occurs on the test. However, a proactive approach that eliminates the disengagement problem before it occurs would be a more desirable solution. To accomplish this goal, the item selection algorithm should consider not only the examinee's interim ability but also their engagement level when selecting the items from the item bank.

Current study

A proactive approach considering test-taking engagement during the test can be a more promising solution for dealing with disengaged responses in low-stakes CATs. However, an important question remains to be addressed before modifying the item selection algorithm: How can we predict whether an examinee would engage with an item before administering it? In other words, is it possible to select the optimal items that would be not only informative but also engaging for each examinee? To address these questions, we propose to conceptualize test-taking engagement as a secondary latent trait. To date, several researchers conceptualized test-taking engagement as a latent trait and found systematic differences among examinees in terms of their engagement levels across the items (e.g., Goldhammer et al., 2016; Setzer et al., 2013). For example, Goldhammer et al., (2017) recoded item responses in the Programme for the International Assessment of Adult Competencies based on the examinees' test-taking engagement levels (i.e., 0 = engaged, 1 = disengaged) and fitted a one-parameter item response theory (IRT) model to the recoded data. The authors found empirical evidence supporting the operationalization of test-taking engagement as a latent examinee characteristic.

In this study, we follow Goldhammer et al.'s (2017) approach to operationalize test-taking engagement as *latent engagement* (LE)—a latent examinee characteristic indicating the degree of test-taking engagement based on examinees' response times. Our ultimate goal is to incorporate LE into the CAT and optimize the item selection process based on both ability and test-taking engagement. Our study consists of two post-hoc simulation studies based on real data from a low-stakes computerized formative assessment focusing on students' automaticity skills in reading (i.e., recognizing, decoding, and reading words rapidly, effortlessly, and accurately). In Study 1, we use the students' existing response time information to assign ranks to the items based on test-taking engagement (i.e., 1=rapid guessers, 2=idle responders, and 3=optimal time users). Then, when

implementing the CAT, we iteratively select the items with the maximum Fisher information at the interim ability level and the highest ranking for test-taking engagement. This item selection method is referred to as *engagement-ranked MFI* hereafter. The results of Study 1 served as a benchmark model to understand the impact of incorporating test-taking engagement into the adaptive item selection.

Study 1 relies on an unusual assumption that each student's response time is known before administering the items, which would not be realistic for a real-time CAT. Therefore, in Study 2, we provide a solution for the absence of response time information in a real-time CAT administration by operationalizing test-taking engagement as a latent examine characteristic (i.e., LE). Then, we modify the item selection algorithm by selecting the most informative items based on the interim estimates of ability and LE. Specifically, we compute two sets of Fisher information values for each examinee: an ability-based information function using item responses and an LE-based information function using response times. Next, we sort the items in descending order based on Fisher information values and select the item with the maximum information in terms of both ability and LE. This item selection method is referred to as *MFI-LE* hereafter. Given the relationship between ability and LE, we proposed two variants of MFI-LE. The first variant (*unconditional MFI-LE*) selects the items iteratively based on interim estimates of both ability and LE. The second variant (*conditional MFI-LE*) considers a threshold for LE to determine whether item selection should involve both ability and LE, or only ability. Interim estimates of ability and LE are computed after each item and if the examinee's LE level is below an engagement threshold, the next item (i.e., the most informative item) is selected based on interim estimates of both ability and LE; otherwise, the conventional MFI approach is used to select the most informative item based on interim ability. That is, unlike *unconditional MFI-LE*, *conditional MFI-LE* incorporates LE into item selection only when the examinee's test-taking engagement level is lower than a predefined LE threshold.

To evaluate the feasibility of the proposed item selection methods introduced in Study 1 and Study 2, we compared their performances with those of the conventional MFI and the effort-guided CAT (Wise, 2020). The conventional MFI selected the most informative items based on interim ability, while the effort-guided CAT removed disengaged responses prior to estimating the final ability value. The research questions underlying the two studies summarized above are as follows:

A) Does selecting the items based on the engagement-ranked information function yield more accurate ability estimates than the MFI and effort-guided CAT method (Study 1)?
B) Does the item selection algorithm utilizing ability and LE together yield more accurate ability estimates than the MFI and effort-guided CAT method (Study 2)?

Method

Participants and study context

The sample of this study included 21,811 students (grades 5 to 12) who participated in a timed computer-based reading assessment in the United States. The teachers used

the assessment to measure students' reading automaticity and monitor their progress throughout the school year (i.e., no direct consequences for students). The assessment consisted of 120 multiple-choice items with four response options focusing on various reading skills (e.g., recognizing words, understanding how words are formed, and vocabulary). Students had approximately 10 seconds to respond to each item. Both students' responses (i.e., 1 = correct and 0 = incorrect) and response times (in seconds) for each item were recorded during the test administration.

In this study, we used the 2-parameter logistic (2PL) IRT model (Birnbaum, 1968) to calibrate the item parameters as item bank and estimate students' ability levels based on the entire test (i.e., 120 items). The 2PL model can be written as:

$$P_j(\theta) = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}} \tag{1}$$

where $P_j(\theta)$ is the probability of answering the item j correctly, a_j is the item discrimination parameter for item j, and b_j is the item difficulty parameter for item j, and θ is the ability level of the student answering the item. Before calibrating the items and estimating the ability levels of students, we identified disengaged responses based on the NT10 method (Wise & Ma, 2012) and recoded them as missing to minimize the negative impact of test-taking disengagement on item calibration and ability estimation. The disengaged response rate ranged from 5 to 15% across the items.

Design and analysis

We designed two post-hoc simulation studies and modified the item selection algorithm to select both informative and engaging items for the students. After administering the first item based on $\theta = 0$, the CAT iteratively selected the most informative and engaging items for students (see the following sections for more details on the item selection algorithms used for each study). The CAT was terminated when the maximum test length (20, 30, or 50 items) was reached or the conditional standard error of measurement (cSEM) of the ability estimate was less than 0.25.[1] Interim and final ability values were estimated using the expected a posteriori (EAP; Bock & Mislevy, 1982) method. The post-hoc simulations were conducted using the mirtCAT package (Chalmers, 2016) in R (R Core Team, 2021).

Simulation study 1: engagement-ranked MFI

In Study 1, we adopted the NT10 method (Wise & Ma, 2012) to identify engaged and disengaged students. Accordingly, we found the median response time for each item and computed 10% and 190% of the median response time to determine the response time thresholds for rapid guessers, idle responders, and effortful responders. Then, we identified students' level of engagement on each item based on the students' response times and the response time thresholds. Using the response time-based polytomous scoring approach introduced by Gorgun and Bulut (2021), responses classified as *rapid guessing* (i.e., response time < 10% of the median response time representing an unrealistically short response

[1] Since the primary goal of the CAT is to measure ability instead of test-taking engagement, cSEM for test-taking engagement was not considered when terminating the CAT administration.

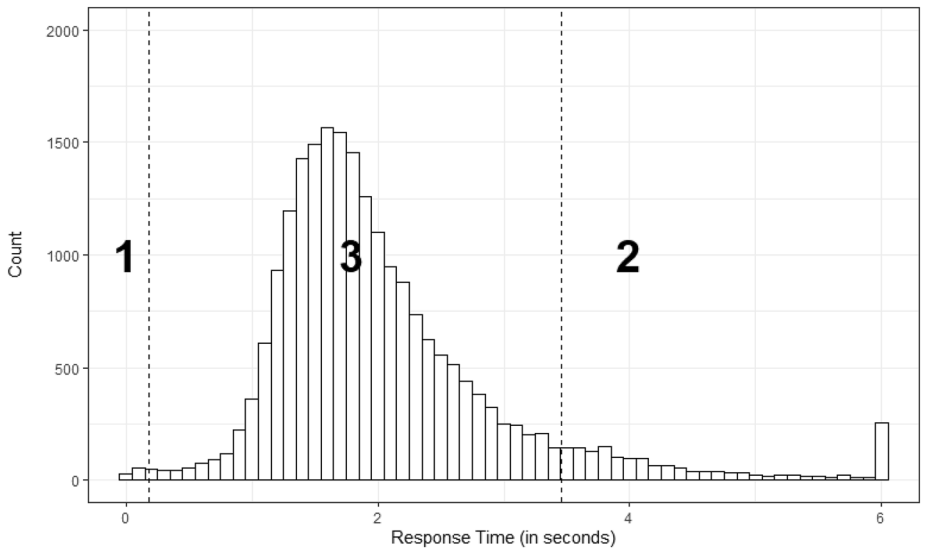

Fig. 1 An illustration of How Engagement Level is Determined based on the Response Time Distribution for Each Item in Simulation Study 1

time) received the rank of 1, and responses classified as *idle* (i.e., response time > 190% of the median response time representing an unrealistically long response time) received the rank of 2, and responses classified as *effortful* (i.e., the middle region in response time distribution representing optimal time use) were assigned a rank of 3 to differentiate the amount of effort that examinees put into answering each item.

Figure 1 demonstrates an example item with the scoring scheme for rapid guessers (left-hand side of the plot), idle responders (right-hand side of the plot), and effortful respondents (middle region of the plot). We used this scoring scheme to rank the information function when selecting items that students are likely to put enough effort. Determining the response time thresholds for each item separately enabled us to take the item difficulty into account as more difficult items tend to have higher median response times. Furthermore, using the NT10 method, rather than NT15 or NT20, enabled us to avoid false positives when classifying students' levels of test-taking engagement (Kong et al., 2007; Wise & Ma, 2012).

We created the engagement-ranked MFI algorithm that considers both item information and rankings of items based on test-taking engagement when selecting the items during the CAT administration. The conventional Fisher information at a given ability level in the 2PL model can be computed as:

$$I_j(\theta_i) = a_j^2 P_j(\theta_i)\left(1 - P_j(\theta_i)\right) \tag{2}$$

where $P_j(\theta_i)$ is the probability of answering item j correctly given the student i's ability θ_i, a_j is the item discrimination parameter for item j, and $I_j(\theta_i)$ is the item information level for student i on item j. When implementing the engagement-ranked MFI algorithm, we compute the Fisher information for each item using the interim ability estimate, rank the items in descending order based on the Fisher information and their engagement levels w_j (i.e., $w_j = 1$ for rapid guessing, $w_j = 2$ for idle responding, or $w_j = 3$ for effortful

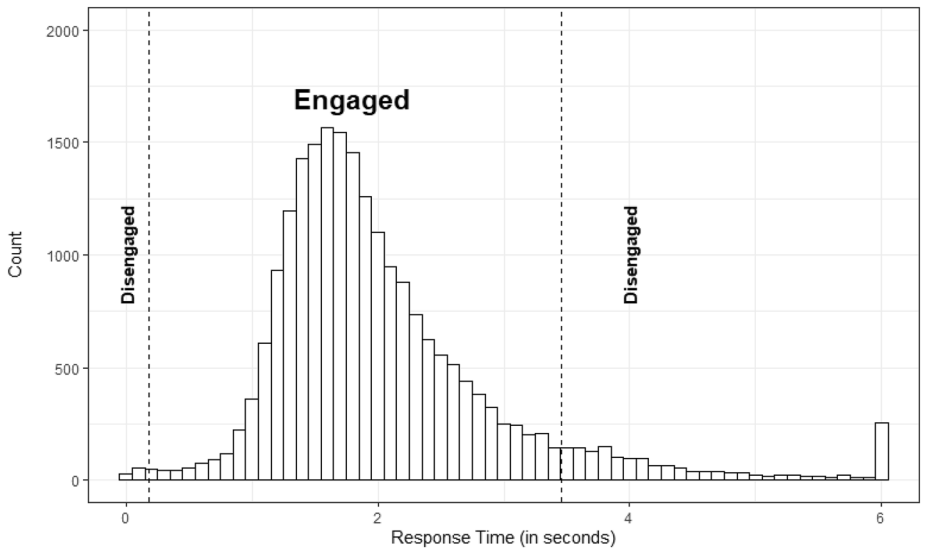

Fig. 2 An Illustration of How Test-Taking Engagement is Determined based on the Response Time Distribution for Each Item in Simulation Study 2

responding in item *j*), and select the item with the maximum Fisher information and the highest ranking for test-taking engagement. That is, the items yielding the highest information while being ranked as the most engaging are prioritized in item selection. Although engagement-ranked MFI cannot be employed in operational testing settings (unless it is a repeated test), the results from engagement-ranked MFI had two important implications. First, the results allowed us to evaluate the utility of incorporating engagement into the item selection in CATs. Second, they served as a benchmark for the second study to assess the feasibility of incorporating LE into item selection. Study 1 represents an ideal but impractical scenario where the examinees' known response times could be utilized to rank the items based on test-taking engagement and select the most informative item iteratively. In Study 2, we use a more practical scenario to evaluate the MFI-LE methods and compare their performance with the performance of engagement-ranked MFI.

Simulation study 2: unconditional and conditional MFI-LE

In Study 2, we conceptualized test-taking engagement as a latent examinee characteristic and estimated the examinees' LE levels (e.g., Goldhammer et al., 2017). We used the NT10 method to identify disengaged and engaged responses for each student. That is, we found the median response time for each item and computed 10% and 190% of the median response time to identify the thresholds to label students' response time data based on their level of test-taking engagement. For this procedure, we considered rapid guesses and idle responses as disengaged responses and effortful responses as engaged responses. Then, we assigned 0 to disengaged responses and 1 to engaged responses (see Fig. 2), yielding a secondary dichotomous dataset based on test-taking engagement. Hence, we created two datasets: one based on item responses (0=Incorrect, 1=Correct) and the second one based on engagement (0=disengaged, 1=engaged) operationalized

through response times. Similar to the calibration of item parameters based on the item responses, we fitted the 2PL IRT model to the engagement dataset and estimated item engagement parameters.[2] The resulting item engagement parameters can be defined as difficulty and discrimination for LE.

To select the most informative items during the CAT administration, we considered two sets of item parameters (one for ability and another for LE). Specifically, two sets of Fisher information values were calculated for each item; one for students' ability level based on the difficulty and discrimination parameters and another for students' LE level based on the engagement parameters. The item information functions for ability and test-taking engagement can be written as:

$$I_{j,\theta}(\theta_i) = a_{j,\theta}^2 P_{j,\theta}(\theta_i)\left(1 - P_{j,\theta}(\theta_i)\right) \tag{3}$$

$$I_{j,e}(\theta_{e.i}) = a_{j,e}^2 P_{j,e}(\theta_{e.i})\left(1 - P_{j,e}(\theta_{e.i})\right) \tag{4}$$

where $I_{j,\theta}(\theta_i)$ is the Fisher information function for person i and item j based on ability (i.e., θ_i) and $I_{j,e}(\theta_{e.i})$ is the Fisher information function for person i and item j based on LE (i.e., $\theta_{e.i}$). During item selection, we ranked the item information values obtained from ability and LE together and selected the item with maximum information based on both ability and LE. The highest rank is determined by multiplying the information functions of both ability and engagement. Using this procedure (i.e., unconditional MFI-LE), the CAT selected the most informative and engaging items from the item bank for all examinees, regardless of their ability and LE levels.

Unlike unconditional MFI-LE, conditional MFI-LE incorporates LE into the item selection process if the examinee's interim estimate of LE is below a predefined threshold (i.e., showing a low level of LE). If, however, the examinee's interim estimate of LE is above the threshold (i.e., showing an acceptable level of LE), then the conventional MFI is applied by selecting the most informative items based on the examinee's interim estimate of ability. An optimal LE threshold can be determined in several ways, such as visual inspection of the test information function (TIF) for LE, norm-based threshold selection based on the population distribution of LE, or an empirical threshold based on the relationship between ability and LE. In this study, we inspected the TIF for LE visually (see Fig. 3) and identified two thresholds (i.e., $\theta = 0$ and $\theta = -1$), after which the items became less informative in terms of LE.

Evaluation criteria

We evaluated the relative performance of the proposed approaches against the conventional MFI and effort-guided CAT methods. The simulation studies were evaluated based on the same criteria: the average number of administered items, the correlation between estimated and true ability levels, average cSEM values, bias, and root-mean-squared error (RMSE). Bias and RMSE values were calculated as follows:

[2] We also used NT15, NT20, and NT25 to identify test-taking disengagement and calibrate the item parameters. The parameters did not change significantly across these different threshold-based methods. We used the NT10 method because it is a more conservative method minimizing the occurrence of false positives.

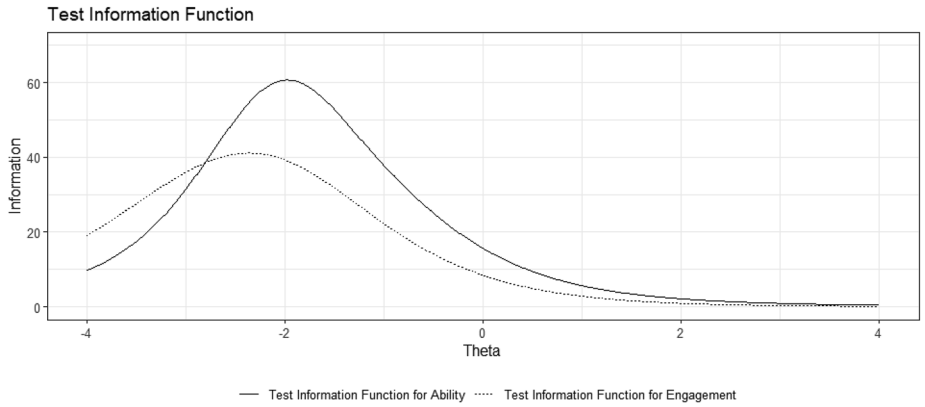

Fig. 3 Test Information Functions for Ability and Latent Engagement

$$Bias = \frac{\sum_{i=1}^{N}\left(\widehat{\theta}_i - \theta_i\right)}{N} \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\widehat{\theta}_i - \theta_i\right)^2}{N}} \tag{6}$$

where $\widehat{\theta}_i$ is student *i*'s final ability estimate from the CAT administration, θ_i is student *i*'s (true) ability level based on the entire test, and *N* represents the sample size. Smaller values of RMSE and smaller absolute values of bias showed more accurate ability estimates. Positive values of bias indicated overestimated ability levels, whereas negative values indicated underestimated ability levels.

Results

Results for simulation study 1

Table 1 summarizes the evaluation indices obtained for each simulation condition in Study 1. The results showed that the engagement-ranked MFI outperformed the conventional MFI and effort-guided CAT based on the correlation between estimated and true ability, bias, and RMSE under the 20-item test length condition. Also, the engagement-ranked MFI yielded the highest correlations between estimated and true ability and the lowest RMSE values across all test length conditions. However, effort-guided CAT in terms of the average number of items administered and the conventional MFI in terms of the mean cSEM performed better than the engagement-ranked MFI. The benefit of ranking Fisher information based on the students' engagement level was more salient when the test length was either 20 or 30. Figures 4, 5 demonstrate bias and RMSE values across the true ability values for the engagement-ranked MFI, conventional MFI, and effort-guided CAT approaches. The engagement-ranked MFI outperformed the conventional MFI and effort-guided CAT for all conditions. However, the difference between the three item selection methods became negligible as the estimated ability values increased, suggesting that considering test-taking engagement in item selection could be more critical for low-ability students.

Table 1 Results for Simulation Study 1 and 2

Test length	Evaluation criteria	Conventional MFI	Effort-guided CAT	Engagement-ranked MFI	Unconditional MFI-LE	Conditional MFI-LE (θ = 0)	Conditional MFI-LE (θ = − 1)
20	*r*	0.889	0.889	**0.911**	0.903	0.901	0.906
	Bias	− 0.011	− 0.015	**− 0.001**	0.004	0.026	0.036
	RMSE	0.450	0.451	**0.339**	0.426	0.433	0.422
	Mean cSEM	**0.388**	0.390	0.393	0.389	0.391	0.391
	n	19.71	19.71	19.74	19.72	**19.57**	**19.57**
30	*r*	0.931	0.930	**0.945**	0.933	0.932	0.932
	Bias	0.030	**0.025**	0.036	0.035	0.058	0.058
	RMSE	0.357	0.357	**0.316**	0.355	0.364	0.364
	Mean cSEM	**0.347**	0.349	0.353	0.349	0.350	0.349
	n	27.95	27.95	28.26	28.09	**27.84**	**27.84**
50	*r*	0.952	0.952	**0.958**	0.955	0.953	0.953
	Bias	− 0.079	**0.061**	− 0.069	0.073	0.091	0.092
	RMSE	0.303	0.302	**0.282**	0.296	0.313	0.313
	Mean cSEM	**0.316**	0.317	0.322	0.355	**0.316**	0.317
	n	40.80	40.80	42.06	40.95	**40.63**	**40.63**

Bold values indicate the best results for each evaluation criterion

MFI: Maximum Fisher Information, *Effort-Guided CAT*: Removing disengaged items when estimating ability levels after the CAT is completed, *Engagement-Ranked MFI*: MFI that considers item engagement ranking in item selection, *Unconditional MFI-LE*: MFI that considers both ability and latent engagement in item selection, *Conditional MFI-LE*: MFI that considers both ability and latent engagement in item selection only when latent engagement is below a certain threshold, *r* : Correlations between true and estimated theta values, *RMSE*: Root-mean-squared error, *cSEM*: Conditional standard error of measurement, *n*: The average number of items administered

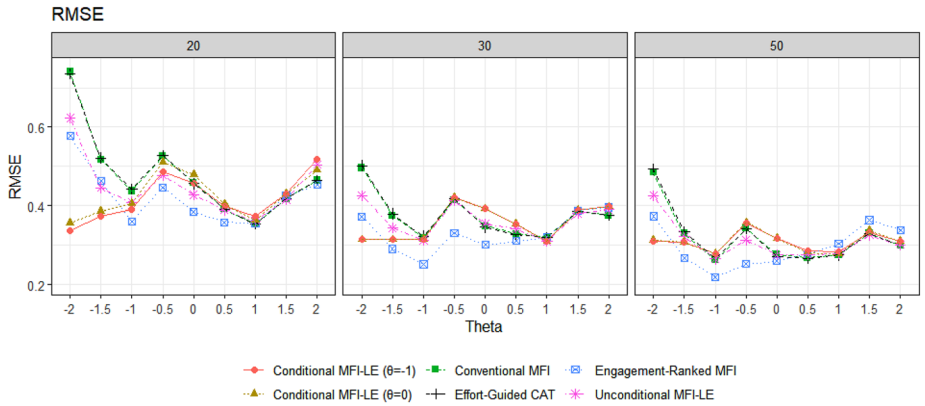

Fig. 4 Estimated RMSE across the Item Selection Methods

The bias results in Fig. 5 also suggest that the engagement-ranked MFI yielded slightly overestimated values for higher ability levels (i.e., $\theta > 0.5$) under the 50-item condition. This finding appears to be a result of selecting nearly half of the items from a small item bank. In a longer test, the students with high ability levels could not benefit from adaptive item selection that considered the engagement level. Another interesting finding is that the performances of the conventional MFI and effort-guided CAT were very similar in terms of bias and RMSE across all ability levels. This finding underscored the fact that the effort-guided CAT, as a reactive approach, may not be able to alleviate the negative impact of disengagement on item selection and ability estimation in low-stakes CATs. Overall, the findings of Study 1 suggest that taking test-taking engagement into account

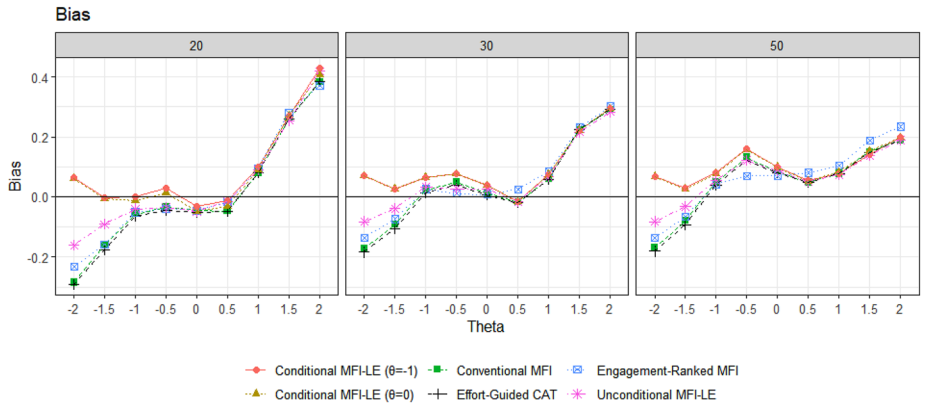

Fig. 5 Estimated Bias across the Item Selection Methods

could improve the item selection process and consequently the accuracy of estimated ability values.

Results for simulation study 2

In Study 2, we evaluated the performances of the unconditional and conditional MFI-LE, conventional MFI, and effort-guided CAT by comparing the correlations between true and estimated ability, bias, RMSE, and average cSEM values. The simulation results of Study 2 showed that on the aggregate level, both unconditional and conditional MFI-LE outperformed the conventional MFI and effort-guided CAT approaches under most conditions (see Table 1). The superior performance of the MFI-LE methods was more salient under shorter test length conditions (i.e., 20 or 30 items). Regardless of the engagement level threshold, the conditional MFI-LE methods yielded smaller values for the average number of items administered in the CAT, compared with the conventional MFI and effort-guided CAT.[3] Like in Study 1, the conventional MFI was the best performing method in terms of the mean cSEM value. This was an expected finding because the conventional MFI maximizes the TIF, which is inversely related to the cSEM value of the estimated ability.

The results in Table 1 also show that both effort-guided CAT and MFI-LE performed better than the conventional MFI, suggesting that considering test-taking engagement in item selection is likely to yield more accurate results in low-stakes CATs. However, between the two item selection methods, effort-guided CAT might be less desirable than MFI-LE due to decreased efficacy of effort-guided CAT for testing situations where the test length is short and disengaged response behaviors (e.g., rapid guessing and idle responding) are very prevalent. With effort-guided CAT as a reactive approach, excluding many disengaged responses from the ability estimation at the end of a CAT administration might lead to less accurate estimates of ability. Unlike effort-guided CAT, the MFI-LE as a proactive approach is likely to yield more accurate ability estimates since this method can regulate the item selection process during the CAT administration.

[3] Since effort-guided CAT uses the conventional MFI method to select the items, both methods yielded the same average number of items administered.

In addition to the aggregated results from Simulation Study 2 (see Table 1), Figs. 4 and 5 also demonstrate RMSE and bias values across different ability levels for all item selection methods utilized in this study, respectively. Especially for lower ability levels (i.e., $\theta < -1$), conditional MFI-LE performed better than the other item selection methods, including engagement-ranked MFI. There was only a negligible difference between the two versions of conditional MFI-LE (i.e., conditional MFI-LE using either $\theta = 0$ and $\theta = -1$ as a latent engagement threshold). For lower ability levels (i.e., $\theta < 0$), the performance of unconditional MFI-LE was better than the conventional MFI and effort-guided CAT but worse than the conditional MFI-LE methods. Also, compared with engagement-ranked MFI, both unconditional and conditional MFI-LE performed better for higher ability levels (i.e., $\theta > 0.5$). This finding suggests that the MFI-LE methods could utilize test-taking engagement more effectively for students with higher ability levels who are less likely to experience disengagement during the CAT administration. As the ability level and test length increased, the performance difference between the item selection methods became very negligible.

Discussion

CATs offer a personalized testing experience to each student by iteratively selecting items based on the student's interim ability levels (Eggen, 2012; Lord, 1980; Veldkamp, 2003). In this study, we proposed alternative item selection methods for low-stakes CATs where some students are likely to demonstrate aberrant response behaviors (e.g., rapid guessing and idle responding) due to the lack of test-taking engagement (e.g., Finn, 2015; Rose et al., 2010; Ulitzsch et al., 2020; Wise & DeMars, 2005; Wise, 2006, 2019). Using the conventional MFI algorithm as a baseline, the proposed item selection methods aim to find the most optimal item by ranking the items based on their engagement levels. In Simulation Study 1, we assumed a hypothetical scenario where students' response times for all items in the item bank were available prior to participating in the CAT. Using the existing response times, we categorized the items into one of the three engagement categories; namely, 1: rapid guessing, 2: idle responding, and 3: effortful responding. For each student, we ranked the items based on both the Fisher information using interim ability and the engagement level using response times, and then selected the most informative item with the highest engagement and ability ranking. The results of Study 1 showed that engagement-ranked MFI could provide more accurate ability estimates than the conventional MFI and effort-guided CAT (Wise, 2020; Wise & Kingsbury, 2016). The utility of using engagement-ranked MFI, instead of traditional item selection methods, was more salient for lower ability levels.

In Simulation Study 2, we proposed another item selection method that could be more suitable for operational CAT settings since engagement-ranked MFI relies on the unusual assumption that students' response times are known prior to participating in the test. Following Goldhammer et al.'s (2017) approach of defining test-taking engagement as a latent examinee characteristic, we calibrated the items in the item bank in two different ways. We estimated item difficulty and discrimination parameters based on dichotomous item responses and engagement parameters based on dichotomized response times (i.e., 1=Engaged, 0=Disengaged). The goal of the proposed method, MFI-LE, was to harness test-taking engagement as a secondary latent trait (i.e., LE) in

the item selection process. For each student, we selected the items that maximized the Fisher information based on both item parameters and engagement parameters. That is, MFI-LE identified the most informative item based on interim estimates of ability and LE. The results of Study 2 indicated that MFI-LE outperformed the conventional MFI and effort-guided CAT wherein only the students' interim ability levels were considered when selecting the items.

The simulation studies also showed that effort-guided CAT, which was also considered a reactive approach for dealing with disengaged examinees in CATs, was not as effective as the proposed MFI-LE methods. This finding emphasized the need for developing proactive solutions for tackling the lack of test-taking disengagement in low-stakes CATs. Furthermore, the findings of this study showed that incorporating test-taking engagement into the item selection process could be more essential for students with low ability, especially in a low-stakes CAT setting. This is congruent with previous studies that reported a significant relationship between effort and ability (e.g., Lindner et al., 2019). The utility of considering test-taking engagement in item selection appears to decrease as the student's ability level increases. Therefore, the conditional MFI-LE method could be a more suitable solution for tackling disengaged responses by regulating the item selection process depending on the student's engagement level in low-stakes CATs.

To date, researchers recommended various ways to improve students' test performances in low-stakes assessments, such as encouraging students to take a low-stakes assessment more seriously by expending maximal effort through proctors (Wise et al., 2019), response time-based scoring approaches (Gorgun & Bulut, 2021), or engagement monitoring such as auto-pauses (Wise et al., 2022). As a more proactive approach, the use of the engagement-ranked MFI and MFI-LE methods for item selection may also offer several benefits for low-stakes CATs. First, these methods could help promote optimal time use among students (Gorgun & Bulut, 2021; Tijmstra & Bolsinova, 2018), curtailing the occurrence of not-reached items, missing responses, effortless or idle responses, or rapid guesses in low-stakes CATs with a time limit. Second, finding the optimal items based on both ability and test-taking engagement could maximize the information about one's interim ability, while reducing the effect of construct-irrelevant variance in ability estimation due to the presence of aberrant responses such as rapid guesses (Haladyna & Downing, 2004). Third, incorporating test-taking engagement into item selection in low-stakes CATs can help practitioners evaluate whether insufficient test-taking engagement interferes with how students respond to the items (Wise & Kingsbury, 2016).

As explained earlier, the engagement-ranked MFI method requires a pre-knowledge of students' response time for each item and thus it may not be feasible for low-stakes CATs where each student responds to a different set of items selected from a large item bank. However, this method could still be employed in low-stakes, computerized assessments where the same items are often used repeatedly for tracking students' learning progress over time. For low-stakes CATs, the MFI-LE method could be a more feasible approach for diminishing the negative impact of disengagement on item selection and ability estimation. Especially with the conditional MFI-LE method, the flexibility of setting an LE threshold could help accommodate different scenarios in low-stakes CATs (e.g., test-taking engagement issues only among low-ability students). In addition, the

MFI-LE methods proposed in this study could pave the way for more advanced CAT applications taking other types of test-taking behaviors into account during the test administration and thereby increasing the reliability and validity of low-stakes CATs (Wise, 2020).

Limitations and future direction

This study has several limitations. First, this study utilized response time as a proxy for test-taking engagement (e.g., rapid guesses, idle responses, and effortful responses). However, it is also possible to use additional variables (e.g., process data extracted from logfiles captured by the computer, students' self-reports of test-taking engagement, and students' previous test scores) to account for test-taking engagement in low-stakes CATs. Thus, future research can investigate how additional indicators of test-taking engagement could be incorporated into item selection to render low-stakes CATs more adaptable to different testing situations and students with different test-taking behaviors. Second, we demonstrated the engagement-ranked MFI and MFI-LE algorithms using real data from a cross-sectional assessment. However, test-taking engagement levels may change from one assessment to another, affecting decisions to be made based on the change between the test scores, such as identifying academic growth (Yildirim-Erbasli & Bulut, 2021). Future research is needed to better understand the impact of test-taking engagement on low-stakes CATs when the assessments are used to monitor performance and progress over time. Third, we operationalized ability and LE as distinct latent traits and used both when finding the most optimal items during the item selection process, instead of jointly modeling ability and engagement within a single model. Thus, researchers may consider using a multidimensional IRT (MIRT) framework to model ability and engagement jointly and then use the MIRT model to implement a multidimensional CAT.

Conclusion

This study proposed new item selection algorithms that could minimize the negative influence of test-taking disengagement on item selection and ability estimation in low-stakes CATs. The MFI-LE method showed great promise for enhancing the operational low-stakes CATs by optimally selecting items that are informative in terms of both ability and engagement. The utility of MFI-LE was evident, especially for low-ability students since these students are more likely to show disengaged response behaviors in low-stakes assessments. Also, this study underscored the need for developing proactive remedies to deal with disengaged responses in real-time to optimize CAT administrations more effectively.

Abbreviations

CAT	Computerized adaptive test
MFI	Maximum fisher information
LE	Latent engagement
IRT	Item response theory
NT	Normative threshold
RMSE	Root-mean-squared error
cSEM	Conditional standard error of measurement
TIF	Test information function

Acknowledgements
Not applicable.

Author contributions
GG: Conceptualization, methodology, formal analysis, software, writing—original draft preparation. OB: Conceptualization, methodology, writing—review and editing.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials
The dataset used in this study comes from a commercial computer-based large-scale assessment and cannot be shared publicly. The dataset is available from the corresponding author, Guher Gorgun, upon reasonable request.

Declarations

Ethics approval and consent to participate
Ethical approval was not sought for the present study because of utilizing secondary data analysis retrospectively.

Consent for publication
All authors read and approved the final manuscript.

Competing interests
The authors have no competing interest to declare that are relevant to the content of this article.

Received: 25 June 2022 Accepted: 12 July 2023
Published online: 22 July 2023

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the difficulty in computer adaptive testing. *Applied Measurement in Education, 5*, 137–149. https://doi.org/10.1207/s15324818ame0502_4

Betz, N. E., & Weiss, D. J. (1976). Effects of immediate knowledge of results and adaptive ability testing on ability test performance. *Applied Psychological Measurement*. https://doi.org/10.1177/014662167700100212

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444. https://doi.org/10.1177/014662168200600405

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331–348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x

Braun, H., Kirsch, I., Yamamoto, K., Park, J., & Eagan, M. K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309–2344.

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research, 49*, 61–80.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika, 73*, 209–230. https://doi.org/10.1007/S11336-007-9045-9

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*(5), 1–39. https://doi.org/10.1863/jss.v071.i05

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 34*, 438–452. https://doi.org/10.1177/014662169602000303

Choe, E. M., Kern, J. L., & Chang, H. H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 43*(2), 135–158. https://doi.org/10.3102/1076998617723642

Davey, T. (2011). *A guide to computer adaptive testing systems*. Council of Chief State School Officers.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, 102. https://doi.org/10.3389/fpsyg.2019.00102

Eggen, T. J. (2012). Computerized adaptive testing item selection in computerized adaptive learning systems. In T. J. H. M. Eggen, C. A. W. Glas, T. J. H. M. Eggen, A. Beguin, B. P. Veldkamp, Q. He, M. Paap, M. Hiske Feenstra, G. Marsman, T. B. Maris, S. Wools, M. Hubregtse, M. van Groen, S. Klerk, J. A. Vermeulen, & F. M. van der Kleij (Eds.), *Psychometrics in practice at RCEC*. Enschede: University of Twente.

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

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*, 311–326. https://doi.org/10.1080/15305050701438074

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

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*, 655–670. https://doi.org/10.3102/1076998611422912
Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series, 2015*(2), 1–17. https://doi.org/10.1002/ets2.12067
Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732
Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC*. Paris: OECD Publishing.
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), 1–25. https://doi.org/10.1186/s40536-017-0051-9
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
Guo, X., Luo, Z., & Yu, X. (2020). A speed-accuracy tradeoff hierarchical model based on a cognitive experiment. *Frontiers in Psychology, 10*, 2910. https://doi.org/10.3389/fpsyg.2019.02910
Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x
Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quarterly, 50*, 75–87.
Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The effort paradox: effort is both costly and valued. *Trends in Cognitive Sciences, 22*(4), 337–349. https://doi.org/10.1016/j.tics.2018.01.007
Jin, K., & Wang, W. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement, 51*, 178–200. https://doi.org/10.1111/jedm.12041
Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*(4), 606–619. https://doi.org/10.1177/0013164406294779
Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. (MESA Memorandum No. 69). University of Chicago: MESA Psychometric Laboratory.
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
Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement, 41*(7), 495–511. https://doi.org/10.1177/0146621617707556
Liu, O. L., Rios, J. A., & Borden, V. (2014). *The effect of motivational instruction on college students 'performance on low-stakes assessment*. Philadelphia: Paper presented at the American Educational Research annual meeting.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.
Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive testing conditions. *Journal of Educational Measurement, 31*, 251–263. https://doi.org/10.1111/j.1745-3984.1994.tb00446.x
Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology, 110*(1), 27–45. https://doi.org/10.1037/edu0000205
Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychological Bulletin, 114*, 449–458. https://doi.org/10.1037/0033-2909.114.3.449
Novikov, N. A., Nurislamova, Y. M., Zhozhikashvili, N. A., Kalenkovich, E. E., Lapina, A. A., & Chernishev, B. V. (2017). Slow and fast responses: two mechanisms of trial outcome processing revealed by EEG oscillations. *Frontiers in Human Neuroscience, 11*, 218. https://doi.org/10.3389/fnhum.2017.00218
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., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences, 42*, 27–35. https://doi.org/10.1016/j.lindif.2015.08.002
Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74*(3), 423–452. https://doi.org/10.1177/0013164413504926
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. (2021). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
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), 1–25. https://doi.org/10.1186/s40536-021-00110-8
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*(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193
Rios, J. A., & Soland, J. (2021). Investigating the impact of noneffortful responses on individual-level scores: can the effort-moderated IRT model serve as a solution? *Applied Psychological Measurement, 45*(6), 391–406. https://doi.org/10.1177/01466216211013896
Rose, N., von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report*. https://doi.org/10.1002/j.2333-8504.2010.tb02218.x
Ross, B., Chase, A. M., Robbie, D., Oates, G., & Absalom, Y. (2018). Adaptive quizzes to increase motivation, engagement and learning outcomes in a first-year accounting unit. *International Journal of Educational Technology in Higher Education, 15*(1), 15–30. https://doi.org/10.1186/s41239-018-0113-2
Samuels, S. J., & Flor, R. F. (1997). The importance of automaticity for developing expertise in reading. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 13*(2), 107–121. https://doi.org/10.1080/1057356970130202

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: insights gained from response-time analyses. *Computer-Based Testing: Building the Foundation for Future Assessments, 25*(1), 237–266.
Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453
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
Sundre, D. L. (1999). Does examinee motivation moderate the relationship between test consequences and test performance? Paper presented at the annual meeting of the American Educational Research Association. Montreal
Sundre, D. L., & Wise, S. L. (2003). 'Motivation filtering' An exploration of the impact of low examinee motivation on the psychometric quality of tests. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago.
Sundre, D. L., & Moore, D. L. (2002). Assessment measures: the student opinion scale—a measure of examinee motivation. *Assessment Update, 14*(1), 8–9.
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*, 162–188. https://doi.org/10.1080/08957347.2011.555217
Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: using the student opinion scale to make valid inferences about student performance. *Journal of General Education, 58*(3), 129–151.
Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*, 1–9. https://doi.org/10.7275/wqzt-9427
Tijmstra, J., & Bolsinova, M. (2018). On the importance of the speed-ability trade-off when dealing with not reached items. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.00964
Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: the impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*(2), 320–332. https://doi.org/10.1037/0021-9010.87.2.320
Ulitzsch, E., Penk, C., von Davier, M., & Pohl, S. (2021). Model meets reality: validating a new behavioral measure for test-taking effort. *Educational Assessment, 26*(2), 104–124. https://doi.org/10.1080/10627197.2020.1858786
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*, 83–112. https://doi.org/10.1111/bmsp.12188
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308. https://doi.org/10.1007/s11336-006-1478-z
van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*(1), 5–20. https://doi.org/10.3102/1076998607302626
van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365–384. https://doi.org/10.1007/S11336-007-9046-8
Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203–226. https://doi.org/10.3102/10769986022002203
Veldkamp, B. P. (2003). Item selection in polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics*. Tokyo: Springer.
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
Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473–492. https://doi.org/10.1177/014662168200600408
Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement. *Measurement and Evaluation in Counseling and Development, 37*, 70–84. https://doi.org/10.1080/07481756.2004.11909751
Weiss, D. J., & Betz, N. E. (1973). *Ability measurement: conventional or adaptive? (Research Report 73–1)*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x
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. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing, 2*(3), 1–17. https://doi.org/10.7333/1401-02010001
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. (2020). An intelligent CAT that can deal with disengaged test taking. In H. Jiao & R. W. Lissitz (Eds.), *Application of artificial intelligence to assessment* (pp. 161–174). Information Age Publishing Inc.
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., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement, 53*(1), 86–105. https://doi.org/10.1111/jedm.12102
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., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education, 32*(2), 183–192. https://doi.org/10.1080/08957347.2019.1577248
Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Vancouver: Annual meeting of the National Council on Measurement in Education.

Wise, S. L., Ma, L., & Theaker, R. A. (2014). Identifying non-effortful student behavior on adaptive tests: implications for test fraud detection. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud* (pp. 175–185). Routledge.
Wise, S., Pastor, D. A., & Kong, X. (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
Yildirim-Erbasli, S. N., & Bulut, O. (2021). The impact of students' test-taking effort on growth estimates in low-stakes educational assessments. *Educational Research and Evaluation*. https://doi.org/10.1080/13803611.2021.1977152
Soland, J., & Kuhfeld, M. (2019). Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment*, *24*(4), 327-342. https://doi.org/10.1080/10627197.2019.1645592.
Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68-81. https://doi.org/10.1006/ceps.1999.1015.
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
Wise, S. L., Kuhfeld, M. R., & Cronin, J. (2022). Assessment in the time of COVID-19: Understanding patterns of student disengagement during remote Low-Stakes testing. *Educational Assessment*, *27*(2), 136-151. https://doi.org/10.1080/10627197.2022.2087621

Publisher's Note

Guher Gorgun Guher Gorgun is a PhD candidate in Measurement, Evaluation, and Data Science program at the University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5 CANADA. Her research interests include computerized assessments, student modelling, educational data mining, human-centered AI applications in education, learning analytics, evaluation of automatically generated items.

Okan Bulut Okan Bulut is an Associate Professor at the University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5 CANADA. His current research interests include educational data mining, big data modeling, computerized/digital assessments, psycho-educational assessments, and statistical programming using the R programming language.