# *Journal of Computerized Adaptive Testing*

***Volume 13 Number 1***
***February 2026***

# Optimizing Test Completion in Time-Constrained Adaptive Tests: A Post-Hoc Simulation Study

**Joyce Xinle Liu, Nour Armoush, and Okan Bulut**
**University of Alberta**

**The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing**
***www.iacat.org/jcat***

**ISSN: 2165-6592**

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

*Journal of Computerized Adaptive Testing*

*Volume 13, Number 1, February 2026*

DOI 10.7333/2602-1301001

ISSN 2165-6592

# Optimizing Test Completion in Time-Constrained Adaptive Tests: A Post-Hoc Simulation Study

**Joyce Xinle Liu, Nour Armoush, and Okan Bulut**
**University of Alberta**

In operational computerized adaptive tests (CATs), time limits are often imposed to manage logistical constraints, maintain test security, or preserve classroom instructional time. However, time constraints in CATs can lead to inaccurate ability estimates as some examinees might struggle to complete enough items within the allotted time due to differences in working speed and time intensities of individual items. This study proposes a simple dynamic approach to incorporate item response time (RT) data into the item selection algorithm, with the goal of optimizing test completion in a time-constrained setting. Linear optimization techniques were employed to balance information and RT when selecting subsequent items when needed. A post-hoc simulation study was conducted using real data from a 244-item digital assessment. CATs were simulated for 21,356 students, comparing the proposed (optimized) item selection algorithm with the conventional maximum Fisher information algorithm. The results indicate that the optimized algorithm improved test completion rates while slightly enhancing the accuracy and precision of ability estimates. The benefits were most pronounced under the shortest time constraint condition. Practical implications for implementations of time-constrained CATs are discussed.

*Keywords: computerized adaptive testing, item selection, linear optimization, response time, time limit*

In conventional computerized adaptive tests (CATs), the item selection algorithm relies solely on the examinee's previous responses to choose the following items (Weiss & Şahin, 2024). The examinee's ability estimate is updated after each response, and the most suitable item is then selected. Among the various item selection methods, the maximum Fisher information (MFI) method remains the most widely used in operational CATs (Veldkamp, 2016, Thompson & Weiss, 2011). Using the MFI method, the algorithm selects the item from the item bank that provides the maximum information based on the examinee's interim ability estimate. In recent years, there has been a growing interest in using behavioral data, such as response time (RT), to create more

innovative CATs. These ancillary variables can be incorporated into either the item selection or scoring processes to mitigate the impact of construct-irrelevant factors (e.g., test-taking disengagement), improve measurement precision, and allow CATs to better adapt to individual differences (Gorgun & Bulut, 2023; Weiss & Şahin, 2024; Wise, 2020).

Under optimal testing conditions, examinees should have sufficient time to complete each item, and the CAT will terminate when the conditional standard error of measurement (CSEM) for the examinee falls below a predetermined threshold (Weiss & Şahin, 2024). In practice, however, this criterion is often not used in isolation due to logistical constraints or suboptimal coverage of the item bank. Many operational CATs impose time limits or restrict the maximum number of items presented to examinees to maintain testing efficiency, reduce testing costs, or accommodate scheduling requirements. Examples of CATs with strict time limits include the computerized Armed Services Vocational Aptitude Battery (CAT-ASVAB; Sands et al., 1997) and the STAR assessments (Renaissance Learning, 2024a, 2024b). Other CATs, such as MAP Growth by the Northwest Evaluation Association, are intended to be untimed, but it is common for schools to set loose time limits during administration to minimize disruptions to instructional time (NWEA, 2022).

These time-constrained CATs might be disadvantageous to examinees who need more time to respond to the items, potentially preventing them from completing enough items to obtain an accurate ability estimate. Strict time limits might prevent some examinees from completing items they would otherwise be able to finish successfully, resulting in a systematic underestimation of their ability levels (Kane, 2020). This issue also exists in fixed-item tests, but might be exacerbated in CATs because examinees receive different sets of items with varying time intensities (Bridgeman & Cline, 2004; van der Linden, 2009). As such, the most informative items might sometimes be too time-consuming, particularly for certain ability ranges and for examinees who respond to items at a slower pace (He & Qi, 2023). It is, therefore, worthwhile to investigate whether accounting for examinees' use of time during a timed CAT and incorporating this information into item selection would add value to the implementation of CATs and lead to improved measurement precision and test completion rates.

RT, which refers to the total time duration that an examinee spends on an item, is a widely studied form of behavioral process data that is routinely captured in CATs and other digital assessments (Anghel et al., 2024). Previous studies have explored several different ways of incorporating RT into item selection, often using it as an indicator of test-taking engagement (Gorgun & Bulut, 2023; Wise, 2020) or speed (He & Qi, 2023; Kern & Choe, 2021; Veldkamp, 2016). For the latter, several studies have focused on maximizing the efficiency of a CAT and reducing overall testing time. One popular approach is maximizing Fisher information per unit of expected RT (MIT), a method first proposed by Fan et al. (2012). The authors applied van der Linden's (2007) hierarchical model for responses and RTs to adaptively estimate the examinee's latent ability and speed parameters after each item, and subsequent items are selected to efficiently accrue information about the examinee's ability by considering both the examinee's current parameter estimates and the time intensity of the items in the bank. Cheng et al. (2017) subsequently proposed a simplified version which models the examinee's expected RT for each item using the average log-transformed RT for that item, instead of fitting an RT model to individual-level RT data, making it more computationally efficient. Unsurprisingly, the MIT approach commonly resulted in unbalanced item exposure due to favoring items with high discrimination and low time intensity. To mitigate this issue, researchers have explored

incorporating item exposure control techniques such as time-weighted *a*-stratification (Fan et al., 2012) and experimenting with different information-to-RT ratios (Choe et al., 2018). He and Qi (2023) later extended this idea of maximizing information per unit of time to multidimensional CATs.

To address the challenges posed by time limits more directly, several studies have investigated ways to use RT information to mitigate differential speededness among examinees. Differential speededness refers to the phenomenon where examinees might experience different amounts of time pressure during a time-constrained test (van der Linden, 2009). Because each examinee receives a unique set of items in a CAT, it is important to manage the overall speededness of tests so that all examinees have a fair opportunity to complete the test within the allotted time. In the literature, strategies to address differential speededness in CATs include using RT models to predict the amount of time examinees will spend on remaining items in the bank, then placing specific RT constraints in the item selection algorithm to ensure that the sum of the actual and expected RT does not exceed the time limit (van der Linden, 2009; van der Linden et al., 1999). An alternative approach, proposed by van der Linden and Xiong (2013), involves constructing the CAT based on a reference test form that already possesses an ideal level of test speededness. More recently, Kern and Choe (2021) proposed another method that employs a joint expected a posteriori estimator, instead of the standard maximum likelihood estimator, for ability and speed parameters in van der Linden's (2007) hierarchical model.

Building on these previous studies, this paper proposes a simple alternative approach to incorporate RTs into the item selection algorithm of a CAT, with the goal of optimizing test completion in a time-constrained setting. Facilitating test completion offers psychological benefits to examinees by giving them a fair opportunity to demonstrate their abilities and ensuring that items are selected to allow sufficient time for completion (Bridgeman & Cline, 2004). It also provides operational advantages for test administrators by minimizing missing responses toward the end of the test and allowing more examinees to complete the required minimum number of items. The proposed method monitors the examinee's time use as the CAT progresses and predicts whether they are at risk of running out of time after each item. When necessary, a time-adjusted item selection algorithm is activated, employing linear optimization to balance item information and RT in selecting the next item for the examinee. Once the examinee is no longer predicted to be at risk of not completing the assessment, the item selection method reverts to the standard approach. This method ensures that the test remains efficient in terms of both precision and time, dynamically adjusting to the examinee's progress and the time constraints of the test. It is relatively simple to implement and avoids the computational demands of estimating latent speed parameters in real time. The effectiveness of the proposed method was investigated by comparing it to the conventional MFI method, evaluating both the test completion rates and accuracy of the ability estimates across different test conditions and subgroups.

## Method

A post-hoc simulation study was conducted using real data from a large-scale digital assessment administered to 42,711 Canadian students in grades 8 and 9. The assessment included 244 multiple-choice items covering various subject areas, such as mathematics and science. To create a large item bank for the current study, the items on the assessment were assumed to measure a single (i.e., unidimensional) hypothetical construct, representing students' overall competence.

The data consisted of both students' binary responses (correct/incorrect) and RTs for all questions. The average RT across all items was 62.94 seconds (*SD* = 41.73), and the maximum RT captured for the items was 200 seconds.

In operational CATs, item parameters are typically obtained from pre-calibrated items. To conduct a more realistic CAT simulation, the full dataset was randomly split into two halves to mimic a calibration dataset ($n$ = 21,355) and a simulation dataset ($n$ = 21,356). The calibration dataset was used to estimate the item parameters: item difficulty ($M$ = –0.26, $SD$ = 0.85) and item discrimination ($M$ = 0.87, $SD$ = 0.29). It also served as a bank of students with known RT patterns for each item so that RT could be used to select items in the simulated CATs. Based on the simulation dataset, students' abilities ($M$ = –0.01, $SD$ = 0.98) were estimated using the 2-parameter logistic (2PL) item response theory model (with scaling constant $D$ = 1) based on their responses to all 244 items[1]. These ability estimates were treated as indicators of the students' true abilities, used to evaluate the performance of the two item selection approaches in this study. The maximum a posteriori estimator was used for ability estimation throughout this study.

The mirtCAT package (Chalmers, 2016) in R (R Core Team, 2024) was used to simulate CATs with a fixed time limit (in seconds), denoted as $T_{\text{total}}$, and specified minimum ($L_{\text{min}}$) and maximum ($L_{\text{max}}$) numbers of items to be administered. Two separate CATs were simulated for each student, one using the *conventional* MFI method for item selection, and the other using a modified time-adjusted approach.

## The Modified Method of Item Selection

In the modified approach (hereafter referred to as the *optimized* approach), the MFI method remained the default item selection strategy, but item selection also incorporated RT data. Items were selected by Fisher information for a given interim ability estimate, $\hat{\theta}$. For the 2PL model, the information for item $i$ is given by

$$I_i(\hat{\theta}) = a_i^2 \cdot P_i(\hat{\theta}) \cdot \left(1 - P_i(\hat{\theta})\right), \tag{1}$$

where $P_i(\hat{\theta})$ is the probability of the student with estimated ability $\hat{\theta}$ answering the item correctly. This probability is:

$$P_i(\hat{\theta}) = \frac{\exp\left(a_i(\hat{\theta} - b_i)\right)}{1 + \exp\left(a_i(\hat{\theta} - b_i)\right)}, \tag{2}$$

where $a_i$ and $b_i$ are the discrimination and difficulty parameters of item $i$, respectively.

To dynamically adjust item selection based on a student's time use, the following procedure was used. After each item on the CAT, the student's total time spent up to that point was recorded as $T_{\text{actual}}$, and a prediction was made as to whether the student would have sufficient time to complete the remaining items. First, $T_{\text{remaining}}$ was defined as the amount of time remaining on the test:

[1] In the original assessment, the items were also calibrated with the 2PL model.

$$T_{\text{remaining}} = T_{\text{total}} - T_{\text{actual}} . \tag{3}$$

Suppose the item bank has a total of $N$ items, indexed by $i = 1, 2, \ldots, N$. Let $S_m = \{i_1, i_2, \ldots, i_m\}$ denote the set of indices for the $m$ items that have been administered to the student. Then the indices of the remaining $N - m$ items in the bank can be represented by the set $R_m = S_m^c$, where the complement is taken with respect to $\{1, 2, \ldots, N\}$. The estimated amount of time that the student would need for the remainder of the test, $T_{\text{expected}}$, was computed as:

$$T_{\text{expected}} = (L_{\max} - m) \cdot \frac{1}{|R_m|} \sum_{i \in R_m} T_i , \tag{4}$$

where $L_{\max}$ is the maximum length of the CAT, and $T_i$ is the average (mean) RT for item $i$. In this method, the speed of individual students was not directly estimated to avoid overcomplicating the algorithm. Instead, the expected RT for item $i$ and student $j$, $T_{ij}$, was simply modeled as the average RT for item $i$, based on the students in the calibration dataset. In essence, $T_{\text{expected}}$ represents an estimate of the average time needed for a student to complete the remaining $N - m$ items. After each administered item on the CAT, the values of $T_{\text{actual}}$, $T_{\text{remaining}}$, and $T_{\text{expected}}$ were updated based on the current set of administered and remaining items.

When $T_{\text{expected}}$ was less than $T_{\text{remaining}}$, it was assumed that the student would have sufficient time to complete the test, and items would continue to be selected using the MFI method. However, if $T_{\text{expected}}$ exceeded $T_{\text{remaining}}$, the following time-adjusted item selection algorithm was implemented to select items that could better balance item information and RTs. For each student $j$, the current ability estimate is denoted by $\hat{\theta}_j$, and the remaining time in the test is denoted by $R_j = T_{\text{remaining}}$. Item selection was optimized based on two weighted criteria:

$$C_i = \omega_{\text{info}} \cdot I_i(\hat{\theta}_j) - \omega_{\text{time}} \cdot T_{ij} , \tag{5}$$

where $\omega_{\text{info}}$ is the weight assigned to the information, $I$, of item $i$, and $\omega_{\text{time}}$ is the weight assigned to the expected RT of item $i$. $C_i$ is the combined criterion for item $i$.

For this study, three different combinations of the weights $\omega_{\text{info}}$ and $\omega_{\text{time}}$ were tested: (1) $\omega_{\text{info}} = 0.8$, $\omega_{\text{time}} = 0.2$; (2) $\omega_{\text{info}} = 0.7$, $\omega_{\text{time}} = 0.3$; and (3) $\omega_{\text{info}} = 0.6$, $\omega_{\text{time}} = 0.4$. A higher weight was given to item information to prioritize measurement precision while facilitating test completion. The following linear programming problem was solved to find the optimal balance between information and time:

$$\max\left(\omega_{\text{info}} \cdot I_i(\hat{\theta}_j) - \omega_{\text{time}} \cdot T_{ij}\right) , \tag{6}$$

subject to:

$$\omega_{\text{info}} + \omega_{\text{time}} = 1, \tag{7}$$

$$T_{ij} \leq R_j. \tag{8}$$

This optimization ensures that the selected item maximizes the combined criterion $C_i$ while respecting the remaining time constraints. To compute $C_i$ in this step, $T_{ij}$ was expressed in minutes

rather than in seconds, to keep its scale more comparable to that of information. The item $i$ that maximized the combined criterion $C_i$ was selected as the next item:

$$i^* = \arg\max_i \left(\omega_{\text{info}} \cdot I_i\left(\hat{\theta}_j\right) - \omega_{\text{time}} \cdot T_{ij}\right). \tag{9}$$

The item selection process described above was repeated until either the maximum test length $L_{\max}$ or time limit $T_{\text{total}}$ was reached, or the conditional observed standard error of measurement (CSEM) for the examinee fell below the threshold of 0.3. The final ability estimate was recorded. It should be noted that in both the conventional and optimized approaches, ability estimation was based solely on students' responses. RT information was used only to inform item selection in the optimized method.

### Simulation Conditions

Nine different conditions were examined, each with a unique combination of time limits and information-time weights, as summarized in Table 1. A maximum test length of 15 items was used for all simulation conditions. The time limits were set based on the average RT observed per item (approximately one minute), with additional conditions extending the time in five-minute increments to simulate more realistic testing scenarios. In each condition, the proposed optimized item selection method was compared to the conventional MFI method. Performance was evaluated based on the correlation between $\hat{\theta}_j$ and "true" $\theta$ based on all 244 items, bias, root-mean-squared error (RMSE), and test completion rate. RMSE and bias values were calculated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{n}\left(\hat{\theta}_j - \theta_j\right)^2}{n}}, \tag{10}$$

$$\text{Bias} = \frac{\sum_{j=1}^{n}\left(\hat{\theta}_j - \theta_j\right)}{n}, \tag{11}$$

where $n$ is the total number of students in the dataset. Test completion rate was defined as the proportion of students who either reached the maximum number of items or whose test was terminated early upon meeting the CSEM threshold.

## Results

Table 2 presents the comparison of evaluation criteria for the conventional and optimized CATs across the nine simulation conditions. In general, the optimized CAT conditions slightly outperformed the conventional CAT in terms of the accuracy of ability estimates. Compared with the conventional CAT, the optimized CAT yielded slightly higher correlations between true and estimated abilities and lower RMSE values. Bias values were close to zero and comparable across all conditions, indicating that there was minimal systematic over- or underestimation of ability levels. Test completion rates were higher in the optimized conditions, as expected, and increased

**Table 1. Simulation Conditions**
**($L_{\min} = 5,\ L_{\max} = 15,\ \text{CSEM} = 0.30$)**

| Simulation | $\omega_{\text{info}}$ | $\omega_{\text{time}}$ | $T_{\text{total}}$ |
|---|---|---|---|
| 1 | 0.8 | 0.2 | 900 |
| 2 | | | 1200 |
| 3 | | | 1500 |
| 4 | 0.7 | 0.3 | 900 |
| 5 | | | 1200 |
| 6 | | | 1500 |
| 7 | 0.6 | 0.4 | 900 |
| 8 | | | 1200 |
| 9 | | | 1500 |

as more weight was placed on time. These overall results suggest that the optimized approach, which incorporates RT information, can facilitate test completion without compromising on the accuracy of ability estimates.

**Table 2. Comparison of Evaluation Criteria for Conventional Versus Optimized CAT**

| $T_{\text{total}}$ | Simulation | $\omega_{\text{info}}/\omega_{\text{time}}$ | CAT Type | RMSE | Bias | *r* | TCR |
|---|---|---|---|---|---|---|---|
| 900 | | | Conventional | 0.384 | –0.0033 | 0.920 | 17.6% |
| | 1 | 0.8/0.2 | Optimized | 0.365 | –0.0028 | 0.928 | 42.8% |
| | 2 | 0.7/0.3 | Optimized | 0.365 | –0.0032 | 0.928 | 52.4% |
| | 3 | 0.6/0.4 | Optimized | 0.369 | –0.0043 | 0.926 | 55.7% |
| 1200 | | | Conventional | 0.350 | 0.0012 | 0.933 | 62.2% |
| | 4 | 0.8/0.2 | Optimized | 0.344 | –0.0002 | 0.936 | 73.9% |
| | 5 | 0.7/0.3 | Optimized | 0.345 | –0.0011 | 0.936 | 76.0% |
| | 6 | 0.6/0.4 | Optimized | 0.345 | –0.0004 | 0.935 | 76.9% |
| 1500 | | | Conventional | 0.339 | 0.0013 | 0.938 | 90.9% |
| | 7 | 0.8/0.2 | Optimized | 0.338 | 0.0010 | 0.938 | 93.3% |
| | 8 | 0.7/0.3 | Optimized | 0.339 | 0.0012 | 0.938 | 93.6% |
| | 9 | 0.6/0.4 | Optimized | 0.338 | 0.0010 | 0.938 | 93.7% |

*Note.* $n$ = 21,356. $T_{\text{total}}$ denotes the time limit in seconds and corresponds to 15, 20, and 25 minutes respectively. RMSE: root-mean-squared error. *r:* correlation between true and estimated $\theta$ values. TCR: test completion rate.

The advantages of the optimized CAT were most prominent under the shortest time constraint ($T_{\text{total}} = 900$), with the largest increases in $r$ values and decreases in RMSE compared to the conventional CAT. As the time limit became more generous (i.e., $T_{\text{total}} = 1200$ and $T_{\text{total}} = 1500$), the performance gains in the optimized CAT decreased—understandably, since more students were able to complete the test within the allotted time under MFI. This finding indicates that the RT-based item selection method might be particularly useful for CATs with tighter time constraints. Adjustments to the relative weighting of information and time had little impact on RMSE but did produce noticeable differences in test completion rates. Specifically, test completion rates increased as greater weight was assigned to time, with the effect being most pronounced under the most restrictive time limit. However, the optimized CAT showed a slight loss in ability estimation accuracy as more weight was transferred to time. For this study, subsequent analysis was conducted using the weighting combination $\omega_{\text{info}} = 0.8$ and $\omega_{\text{time}} = 0.2$, which provided the best accuracy along with substantial improvements in test completion rates over the MFI condition.

To further investigate the impact of the optimized approach, the evaluation was also performed on two subgroups of students: those who did not complete the test under the conventional MFI condition and those who did. Table 3 presents a comparison of the evaluation criteria by subgroup, using the 0.8/0.2 weighting combination. When interpreting the results, greater emphasis was placed on the relative performance gains between the conventional and optimized approaches within each subgroup and time limit condition, rather than on the absolute values of the metrics.

**Table 3. Comparison of Evaluation Criteria by Subgroup**
**($\omega_{\text{info}} = 0.8,\ \omega_{\text{time}} = 0.2$)**

| Subgroup | *n* | Time Limit ($T_{\text{total}}$) | CAT Type | RMSE | Bias | *r* | TC |
|---|---|---|---|---|---|---|---|
| Did not complete test under MFI | 17,577 | 900 | Conventional | 0.394 | –0.0075 | 0.906 | |
| | | | Optimized | 0.371 | –0.0077 | 0.917 | 5,439 |
| | 8,063 | 1200 | Conventional | 0.377 | –0.0084 | 0.883 | |
| | | | Optimized | 0.363 | –0.0121 | 0.892 | 2,517 |
| | 1,951 | 1500 | Conventional | 0.375 | 0.0018 | 0.830 | |
| | | | Optimized | 0.370 | –0.0003 | 0.836 | 538 |
| Completed test under MFI | 3,779 | 900 | Conventional | 0.331 | 0.0163 | 0.890 | |
| | | | Optimized | 0.332 | 0.0200 | 0.889 | 3,700 |
| | 13,293 | 1200 | Conventional | 0.333 | 0.0070 | 0.922 | |
| | | | Optimized | 0.333 | 0.0070 | 0.922 | 13,265 |
| | 19,405 | 1500 | Conventional | 0.335 | 0.0012 | 0.936 | |
| | | | Optimized | 0.335 | 0.0012 | 0.936 | 19,394 |

*Note.* TC: test completion (number of students in subgroup who completed the test), shown for the optimized CAT condition)

The results demonstrate that the optimized item selection approach had an overall positive impact on students who struggled to complete the test under the conventional item selection approach. For the subgroup of students who did not complete the test under MFI, approximately one-third were able to complete the test under the optimized approach. This subgroup also showed improved accuracy in ability estimates, as indicated by the reduced RMSE. On the other hand, test completion rates and ability estimation accuracy for students who completed the test under MFI were generally not impacted.

Figure 1 provides a more detailed view of the impact of the optimized CAT on the accuracy and precision of ability estimates for students in each subgroup. In this figure, accuracy is shown on the horizontal axis as the difference between the estimated ability and the student's true ability—points closer to the vertical zero line indicate higher accuracy. Precision is shown on the vertical axis as the final CSEM achieved at the end of the test, with lower values indicating better precision. Again, the benefits of the optimized approach are more apparent in the first subgroup. By using RT information to further customize item selection, the optimized approach can help to improve both measurement accuracy and precision for students who were otherwise unable to complete the test. Due to the short length of the test in this simulation study, the CSEM tended to plateau at approximately 0.340, and no student was able to achieve early termination. A longer CAT design was not explored due to time considerations.

Figure 2 presents a comparison of the total number of items completed by students for the conventional and optimized CATs, to complement the findings in Table 2. Overall, the optimized approach helped students to complete more items on the test within the time limit, which in turn enhanced ability measurement.

## Discussion and Conclusions

CAT requires fewer administered items than a fixed-item assessment to accurately estimate an examinee's ability by selecting the most suitable items for each examinee (Kingsbury & Zara, 1989; van der Linden & Pashley, 2009). As the examinee answers more items, the adaptive testing algorithm progressively improves the precision of the $\theta$ estimate until it reaches a predefined endpoint. CATs have been found to reduce testing time by 50% to 60% (Yasuda et al., 2021) and shorten assessment length by 50% to 90% (Brown & Weiss, 1977; Ebenbeck et al., 2024). Despite these advantages, the effectiveness of a CAT can be reduced when a time limit is imposed. Some examinees might struggle to complete enough items within the given timeframe due to differences in working speed and the time intensities of administered items, potentially impacting the accuracy of their $\theta$ estimates. Strict or loose time limits are still commonly used in operational CATs, either to preserve classroom instructional time, maintain test security, or manage logistical constraints (Kern & Choe, 2021; Weiss & Şahin, 2024). In these time-constrained CATs, there is value in incorporating RT data into the adaptive testing process to select items for examinees that give them the best opportunity to demonstrate their true ability within the allotted time.

This study aimed to evaluate the potential benefits of implementing a time-adjusted item selection algorithm alongside the conventional MFI algorithm. The time-adjusted (optimized) algorithm is implemented only when needed during the CAT administration, allowing examinees to complete more items within the time limit to aid ability estimation. The results from this study indicate that the optimized algorithm led to notable improvements in test completion rates, as well

**Figure 1. Impact of Optimized CAT on Ability Estimation (by Subgroup)**
(Each circle represents a student in the dataset)

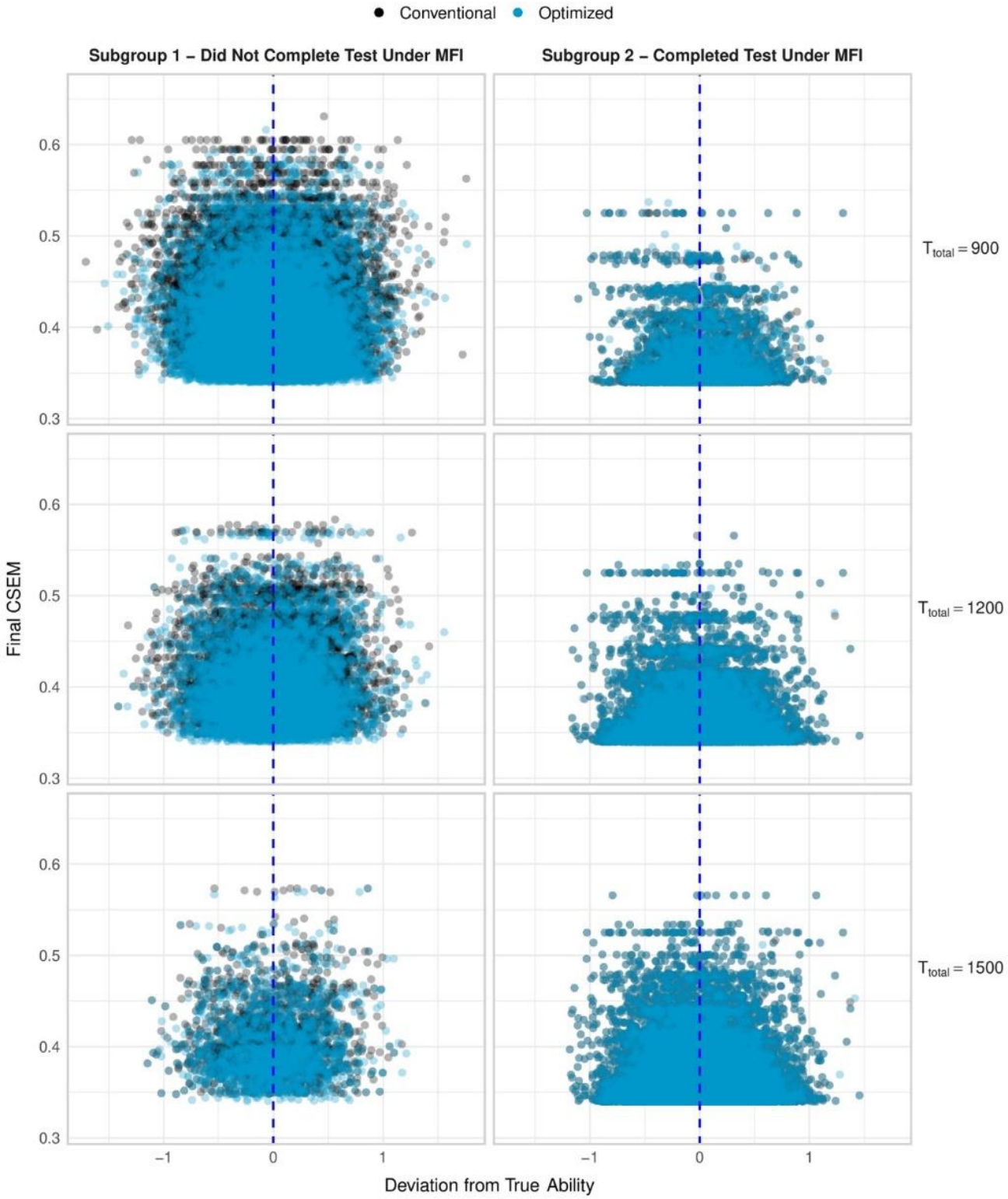
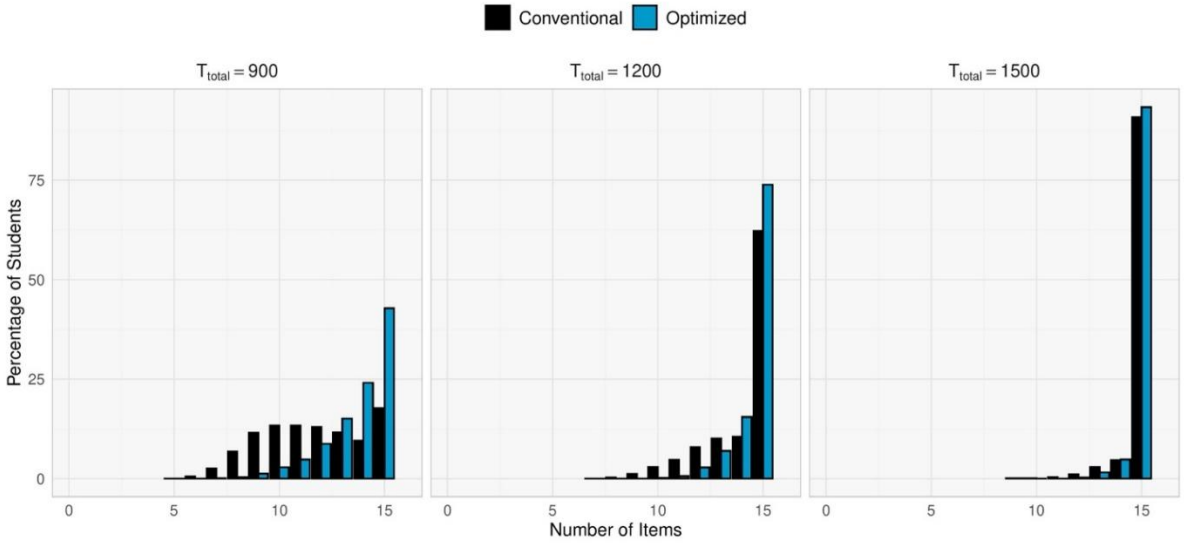


**Figure 2. Distribution of Number of Items Answered**
**($\omega_{\text{info}} = 0.8,\ \omega_{\text{time}} = 0.2$)**




as a modest improvement in the accuracy and precision of ability estimation for examinees who would have otherwise been unable to complete the test. Given the large sample size, this finding is promising and suggests that RT information could add value to CAT administrations by facilitating test completion without compromising measurement quality.

Overall, the results corroborate previous studies that have theorized or demonstrated the added value of RTs in enhancing measurement when considered together with response data (e.g., Molenaar, 2015; Sideridis & Alahmadi, 2022; van der Linden, 2007). RTs can serve as an additional source of information about examinees' ability, though the contribution is typically modest (Liu & Bulut, 2025). In the context of CAT, this study's results highlight the benefits of incorporating RTs into the item selection algorithm, aligning with findings from other studies exploring alternative methodologies. For example, Fan et al. (2012) demonstrated that the MIT method—which iteratively estimates examinees' latent ability and speed parameters during CAT administration and selects items that maximize Fisher information per unit of expected RT—reduces the average testing time while maintaining similar measurement accuracy to a standard MFI approach. More recently, Gorgun and Bulut (2023) found that incorporating test-taking engagement (operationalized as a latent trait through RTs) into item selection improved ability estimation accuracy, especially for low-ability examinees. In this study, the problem was reconsidered within the context of time-constrained CATs, focusing on supporting examinees' test completion in a dynamic manner while keeping the algorithm relatively simple and easily implementable by avoiding the use of latent speed or engagement constructs. The results further support existing literature that RTs can be used to support measurement in contexts where timing or speededness are important (Kern & Choe, 2021; van der Linden & Xiong, 2013).

One interesting finding from this study was that increasing the time weight ($\omega_{\text{time}}$) from 0.2 to 0.4 continued to improve the test completion rate, but the RMSE, bias, and *r* values largely remained stable (see Table 2), maintaining a slight improvement from the conventional MFI condition. As described by Equation 8, the optimized algorithm included an important constraint so that it only considers items with an average RT within the remaining time available on the test.

This constraint alone, because of the pre-filtering step, might cause item selection in the optimized CAT to differ from the conventional CAT, especially toward the end of the test. To investigate its impact, an additional simulation was run for the $T_{\text{total}} = 900$ condition, with all the weight given to information (i.e., $\omega_{\text{info}} = 1.0$, $\omega_{\text{time}} = 0.0$). This condition produced an RMSE of 0.378, bias of –0.024, $r$ of 0.922, and test completion rate of 23.9%. Taken together with the results in Table 2, the findings suggest that the constraint specified by Equation 8 appears to already provide some benefit by itself, but placing an explicit weight on time during item selection adds additional value. Nevertheless, there is likely a point at which assigning too much weight to time begins to compromise accuracy. Hence, it is important in practice for test administrators to adjust the weightings based on their specific testing context to achieve an appropriate balance, for example, to prioritize information or to minimize the effect of noise in RT data.

Some important practical implications can be highlighted for CAT implementations under time-constrained settings. In educational contexts, CATs might be used to quickly determine an examinee's ability level and identify whether they require additional support or intervention (Ebenbeck et al., 2024; Kingsbury & Hauser, 2004). While time limits are often imposed for practical reasons, examinees who are unable to complete enough items within the given time might not be able to receive an accurate ability estimate. This could put them at a disadvantage if important educational decisions are made based on those estimates. This study proposed an approach to mitigate the issue by incorporating RTs into the item selection algorithm and allowing the CAT to select more optimal items that balance information and time. The simulation study showed that using RTs as collateral data in this way can yield practical benefits by enabling examinees to complete more items on the test, thereby improving the measurement of their ability. Another useful benefit to test administrators is that the item bank utilization rate will also improve, as the algorithm becomes less concentrated on a small set of highly informative items.

## Limitations and Future Research

Several limitations of this study are considered, along with directions for future work. First, this study used average item RTs to predict the amount of time that an examinee would need to complete the remaining items after each iteration. While this simplified approach yielded promising performance gains in the optimized algorithm, it does not account for individual differences in working speed. Future research could explore practical ways of incorporating individuals' speed or RT patterns when modeling the expected RT for each item. The feasibility of such approaches in operational testing settings (e.g., real-time computational demands and required resources) should also be considered and investigated. Second, this study only compared the optimized algorithm to the MFI method. Although MFI is the most widely used approach and has shown minimal differences compared to other item selection methods, such as the maximum posterior weighted information (MPWI) and minimum expected posterior variance (MEPV; Murphy et al., 2010), it might still be worthwhile to explore the incorporation of RTs with other conventional algorithms. Future studies could also include a more detailed comparison of different RT-based item selection algorithms, including the MIT method or its variants, in different assessment contexts. Third, it was assumed that all items in the original fixed-item test measured a unidimensional construct, which was necessary to create a large item bank for the post-hoc simulation. This assumption was reasonable in the context of this study, given that the items were drawn from a single test measuring students' academic competence. The dataset also provided

students' actual response and RT data, making the simulation more realistic. However, many educational assessments are inherently multidimensional. Hence, it might be worthwhile to extend the proposed approach to multidimensional contexts, possibly with more comprehensive simulation studies to better cover cases involving rare or atypical response-RT patterns. Content balancing is also a common consideration in operational CAT implementations (Leung et al., 2003). Future studies could incorporate content constraints to ensure that the use of RT-based algorithms does not skew the content distribution of the test.

## References

Anghel, E., Khorramdel, L., & von Davier, M. (2024). The use of process data in large-scale assessments: A literature review. *Large-scale Assessments in Education*, *12*. DOI

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement*, *41*(2), 137–148. DOI

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries (Research Report 77-6)*. University of Minnesota, Department of Psychology, Psychometric Methods Program. WebLink

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1–38. DOI

Cheng, Y., Diao, Q., & Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research Methods*, *49*. DOI

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, 135–158. DOI

Ebenbeck, N., Bastian, M., Mühling, A., & Gebhardt, M. (2024). Duration versus accuracy—what matters for computerised adaptive testing in schools? *Journal of Computer Assisted Learning*, *40*(6), 3443–3453. DOI

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*(5), 655–670. DOI

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*. DOI

He, Y., & Qi, Y. (2023). Using response time in multidimensional computerized adaptive testing. *Journal of Educational Measurement*, *60*(4), 697–738. DOI

Kane, M. (2020). The impact of time limits and timing information on validity. In M. J. Margolis & R. A. Feinberg (Eds.). *Integrating timing considerations to improve testing practices* (pp. 19–31). Routledge.

Kern, J. L., & Choe, E. (2021). Using a response time–based expected a posteriori estimator to control for differential speededness in computerized adaptive test. *Applied Psychological Measurement*, *45*(5), 361-385. DOI

Kingsbury, G. G., & Hauser, C. (2004, April). *Computerized adaptive testing and No Child Left Behind* [Conference presentation]. Annual Meeting of the American Educational Research Association, San Diego, CA, United States.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359–375. DOI

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2003). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement*, *63*(2), 257–270. DOI

Liu, J. X., & Bulut, O. (2025, April 23–26). Sequential modeling of responses and response times in a CAT with LSTM [Conference presentation]. In H. Kuang (Chair), *Fuse diverse process data in education measurement*. Annual Meeting of the National Council on Measurement in Education, Denver, CO, United States.

Molenaar, D. (2015). The value of response times in item response modeling. *Measurement*, *13*, 177–181. DOI

Murphy, D. L., Dodd, B. G., & Vaughn, B. K. (2010). A comparison of item selection techniques for testlets. *Applied Psychological Measurement*, *34*(6), 424–437. DOI

NWEA. (2022). *Average MAP Growth test durations*. WebLink

R Core Team. (2024). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. WebLink

Renaissance Learning. (2024a). *Star Assessments™ for reading technical manual.* WebLink

Renaissance Learning. (2024b). *Star Assessments™ for math technical manual.* WebLink

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation.* American Psychological Association. DOI

Sideridis, G., & Alahmadi, M. T. S. (2022). The role of response times on the measurement of mental ability. *Frontiers in Psychology*, *13*. DOI

Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests, *Practical Assessment, Research, and Evaluation, 16*(1). DOI

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. DOI

van der Linden, W. J. (2009). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*(1), 25–41. DOI

van der Linden, W. J., & Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In *Elements of adaptive testing* (pp. 3–30). New York, NY: Springer New York.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210. DOI

van der Linden, W. J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 418–438. DOI

Veldkamp, B. P. (2016). On the issue of item selection in computerized adaptive testing with response times. *Journal of Educational Measurement*, *53*(2), 212–228. DOI

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–27. DOI

Weiss, D. J., & Şahin, A. (2024). *Computerized adaptive testing: From concept to implementation*. The Guilford Press.

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. DOI

Yasuda, J., Mae, N., Hull, M. M., & Taniguchi, M. (2021). Optimizing the length of computerized adaptive testing for the force concept inventory. *Physical Review Physics Education Research*, *17*(1). DOI

## External Support

This research received no specific grant from any funding agency
in the public, commercial, or not-for-profit sectors.

## Author Address

xinle4@ualberta.ca

## Citation

Liu, J. X., Armoush, N., & Bulut, O. (2026). Optimizing test completion
in time-constrained adaptive tests: A post-hoc simulation study.
*Journal of Computerized Adaptive Testing, 13*(1), 1-14.
DOI: https://10.7333/2602-1301001