Journal of Computerized Adaptive Testing

Volume 12 Number 2
June 2025

A Score-Based Method for Detecting Item Compromise and Preknowledge in Computerized Adaptive Testing

Kylie Gorney, Michigan State University

Chansoon Lee, American Board of Internal Medicine

Jianshen Chen, College Board

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 12, Number 2, June 2025

DOI 10.7333/2506-1202123

ISSN 2165-6592

A Score-Based Method for Detecting Item Compromise and Preknowledge in Computerized Adaptive Testing

Kylie Gorney
Michigan State University

Chansoon Lee
American Board of Internal Medicine

Jianshen Chen
College Board

In recent years, several methods have been proposed to detect compromised items (CI) and examinees with preknowledge (EWP) simultaneously. However, most of these methods are limited in one of two ways: (1) The method was specifically designed for non-adaptive tests and might therefore not be suitable for adaptive tests, or (2) the method involves the analysis of more than just the item scores. In this paper, a new method is developed for computerized adaptive tests that requires only an analysis of the item scores. The performance of the method is evaluated using simulations in which several factors are manipulated, such as test length, the percentage of CI, the percentage of EWP, and EWP ability distribution. Across most conditions, the method is shown to produce reasonable false positive rates and favorable true positive rates for both items and examinees.

Keywords: computerized adaptive testing, item compromise, item preknowledge, test fraud, test security

In educational and psychological testing, validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11), while fairness requires the removal of "construct-irrelevant barriers to maximal performance for any examinee" (American Educational Research Association et al., 2014, p. 63). Item preknowledge—which refers to a situation in which examinees have had access to items and/or answers prior to taking a test—poses a direct threat to both test fairness and validity. Consequently, there has been growing interest in the development of methods for detecting compromised items (CI; e.g., Choe et al., 2018; van der Linden & Belov, 2023; Zhang, 2014; Zhang & Li, 2016), examinees with preknowledge (EWP; e.g., Man & Harring, 2023; Sinharay, 2017; Sinharay & Johnson, 2020), and CI and EWP simultaneously (e.g., Belov, 2014; Boughton et al., 2017; Chen et al., 2022; Gorney et al., 2023; O'Leary & Smith, 2017; Pan et al., 2022; Pan & Wollack, 2023).

This paper focuses on the problem of simultaneously detecting CI and EWP, which is easily the most challenging—yet arguably most realistic—scenario. Although several methods have been proposed for this purpose, most are limited in one of two ways: (1) The method was specifically designed for non-adaptive tests, or (2) the method involves the analysis of more than just the item scores. The first limitation is problematic because methods designed for non-adaptive tests might not be suitable for adaptive tests, where much of the data are missing since different examinees are administered different items. The second limitation can be problematic because information beyond the item scores is not always trustworthy or available. For example, the method of Gorney et al. (2023) involves the analysis of item distractors, which are not available for all item types. Meanwhile, the methods of Boughton et al. (2017), Chen et al. (2022), Pan et al. (2022), and Pan and Wollack (2023) involve the analysis of item response times. Item response times are increasingly available in computer-based tests; however, they might not always be trustworthy due to the potential for examinees to "fake" realistic response times by altering their behavior. Therefore, the purpose of this paper was to develop a new method for detecting CI and EWP that addresses both limitations of previous research. The new method is specifically designed for computerized adaptive tests (CATs) and requires only an analysis of the item scores.

Method

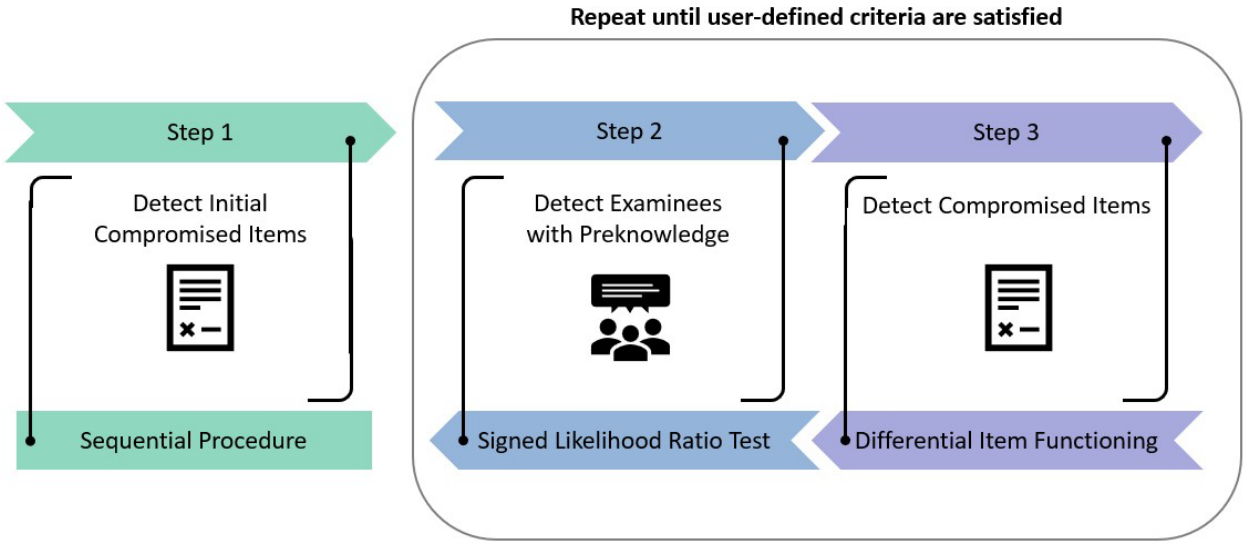
The proposed method follows the general framework of Gorney et al. (2023) for detecting CI and EWP. The framework consists of three steps and is illustrated in Figure 1.

1. **Initial item flagging:** Identify an initial set of items that are suspected of being compromised.
2. **Examinee flagging:** Use the set of compromised items to identify the set of examinees who are suspected of having preknowledge.
3. **Item flagging:** Use the set of examinees with preknowledge to refine the set of items that are suspected of being compromised.

The procedure is iterative in that Steps 2 and 3 are repeated until user-defined criteria are satisfied.

The procedure is terminated when one of the following conditions is satisfied: (1) Too few items are flagged, (2) too many items are flagged, or (3) the exact set of items has already been flagged in a previous iteration. Following Gorney et al. (2023), an item set is defined to be of "reasonable size" when it consists of at least 4 items. Thus, condition 1 is satisfied if fewer than 4

Figure 1. Proposed Method to Detect CI and EWP



items are flagged, and condition 2 is satisfied if more than $I - 4$ items are flagged, where I is the size of the item bank. Finally, note that if the procedure is terminated due to condition 1, then no solution is returned, meaning that no items or examinees are flagged. However, if the procedure is terminated due to condition 2 or 3, then the most recent set of flagging results is returned and reported as the solution.

Step 1: Initial Item Flagging

To identify the initial set of items that are suspected of being compromised, variations of the sequential procedures of Zhang (2014) and Zhang and Li (2016) are employed. Sequential procedures are appropriate for monitoring changes in the properties of an item over time—in this case, the question is whether an item appears to be becoming easier over time, which might occur if the item has been compromised.

The classical test theory (CTT)-based sequential procedure (Zhang, 2014) monitors the difficulty of an item over time by comparing the observed proportion of correct responses for a target sample of examinees to the observed proportion of correct responses for a reference sample of examinees for whom the item is assumed *not* to have been compromised. The target sample is taken as the most recent m examinees who were administered the item, while the reference sample is taken as the first $n - m$ examinees who were administered the item, where n is the total number of examinees to whom the item has been administered. Then, for item i, the test statistic

$$Z_i^{(n)} = \frac{p_i^{(n-m+1,n)} - p_i^{(1,n-m)}}{\sqrt{p_i^{(1,n-m)}[1 - p_i^{(1,n-m)}]}} \sqrt{\frac{m(n-m)}{n}} \tag{1}$$

is constructed, where

$$p_i^{(n-m+1,n)} = \frac{1}{m}\sum_{j=n-m+1}^{n} X_{ji} \tag{2}$$

is the proportion of correct responses in the target sample,

$$p_i^{(1,n-m)} = \frac{1}{n-m}\sum_{j=1}^{n-m} X_{ji} \tag{3}$$

is the proportion of correct responses in the reference sample, and $X_{ji} \in \{0,1\}$ is the score of examinee j on item i.

The item response theory (IRT)-based sequential procedure (Zhang & Li, 2016) is similar to the CTT-based sequential procedure in that it monitors the difficulty of an item over time. However, for the IRT-based sequential procedure, difficulty is monitored by comparing the observed number of correct responses to the expected number of correct responses for a target sample of examinees, where the expected number is estimated using an IRT model. Then, for item i, the test statistic

$$Y_i^{(n)} = \frac{\sum_{j=n-m+1}^{n}[X_{ji} - p_i(\hat{\theta}_j)]}{\sqrt{\sum_{j=n-m+1}^{n} p_i(\hat{\theta}_j)\,[1 - p_i(\hat{\theta}_j)]}} \tag{4}$$

is constructed, where $p_i(\hat{\theta}_j)$ is the estimated probability of a correct response for examinee j on item i given the examinee ability estimate $\hat{\theta}_j$. For example, under the Rasch model, the probability of a correct response is assumed to be

$$p_i(\theta_j) = P(X_{ji} = 1|\theta_j) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}, \tag{5}$$

where b_i is the difficulty parameter of item i and θ_j is the ability parameter of examinee j. Thus, $p_i(\hat{\theta}_j)$ is obtained by inserting $\hat{\theta}_j$ into Equation 5.

The $Z_i^{(n)}$ and $Y_i^{(n)}$ statistics are appropriate for testing whether an item is easier than expected at a specific time point; however, in a CAT program, items are often administered to examinees continuously and at different time points. Therefore, to monitor the difficulty of an item over time, the $Z_i^{(n)}$ and $Y_i^{(n)}$ statistics can be computed at multiple time points. Large positive values of $Z_i^{(n)}$ and $Y_i^{(n)}$ indicate that the item was easier than expected and might therefore have been compromised.

When applying the CTT- or IRT-based sequential procedures, it is necessary to define the critical values that determine what constitute "large positive values" of $Z_i^{(n)}$ and $Y_i^{(n)}$, respectively.

Zhang (2014) and Zhang and Li (2016) compared the use of several predetermined critical values, and in doing so applied the same critical value to every item. In contrast, in the proposed method, monte-carlo simulations are used to obtain separate critical values for each item (e.g., Kang, 2023). Specifically, for a given item bank, a null dataset (i.e., a dataset with no unusual behavior) is simulated, and the $Z_i^{(n)}$ and $Y_i^{(n)}$ statistics are computed using a target sample size of $m = 50$ and an initial monitoring point of $n_0 = 60$, meaning that each item starts being monitored once it has been administered to 60 examinees. Then, for each item, the maximum values of $Z_i^{(n)}$ and $Y_i^{(n)}$ across all time points are recorded. This process of simulating and analyzing null data is repeated 1,000 times, and the 95th percentiles of the resulting distributions are taken as the empirical critical values.

Step 2: Examinee Flagging

To identify the set of examinees who are suspected of having preknowledge, the signed likelihood ratio test statistic proposed by Sinharay (2017) is employed. This statistic has attractive theoretical properties (e.g., its asymptotic null distribution is known) and has been shown to perform well against competing statistics (e.g., Sinharay, 2017). It is appropriate for testing the equality of examinee ability over two sets of items—in this case, the set of compromised items (C) and the set of secure (or non-compromised) items (S).

Let $\hat{\theta}_j^{(C)}$, $\hat{\theta}_j^{(S)}$, and $\hat{\theta}_j$ denote the ability estimates of examinee j based on C, S, and all items, respectively. The signed likelihood ratio test statistic for testing $H_0: \theta_j^{(C)} = \theta_j^{(S)}$ against $H_1: \theta_j^{(C)} > \theta_j^{(S)}$ is given by

$$L_j = sgn(\hat{\theta}_j^{(C)} - \hat{\theta}_j^{(S)}) \sqrt{2[\ell(\hat{\theta}_j^{(C)}, \hat{\theta}_j^{(S)}) - \ell(\hat{\theta}_j)]}\,, \tag{6}$$

where $\ell(\hat{\theta}_j^{(C)}, \hat{\theta}_j^{(S)})$ is the log-likelihood of the item scores at $\hat{\theta}_j^{(C)}$ and $\hat{\theta}_j^{(S)}$, and $\ell(\hat{\theta}_j)$ is the log-likelihood of the item scores at $\hat{\theta}_j$.

Under the null hypothesis of no preknowledge, the L_j statistic has an asymptotic $\mathrm{N}(0,1)$ distribution. A large positive value of L_j indicates that the examinee performed better on the set of compromised items than on the set of secure items and might therefore have benefitted from item preknowledge.

Step 3: Item Flagging

To refine the set of items that are suspected of being compromised, logistic regression is employed. Logistic regression is commonly used to detect differential item functioning (DIF) across two groups of examinees (e.g., Swaminathan & Rogers, 1990; Zumbo, 1999)—in this case, the EWP (i.e., the focal group) and the non-EWP (i.e., the reference group). Unlike some other DIF detection procedures, logistic regression does not require matching examinees based on total score (which is inappropriate in CAT since different examinees are administered different items), nor does it require very large sample sizes.

The logistic regression model used in this paper[1] can be expressed as

$$\ln\left(\frac{p_{ji}}{1-p_{ji}}\right) = \beta_0 + \beta_1\theta_j + \beta_2 g_j\,, \tag{7}$$

where p_{ji} is the probability of examinee j responding to item i correctly given θ_j and g_j, and g_j represents group membership and is equal to 1 if the examinee is suspected of having pre-knowledge and 0 otherwise. An item is said to exhibit DIF if examinees who have the same ability, but belong to different groups, have different probabilities of answering the item correctly. To test for DIF, the compact model is first specified as the logistic regression model with only β_0 and β_1 and the augmented model is specified as the logistic regression model with β_0, β_1, and β_2. An item is statistically significant if the product of -2 and the difference between the log-likelihoods of the compact and augmented models exceeds the critical value corresponding to a χ^2 distribution with one degree of freedom. Jodoin and Gierl (2001) further provided an effect size measure, $R^2\Delta$, for assessing practical significance. They also developed the following criteria:

1. *Negligible DIF* if the item is *not* statistically significant or $R^2\Delta < 0.035$.
2. *Moderate DIF* if the item is statistically significant and $0.035 \leq R^2\Delta < 0.070$.
3. *Large DIF* if the item is statistically significant and $R^2\Delta \geq 0.070$.

In the proposed method, a significance level of $\alpha = 0.05$ is used, and moderate and/or large DIF items are flagged as potentially compromised.

Unique Features of the Proposed Method

Although the proposed method is similar to that of Gorney et al. (2023), there are important differences in the ways the two methods are implemented. First, the method of Gorney et al. involves the analysis of item scores and distractors and is thus only suitable for certain item types. In contrast, the present method requires only the analysis of item scores and is thus suitable for a wider variety of item types. Second, the statistics used in Step 1 of the method of Gorney et al. are only computed at a single time point and are thus not intended to monitor changes over time. In contrast, the statistics used in Step 1 of the present method are computed at multiple time points, which might be useful in situations where the items are administered continuously. Finally, the statistics used in Step 3 of the method of Gorney et al. require matching examinees based on total score. In contrast, the statistics used in Step 3 of the present method do not require such matching and are thus more appropriate for CAT data.

Study 1: Comparing Settings

Design and Analysis

To examine the performance of the proposed method, two simulation studies were conducted. The purpose of the first study was to compare the performance of the proposed method when different settings were used in each of the three steps. Conditions were created by manipulating

[1]It is possible to add an interaction term to the logistic regression model in Equation 7 to test for non-uniform DIF. However, in the proposed method, the simplifying assumption is made that preknowledge tends to benefit—rather than harm—examinees.

the following factors: test length (20, 40, 80), sequential procedure used in Step 1 (IRT, CTT), examinee significance level used in Step 2 (0.010, 0.025, 0.050), and type of DIF flagged in Step 3 (moderate/large, large). The four factors were fully crossed, resulting in a total of 36 (3 × 2 × 3 × 2) conditions that were then replicated 100 times.

Item scores were generated using the Rasch model. The item difficulty parameters were sampled such that $b_i \sim \mathrm{N}(0,1)$ and the examinee ability parameters were sampled such that $\theta_j \sim \mathrm{N}(0,1)$. The size of the item bank was 10 times the length of the test. For example, for a test length of 40 items, the item bank was comprised of 400 items.

For each replication, 5,000 examinees were simulated, where 10% of the examinees had preknowledge of 20% of the items in the item bank—the extent to which these percentages affect the results is investigated in Study 2. The EWP were selected with probabilities proportional to the order in which they were administered the test. Thus, the first examinee to take the test had the smallest probability of having preknowledge, while the last examinee to take the test had the largest probability of having preknowledge. The CI were randomly selected from all items in the item bank and were assumed to be the same for all EWP. When an EWP was administered a CI, the probability of a correct response was 0.9, as in, for example, Gorney et al. (2023), Pan et al. (2022), and Pan and Wollack (2023).

For each CAT administration, the initial item was randomly selected as one of 10 maximally informative items at $\theta = 0$. Subsequent items were selected using the maximum Fisher information criterion at the interim θ estimate, subject to a maximum item exposure rate of 0.20. Examinee ability parameters were estimated using maximum likelihood estimation, where estimates were bounded between –4 and 4. Thus, examinees with all correct scores received a θ estimate of 4, while examinees with all incorrect scores received a θ estimate of –4.

To evaluate the performance of the proposed method, the false positive rate (FPR) and true positive rate (TPR) were computed for both items and examinees. For items, the FPR represents the proportion of secure items that were incorrectly flagged as compromised, while the TPR represents the proportion of CI that were correctly flagged as such. For examinees, the FPR represents the proportion of non-EWP who were incorrectly flagged as having preknowledge, while the TPR represents the proportion of EWP who were correctly flagged as such.

Results

Tables 1 and 2 display flagging rates for the items and examinees, respectively. Each row corresponds to a different combination of sequential procedure, type of DIF, and examinee significance level. Each column corresponds to a different combination of test length and outcome measure.

Table 1 reveals that the FPRs for the items were consistently smaller than the item significance level that was used in Step 3 ($\alpha = .05$). This result is not entirely surprising, given that an item needed to display both statistical and practical significance in order to be flagged as compromised. Table 2 reveals that the FPRs for the examinees tended to be smaller, but were occasionally larger, than the examinee significance level that was used in Step 2. For example, when the examinee significance level was $\alpha = .05$, moderate/large DIF was flagged, and the test length was 40 or 80 items, the FPR was consistently larger than .05. One possible explanation for this result is as follows. Consider that the use of a larger examinee significance level typically produces a flagged sample of examinees with a smaller EWP to non-EWP ratio. Similarly, the flagging of items that

exhibit moderate/large DIF (rather than just large DIF) typically produces a flagged sample of items with a smaller CI to secure items ratio. In such samples, the preknowledge signal is obscured and is therefore more difficult to detect. Thus, it is not surprising that these conditions tended to produce suboptimal results.

Table 1. Item Flagging Rates for Study 1

			20 Items		40 Items		80 Items	
Seq. Proc.	DIF	Examinee Sig. Level	FPR	TPR	FPR	TPR	FPR	TPR
IRT	Moderate/Large	.010	.001	.100	.001	.199	.000	.396
		.025	.011	.210	.012	.318	.006	.478
		.050	.040	.295	.029	.341	.014	.446
	Large	.010	.000	.038	.000	.071	.000	.094
		.025	.000	.083	.000	.113	.000	.137
		.050	.012	.189	.013	.236	.002	.190
CTT	Moderate/Large	.010	.000	.002	.000	.017	.000	.021
		.025	.002	.011	.008	.060	.011	.078
		.050	.047	.096	.045	.109	.035	.099
	Large	.010	.000	.000	.000	.002	.000	.001
		.025	.000	.000	.000	.006	.000	.003
		.050	.008	.028	.019	.056	.005	.023

Tables 1 and 2 further reveal that the TPRs for both items and examinees were larger when the test was longer, the IRT-based sequential procedure was used in Step 1, a larger examinee significance level was used in Step 2, and moderate/large DIF was flagged in Step 3. The superior performance of the IRT-based sequential procedure is not surprising, given that (1) the null data were simulated using an IRT model and (2) the use of the CTT-based sequential procedure involves the assumption that the data of the reference group are uncontaminated by preknowledge. In reality—as in the simulations—it is possible for this assumption to be violated.

The larger TPRs associated with larger examinee significance levels and moderate/large DIF are also not surprising, given that these settings apply less stringent criteria when flagging examinees and items. However, it is important to balance the reward of a large TPR with the risk of an increased FPR. Therefore, in the next set of simulations, only the IRT-based sequential procedure is used with an examinee significance level of $\alpha = .025$ and moderate/large DIF, as this combination of settings was found to produce the largest TPRs while also maintaining reasonable FPRs.

Table 2. Examinee Flagging Rates for Study 1

			20 Items		40 Items		80 Items	
Seq. Proc.	DIF	Examinee Sig. Level	FPR	TPR	FPR	TPR	FPR	TPR
IRT	Moderate/Large	.010	.001	.101	.003	.263	.006	.572
		.025	.009	.202	.020	.403	.028	.706
		.050	.038	.294	.062	.485	.066	.735
	Large	.010	.000	.057	.001	.147	.002	.255
		.025	.003	.135	.006	.252	.010	.392
		.050	.026	.260	.051	.436	.043	.549
CTT	Moderate/Large	.010	.000	.001	.000	.027	.000	.033
		.025	.001	.011	.007	.082	.016	.125
		.050	.028	.068	.052	.141	.071	.193
	Large	.010	.000	.000	.000	.006	.000	.005
		.025	.000	.000	.000	.016	.000	.012
		.050	.009	.035	.033	.106	.020	.083

Study 2: Preknowledge Characteristics

Design and Analysis

The purpose of the second study was to apply the best combination of settings from Study 1 to examine their performance for different types of simulated preknowledge. Thus, the IRT-based sequential procedure was used in Step 1, the examinee significance level of $\alpha = .025$ was used in Step 2, and moderate/large DIF was flagged in Step 3. Conditions were created by manipulating the following factors: test length (20, 40, 80), percentage of CI (10, 20, 40), percentage of EWP (10, 30), and EWP ability distribution (same as non-EWP, lower than non-EWP). The four factors were fully crossed, resulting in 36 (3 × 3 × 2 × 2) preknowledge conditions. A null condition in which no preknowledge was simulated was also included for each of the three test lengths. Thus, 39 conditions were studied in total, each of which was replicated 100 times.

The data generating process was nearly identical to that of Study 1. The only differences were as follows:

1. The percentage of CI and the percentage of EWP varied depending on the condition.
2. When the EWP ability distribution was simulated to be the same as that of the non-EWP, the EWP were selected exactly as in Study 1. However, when the EWP ability distribution was simulated to be lower than that of the non-EWP, the EWP were selected from those examinees with $\theta \leq 0$ with probabilities proportional to the order in which they took the test. Thus, the first examinee to take the test with $\theta \leq 0$ had the smallest non-zero probability of having preknowledge, the last examinee to take the test with $\theta \leq 0$ had the largest

probability of having preknowledge, and the examinees with $\theta > 0$ had zero probability of having preknowledge.

Results

Tables 3 and 4 display flagging rates for the items and examinees, respectively. Each row corresponds to a different combination of EWP ability distribution, percentage of CI, and percentage of EWP. Each column corresponds to a different combination of test length and outcome measure.

Table 3 reveals that, as in Study 1, the FPRs for the items were consistently smaller than the item significance level that was used in Step 3 ($\alpha = .05$). This result is encouraging, as it shows that the method successfully limited the proportion of secure items that were incorrectly flagged as compromised. Table 4 reveals that the FPRs for the examinees tended to be smaller, but were sometimes larger, than the examinee significance level that was used in Step 2 ($\alpha = .025$). Larger FPRs were associated with longer tests and larger percentages of EWP.

Table 3. **Item Flagging Rates for Study 2**

			20 Items		40 Items		80 Items	
Condition	% CI	% EWP	FPR	TPR	FPR	TPR	FPR	TPR
Null	0	0	.003	–	.012	–	.022	–
EWP same ability	10	10	.009	.127	.020	.196	.015	.229
		30	.038	.804	.027	.924	.005	.986
	20	10	.011	.210	.012	.318	.006	.478
		30	.043	.834	.011	.944	.002	.988
	40	10	.005	.253	.004	.323	.004	.366
		30	.031	.728	.005	.834	.002	.819
EWP lower ability	10	10	.012	.166	.025	.314	.010	.380
		30	.036	.874	.021	.857	.006	.800
	20	10	.015	.259	.016	.455	.002	.688
		30	.035	.970	.009	.972	.002	.944
	40	10	.009	.375	.003	.526	.001	.596
		30	.032	.954	.008	.981	.003	.958

Tables 3 and 4 further reveal that the TPRs for both items and examinees were larger when the test was longer, the percentages of CI and EWP were larger, and the EWP ability distribution was lower than that of the non-EWP. These results are not surprising, because (1) Increasing the percentages of CI and EWP produces a stronger preknowledge signal that is therefore easier to detect, (2) simulating preknowledge exclusively in lower ability examinees also produces a stronger preknowledge signal since these examinees should have the most to gain, and (3) similar

findings have been reported in previous research on both CATs (e.g., Pan et al., 2022) and non-adaptive tests (e.g., Gorney et al., 2023).

Table 4. Examinee Flagging Rates for Study 2

			20 Items		40 Items		80 Items	
Condition	% CI	% EWP	FPR	TPR	FPR	TPR	FPR	TPR
Null	0	0	.001	–	.008	–	.027	–
EWP same ability	10	10	.006	.065	.019	.117	.032	.242
		30	.021	.198	.035	.395	.030	.720
	20	10	.009	.202	.020	.403	.028	.706
		30	.029	.411	.029	.728	.026	.942
	40	10	.009	.512	.014	.740	.020	.860
		30	.021	.658	.021	.903	.023	.978
EWP lower ability	10	10	.007	.068	.026	.180	.032	.448
		30	.020	.260	.033	.441	.031	.721
	20	10	.013	.174	.031	.431	.031	.880
		30	.031	.522	.029	.781	.028	.960
	40	10	.014	.533	.019	.773	.023	.923
		30	.027	.796	.025	.965	.025	.999

Discussion

In recent years, several methods have been proposed to detect CI and EWP simultaneously. However, most of these methods are limited in one of two ways: (1) The method was specifically designed for non-adaptive tests and might therefore not be suitable for adaptive tests or (2) the method involves the analysis of more than just the item scores. In this paper, both limitations of previous research were addressed, and a new method was developed specifically for CATs that requires only an analysis of the item scores.

To examine the performance of the proposed method, two simulation studies were conducted. In Study 1, performance was examined when different settings were used in each of the three steps. Results showed that for most settings, the method produced small and reasonable FPRs for both items and examinees. In addition, the TPRs were largest when the IRT-based sequential procedure was used in Step 1, a larger examinee significance level was used in Step 2, and moderate/large DIF was flagged in Step 3.

In Study 2, the best combination of settings from Study 1 was taken and applied to different types of simulated preknowledge. Results again showed that the method tended to produce small and reasonable FPRs for both items and examinees. Furthermore, the TPRs were largest when the test was longer, the percentages of CI and EWP were larger, and the EWP ability distribution was lower than that of the non-EWP. Each of these findings paralleled those found in previous research on both CATs (e.g., Pan et al., 2022) and non-adaptive tests (e.g., Gorney et al., 2023).

There are several limitations to this research, providing many opportunities for future study. First, although the simulation results suggest that the new method is promising, it is important to study its performance using real CAT data, especially data for which the CI and EWP are known. Unfortunately, as noted by Gorney et al. (2024), such data are typically difficult—if not impossible—to obtain. Therefore, it might be sufficient to apply the method to a dataset for which some information is known regarding the CI and EWP prior to starting the analysis.

Second, although the simulation studies were detailed, it is possible to study additional simulation conditions, including those that involve the use of different item selection methods, different item exposure control methods, content balancing, and different ability estimation methods. It is also possible to perform a more detailed analysis of the simulation results. For example, although aggregate flagging rates were reported, it would be interesting to study the flagging rates of easy vs. difficult items or low- vs. high-ability examinees. It would also be interesting to study the flagging rates of examinees who were administered different numbers of CI.

Third, it would be useful to compare the performance of the proposed method to that of existing methods for detecting CI and EWP simultaneously in CAT. Two examples of such methods are the information theory and combinatorial optimization approach of Belov (2014) and the machine learning-based approach of Pan et al. (2022).

Finally, even though different settings were compared in each of the three steps, it is possible to use other settings or statistics. For example, recent researchers have shown that the use of item distractors (Gorney et al., 2023) and item response times (Choe et al., 2018; Sinharay & Johnson, 2020; van der Linden & Belov, 2023) can lead to improved detection results for both CI and EWP. It seems reasonable to believe that these additional sources of information would lead to improved detection results in the present context, as well, provided that such information is trustworthy and available.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (5th ed.). American Educational Research Association.

Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, *2*(3), 37–58. DOI

Boughton, K. A., Smith, J., & Ren, H. (2017). Using response time data to detect compromised items and/or people. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 177–190). Routledge.

Chen, Y., Lu, Y., & Moustaki, I. (2022). Detection of two-way outliers in multivariate data and application to cheating detection in educational tests. *Annals of Applied Statistics*, *16*(3), 1718–1746. DOI

Choe, E. M., Zhang, J., & Chang, H.-H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. *Psychometrika*, *83*(3), 650–673. DOI

Gorney, K., Chen, J., & Bay, L. (2024). The impact of generating model on preknowledge detection in CAT. In M. Wiberg, J.-S. Kim, H. Hwang, H. Wu, & T. Sweet (Eds.), *Quantitative psychology: The 88th annual meeting of the Psychometric Society, Maryland, USA, 2023* (pp. 373–381). Springer. DOI

Gorney, K., Wollack, J. A., Sinharay, S., & Eckerly, C. (2023). Using item scores and distractors to detect item compromise and preknowledge. *Journal of Educational and Behavioral Statistics*, *48*(5), 636–660. DOI

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. DOI

Kang, H.-A. (2023). Sequential generalized likelihood ratio tests for online item monitoring. *Psychometrika*, *88*(2), 672–696. DOI

Man, K., & Harring, J. R. (2023). Detecting preknowledge cheating via innovative measures: A mixture hierarchical model for jointly modeling item responses, response times, and visual fixation counts. *Educational and Psychological Measurement*, *83*(5), 1059–1080. DOI

O'Leary, L. S., & Smith, R. W. (2017). Detecting candidate preknowledge and compromised content using differential person and item functioning. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 151–163). Routledge.

Pan, Y., Sinharay, S., Livne, O., & Wollack, J. A. (2022). A machine learning approach for detecting item compromise and preknowledge in computerized adaptive testing. *Psychological Test and Assessment Modeling*, *64*(4), 385–424.

Pan, Y., & Wollack, J. A. (2023). A machine learning approach for the simultaneous detection of preknowledge in examinees and items when both are unknown. *Educational Measurement: Issues and Practice*, *42*(1), 76–98. DOI

Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*(1), 46–68. DOI

Sinharay, S., & Johnson, M. S. (2020). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 397–419. DOI

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. DOI

van der Linden, W. J., & Belov, D. I. (2023). A statistical test for the detection of item compromise combining responses and response times. *Journal of Educational Measurement*, *60*(2), 235–254. DOI

Zhang, J. (2014). A sequential procedure for detecting compromised items in the item bank of a CAT system. *Applied Psychological Measurement*, *38*(2), 87–104. DOI

Zhang, J., & Li, J. (2016). Monitoring items in real time to enhance CAT security. *Journal of Educational Measurement*, *53*(2), 131–151. DOI

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense.

Author Address

"Kylie Gorney" <kgorney@msu.edu>

Citation

Gorney, K., Lee, C., & Chen, J. (2025). A score-based method for detecting item compromise and preknowledge in computerized adaptive testing. *Journal of Computerized Adaptive Testing, 12(2),* 123–136. https://doi.org/10.7333/2506-1202123