ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 138-147

Robustness of Computer Adaptive Tests to the Presence of Item Preknowledge: A Simulation Study

Hakan KARA* Nuri DOĞAN** Başak ERDEM KARA***

Abstract

Item preknowledge describes a scenario where candidates may have access to some of the test items prior to the test administration. This involves sharing test materials and/or answers and it is difficult to identify the individuals with item preknowledge or the shared materials of the test. Nevertheless, it is essential to investigate the 'item preknowledge' problem because it can significantly affect the validity of the test results. It is believed that traditional linear tests are more robust to this type of aberrant response behavior than adaptive tests. In this context, the aim of this study is to examine the effect of item preknowledge on computer adaptive tests and identify the conditions under which adaptive tests are most resistant to the item preknowledge. For this purpose, a Monte Carlo simulation study was performed and 28 different conditions were examined. The results of the study indicated that the EAP estimation method provided better measurement precision than ML over all conditions. When 2PL and 3PL IRT models were compared, it was observed that 2PL had higher precision at most of the conditions. However, when the aberrancy ratio increased and reached 20% for both individuals and items, 3PL outperformed the 2PL model and gave the best results with the EAP combination. The results were discussed in line with the literature on item preknowledge and CAT and implications for practitioners and further research were provided.

Keywords: item preknowledge, aberrant responses, computer adaptive tests, test security

Introduction

Test scores from any assessment tool are used to obtain information about the proficiency level of the examinees. The main assumption in using these scores is that examinees' responses reflect their actual level of proficiency and are not influenced by factors other than the latent trait (Meijer, 1996; Wan & Keller, 2023). However, this assumption is often violated and some other factors such as lucky guessing, cheating, careless responding, creative responding and random responding (Meijer, 1996) are involved in the process. The mentioned undesirable factors may cause responses that are inconsistent with the respondent's ability level, and these unexpected responses are referred to as aberrant responses (Clark, 2010). Aberrant responses occur when the observed patterns of response do not align with the expected ones (Meijer, 1996; Meijer & Sijtsma, 2001) and they are commonly encountered in practical testing situations (Wan & Keller, 2023; Yen et al., 2012).

When aberrant responses are included in the testing process, the test score does not reflect the 'true' level of ability estimate (Magis, 2014). The validity of test scores may suffer from the inclusion of such responses, as they prevent test takers from demonstrating their accurate level of measured latent trait (Rios et al., 2017). Therefore, it is crucial to monitor test results to detect aberrant responses in order to reduce their negative impact on the validity of test scores (Tendeiro & Meijer, 2014; Wan & Keller, 2023).

* PhD Student., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: hakankaraodtu@gmail.com, ORCID ID: 0000-0002-2396-3462
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016
*** Assist. Prof. Dr., Anadolu University, Faculty of Education, Eskişehir-Turkey, e-mail: basakerdem@anadolu.edu.tr, ORCID ID: 0000-0003-3066-2892

To cite this article:
Kara, H., Doğan, N., & Erdem-Kara, B. (2024). Robustness of computer adaptive tests to the presence of item preknowledge: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology, 15*(2), 138-147. https://doi.org/10.21031/epod.1470949

Received: 19.04.2024
Accepted: 6.06.2024

Aberrant responses may arise from a variety of sources. Yen et al. (2012) classified examinee responses in a selected response test into three groups: (a) responses that reflect the examinee's true ability, (b) correct responses made by lucky guesses, and (c) incorrect responses due to anxiety, carelessness, or distraction. The latter two types of response behavior are aberrant response types because they differ from what is expected and do not reflect the examinee's actual knowledge. Meijer (1996) proposed that there are at least five different factors that can cause aberrant responses: lucky guessing, cheating, careless responding, creative responding and random responding. In this paper the focus is on cheating behavior, specifically the item preknowledge.

Theoretical Framework

Item preknowledge occurs when examinees may have access to test items prior to taking the exam (Eckerly, 2017). As Tendeiro and Meijer (2014) stated, cheating often enables an examinee to perform better than their actual ability and this could be a result of preknowledge of the items before the test. Belov (2016) stated that item preknowledge describes a scenario where a group of examinees (referred to as aberrant) have access to a subset of items (referred to as compromised items) prior to the administration of the test. Aberrant test takers exhibit improved performance on compromised items compared to non-compromised ones. As a result of item preknowledge, examinees unfairly get the right answers on test items that they would not normally get right. Thus, these items no longer effectively distinguish between examinees (Kim & Moses, 2016). As the percentage of compromised items and individuals with prior knowledge increases, the error in parameter estimates increases because the scores of aberrant examinees are invalid (Belov, 2016; Eckerly, 2017). Given the negative impact of item disclosure on test scores, item preknowledge should be investigated.

Item preknowledge could be defined as a special case of test collusion, which can be defined as the large-scale sharing of test materials or answers to questions. The shared information may come from different sources such as teachers, testing companies, the Internet or communication between examinees (Wollack & Maynes, 2017). It is hard to detect the aberrant examinees or items because there are multiple unknowns, such as the unknown group of cheating examinees accessing the unknown group of compromised items (Belov, 2016). However, it is essential to investigate since it affects the validity of the test results (Eckerly, 2017).

It is generally assumed that adaptive tests might suffer more from aberrant responses compared to traditional linear tests (Kim & Moses, 2016). Traditional linear tests are generally based on classical test theory (CTT), and the effect of aberrant responses on ability estimation in a traditional paper-pencil test might be little if items are equally weighted (Yen et al., 2012). However, item response theory models (IRT) are highly sensitive to these kinds of changes in response patterns (Magis, 2014). IRT models often struggle to accurately calculate true individual response probabilities due to various factors like guessing and cheating, leading to the presence of response disturbances, and IRT can return a strongly biased ability estimation when aberrant responses occur (Jia et al., 2019). As computer adaptive testing (CAT) applications are generally based on IRT models, they become more vulnerable to the biased estimation and measurement errors that aberrant responses may cause (Yen et al., 2012; Zheng & Chang, 2014). CAT is designed to select and administer items in accordance with test takers' proficiency level during the testing process. Ability estimation, whereas IRT models are used, is performed continuously after the administration of each item and the next item is selected based on the estimated ability (Yan et al., 2014; Zheng & Chang, 2014). Therefore, aberrant responses might not only cause the ability estimation error but also affect the item selection (Yen et al., 2012).

In a specific manner, CATs can be administered to small groups of test takers at different times, frequently and consecutively. This approach is referred to as continuous testing and offers flexibility in test scheduling. However, as with other forms of continuous testing, it can raise concerns about test security. Examinees who have taken the test earlier could share information about the test with those who have taken it later, and many items are at risk of disclosure prior to the test. Students may memorize and share test information with others, and this may artificially inflate the scores of those who have obtained advanced knowledge of the material, therefore posing a threat to its validity (Zhang et al., 2012). CAT applications are, as a consequence, open to the threat of item preknowledge. Similar to its paper-pencil counterpart, CAT may award a higher score to a test taker as a result of his/her prior knowledge of the answers to compromised items. Unlike traditional paper-pencil tests, CATs customize each test for each individual examinee. If certain compromised items are answered correctly due to pre-existing knowledge of the answers, the CAT algorithm can recover the true ability through subsequent item selection based on factors such as the location and number of compromised items (Guo, 2009). CAT applications differ in several aspects, including item bank characteristics, item selection and stopping algorithms, exposure control and IRT model. These aspects influence the way in which compromised items affect test performance. Accordingly, it is important to see the performance of CAT applications under different conditions when item preknowledge exists.

The presence of compromised items is problematic for the reasons mentioned before. Several studies have been conducted on the performance of several detection methods of aberrant responses caused by item preknowledge in CAT environments (e.g., Belov, 2014; Liu, et al., 2019; McLeod et al., 2003; Pan et al., 2022; Qian et al., 2016). However, there is no single best way to detect item preknowledge and it is difficult to detect (Belov, 2014). Therefore, it is also important to understand the conditions that are more or less robust to the presence of item preknowledge. Several studies have been conducted on the impact of compromised items for different purposes. Yi et al. (2008) investigated compromised items under the‘item theft’ context in the CAT environment. They investigated the potential damage that item theft can cause on CAT under two item selection algorithms (maximum item information and a-stratified methods). The findings suggested that although ‘item theft’ could result in significant harm to CAT using either item selection approach, the maximum item information method was more susceptible to organized item theft simulation than the a-stratified method. In another study, Guo et al. (2009) investigated the resistance of CAT to small-scale cheating under different item selection methods and compared the results with a traditional paper-pencil test. They indicated that CAT is better at giving resistant results than conventional tests at the presence of small-scale cheating. Lengthier tests and more test forms provided much more secure conditions for conventional tests. In addition, six-item selection methods were compared and ‘*a*-stratified with *b* blocking’ (ASBB) and maximum information (MI) methods had better resistance to small-scale cheating under 30 item test length but they gave similar results to the other four methods in the 60 item test. Lastly, Zhang et al. (2012) investigated the phenomenon of ‘item preknowledge’ under the name of ‘item sharing’ context and compared the use of single and multiple item pools under different item selection and exposure control methods. This study suggested that two-pool design provided a higher resistance to item sharing compared to single-pool designs resulting in greater precision in measuring ability using the Maximum Item Information Method with Sympson-Hetter item exposure control method. Although the mentioned studies demonstrated the serious and negative effects of the presence of compromised items, they were limited in some respects. The current study aimed to approach the problem from an expanded perspective by including the ability estimation method, the IRT model, the percentage of aberrant items and the percentage of aberrant individuals. Therefore, it is thought that the results of the study will contribute to the literature related to item preknowledge on CAT.

The Purpose of the Study

For the reasons discussed above, it is important to understand the possible effects of the presence of item preknowledge, how CAT applications were affected and how the resistance of CAT changes under different conditions. Thus, the present study aims to investigate the performance of CAT in the presence of item preknowledge and to examine the conditions under which it is most resistant to prior knowledge. In this context, the following research question were addressed.

- How does the test performance of the CAT change in case of the presence of item preknowledge under different ability estimation methods (Maximum Likelihood (ML) and Expected a Priori (EAP)), IRT model (2 PL and 3 PL), percentage of aberrant items (10%, 20% and 30%) and percentage of aberrant individuals (10% and 20%)?

Methods

Within the scope of the research, it is aimed to investigate the effect of the inclusion of compromised items on the effectiveness of different CAT conditions. The data used in the research were generated by the simulation method using Monte Carlo approach and 28 different conditions were compared in a controlled manner. Simulation data was preferred because it is difficult to meet all the conditions discussed in the study simultaneously in real data.

Design Overview

To demonstrate the effect of item preknowledge on CAT, normal response (no aberrancy) and item preknowledge conditions were simulated under different conditions (Ability estimation method; MLE and EAP – Percentage of aberrant items; 10%, 20% and 30% - Percentage of aberrant persons; 10% and 20% – IRT model; 2PL and 3PL). The test length was fixed at 30 items for each condition. All manipulated conditions were fully crossed with each other. There were a total of 24 conditions (3 aberrant item ratio x 2 aberrant person ratio x 2 ability estimation method x 2 IRT model) resulting from those manipulated conditions. In addition, response data of no aberrancy were generated for two ability estimation methods and two IRT models, resulting in four extra conditions. For each condition, 20 replications were executed. All procedures were carried out in R Statistical Software (v4.1.2; R Core Team, 2021)

Data Generation

To see the effect of item preknowledge on CAT performance, we simulated normal responses and responses with item preknowledge for a 30-item test. Two item pools of 300 items were generated using 2PL and 3PL. Item difficulty parameters were generated based on a standard normal distribution N (0, 1) and item discrimination parameters were sampled from log-normal distribution L (0, 0.25). In addition to these, the c parameters were set at .20 (indicating a guessing parameter for a five-option multiple-choice test) for the 3 PL item pool. The ability parameters of 1000 examinees were randomly generated based on standard normal distribution N (0, 1). After the generation of ability parameters and item pools, normal response patterns of 1000 examinees on 300 items were generated as a base condition and CAT simulations were performed on that dataset (catR package; Magis et al., 2022).

For each condition, the ability level for the starting rule was set to '0' and the Maximum Fisher Information (MFI) method was used as the item selection method. MFI is one of the most commonly used methods for item selection in computer adaptive testing and was preferred because it selects the item that provides the maximum information each time tests (Wang, 2017; Weiss & Kingsbury, 1984). In order to avoid the same item being taken by each individual, the randomesque method was used with a five-item group. A value of 0.40 was used for item exposure.

While generating responses for item preknowledge behavior:

1. Firstly, items with the highest item exposure values were taken from the CAT simulation conducted under normal response conditions and these items were coded as compromised items.
2. The dataset was then updated for aberrant responders and items. Examinees with item preknowledge were randomly selected from individuals with low ability levels (th<0), and compromised items were randomly selected from the items identified in the previous stage.
3. Responses of specified individuals on those specified items were simulated based on the Bernoulli random variable with a success probability of .90.
4. The dataset generated in the first step was replaced with the newly generated aberrant dataset for aberrant individuals and items.

Evaluation Criteria

In order to assess the impact of item preknowledge on CAT applications, RMSE, average bias, mean absolute error (MAE) and correlation values were calculated for each replication. Values were then averaged over 20 replications.

Root mean square error (RMSE) was calculated with the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_\iota} - \theta_i)^2}{N}} \tag{1}$$

Bias indicates the mean difference between individuals' true and estimated ability level and was calculated by using the following formula:

$$Bias = \frac{\sum_{i=1}^{n}(\widehat{\theta_\iota} - \theta_i)}{N} \tag{2}$$

Mean absolute error (MAE) represents the mean average difference between individuals' estimated and actual ability level and was calculated with the following formula:

$$MAE = \frac{\sum_{i=1}^{n}\left|\widehat{\theta_\iota} - \theta_i\right|}{N} \tag{3}$$

Lastly, correlation value was obtained by the following formula:

$$\rho_{\widehat{\theta_J},\theta_j} = \frac{cov(\widehat{\theta_J},\theta_j)}{\sigma_{\widehat{\theta_J}}\sigma_{\theta_j}} \tag{4}$$

$\widehat{\theta}_j$ represents the estimated ability parameter, θ_j represents the true ability parameter, and N represents the total number of individuals. Besides, ($\sigma_{\widehat{\theta}_j}$) and ($\sigma_{\theta_j}$) stand for the standard error values of the estimated and true ability parameters, respectively.

Results

In the current study, the performance of CAT is investigated under different ability estimation methods, aberrant item ratio and aberrant person ratio. RMSE, bias, correlation and mean absolute error (MAE) values across all conditions are provided in Table 1. In addition, these values were visualized and presented in Figure 1. Results were interpreted with the help of both Table 1 and Figure 1.

According to Table 1 and Figure 1, the outcomes had the lowest RMSE, bias, MAE and highest correlation at the base (normal) condition regardless of estimation methods for both 2PL and 3PL models. The inclusion of compromised items reduced the measurement precision of the test, as expected, in both MLE and EAP conditions and 2PL and 3PL models. Comparing MLE and EAP estimation methods, EAP demonstrated the highest measurement precision (lowest RMSE, MAE and highest correlation values) across all conditions. In addition, the correlation between true and estimated ability values was high (>.871) across all conditions. However, it was higher for normal conditions compared to aberrant response conditions decreasing with the increasing percentage of aberrant items and persons. As the percentage of individuals with item preknowledge increased, the RMSE, bias, and MAE values increased and correlation decreased for both MLE and EAP estimation methods. Hence, increasing the percentage of aberrant responders resulted in a decline in measurement precision. The same situation held for the increment of aberrant item percentage as well.

Table 1.

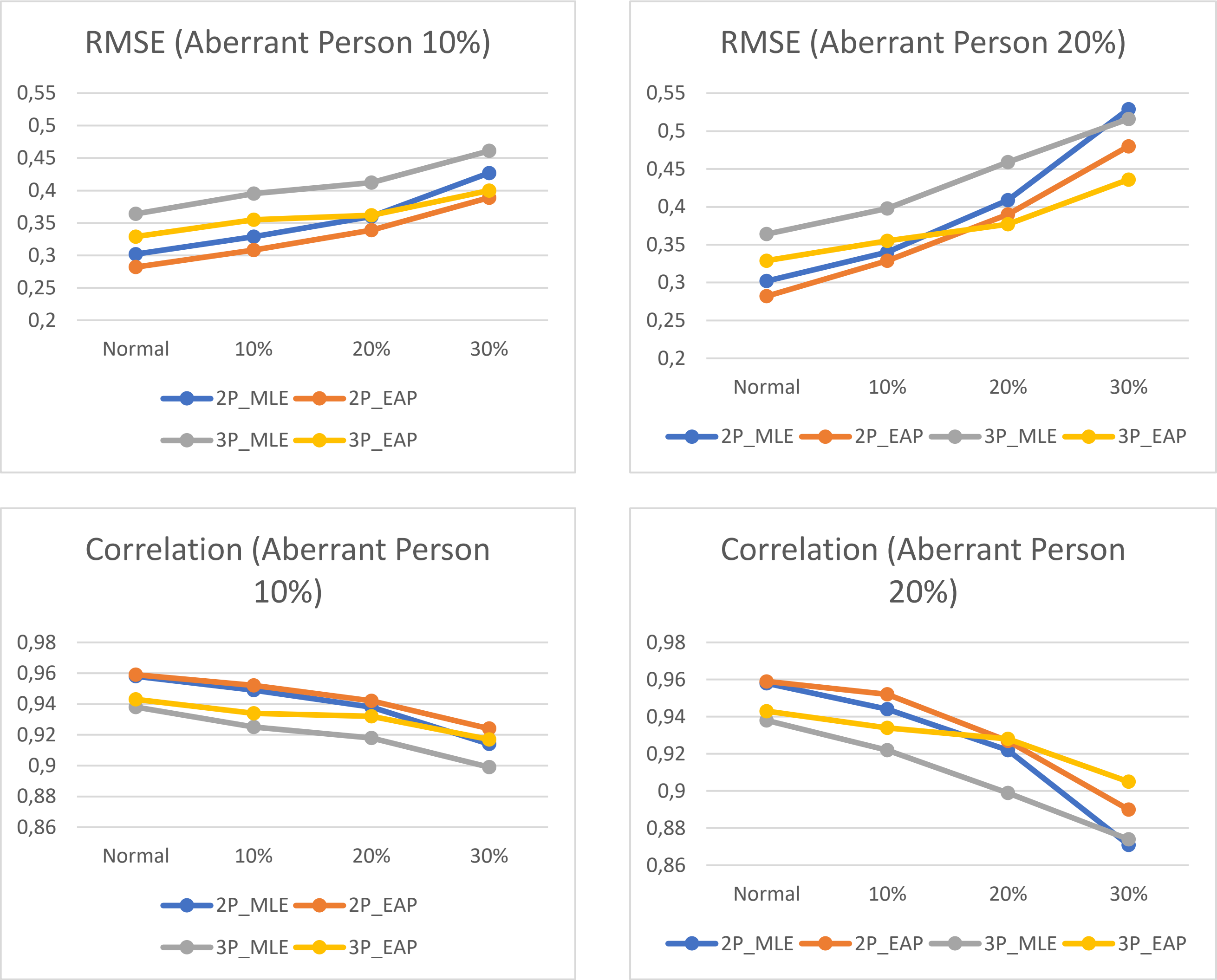
RMSE, Bias, Correlation and MAE Values under Different Conditions

	Ability Estimation	Aberrant Person	Aberrant Item	RMSE	Bias	Correlation	MAE
2PLM	MLE	10% person	Normal	0.302	-0.001	0.958	0.239
			10%	0.329	-0.032	0.949	0.260
			20%	0.360	-0.059	0.938	0.274
			30%	0.427	-0.090	0.914	0.306
		20% person	10%	0.340	-0.060	0.944	0.268
			20%	0.409	-0.118	0.922	0.312
			30%	0.529	-0.185	0.871	0.373
	EAP	10% person	Normal	0.282	-0.003	0.959	0.223
			10%	0.308	-0.026	0.952	0.244
			20%	0.339	-0.053	0.942	0.258
			30%	0.389	-0.073	0.924	0.280
		20% person	10%	0.329	-0.053	0.952	0.259
			20%	0.390	-0.101	0.927	0.296
			30%	0.480	-0.153	0.890	0.343
3 PLM	MLE	10% person	Normal	0.364	-0.024	0.938	0.288
			10%	0.395	-0.018	0.925	0.306
			20%	0.412	-0.052	0.918	0.314
			30%	0.461	-0.096	0.899	0.339
		20% person	10%	0.398	-0.049	0.922	0.309
			20%	0.459	-0.117	0.899	0.340
			30%	0.516	-0.151	0.874	0.379
	EAP	10% person	Normal	0.329	-0.003	0.943	0.259
			10%	0.355	-0.004	0.934	0.283
			20%	0.362	-0.0277	0.932	0.288
			30%	0.400	-0.053	0.917	0.302
		20% person	10%	0.355	-0.014	0.934	0.280
			20%	0.377	-0.056	0.928	0.297
			30%	0.436	-0.102	0.905	0.329

Besides, 2PL model had a higher correlation and lower RMSE and MAE values for most of the conditions compared to 3PL model. 2PL with EAP estimation was the best combination at all aberrant items for 10% aberrant person and 2PL with ML combination was better than 3PL mostly. However, 3PL_EAP combination outperformed 2PL_MLE only for the 30% of aberrant item condition, whereas 2PL performed better in all other conditions (Figure 1). The result obtained can be interpreted as that EAP is more resistant to the increase of the percentage of preknowledge items than MLE. For the 20% aberrant person condition, 2PL_EAP had the highest correlation and the lowest RMSE again. But the difference here was that, as the percentage of aberrant items increased (≥.20), the performance of 3PL_EAP became better than 2PL model (Figure 1). This result, again, can be interpreted as the robustness of 3PL model and EAP estimation method to the high percentage of aberrant items and persons.

Figure 1.

Correlation and RMSE Values under Different Conditions



Overall, the main finding of the study is that the presence of item pre-knowledge impacts CAT test results and the severity of that impact depends on the percentage of the aberrant item and person. However, this impact is not even comparable with the base condition that item preknowledge was not present.

Discussion

The purpose of the current study was to investigate the robustness of CAT results in the presence of item preknowledge under different conditions. We observed that the presence of compromised items was a threat to the performance of CAT because this presence led to a decline in the measurement precision of the CAT applications. The specific results were stated and discussed in the light of literature in this section.

Firstly, we observed that the increase in the percentage of aberrant items and persons resulted in a decrease in measurement precision as observed in the literature (Belov, 2016). Specifically, base (no aberrance) condition had the highest measurement precision (lowest RMSE, MAE and highest correlation) and 20% aberrant person condition had the lowest precision, decreasing with the increment of aberrant item percent. It is an expected result since item preknowledge is an important threat for test scores and the number of compromised items and aberrant persons has an impact on the magnitude of this threat. Since CAT is mostly based on IRT models and these models are highly sensitive to the aberrancy of response patterns, ability estimations can be strongly biased (Jia et al., 2019; Kim & Moses, 2016; Magis, 2014).

The examination of the robustness of estimation methods to the presence of item preknowledge indicated that using EAP estimation method provided more measurement precision than MLE through all conditions. To our knowledge, there is no CAT-specific study comparing these estimation methods in the presence of aberrant responses. However, Kim & Moses (2016) compared different ability estimation methods at different aberrancy conditions in a multi-stage testing (MST) context, which is also adaptive. Consistent with the current study, they found that EAP yielded smaller RMSE than did MLE, especially at the highest and lowest ability regions under preknowledge condition.

Another result observed in our study is that 2PL model had higher measurement precision than 3PL at most of the conditions. 2PL with EAP ability estimation was the best combination for 10% person condition at all aberrant item percentages; but, 3PL-EAP combination outperformed the 2PL counterpart when aberrant person was 20% and the percentage of the compromised item was high (≥.20). It means that at a high amount of aberrancy, 3PL and EAP becomes more resistant to the threat of item preknowledge than 2PL. Although 2PL model is operationally used at large-scale testing programs such as GRE, TOEFL and PISA, it should be used carefully because of ignoring the ‘guessing effect’ (Hambleton, et al., 1991; Kim, et al., 2016). Haberman (2006) stated that the advantage of employing a 3PL model over a 2PL model seemed to be small, considering the much greater computational difficulties associated with 3PL. However, it should be carefully considered when ‘guessing effect’ is probably present. ‘Guessing effect’ causes individuals who do not know the answer to the question to answer the question correctly and it is one of the types of aberrant responses. In the current study, results indicated that ‘guessing effect’ had become more important at higher levels of item preknowledge (both item and person level). That is an expected result since item preknowledge poses an advantage for low-ability individuals and its impact increases with the level of aberrant persons and items. ‘Guessing effect’ also poses an advantage for the ones who do not have the knowledge to answer the question accurately. 2 PL model was able to compensate for the effect of both item preknowledge and guessing effect up to a certain point, but at some point, 3PL took the lead. Hence, the use of 3PL with EAP estimation method could be suggested especially in situations where item preknowledge is considered to pose a significant threat.

As a limitation of our study, we fixed the test length at 30 and used the Maximum Fischer Information method only as the item selection method. Further research studying different fixed or variable length conditions and different item selection methods can be conducted. Besides, different item exposure control methods can also be addressed in further studies. Additional research should be undertaken to examine different types of aberrant behaviors and under different conditions (such as item pool, ability parameters and number of individuals). Besides, the current study is limited to unidimensional IRT models. However, many educational and psychological tests are multidimensional (Ackerman, et al., 2003) and aberrant response behaviors may affect the statistical biases of the latent traits (Wang, 2015). Therefore, same problem considered in this research can be looked at in MIRT framework in further research.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.
Author Contribution: Hakan KARA: conceptualization, investigation, methodology, writing - review & editing. Nuri DOĞAN: Conceptualization, methodology, supervision, writing - review & editing. Başak ERDEM-KARA: Conceptualization, methodology, data analysis, visualization, writing - review & editing.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as data has been simulated in this study.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Ackerman T., Gierl M. J., Walker C. M. (2003). Using multidimensional item response theory to evaluate educational psychological tests. Educational Measurement Issues and Practice, 22(3), 37–51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

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. https://doi.org/10.7333/jcat.v2i0.36

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*(2), 83-97. https://doi.org/10.1177/0146621615603327

Clark, J. M. (2010). Aberrant response patterns as a multidimensional phenomenon: using factor-analytic model comparison to detect cheating. [Unpublished doctoral dissertation, University of Kansas]. ProQuest Dissertations and Theses Global.

Eckerly, C. A. (2017). Detecting item preknowledge and item compromise: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), Handbook of detecting cheating on tests (pp. 101-123). Routledge.

Guo, F. (2009). Quantifying the impact of compromised items in CAT. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. www.psych.umn.edu/psylabs/CATCentral/

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. International Journal of Testing, 9(4), 283–309. https://doi.org/10.1080/15305050903351901

Jia, B., Zhang, X., & Zhu, Z. (2019). A short note on aberrant responses bias in item response theory. Frontiers in Psychology, 10. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00043

Kim, S., & Moses, T. (2016). Investigating robustness of item response theory proficiency estimators to atypical response behaviors under two-stage multistage testing. ETS Research Report Series, 2016(2), 1–23. https://doi.org/10.1002/ets2.12111

Liu, T. , Sun, Y., Li, Z. & Xin, T. (2019) The impact of aberrant response on reliability and validity, measurement. Interdisciplinary Research and Perspectives, 17(3), 133-142, https://doi.org/10.1080/15366367.2019.1584848

Magis, D. (2014). On the asymptotic standard error of a class of robust estimators of ability in dichotomous item response models. British Journal of Mathematical and Statistical Psychology, 67(3), 430–450. https://doi.org/10.1111/bmsp.12027

Magis, D. , Raiche, G. & Barrada, J. R. (2022). catR: Generation of IRT response patterns under computerized adaptive testing (package version 3.17). https://cran.r-project.org/web/packages/catR

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. Applied Psychological Measurement, 27(2), 121-137. https://doi.org/10.1177/0146621602250534

Meijer, R. R. (1996). Person-Fit research: An introduction. Applied Measurement in Education, 9, 3–8. https://doi.org/10.1207/s15324818ame0901_2

Meijer, R. & Sijtsma K. (2001). Methodology review: Evaluating person fit. Applied Psychological Measurement, 25(2), 107–135. https://doi.org/10.1177/01466210122031957

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. https://doi.org/10.31234/osf.io/hk35a

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. Educational Measurement: Issues and Practice, 35(1), 38-47. https://doi.org/10.1111/emip.12102

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rios J. A., Guo H., Mao L., Liu O. L. (2017). Evaluating the impact of noneffortful responses 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

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. Journal of Educational Measurement, 51(3), 239–259. https://doi.org/10.1111/jedm.12046

Wan, S., & Keller, L. A. (2023). Using cumulative sum control chart to detect aberrant responses in educational assessments. Practical Assessment, Research and Evaluation, 28(2). https://doi.org/10.7275/pare.1257

Wang, K. (2017). A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing (Publication No. 10273809). [Doctoral Dissertation, Michigan State University]. ProQuest Dissertations & Theses.

Weiss, D.J., & Kingsbury, G.G. (1984). Application of computer adaptive testing to educational problems. Journal of Educational Measurement, 21 (4), 361-375. https://doi.org/10.111 1/j.1745-3984.1984.tb01040.x

Wollack, J. A., & Maynes, D. D. (2017). Detection of test collusion using cluster analysis. In Handbook of Quantitative Methods for Detecting Cheating on Tests (pp. 124-150). Routledge.

Yan, D., von Davier, A. A., & Lewis, C. (2014). Computerized multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), Computerized Multistage Testing. CRC Press.

Yen, Y. C., Ho, R. G., Laio, W.W., Chen, L.J., & Kuo, C. C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. Applied Psychological Measurement, 36(2), 75–87. https://doi.org/10.1177/0146621611432862

Yi, Q., Zhang, J., & Chang, H. H. (2008). Severity of organized item theft in computerized adaptive testing: A simulation study. Applied Psychological Measurement, 32(7), 543-558. https://doi.org/10.1177/0146621607311336

Zhang, J., Chang, HH. & Yi, Q. (2012). Comparing single-pool and multiple-pool designs regarding test security in computerized testing. Behavior Research Methods, 44, 742–752. https://doi.org/10.3758/s13428-011-0178-5

Zheng, Y., & Chang, H.H. (2014). On-the-fly assembled multistage adaptive testing. Applied Psychological Measurement, 39 (2), 104-118. https://doi.org/10.1177/0146621614544519