International Journal of Assessment Tools in Education
2025, Vol. 12, No. 3, 681–700
https://doi.org/10.21449/ijate.1598330
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Detection of Aberrant Testing behaviour in unproctored CAT via a verification test

Ebru Balta[1*], **Arzu Ucar**[2]

[1]Agri Ibrahim Cecen University, Faculty of Education, Department of Educational Sciences, Agri, Türkiye
[2]Hakkari University, Faculty of Education, Department of Educational Sciences, Hakkari, Türkiye

ARTICLE HISTORY

Received: Dec. 8, 2024

Accepted: June 18, 2025

Keywords:

Aberrant testing behaviour,
l_z person-fit statistic,
Divergence measure,
Unproctored CAT,
Verification test.

Abstract: Unproctored Computerized Adaptive Testing (CAT) is gaining traction due to its convenience, flexibility, and scalability, particularly in high-stakes assessments. However, the lack of proctor can give rise to aberrant testing behavior. These behaviors can impair the validity of test scores. This paper explores the use of a verification test to detect aberrant testing behavior in unproctored CAT environments. This study aims to use multiple measures to detect aberrant response patterns in CAT via a paper-and-pencil (P&P) test as well as to compare the sensitivity and specificity performances of the l_z person-fit statistic (PFS) using no-stage and two-stage (l_z is used after the Kullback–Leibler divergence (*KLD*) measure) methods in different conditions. Three factors were manipulated – the aberrance percentage, the aberrance scenario, and the aberrant examinee's ability range. The study found that in all scenarios, the specificity performance of l_z in classifying examinees was higher than its sensitivity performance in no-stage and two-stage analyses. However, the sensitivity performance of l_z was higher in two-stage analysis.

1. INTRODUCTION

With globalization, technology has significantly transformed educational environments. Unlike traditional paper-and-pencil (P&P) testing applications, computerized adaptive testing (CAT) provides higher measurement precision, lower test time, and flexible applications by using the invariance feature of item response theory (IRT) compared to traditional applications. CAT, which centres on examinee differences in the field of psychometrics, allows the examinees to receive tests optimised for themselves (Eggen, 2004). The CAT algorithm primarily involves ability estimation and item selection, largely based on the examinee's item response. Thus, a large item pool consisting of items that are grouped according to subject areas and difficulty levels (whose item information functions have been previously determined) and that provide information in all ranges of the examinee's ability level (θ) is created, and the test starts by selecting the item that will give the best information about the examinee. Large-scale, item-level adaptive test applications such as the Educational Record Bureau (ERB), the Graduate Management Admission Test (GMAT), the Graduate Record Examination (GRE), the National

*CONTACT: Ebru BALTA ebrubalta2@gmail.com Agri Ibrahim Cecen University, Faculty of Education, Department of Educational Sciences, Agri, Türkiye

e-ISSN: 2148-7456

Assessment of Educational Progress (NAEP), the Law School Admissions Test (LSAT), the Test of English as a Foreign Language (TOEFL), the National Council Licensure Examination (NCLEX), the Smarter Balanced Assessment System (SBAC), and the United States Medical Licensing Examination (USMLE) are conducted via computers over the internet (Armstrong *et al.*, 2010; Cui, 2022; Wise, 2023; Yan, 2020). The fact that these exams are administered on a large scale and without proctors makes them vulnerable to test fraud. For example, the Educational Testing Service (ETS) stated that examinees taking the GRE in Asian countries had a high rate of anormal response patterns and that it suspended the administration of the exam because of damage to test security (Sarı, 2019). Thus, test security in CAT applications cannot be ensured at a high level because the items are selected from an item pool and some items in this pool are reused and shared among examinees in future test applications (Guo *et al.*, 2009; Segall, 2004). It seems necessary to constantly add new items to the item pool by creating a large item pool considering the item exposure rate to prevent situations that could compromise test security (Glas & van der Linden, 2003; Magis & Raîche, 2012; Veldkamp & van der Linden, 2010). However, despite these precautions during the test progress, if aberrant test behaviour occurs, inappropriate items may be administered to examinees, resulting in inaccurate ability estimates.

Aberrant testing behaviours can impair the validity of test scores in CAT. Therefore, aberrant response patterns should be identified (Liu, 2019). Various aberrant testing behaviours need to be detected, including answer-copying, pre-knowledge cheating, careless answering, creative thinking, lucky guessing, plodding, random responding, and sleeping behaviour (Cizek & Wollack, 2017; Haberman & Lee, 2017; Kingston & Clark, 2014; Lee & Chen, 2011; Lee & Haberman, 2016; Sinharay, 2017b, 2020; van der Linden & Guo, 2008; Wang *et al.*, 2018). The literature mentions several methods such as similarity analysis and person-fit statistics (PFSs) for fixed tests to detect aberrant response patterns at the examinee and group levels (e.g. Cizek & Wollack, 2017; Karabatsos, 2003; van Krimpen-Stoop & Meijer, 2001; Maynes, 2005; Meijer & Sijtsma, 2001; Meijer & Tenderio, 2014; Thissen, 2008; van der Linden & Sotaridona, 2006). A common strategy is to flag the examinees or items with aberrant patterns (e.g. Belov & Armstrong, 2011; Belov *et al.*, 2007; Choe *et al.*, 2018; Drasgow *et al.*, 1985; Liu *et al.*, 2019; McLeod *et al.*, 2003; Shu *et al.*, 2013; Sinharay, 2017a, 2017b; Zhang, 2014; Zhang & Li, 2016). Based on IRT, a number of PFSs have been proposed to identify aberrant response patterns (Drasgow *et al.*, 1985; Molenaar & Hoijtink, 1990). Many PFSs have been improved for dichotomous items developed based on IRT, such as *U* (Wright & Stone, 1979), l_0 (Levine & Rubin, 1979), *W* (Wright & Masters, 1982), *D(θ)* (Trabin & Weiss, 1983), *ECI* (Tatsuoka, 1984), *UB* and *UW* (Smith, 1985), l_z (Drasgow *et al.*, 1985), *JK*, *O/E* (Drasgow *et al.*, 1987), *c* (Levine & Drasgow, 1988), l_{zm} (Drasgow *et al.*, 1991), *M* (Molenaar & Hoijtink, 1990), ${\chi_{SC}}^2$ (Klauer & Retting, 1990), *T(X)* (Klauer, 1991), and ${l_z}^*$ (Snijders, 2001). Some PFSs are based on division into two sets of items in the test, such as the Kullback–Leibler divergence (*KLD*) measure (Belov, 2007; Belov & Armstrong, 2010), the *Z* statistic (Guo & Drasgow, 2010; Maynes, 2014b), matched percentile (MPI; Kolen & Brennan, 2008), and the Irregularity Index (Li *et al.*, 2014).

Several studies (e.g. Armstrong & Shi, 2009; Belov, 2014, 2016; Chang & Zhang, 2002, 2003; Chao *et al.*, 2011; Choe *et al.*, 2018; Davey & Nering, 2002; Egberink *et al.*, 2010; Goren *et al.*, 2022; Guo *et al.*, 2009; Liu, 2019; Liu *et al.*, 2019; McLeod *et al.*, 2003; Pan *et al.*, 2022; Rizavi, 2001; Shu, 2010; Tendeiro & Meijer, 2012; van der Linden & Guo, 2008; van der Linden & van Krimpen-Stoop, 2003; van Krimpen-Stoop & Meijer, 2002; Yi *et al.*, 2006; Zhang, 2014; Zhang & Li, 2016; Zhong, 2022) discuss statistical methods to detect aberrant testing behaviours in CAT applications. When the studies are examined, CAT applications to detect pre-knowledge cheating have been suggested, such as the final log-odds ratio (*FLOR*) index (McLeod *et al.*, 2003), response time (RT) modelling (such as the Bayesian lognormal RT model) (van der Linden, 2006), the hierarchical latent variable model (van der Linden, 2007;

van der Linden & Guo, 2008), the mixture model (Lee & Wollack, 2017; von Davier & Rost, 2007; Wang & Xu, 2015; Wang *et al.*, 2018; Zhan *et al.*, 2018), machine-learning approaches (such as supervised, unsupervised, and reinforcement learning) (Bishop, 2006; Murphy, 2012), cluster analysis (Wollack & Maynes, 2011), factor analysis (Zhang *et al.*, 2011), the cumulative sum (CUSUM) method (Armstrong & Shi, 2009; Egberink *et al.*, 2010; van Krimpen-Stoop & Meijer, 2002), PFSs (Z_c (McLeod & Lewis,1999), K (Bradlow *et al.*, 1998), T (van Krimpen-Stoop & Meijer,2000), *lz* (Karabatsos, 2003; Shu *et al.*, 2013), and KLD (Belov, 2011, 2013; Chao *et al.*, 2011)). In many studies (Armstrong *et al.*, 2007; Drasgow *et al.*, 1991; Li & Olejnik, 1997; Meijer & Sijtsma, 2001; Nering, 1995, 1997; Nering & Meijer, 1998; Reise, 1995; Reise & Due, 1991; Shu *et al.*, 2013; St-Onge *et al.*, 2011; Zopluoglu & Davenport, 2012), l_z has been determined to be the most powerful PFS for fixed tests in detecting aberrant response patterns. Considering this, in related studies (Balta & Dogan, 2024; Belov, 2013, 2014; Belov *et al.*, 2007; Belov & Armstrong, 2010; Chao *et al.*, 2011; Man *et al.*, 2018; Marianti *et al.*, 2014; Ucar, 2021; Ucar & Dogan, 2021), one can observe that divergence measure approaches such as *KLD* exhibit high performance to determine aberrant response and response time patterns in both fixed tests and CAT applications. In addition, pre-knowledge cheating is largely investigated, with lesser focus on other aberrant test behaviours, hence a greater need to investigate several aberrant testing behaviours in CAT applications.

The use of unproctored computer-based testing (CBT) and CAT applications is becoming more widespread. Several researchers (e.g. Chapman & Webster, 2003; Lievens & Burke, 2011; Naglieri *et al.*, 2004; Nye *et al.*, 2008; Pearlman, 2009; Tippins *et al.*, 2006; Wunder *et al.*, 2010; Wright *et al.*, 2014) have cited the benefits of unproctored testing in terms of lower cost. However, in these applications, situations that facilitate security violations such as test theft and cheating caused by uncontrolled exam management may arise. Unproctored CAT, on the other hand, allows examinees to take the test without proctor, potentially introducing risks related to the validity of the data collected (Ryan *et al.*, 2015; Tippins *et al.*, 2006). Therefore, in unproctored CBT and CAT applications, psychometric identification such as a two-stage exam administration mode has been proposed by making the examinees undergo proctored verification tests (Nye *et al.*, 2008; Lievens & Burke, 2011; Coyne & International Test Commission, 2006). The use of verification tests allows for continuous monitoring of test-taker behavior, providing an additional layer of security in unproctored testing environments. The aim of this paper is to address this issue. There are few studies (Aguado *et al.*, 2018; Guo & Drasgow, 2010; Sanz *et al.*, 2020; Segall, 2001) on how psychometric identification should be performed. Segall (2001) proposed a Bayesian approach to detect the consistency of test performance across the CBT as well as verification testing approaches such as score-based and Bayesian methods. Guo and Drasgow (2010) detected aberrant response behaviour in CAT via a proctored verification test with a Z-test and a likelihood ratio (LR) test. Aguado *et al.* (2018) conducted psychometric identification by applying a Z-test and using RTs. Sanz *et al.* (2020) compared five statistics used to detect cheating in CATs Z-test, the Adaptive Measure of Change (AMC), LR, Score Test, and Modified Signed Likelihood Ratio Test (MSLRT). There is no general acceptance regarding which of the indices and statistics used to determine aberrant response patterns has high performance due to the many variables that impair test security. The performance of the methods is investigated by simulating various scenarios considering the common response patterns and testing conditions in real life. In addition, in several studies, it is seen that two-stage analyses are performed in which the PFSs and answer copying indices together with the divergence measure approaches are used together in order to increase the available evidence in determining aberrant response patterns. Belov (2013) proposed a two-stage method was made using PFSs and *KLD* to detecting test collusion in CAT and P&P test Similarly, Belov and Armstrong (2010) and Ucar and Dogan (2021) stated that the two-stage approach performed better in detecting answer copying in P&P test. So far, there is no study in which l_z and *KLD* are considered together in the detection of aberrant response patterns in CAT via a P&P test. However, the purpose of this study is to use multiple measures to detect potential aberrant examinees involved in aberrant testing behaviour in CAT via a P&P test.

A study was conducted to determine the performance of the l_z and *KLD* measures in identifying simulated aberrant testing behaviour under various conditions. In CAT applications and fixed tests, methods to identify aberrant response and RT patterns may mistakenly flag a non-aberrant as a suspected cheater. In the literature, several studies that investigate aberrant testing behaviour use power and Type I error rates as measures of the performance of these methods. The Type I error rate is when the method considers examinees who do not actually cheat. The benefit of this method is that it can accurately identify examinees who cheat. In this study, two indices – sensitivity and specificity – are used to evaluated the performance of these methods. Sensitivity is the rate of examinees who are correctly flagged as aberrant, and specificity is the rate of examinees who are correctly flagged as non-aberrant (Shu, 2010; Yormaz, 2019). Test validation is the process of verifying, based on evidence, whether the test development stages (e.g. overall plan, test blueprint, item development, test design and assembly, test administration, scoring test responses, standard setting, item bank management) have been fulfilled (Haladyna, 2011; Messick, 1994). The aberrant test behaviour of examinees in responding to items, those acting on behalf of the examinees (the proctor or test administrator), or aberrant behaviour such as cheating are among the factors that cause aberrant response and test scores (Karabatsos, 2003; Thiessen, 2008). In CAT applications, providing test management and controlling aberrant testing behaviour greatly increases the validity of test scores (Foster, 2013). Thus, it is important to recommend several methods and approaches to provide more evidence to increase the validity of test scores in unproctored CAT applications. In this study, to increase the available evidence in identifying aberrant examinees, a two-stage method was made using l_z and *KLD*. We calculated the sensitivity and specificity values using both no-stage and two-stage analyses.

We aim to compare the performances of the PFS and divergence measures (l_z and *KLD*) using no-stage and two-stage methods in different conditions. The research questions are as follows:

1) What are the sensitivity and specificity performances of various factors of l_z used in the no-stage method?
2) What are the specificity and sensitivity performances of various factors of l_z (used after the *KLD* measure) in the two-stage method?

2. METHOD

2.1. Research Design

In this study, a Monte Carlo simulation was conducted using simulation data to detect aberrant testing behaviour in CAT via a P&P test. Simulation data were used because all the conditions discussed in the study could not be met with real data. When deciding on simulation design conditions and levels, studies investigating aberrant response patterns in fixed tests and unproctored CAT applications were considered.

In several studies (Balta & Dogan, 2024; Li, 2019; Shu *et al.*, 2013; Steinkamp, 2017; Ucar, 2021, Ucar & Dogan, 2021), the aberrant examinee's percentage is manipulated as 5%, 10%, 15%, 20%, 35%, and 70%. Belov (2014), in his study which investigated aberrant response patterns at the group level in CAT applications, changed the percentage of aberrant examinees to 10% and 20% in each test centre. Karabatsos (2003) stated that when the number of copiers increases, the performance of PFSc to identify suspected copiers decreases. For this reason, the aberrant examinee's percentage in the cheating scenario was fixed at 5%, which was considered the minimum percentage in previous studies.

In the CAT literature, the aberrance percentage and the aberrant examinee's ability range are seen as important factors in determining aberrant response patterns; the latter, for instance, might affect the power of the methods to detect aberrant response patterns (Sotaridona &

Meijer, 2002; Steinkamp, 2017; Sunbul & Yormaz, 2018; Ucar, 2021; Ucar & Dogan, 2021; van der Linden & Sotaridona, 2006; Yormaz & Sunbul, 2017). Sunbul and Yormaz (2018) determined the ability range of the aberrant examinees as (−3, −1.5), (−1.51, 0), (0.01, 1.5), and (1.51, 3); Ucar (2021) changed this to (−3, −1.5) and (−1.51, 0) in his study. Aguado et al. (2018), in the cheating scenario, simulated 1,000 examinees for each of the 15 (θ_u: the ability levels for the unproctored test conditions; θ_v: the ability level in the verification test conditions) pairs: (−2, −2), (−1, −2) … (2, 2). In Belov's (2014) study, aberrant examinees were simulated with abilities drawn from U (−3, −2), U(−2, −1), and U(−1, 0). In this study, to evaluate the ability level effects, the ability range of the aberrant examinees was divided into two categories: (−3 to −1.5) (low ability level) and (−1.5 to 1.5) (medium ability level). In several studies (Belov, 2014, 2016; Liu, 2019; Pan *et al.*, 2022; Rizavi & Swaminathan, 2001; Shu *et al.*, 2013), in the CAT applications, the percentage of aberrance varied – 5%, 10%, 20%, 25%, 30%, 50%, 70%, 75%, and 90%. This study assumes a large percentage aberrance, such as the lower bounds of 60% and 70% considering the unproctored CAT applications.

In studies which determining aberrant response and response time patterns in CBT, CAT and P&P test applications (Belov, 2013, 2014, 2016; Fox & Marianti, 2017; Marianti *et al.*, 2014; Lee, 2018; Liu, 2019; Liu *et al.*, 2019; McLeod *et al.*, 2003; Pan *et al.*, 2022; Rizavi, 2001;Shu, 2010; Sotaridona & Meijer, 2002; van der Linden & Guo, 2008; van der Linden & Krimpen-Stoop, 2003; Wollack, 2006; Yi *et al.*, 2008; Zopluoğlu, 2016), it is seen that the sample size varies as 100, 500, 1,000, 2,000, 2,500, 10,000 and 50,000. In addition, in studies examining cheating behavior in unproctored CAT applications through a verification test (Aguado *et.al.*, 2018; Guo & Drasgow, 2010), 3,486 canditates participated in the unproctored CAT application and, 1,000 test takers were simulated in the CAT application. The sample size factor was not changed in this study. Considering the requirement of test takers participating in the unproctored CAT application to also take the P&P verification test, the current capacity of the exam halls, and the item parameter estimation, the sample size was determined as 1,000. In this study, 1,000 examinees were simulated with abilities drawn from N(0,1).

The test length was changed to 30, 40, 50, and 75, in studies which detected aberrant response patterns in CAT (Aguado *et al.*, 2018; Belov, 2013, 2014, 2016; Guo & Drasgow, 2010; Liu, 2019; Liu *et al.*, 2019; McLeod *et al.*, 2003; Pan *et al.*, 2022; Rizavi, 2001; Yi *et al.*, 2008). Balta and Ucar (2022) concluded that in CAT applications, when the starting rule was zero (θ=0) and, the test was terminated with the most 40 items and the highest fidelity value was obtained under this condition. Therefore, in this study, the test length was fixed at 50 items provide more accurate ability estimation considering the unproctored CAT application conditions.

Fifty aberrant examinees were selected at random from low– and medium–ability level examinees, obtained using the ability estimations in CAT application; 60% and 70% of the response patterns of these 50 aberrant examinees in the P&P test were manipulated. For these response patterns, if the difficulty level of the item is greater than the level of ability of the examinee (θ > b), the correct responses (1) have been converted to the wrong response (0). In another condition, the response patterns were determined randomly, and the correct answers were changed to be incorrect. After these changes were made to both conditions, the abilities of the examinees were re-estimated using the modified P&P test data.

To analyze the sensitivity and specificity performances of the methods, there were eight conditions (aberrance percentage (2) × aberrance scenario (2) × aberrant examinee's ability range (2) = 8). In Table 1, the simulation design conditions and levels are presented.

Table 1. *Simulation design conditions and levels.*

Condition	Level values	Number of levels
Aberrance percentage	60%–70%	2
Aberrance scenario	θ > b-random	2
Aberrant examinee's ability range	(−3.00 to −1.50)- (−1.50 to 1.50)	2
Sample size*	1,000	1
Test length*	50	1
Aberrant examinee's percentage*	5%	1

*fixed variable

2.2. Data Simulation

Data generation had been performed using the '*irtoys*' package (Partchev, 2017) for the P&P test and the '*catR*' package (Magis & Barrada, 2017) for CAT in the R software. A CAT simulation was carried out using the disclosed logical reasoning (LR) items of the LSAT (information about the items was obtained from www.LSAC.org). Thus, the three-parameter logistic (3PL) IRT model is used to describe the response probability for items. The means of the (a) discrimination, (b) difficulty, and (c) guessing parameters were 0.75, 0.49, and 0.17, with variances of 0.24, 1.13, and 0.25, respectively.

CAT applications consisting of large numbers of items with difficulty levels appropriate to each ability level and high levels of discrimination give better results (Embretson & Reise, 2000; Magis & Raîche, 2012; Veldkamp & van der Linden, 2010; Weiss, 2004). However, it has been stated that to create an effective ability estimation in CAT applications, the item pool size should be at least 100 items and contain at least 6 to 12 times more items than the test length (Stocking, 1992). The CAT item pool contained 500 items, 10 times the test length (50), similar to Belov (2014) and Belov (2016).

IRT based cut-off score based methods such as Maximum Fisher Information (MFI), Kullback Leibler Information, and log-odds ratio select the items that provide the highest information at the cut-point (Thompson, 2007b). MFI uses the measure of information (local information) around a certain ability level and the level of information it provides increases as the item discrimination level increases (Han, 2009; Ho, 2010). Thus, the MFI method was chosen as the item selection method in the CAT algorithm because the MFI item selection method selects items with high discrimination levels to provide maximum test information for the examinees. The disadvantage of the MFI is that leads to biased use of the item pool and the re-selection of the same items leads to the item exposure problem (van der Linden & Pashley, 2010; Wang, 2017). Thus, when MFI method is used, the item exposure should be controlled to ensure test security because the probability of selecting items with high discrimination levels is high (Barrada *et al.*, 2006). Barrada *et al.* (2009) stated that the restricted method is the best method to control maximum exposure rates in CAT applications. In this study, the item exposure rate was fixed at 0.25, as in the studies of Barrada *et al.* (2009) and Erdem-Kara and Dogan (2022).

In CAT applications, variables such as item selection methods, content balance in item selection, and item exposure rate play an important role in deciding which of the ability estimation methods is better (Embretson & Reise, 2000; Ho, 2010). After the selection of the first item, Maximum Likelihood Estimation (MLE), Weighted Likelihood Estimation (WLE), Marginal Maximum Likelihood Estimation (MMLE), and Bayesian based ability estimation methods such as Expected a Posteriori (EAP) and Maximum a Posteriori (MAP) are frequently used to estimate the ability (Baker & Kim, 2004; Embretson & Reise, 2000). van der Linden (2008) and van der Linden and Pashley (2010) suggested the EAP estimation method, which makes a finite estimate for ability levels when MFI is used as the item selection method and thus performs an ability estimation even when all of the examinee's responses are correct or incorrect. Thus, in this study, EAP method with a uniform prior over [−4, 4], which does not involve an iterative process while making a finite estimate for all ability levels, was used.

At the beginning of the test, the aim is to determine the most appropriate item for the examinee's true ability level. For this reason, CAT applications usually start with an item suitable for examinees with 0 ability level (Magis *et al.*, 2017). However, initiating the test in CAT applications, depending on the prior knowledge about the examinee's ability, can be achieved with different approaches, such as starting with easy items or medium-difficulty items (Hambleton & Xing, 2006; Thompson & Weiss, 2011). Thus,in this study, the ability estimate was initialised at θ = 0 to start with medium-difficulty items. The P&P test data (50 items) had been simulated based on the ability estimates and item parameters obtained from CAT simulations.

2.3. Analysis

In this study, to increase the available evidence in identifying aberrant examinees, no-stage and two-stage methods were employed using l_z and *KLD*. The l_z measure is the standardised log likelihood of l_0 and is given by the following:

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}} \tag{1}$$

Since the standard normal distribution is observed, high negative values of l_z (less than −2) are interpreted as indicating that the examinee's response patterns are not appropriate, while high positive values (greater than +2) can be interpreted as indicating that the responses fit well with the model (Dimitrov & Smith, 2006; Karabatsos, 2003). The *KLD* measure is a measure divergence between two posterior distributions for examinees based on responses and is given by the following:

$$KL = D(R \| S) = \int_{-\infty}^{+\infty} R(\theta_j) \log \frac{R(\theta_j)}{S(\theta_j)} d\theta_j \tag{2}$$

where $R(\theta_j)$ and $S(\theta_j)$ are the posterior distributions of ability for examinee j based on responses to two parts of test items. Large values for the *KLD* measure indicate a significant difference in the examinee's performance between the two parts (Kullback & Leibler, 1951). This difference may point to aberrant testing behaviours. We calculated the sensitivity and specificity values using both no-stage and two-stage analyses. The calculation of these values can be easily done with the help of the table prepared below:

Table 2. *Quota table.*

		Real	
		Aberrant	Non-aberrant
Decision	Aberrant	A	B
	Non-aberrant	C	D

According to Table 2, the number of examinees who were aberrant and were found to have cheated according to the analysis result is A, and this value is called 'true positive'. The value B is the 'false positive' value, which is the number of examinees who were not actually aberrant but were predicted to cheat as a result of the analysis. The number of examinees who were actually aberrant but were found not to be aberrant as a result of the analysis, C, is the 'false negative' value. D is the number of examinees who were correctly identified as not aberrant, and this value is called 'true false'.

Sensitivity is the method's power to distinguish true aberrant examinees.

$$\textit{Sensitivity} = A / (A+C)$$

Specificity is the method's power to identify true non-aberrant examinees.

$$Specificity = D / (B + D)$$

In the no-stage analysis, only l_z was used, and we calculated the probability value of l_z using the '*PerFit*' package (Tenderio *et al.*, 2016) to obtain sensitivity and specificity values. We compared the probability values of the PFS using the P&P test data with α = 0.05. In the two-stage analysis, we calculated l_z and the *KLD* measure with the '*PerFit*' package (Tenderio *et al.*, 2016) and the '*LaplaceDemon*' package (Statisticat, 2016) included in the R program. We have chosen to use *KLD* because it is the expected value of an LR. The examinee's ability levels in both tests (P&P (posterior)–CAT (prior)) had been compared via the *KLD* measure. The receiver operating characteristic (ROC) curve analysis method was used to obtain the cutoff scores for the *KLD* measure function values. To determine cutoff scores at α = 0.05 using the Youden Index, we used the '*OptimalCutpoints*' package (Raton-Lopez *et al.*, 2014). Then to detect previously marked aberrant examinees, l_z had been used, and the sensitivity and specificity values of l_z in identifying the aberrant examinees had been calculated. The *KLD* measure and l_z used to determine the aberrant examinees were repeated 100 times, and the results were reported as the average of 100 replications.

3. RESULTS

In the no-stage method, the sensitivity and specificity performances of l_z under various conditions were determined. Table 3 shows the results for no-stage analysis for different conditions.

Table 3. *Results for no-stage analysis.*

Aberrance percentage	Aberrant examinee's ability range	Aberrance scenario	Real aberrance decision	Simulated aberrance decision	
				Yes	No
60%	Low ability level	θ > b	Yes	2	14
			No	48	936
		Random	Yes	1	16
			No	49	934
	Medium ability level	θ > b	Yes	2	19
			No	48	931
		Random	Yes	1	21
			No	49	929
70%	Low ability level	θ > b	Yes	1	18
			No	49	932
		Random	Yes	2	19
			No	48	931
	Medium ability level	θ > b	Yes	4	17
			No	46	933
		Random	Yes	2	17
			No	48	933

Table 3 shows that in scenarios where the aberrance percentage is 60%, the sensitivity performance of l_z in classifying examinees is higher in the θ > b aberrance condition. The sensitivity performance in the scenarios where the aberrance percentage was 70% showed the highest classification performance in the aberrance condition involving the aberrant examinees with a medium ability level and θ > b aberrance condition. When the sensitivity performance of l_z is examined, it shows low performance in identifying aberrant examinees. However, in scenarios where the aberrance percentage is high, the examinee cannot distinguish between the normal response pattern and the anormal response pattern and cannot identify the examinee as an aberrant examinee. The specificity performance of l_z was found to be higher in the scenarios where the aberrance percentage was 60% among the aberrant examinees with a medium ability level and θ > b aberrance condition as well as higher in the scenarios where the aberrance percentage was 70% among the aberrant examinees with a medium ability level. In addition, the rate of identifying aberrant examinees is higher in scenarios where the aberrance percentage is low (60%) and given the θ > b aberrance condition compared to other conditions. As a general result, it was observed that in all scenarios, the specificity performance of l_z in classifying examinees was higher than its sensitivity performance.

In the two-stage method, the sensitivity and specificity performances of various factors of l_z and the *KLD* measure were determined. Table 4 shows the results for two-stage analysis for different conditions.

Table 4. *Results for two-stage analysis.*

Aberrance percentage	Aberrant examinee's ability range	Aberrance scenario	Real aberrance decision	Simulated aberrance decision	
				Yes	No
60%	Low ability level	θ > b	Yes	14	4
			No	36	946
		Random	Yes	8	5
			No	42	945
	Medium ability level	θ > b	Yes	12	8
			No	38	942
		Random	Yes	8	6
			No	42	944
70%	Low ability level	θ > b	Yes	12	7
			No	38	943
		Random	Yes	6	4
			No	44	946
	Medium ability level	θ > b	Yes	10	5
			No	40	945
		Random	Yes	7	6
			No	43	944

According to Table 4, as a result of the two-stage analysis, in scenarios where the aberrance percentage is 60%, the sensitivity performance of l_z in classifying examinees was found to be higher in the θ > b aberrance condition. In scenarios where the aberrance percentage is 70%, the sensitivity measure showed the highest classification performance in aberrant examinees with a medium ability level and the θ > b aberrance condition. One can see that the rate of identifying aberrant examinees is higher in conditions where the aberrance percentage is low (60%) and given the θ > b aberrance condition compared to other conditions. The specificity performance of l_z is higher in scenarios where the aberrance percentage is 60%, for the aberrant examinees with a low ability level, and in the θ > b aberrance condition as well as higher in scenarios where the aberrance percentage is 70% for the aberrant examinees with a medium ability level. This will reduce the risk of an examinee who does not have an aberrant response pattern being mistakenly identified/marked as having an aberrant response.

4. DISCUSSION and CONCLUSION

As the world increasingly adopts digital platforms for assessment, unproctored CAT systems provide a flexible and efficient method of delivering high-stakes tests to a large population. Unproctored CAT applications are being carried out by institutions and organisations, especially some universities performing large-scale assessment. However, aberrant testing behaviour is still a primary concern in unproctered internet testing (Tippins *et al.*, 2006; Wright *et al.*, 2014). Therefore, in such applications, two-stage exam administration by testing the examinees with proctored verification tests is important. The challenge here is the detection of aberrant testing behaviour based on the data of these two tests, along with the proctored verification test parallel to the CAT taken online. In this study, scenarios were produced by considering the unproctored CAT and proctored P&P test application processes and considering frequently encountered response patterns or situations. Based on these scenarios, the performance of l_z in identifying possible aberrant examinees in CAT applications was examined via a P&P verification test.

Aguado *et al.* (2018), Lievens and Burke (2011), Nye *et al.* (2008), and Tippins *et al.* (2006) have proposed a two-stage exam administration procedure in unproctored CBT and CAT applications, where candidates undergo proctored verification tests. In light of the findings obtained from the study as a general result, it is seen that the use of a verification test in unproctored CAT environments provides a robust solution for detecting aberrant testing behavior. Thus, these findings of the study are consistent with the relevant literature.

In the no-stage analysis, the sensitivity performance of l_z was higher in the simulation conditions where the aberrance percentage was 60%, for the aberrant examinees with a low ability level, and given the θ > b aberrance condition than in the simulation conditions where the aberrance percentage was 70%. This finding is parallel to that in the study of Zopluoglu and Davenport (2012), who reported that the performance of l_z in identifying aberrant examinees decreased as the aberrance percentage increased. However, in the θ > b aberrance conditions, the sensitivity performance of l_z was generally higher than that of the random aberrance conditions. In two-stage analyses where l_z was used together with the *KLD* measure, the sensitivity performance of l_z in classifying examinees was higher in simulation cases where the aberrance percentage was 60%. Additionally, the findings regarding the two-stage use of l_z show that the rate of identifying a suspicious aberrant examinee increases with the aberrance rate. However, the sensitivity performance of l_z in classifying examinees in no-stage analyses was lower than the sensitivity performance of the two-stage analysis. This is because *KLD* is a sensitive measurement against the differences between the distributions of ability (Pardo, 2006). Therefore, the sensitivity of l_z increased using two-stage analysis. In the no-stage and two-stage analyses, the specificity performances of l_z were high in all scenarios except for the condition where the aberrance percentage was 70%, for the aberrant examinees with a medium ability level, and given the θ > b aberrance condition. In the two-stage analyses where l_z was used together with the *KLD* measure, the specificity performance in classifying examinees was high under the condition where the aberrance percentage was 60%, for the aberrant examinees with a low ability level, and given the θ > b aberrance condition. Zhong (2022) stated that aberrant examinees can be identified by PFSs, but aberrant response behaviour types cannot be identified using these PFSs. In this study, both in two-stage and no-stage analyses, l_z had high specificity performance regardless of the aberrance scenario among the aberrant examinees with a medium ability level. In other words, whether the aberrance scenario was θ > b or random did not affect the specificity performance of l_z.

When the specificity and sensitivity performances of l_z were compared, the former was considerably higher than the latter. However, the sensitivity performance of l_z was higher in two-stage analysis. This finding is similar to the studies of Belov (2013), Belov and Armstrong (2010) and Ucar and Dogan (2021). In addition to the results of the study,the two-stage Type I error, power rates, or sensitivity and specificity performances of l_z for identifying examinees with aberrant response patterns in CAT applications can be examined under different conditions (sample size, test length, aberrant examinee’s percentage, aberrant response types, aberrance percentage (percentages lower than 60%)) via a verification test. In addition, simulation studies can be conducted to compare the performance of l_z with other divergence measures via two-stage analysis. Thus, the conditions under which l_z performs better in determining aberrant examinees can be investigated, and contributions can be made regarding its use in real-life applications. In addition, studies on the specificity and sensitivity performances or Type I error and power rates of several divergence measures used in simulation studies to be conducted on several PFSs' performance can be conducted via two-stage analyses. Similar studies can be conducted using real data.

Future research studies should focus on optimizing the design of verification tests and exploring machine learning techniques to improve the accuracy and efficiency of aberrant behavior detection in CAT applications. Additionally, similar studies could be conducted by including a verification set of items that partially overlaps with the CAT to help cross-check responses. Thus, the cheating detection performances of methods such as Z-test, the Adaptive Measure of Change (AMC), Likelihood Ratio Test (LRT), Score Test, and Modified Signed Likelihood Ratio Test (MSLRT) in CAT applications can be tested under several conditions. Moreover, aberrant testing behaviours in unproctored CAT administrations can be explored using response times via a proctored CBT verification test.

Acknowledgments

This study was presented on June 10-13, 2019 in International Association for Computerized Adaptive Testing (IACAT) Conference in Minneapolis, Minnesota, USA and published only as an abstract paper.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. The authors would like to thank Assoc. Prof. Dr. Alper Şahin, who works at Atılım University.

Contribution of Authors

Ebru Balta: Literature Review, Investigation, Data Simulation, Visualization, Statistical Analysis, and Writing-original draft. **Arzu Ucar**: Literature Review, Visualization, Statistical Analysis, Writing-original draft, and Validation

Orcid

Ebru Balta https://orcid.org/0000-0002-2173-7189
Arzu Ucar https://orcid.org/0000-0002-0099-1348

REFERENCES

Aguado, D., Vidal, A., Olea, J., Ponsoda, V., Barrada, J.R., & Abad, F.J. (2018). Cheating on unproctored internet test applications: An analysis of a verification test in a real personnel selection context. *The Spanish Journal of Psychology*, *21*, E62. https://doi.org/10.1017/sjp.2018.50

Armstrong, R.D., Kung, M.T., & Roussos, R.A. (2010). A method to determine targets for multi-stage adaptive tests using integer programming. *European Journal of Operatinal Research, 3,* 709-718. https://doi.org/10.1016/j.ejor.2009.12.009

Armstrong, R., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, *33*(5), 391-410. https://doi.org/10.1177/0146621609331961

Armstrong R.D., Stoumbos, Z.G., Kung, M.T., & Shi, M. (2007). On the performance of the l_z person-fit statistic. *Practical Assessment Research & Evaluation, 12*(16). https://doi.org/10.7275/xz5d-7j62

Baker, F.B., & Kim, S.H. (2004). *Item response theory: Parameter estimation techniques*. Marcel Bekker Inc

Balta, E., & Dogan, C. D. (2024). Investigation of preknowledge cheating via joint hierarchical modeling patterns of response accuracy and response time. *SAGE Open*, *14*(4), 1-15. https://doi.org/10.1177/21582440241297946

Balta, E., & Ucar, A. (2022). Bilgisayar ortamında bireye uyarlanmış test uygulamalarında ölçme kesinliğinin ve test uzunluğunun farklı koşullar altında incelenmesi [Investigation of measurement precision and test length in computerized adaptive testing under different conditions]. *E-International Journal of Educational Research*, *13*(1), 51-68. https://doi.org/10.19160/e-ijer.1023098

Barrada J.R., Abad F.J., & Veldkamp B.P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema*, *21*(2), 313-320.

Barrada, J.R., Mazuela, P., & Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema*, *18*(1), 156- 159.

Belov, D.I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement*, *35*(7), 495-517. https://doi.org/10.1177/0146621611420705

Belov, D.I. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement*, *50*(2), 141–163. https://doi.org/10.1111/jedm.12008

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/10.7333/1410-0203037

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

Belov, D.I., & Armstrong, R.D. (2010). Automatic detection of answer copying via Kullback–Leibler divergence and *K*-index. *Applied Psychological Measurement*, *34(*6*)*, 379–392. https://doi.org/10.1177/0146621610370453

Belov, D., Pashley, P., Lewis, C., & Armstrong, R. (2007). Detecting aberrant responses with Kullback–Leibler distance. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 7–14). Universal Academy Press.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459. https://link.springer.com/article/10.1007/BF02293801

Bradlow, E.T., Weiss, R.E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive testing. *Journal of the American Statistical Association*, *93*(443), 910-919. https://doi.org/10.1080/01621459.1998.10473747

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Chapman, D.S., & Webster, J. (2003). The use of technologies in the recruiting, screening, and selection processes for job candidates. *International Journal of Selection and Assessment*, *11*(2), 113–120. https://doi.org/10.1111/1468-2389.00234

Chang, H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, *67* (3), 387-398. https://doi.org/10.1007/BF02294991

Chang, H., & Zhang, J. (2003, December, 3-5). *Assessing CAT security breaches by the item pooling index* [Oral presentation]. The Annual Meeting of National Council on Measurement in Education, Chicago, IL, USA.

Chao, H.Y., Chen, J.H., & Chen, S.Y. (2011, July,19-22). *Applying Kullback-Leibler divergence to detect examinees with item pre-knowledge in computerized adaptive testing* [Oral presentation]. The 17th International Meeting of the Psychometric Society, Hong Kong.

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. https://doi.org/10.1007/s11336-017-9596-3

Cizek, G., & Wollack, J. (2017). Identification of item preknowledge by the methods of information theory and combinatorial optimization. In G. Cizek, & J. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp.217–233). R outledge.

Coyne, I., & International Test Commission. (2006). International Guidelines on Computer-Based and Internet-Delivered Testing. *International Journal of Testing*, *6*(2), 143–171. https://doi.org/10.1207/s15327574ijt0602_4

Cui, Z. (2022). On measuring adaptivity of an adaptive test. *Measurement: Interdisciplinary Research and Perspectives,20*(1),21-33. https://doi.org/10.1080/15366367.2021.1922232

Davey, T., & Nering, N. (2002). Controlling item exposure and maintaining item security. In C.N. Mills, M.T. Potenza, J.J. Fremer., & W.C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165-191). Lawrence Erlbaum Associates.

Deng, H., Ansley, T., & Chang, H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement, 47*(2), 202-226. https://www.jstor.org/stable/20778948

Dimitrov, D.M., & Smith, R.M. (2006). Adjusted rasch person-fit statistics. *Journal of Applied Measurement*, *7*(2), 170-183.

Drasgow, F., Levine, M.V., & McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*(1), 59–79. https://doi.org/10.1177/014662168701100105

Drasgow, F., Levine, M., & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67-86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Egberink, I., Meijer, R., Veldkamp, B., Schakel, L., & Smid, N. (2010). Detection of aberrant item score patterns in a computerized adaptive test: An empirical example using the CUSUM. *Personality and Individual Differences*, *48*(8), 921-925. https://doi.org/10.1016/j.paid.2010.02.023

Eggen, T. (2004). *Contributions to the theory and practice of computerized adaptive testing* (Publication No. 305136454) [Doctoral dissertation, University of Twente]. ProQuest Dissertations Publishing. https://www.proquest.com/docview/305136454/C97B190BA46B4519PQ/1?accountid=135193&sourcetype=Dissertations%20&%20Theses

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Erdem-Kara, B., & Dogan, N. (2022). The effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests. *International Journal of Assessment Tools in Education, 9*(3), 682-696. https://doi.org/10.21449/ijate.1105769

Foster, D. (2013). Security issues in technology-based testing. In J.A. Wollack, & J.J. Fremer (Eds.), *Handbook of test security* (pp. 39–83). Routledge.

Fox, J.-P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement, 54*(2), 243–262. https://www.jstor.org/stable/45148424

Glas, C.A., & Linden, W. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*(4), 247-261. https://doi.org/10.1177/0146621603027004001

Goren, S., Kara, H., Erdem-Kara, B., & Kelecioglu, H. (2022). The effect of aberrant responses on ability estimation in computer adaptive tests. *Journal of Measurement and Evaluation in Education and Psychology, 13*(3), 256-268. https://doi.org/10.21031/epod.1067307

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored internet tests: The Z-test and the likelihood ratio test. *International Journal of Selection and Assessment*, *18*(4), 351-364. https://doi.org/10.1111/j.1468-2389.2010.00518.x

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

Haberman, S., & Lee, Y. (2017). *A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses.* Educational Testing Service, (Research Report No:RR-17-23). *https://www.ets.org/research/policy_research_reports/publications/report/2017/jxrq.html*

Haladyna M.T. (2011). *Handbook of Test Development.* Taylor and Francis Press.

Hambleton, R.K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Hambleton, R.K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education, 19*(3), 221–239. https://doi.org/10.1207/s15324818ame1903_4

Han, K.T. (2009, June,2-3). *A gradual maximum information ratio approach to item selection in computerized adaptive testing* [Oral presentation]. The 2009 Conference on Computerized Adaptive Testing, Minnesota, USA.

Ho, T. (2010). *A comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on generalized partial credit model* (Publication No.3428993) [Doctoral dissertation, The State University of Texas]. ProQuest Dissertations Publishing. https://www.proquest.com/docview/760034367/CBFB3EA847AA4FB3PQ/1?accountid=135193&sourcetype=Dissertations%20&%20Theses

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298. https://doi.org/10.1207/S15324818AME1604_2

Kolen, M.J., & Brennan, R.L. (2008). *Test equating, scaling, and linking: Methods and Practices.* Springer.

Klauer, K.C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika*, *56*(2), 213-228. https://doi.org/10.1007/BF02294459

Klauer, K.C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology*, *43*(2), 193–206. https://doi.org/10.1111/j.2044-8317.1990.tb00935.x

Kingston, N., & Clark, A. (2014). *Test fraud: Statistical detection and methodology*. Routledge.

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Lee, S.Y. (2018). *A mixture model approach to detect examinees with item Preknowledge* (Publication No.10830593) [Doctoral dissertation, The University of Wisconsin-Madison]. University of Wisconsin-Madison Library. https://asset.library.wisc.edu/1711.dl/FJW23RSLFRKJK8X/R/file-e9109.pdf

Lee, Y.H., & Chen, H. (2011). A review of recent response-time analyses in educationaltesting. *Psychological Test and Assessment Modeling*, *53*(3), 359–379.

Lee, Y., & Haberman, S. (2016). Investigating test-taking behaviors using timing and processdata. *International Journal of Testing*, *16*(3), 240-267. https://doi.org/10.1080/15305058.2015.1085385

Lee S.Y., & Wollack J. (2017, September, 6-8). *A mixture model to detect item preknowledge using item responses and response times* [Oral presentation]. The 2017 Conference on Test Security, Madison, USA.

Levine, M.V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, *53*(2), 161–176. https://doi.org/10.1007/BF02294130

Levine, M.V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*(4), 269–290. https://doi.org/10.2307/1164595

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.

Li, X., Huang, C., & Harris, D. (2014). *Examining individual and cluster test irregularities in mixed-format testing* [Oral presentation]. The 2014 Conference on Test Security. Iowa City, USA.

Li, M.F., & Olejnik, S. (1997). The power of Rasch person–fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, *21*(3), 215-231. https://doi.org/10.1177/01466216970213002

Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology*, *84*(4), 817-824. https://doi.org/10.1348/096317910X522672

Liu, X. (2019, June, 10-13). *Detecting aberrant behavior in CAT: The lognormal response time model* [Oral presentation]. The Annual Meeting of the International Association for Computerized Adaptive Testing, Minnesota, USA.

Liu, C., Han, K.T., & Li, J. (2019). Compromised item detection for computerized adaptive testing. *Frontiers in Psychology*, *10* (829), 1-16. https://doi.org/10.3389/fpsyg.2019.00829

Magis, D., & Barrada, J.R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software*, *76*(1), 1-19. https://doi.org/10.18637/jss.v076.c01

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1-31. https://www.jstatsoft.org/article/view/v048i08

Magis, D., Yan, D., & Von Davier, A.A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR.* Springer.

Man, K., Harring, J.R., Ouyang, Y., & Thomas, S.L. (2018) Response time based nonparametric Kullback-Leibler Divergence Measure for detecting aberrant test-taking behavior. *International Journal of Testing*, *18*(2), *155-177*. https://doi.org/10.1080/15305058.2018.1429446

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B.P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39* (6), 426–451. https://doi.org/10.3102/1076998614559412

Maynes, D.D. (2005). *M_4: A new answer-copying index*. Caveon Test Security, Midvale, UT. https://www.caveon.com/

Maynes, D.D. (2014b). A method for measuring performance inconsistency by using score differences. In N. M. Kingston, & A.K., Clark, (Eds.), *Test Fraud: Statistical detection and methodology*, (pp 186-199). Routledge.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23. https://doi.org/10.3102/0013189X023002013

McLeod, L.D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement*, *23*(2), 147-160. https://doi.org/10.1177/01466219922031275

McLeod, L.D., 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., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A Review of recent developments. *Applied Measurement in Education*, *8*(3), 261-272. https://doi.org/10.1207/s15324818ame0803_5

Meijer, R. (2002). Outlier detection in high-stakes certification testing. J*ournal of Educational Measurement*, *39*(3), 219-233. https://doi.org/10.1111/j.1745-3984.2002.tb01175.x

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

Meijer, R.R., & Tendeiro, J.N. (2014). *The use of person-fit scores in high stakes educational testing: How to use them and what they tell us*. Law School Admission Council. (LSAC Research Report 14-03). https://www.lsac.org/data-research/research/use-person-fit-scores-high-stakes-educational-testing-how-use-them-and-what

Meyer, J.P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521-538. https://doi.org/10.1177/0146621609355451

Molenaar, I.W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*(1), 75–106. https://doi.org/10.1007/BF02294745

Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective.* The MIT Press.

Naglieri, J.A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the internet: New problems, old issues. *The American Psychologist*, *59*(3), 150–162. https://doi.org/10.1037/0003-066X.59.3.150

Nering, M.L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, *19*(2), 121-129. https://doi.org/10.1177/014662169501900201

Nering, M.L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, *21*(2), 115-127. https://doi.org/10.1177/01466216970212002

Nering, M.L., & Meijer, R.R. (1998). A comparison of the person response function and the lz person fit statistic. *Applied Psychological Measurement*, *22*(1), 53-69. https://doi.org/10.1177/01466216980221004

Nye, C.D., Do, B., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment*, *16*(2), 112–120. https://doi.org/10.1111/j.1468-2389.2008.00416.x

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

Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*. Chapman & Hall.

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). *Practical considerations in computer-based testing (statistics for social and behavioral sciences)*. Springer.

Partchev, I. (2017). '*irtoys: Simple interface to the estimation and plotting of IRT Models'* (R package version 0.2.1). https://cran.rproject.org/web/packages/irtoys/irtoys.pdf

Pearlman, K. (2009). Unproctored internet testing: Practical, legal, and ethical concerns. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*(1), 14–19. https://doi.org/10.1111/j.1754-9434.2008.01099.x

Raton-Lopez, M., Rodriquez-Alvarez, X.M., Suarez- Cadarso, C., & Sampedro-Gude, F. (2014). OptimalCutpoints: Computing optimal cutpoints in diagnostic tests. (R package version 1.1.5). https://cran.rproject.org/web/packages/OptimalCutpoints/OptimalCutpoints.pdf

Reise, S.P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19*(3), 213-229. https://doi.org/10.1177/014662169501900301

Reise, S.P., & Due, A.M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, *15*(3), 217-226. https://doi.org/10.1177/014662169101500301

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271-282. https://doi.org/10.1177/014662169001400305

Rizavi, S.M. (2001). *The effect of test characteristics on aberrant response patterns in computer adaptive testing* (Publication No.3027247) [Doctoral dissertation, University of Massachusetts Amherst]. ProQuest Dissertations Publishing. https://www.proquest.com/docview/304699823/1BC249C4F0834BF2PQ/1?accountid=135193&sourcetype=Dissertations%20&%20Theses

Rizavi, S., & Swaminathan, H. (2001, April, 10-14). *The effect of test and examinee characteristics on the occurrence of aberrant response patterns in a computerized adaptive test* [Oral presentation]. The Annual Meeting of the American Educational Research Association, Seattle, USA.

Ryan, A.M., Inceoglu, I., Bartram, D., Golubovich, J., Grand, J., Reeder, M., Derous, E., Nikolaou, I., & Yao, X. (2015). Trends in testing: Highlights of a global survey. In I. Nikolaou., & J.K. Oostrom (Eds.). *Employee Recruitment, Selection, and Assessment: Contemporary Issues for Theory and Practice*, (pp. 136–153). Routledge.

Sanz, S., Luzardo, M., García, C., & Abad, F.J. (2020). Detecting cheating methods on unproctored internet tests. *Psicothema, 32*(4), 549-558. https://doi.org/10.7334/psicothema2020.86

Sarı, H.I. (2019). Investigating consequences of using item pre-knowledge in computerized multistage testing. *Gazi University Journal of Gazi Educational Faculty*, *39*(2), 1113-1134. https://doi.org/10.17152/gefad.535376

Schnipke, D.L., & Scrams, D.J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Segall, D.O. (2001, April, 10-14). *Detecting test compromise in high-stakes computerized adaptive testing: A verification testing approach* [Oral presentation]. The Annual Meeting of the National Council on Measurement in Education, Seattle, USA.

Segall, D.O. (2004). Computerized adaptive testing. In Kempf-Leanard (Eds.), *The encyclopedia of social measurement* (pp. 429–438). Academic Press.

Shu, Z. (2010). *Detecting test cheating using a deterministic, gated item response theory model*, (Publication No. 3434164) [Doctoral dissertation, The University of North Carolina at Greensboro]. ProQuest Dissertations Publishing. https://www.proquest.com/docview/845237696/4015D86F0434446BPQ/1?accountid=135193&sourcetype=Dissertations%20&%20Theses

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, *78*(3), 481-497. https://doi.org/10.1007/s11336-012-9311-3

Smith, R.M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45*(3), 433-444. https://doi.org/10.1177/001316448504500

Sinharay, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*(1), 46-68. https://doi.org/10.3102/1076998616673

Sinharay, S. (2017b). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement*, *41*(6), 403–421. https://doi.org/10.1177/0146621617698453

Sinharay, S. (2020). Detection of item preknowledge using response times. *Applied Psychological Measurement*, *44*(5), 376–392. https://doi.org/10.1177/0146621620909893

Snijders, T.A.B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*(3), 331-342. https://doi.org/10.1007/BF02294437

Sotaridona, L.S., & Meijer, R.R. (2002). Statistical properties of the K-index for detecting answer copying in a multiple-choice test. *Journal of Educational Measurement, 39*(2), 115–132. https://www.jstor.org/stable/1435251

Statisticat, L.L.C. (2016). *'LaplacesDemon: Complete environment for bayesian inference'* (R package version 16.0.1) https://cran.r-project.org/web/packages/LaplacesDemon/vignettes/LaplacesDemonTutorial.pdf

Steinkamp, S. (2017). *Identifying aberrant responding: Use of multiple measures* [Doctoral dissertation, The University of Minnesota]. University Digital Conservancy. https://hdl.handle.net/11299/188885.

Stocking, M.L. (1992). *Controlling item exposure rates in a realistic adaptive testing paradigm.* Educational Testing Service. (Research Report No. 93-2). https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1993.tb01513.x

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement, 35*(6), 419–432. https://doi.org/10.1177/0146621610391777

Sunbul, O., & Yormaz, S. (2018). Investigating the performance of omega index according to item parameters and ability levels. *Eurasian Journal of Educational Research*, *74*, 207–226. https://doi.org/10.14689/ejer.2018.74.11

Tatsuoka, K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*(1), 95–110. https://doi.org/10.1007/BF02294208

Tendeiro, J.N., Meijer, R.R., & Niessen, A.S.M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*(5), 1-27. https://doi.org/10.18637/jss.v074.i05

Thompson, N.A. (2007b, June,7). *Computerized classification testing with composite hypotheses* [Oral presentation]. The GMAC Conference on Computerized Adaptive Testing, Minneapolis, USA.

Thompson, N.A., & Weiss, D.A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1),1-9. http://pareonline.net/getvn.asp?v=16&n=1

Thiessen, B. (2008). *Relationship between test security policies and test score manipulations* (Publication No.3347249) [Doctoral dissertation, University of Iowa]. ProQuest Dissertations Publishing. https://www.proquest.com/docview/304633912/793D439F6D09431APQ/1?accountid=135193&sourcetype=Dissertations%20&%20Theses

Tippins, N.T., Beaty, J., Drasgow, F., Gibson, W.M., Pearlman, K., Segall, D.O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology, 59* (1), 189–225. https://doi.org/10.1111/j.1744-6570.2006.00909.x

Trabin, T.E., & Weiss, D.J. (1983). The person response curve: Fit of individuals to item response theory models. In D.J. Weiss (Eds.), *New horizons in testing*, (pp. 83–108). Academic Press.

Ucar, A. (2021). *Kopya belirlemede benzerlik indeklerinin birey-uyum istatistikleri aracılığıylaaşamalı kullanımının I.tip hatalarının ve gücünün belirlenmesi [Investigation of type-I-error and power of similarity indices by using two-stage analysis via person-fit statistics]* [Doctoral dissertation, Ankara University]. National Thesis Center. *https://tez.yok.gov.tr/UlusalTezMerkezi/tarama.jsp*

Ucar, A., & Dogan, C. D. (2021). Defining cut point for Kullback-Leibler divergence to detect answer copying. *International Journal of Assessment Tools in Education, 8*(1), 156–166. https://doi.org/10.21449/ijate.864078

van der Linden, W.J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31(2)*, 181-204. https://doi.org/10.3102/10769986031002181

van der Linden, W.J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72* (3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W.J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*(1), 5-20. https://doi.org/10.3102/1076998607302626

van der Linden, W.J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365-384. https://doi.org/10.1007/s11336-007-9046-8

van der Linden, W.J., Klein Entink, R.H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327–347. https://doi.org/10.1177/0146621609349800

van der Linden, W.J., & Pashley, P.J. (2010). Item selection and ability estimation in adaptivetesting. In W.J. van der Linden, & C.A.W. Glas (Ed.), *Elements of adaptive testing* (pp. 429 – 438). Springer.

van der Linden, W.J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, *31*(3), 283-304. https://www.jstor.org/stable/4122441

van der Linden, W.J., & van Krimpen-Stoop, E.M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251-265. https://doi.org/10.1007/BF02294800

van Krimpen-Stoop, E.M.L.A., & Meijer, R.R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement, 26*(2)**,** 164-180. https://doi.org/10.1177/01421602026002004

Veldkamp, B.P. (2012). Ensuring the future of computerized adaptive testing. In T.J.H.M. Eggen., & B.P. Veldkamp (Eds.). *Psychometrics in practice at RCEC*, (pp.39-50). RCEC.

Veldkamp, B.P., & van der Linden. W.J. (2010). Designing item pools for adaptive testing. In W.J. van der Linden., & C.A.W. Glas (Eds.). *Computerized adaptive testing: Theory and practice*, (pp.149-162). Springer.

von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C.R. Rao., & S. Sinharay (Eds.) *Handbook of Statistics*, (pp.643-661). Elsevier.

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 Publishing. https://www.proquest.com/docview/1901897901/

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456-477. https://doi.org/10.1111/bmsp.12054

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics, 43*(4), 469-501. https://doi.org/10.3102/1076998618767123

Wollack, J.A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement, 21*(4), 307-320. https://doi.org/10.1177/01466216970214002

Wollack, J.A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, *19*(4), 265-288. https://doi.org/10.1207/s15324818ame1904_3

Wollack, J.A., & Maynes, D. (2011, April, 7-11). *Detection of test collusion using item response data* [Oral presentation]. The 2011 Annual Meeting of the National Council on Measurement in Education, New Orleans, USA.

Wright, B., & Masters, G. (1982). *Rating Scale Analysis: Rasch Measurement*. MESA Press.

Wright, N.A., Meade, A.W., & Gutierrez, S.L. (2014). Using invariance to examine cheating in unproctored ability tests. *International Journal of Selection and Assessment*, *22*(1), 12–22. https://doi.org/10.1111/ijsa.12053

Wright, B., & Stone, M. (1979). *Best test design: Rasch measurement.* MESA Press.

Wise, S. (2023). Expanding the meaning of adaptive testing to enhance validity. *Journal of Computerized Adaptive Testing, 10*(2), 22-31. https://doi.org/10.7333/2305-1002022

Wunder, R.S., Thomas, L.L., & Luo, Z. (2010). Administering assessments and decision-making. In J.L. Farr., & N.T. Tippins (Eds.). *Handbook of Employee Selection*, (pp. 377–398). Routledge.

Yan, D. (2020). Multistage testing in practice. In H. Jiao., & R.W. Lissitz (Eds.). *Application of Artificial Intelligence to Assessment*, (pp.141-160). Information Age Publications.

Yormaz, S. (2019). *Test güvenliği açısından bireyler arasındaki olası iş birliğinin incelenmesi [Investigation of possible collusion between examinees in terms of test securtiy]* [Doctoral dissertation, Mersin University]. National Thesis Center. *https://tez.yok.gov.tr/UlusalTezMerkezi/tarama.jsp*

Yormaz, S., & Sunbul, O. (2017). Determination of type I error rates and power of answer copying ındices under various conditions. *Educatıonal Sciences: Theory & Practıce*, *17*(1), 5-26. https://doi.org/10.12738/estp.2017.1.0105

Yi, Q., Zhang, J., & Chang, H.H. (2006). *Severity of organized item theft in computerized adaptive testing: An empirical study*. Educational Testing Service. (ETS Research Report, RR-06-22). http://dx.doi.org/10.1002/j.2333-8504.2006.tb02028.x

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

Zhan, P., Jiao, H., Wang, W.-C., & Man, K. (2018). A multidimensional hierarchical framework for modeling speed and ability in computer-based multidimensional tests. *arXiv preprintarXiv:1807.04003*. https://doi.org/10.48550/arXiv.1807.04003

Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Applied Psychological Measurement*, *38*(2), 87-104. https://doi.org/10.1177/0146621613510062

Zhang, J., & Li, J. (2016). Monitoring items in real time to enhance CAT security. *Journal of Educational Measurement*, *53*(2), 131-151. https://doi.org/10.1111/jedm.12104

Zhang, Y., Searcy, C.A., & Horn, L. (2011, April, 9-11). *Mapping clusters of aberrant patterns in item responses* [Oral presentation]. The Annual Meeting of the National Council on Measurement in Education, New Orleans, USA.

Zhong, W. (2022). *Using item response theory to detect potential aberrant behaviors in a multi-stage test: An example of the norwegian language test* (Publication No. 304) [Master thesis, The University of Oslo]. CEMO Centre for Educational Measurement. https://www.duo.uio.no/handle/10852/55851/discover?rpp=100&sort_by=dc.date.issued_dt&order=DESC

Zopluoğlu, C. (2016). Classification performance of answer-copying indices under different types of IRT models. *Applied Psychological Measurement*, *40*(8), 592–607. https://doi.org/10.1177/0146621616664724