International Journal of Assessment Tools in Education
2026, Vol. 13, No. 1, 95–107
https://doi.org/10.21449/ijate.1550949
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Detecting suspicious persons in multiple-choice tests: Comparison of performance of ω and GBT indexes

Arzu Uçar[1*], **Celal Deha Doğan**[2]

[1]Hakkari University, Faculty of Education, Department of Educational Sciences, Hakkari, Türkiye
[2]Ankara University, Faculty of Education, Department of Educational Measurement and Evaluation, Ankara, Türkiye

ARTICLE HISTORY
Received: Sep. 16, 2024
Accepted: Sep. 27, 2025

Keywords:
Answer copying,
Similarity indexes,
Person fitting statistics,
Kullback-Leibler index,
Two-stage analysis.

Abstract: The COVID-19 pandemic has led to a widespread shift from traditional, supervised exams to unsupervised online testing environments, which has increased opportunities and motivation for cheating. Such dishonest behaviors threaten the validity and fairness of test results, underscoring the critical importance of robust test security measures. This study focuses on detecting individuals suspected of cheating by using two widely recognized statistical indexes: the ω and Generalized Binomial Test (GBT) indexes. Both indexes were applied within two analytical frameworks no-stage and two-stage methods. In the two-stage approach, the ω and GBT indexes were employed following the detection of potential cheaters using the Kullback-Leibler (KL) divergence index and person fit statistics (l_z and l_z^*). To simulate realistic conditions, we manipulated key variables including test difficulty, ability levels of suspected copiers, and the proportion of copied items. Our findings demonstrate that the GBT index combined with the KL index consistently outperformed other methods across varying scenarios in terms of detection accuracy and control of false positives. These results suggest that integrating person-fit and distribution-based methods enhances the reliability of cheating detection in unsupervised testing environments. The study provides valuable insights for test administrators seeking effective statistical tools to safeguard test integrity, especially in the context of increasing online assessments.

INTRODUCTION

The increasing emphasis on academic performance and credential attainment has heightened test-related pressure, often leading individuals to engage in dishonest behaviors, particularly in high-stakes assessments such as certification and licensure examinations. The transition from traditional proctored environments to unsupervised online testing accelerated notably during the COVID-19 pandemic has further magnified concerns around test integrity and fairness. At the core of the issue lies a critical threat to the validity of score-based inferences, especially when such scores inform consequential decisions about competence, placement, or professional qualification. To maintain the security and integrity of test outcomes, researchers have explored a range of detection techniques aimed at identifying aberrant responding patterns indicative of

*Corresponding author at: Arzu UÇAR arzukapcik@gmail.com Hakkari University, Faculty of Education, Department of Educational Sciences, Hakkari, Türkiye

e-ISSN: 2148-7456

potential cheating behavior. (Krimpen-Stoop & Meijer, 2001; Meijer & Sijtsma, 2001; Meijer & Tendeiro, 2014; Sideridis & Zopluoglu, 2018; Wise & Kong, 2005). Although various detection techniques exist for computer-based and adaptive tests, consensus on the most effective methods that integrate both response patterns and response times remains lacking (Belov, 2014a; Balta & Doğan, 2024; Meijer & Sotaridona, 2006; Sinharay & Johnson, 2020; van der Linden & Guo, 2008; Zhan *et al*., 2021). This study aims to address this need by proposing a rigorous methodological approach designed to improve test security without compromising validity and fairness.

The appropriateness of an individual's response pattern and test score is typically evaluated in relation to the underlying measurement model and benchmarked against the response behaviors and scores of examinees with similar ability levels in the sample or population. When an individual's observed responses deviate significantly from what is expected given their estimated ability level, such irregularities are referred to as aberrant response patterns (Karabatsos, 2003). Aberrant responses may arise from various sources, including disengaged or inconsistent behavior during item response, pre- or post-test exposure to test content via instructors or proctors, or deliberate acts such as cheating (Karabatsos, 2003; Meijer, 1994, 1996; Thiessen, 2008). These patterns often result in unexpectedly inflated or deflated test scores, which can undermine the reliability and validity of score interpretations. As emphasized by Cizek and Wollack (2017), such deviations compromise the intended use of test scores, rendering subsequent inferences questionable. Therefore, identifying and accounting for aberrant response patterns is essential for maintaining the psychometric integrity of assessment outcomes, particularly in high-stakes or unproctored testing environments.

Several behavioral factors can cause test scores to be spuriously high or spuriously low, thereby threatening the validity of score interpretations. These include careless responding (incorrectly answering items within the individual's competence level), lucky guessing (correctly answering unfamiliar items by chance), creative misinterpretation (overthinking simple items, typically among high-ability individuals), and random responding (selecting options without meaningful engagement) (Meijer, 1996). Among these, cheating (or answer copying) constitutes a particularly concerning form of aberrant behavior and is the primary focus of this study. Cheating refers to intentional behaviors in which individuals gain an unfair advantage by illegally accessing test items before the exam, copying answers from other test takers during the exam, or obtaining answer keys from instructors beforehand. These actions lead to correct answers on items that would otherwise be answered incorrectly. Cheating distorts test results systematically rather than being just a random error, which threatens the fairness and validity of the assessment. Therefore, effective detection methods are needed to protect the integrity of the testing process.

Early efforts to detect test cheating primarily focused on individual-level behaviors, but recent research has expanded to investigate group-level cheating that is not confined by geographic proximity, and may be influenced by shared test centers, institutional affiliations, or social networks (Belov, 2014b). Researchers have developed a variety of statistical methods at both individual and group levels to identify aberrant responses and score anomalies (Belov, 2013; Wollack & Maynes, 2011). Individual-level methods include person-fit statistics (Karabatsos, 2003), answer-copying indices (Sunbul & Yormaz, 2018a; Wollack, 1997; Zopluoglu, 2016; Wollack & Cohen, 1998), and similarity-based approaches (Zopluoglu, 2017). These methods aim to detect behaviors such as lucky guessing or answer copying that inflate or distort scores. More advanced techniques include Kullback–Leibler distance index (KL distance) (Kullback & Leibler, 1951), matched percentile index (MPI), and irregularity (IRI) indices, as well as model-based approaches like the deterministic gated IRT model (DG). At the group level, response pattern and score-based analyses, answer change tracking, and multilevel modeling approaches are used (He *et al*., 2018).

With the increasing use of computer-based testing, researchers now have access to detailed response time data, such as the moment a test item appears on the screen, the time spent solving it, and the moment the response is given (van der Linden, 2006, 2007, 2009; van der Linden & van Krimpen-Stoop, 2003; Wang & Xu, 2015; Wise & Kong, 2005). These data help detect whether a test-taker's behavior is appropriate. In particular, they support the identification of unusual behaviors in cases of low effort or deviations from expected response patterns (Hauser *et al.*, 2010; Kingsbury & Houser, 2011; Schnipke & Scrams, 1999; Wise & DeMars, 2006). When considered alongside response accuracy, such information contributes to a better understanding of irregular test-taking behaviors. This study aims to combine different detection methods from the literature to enable more effective identification of suspicious responding behaviors in exams. However, timing-based approaches were not considered, as the focus was on improving detection under realistic testing conditions using traditional methods.

Statistical methods for detecting answer copying must strike a balance between two fundamental goals: minimizing Type I errors (false positives, where a non-cheater is incorrectly flagged as a suspected copier) and maximizing statistical power (the ability to correctly detect actual cheaters). Although numerous studies have examined these performance metrics (Armstrong & Shi, 2009; Belov & Armstrong, 2010; Belov, 2013, 2014a, 2016; Belov *et al*., 2007; Shu, 2010; Sotaridona & Meijer, 2002, 2003; Zopluoglu & Davenport, 2012), findings consistently show that no single statistic consistently offers superior detection accuracy across all testing conditions (Karabatsos, 2003; Krimpen-Stoop & Meijer, 2001; Wollack, 1997; Wollack & Cohen, 1998; Yormaz, 2019). As a result, reliance on a single statistical indicator is often insufficient and may lead to inconsistent or unreliable classifications of suspected copying. This underscores the importance of multi-method approaches or comparative evaluations that can more robustly detect aberrant response behavior while maintaining acceptable error rates.

In this context, the GBT and *ω* indices based on Item Response Theory (IRT) have emerged as prominent tools in cheating detection studies, exhibiting superior performance in terms of Type I error control and statistical power compared to traditional methods (Zopluoglu & Davenport, 2012; Wollack, 2003, 2006). While most existing studies primarily employ no-stage analyses, this study comprehensively evaluates the performance of these indices using both no-stage and two-stage analysis designs across various testing environments.

In real testing conditions whether supervised or unsupervised, computer-based or paper-and-pencil-individuals may employ a wide range of cheating strategies. These include prior access to test items, collaboration with peers during the test, or the use of technological devices such as camera-equipped glasses or wireless earpieces. Furthermore, online and remote testing introduces new security vulnerabilities, such as cheating through mobile applications, social media groups, or outsourcing responses to others. Cheaters may attempt to minimize suspicious response patterns by limiting the number of collaborators and the number of items copied to avoid detection.

To reflect these realistic and varied cheating strategies, a scenario-based simulation encompassing different test modalities and plausible cheating behaviors was developed. Additionally, the spatial distribution and proximity of suspected cheaters were considered to enhance detection efficiency and strengthen the credibility of the evidence by identifying likely sources of copied answers. Using this scenario, we calculated the Type I error rates and power values for the GBT and *ω* indices, then compared their performances across different conditions to thoroughly examine their robustness in realistic testing environments.

In conclusion, this study responds to a pressing need in the literature on test security by emphasizing the importance of multiple statistical methods and scenario-based multi-stage analysis approaches in detecting cheating. It provides significant contributions to cheating detection in today's evolving testing environments.

Purpose of the study

The primary aim of this study is to compare the Type I error rates and statistical power of two answer-copying and similarity indices, the ω and GBT indices, under various conditions using both no-stage and two-stage methods. In the two-stage procedure, person-fit statistics l_z and l_z^* (Levine, & Rubin, 1979; Magis *et al.*, 2012; Molenaar & Hoijtink, 1990) and the KL index are employed in the first stage to identify potentially aberrant test-takers, followed by the application of the ω or GBT index in the second stage. The research questions addressed in this study are as follows:

1.What is the common effect of various factors on the type I error of the ω and GBT indexes that are used in no-stage and two-stage methods (used after person fitting statistics (l_z / l_z^*) and the KL index)?

2.What is the common effect of various factors on the power of the ω and GBT indexes that are used in no-stage and two-stage methods (used after person fitting statistics (l_z / l_z^*) and the KL index)?

METHOD

In this study, item response data were simulated for a sample of 150 examinees and 30 test items using the unidimensional Rasch model framework (Şahin, 2012). To realistically simulate aberrant behaviors such as answer copying while limiting their effect on parameter estimation accuracy, a fixed manipulation rate of 5% was applied. This rate was chosen based on prior simulation studies that balanced realism and analytical reliability (Belov, 2016; Zopluoglu & Davenport, 2012).

The proportion of copied items, or the number of cheating items, plays a critical role as it directly influences both the Type I error rate and the detection power of answer copying indices. Previous research has examined this factor at varying levels. For instance, Zopluoglu (2016) manipulated copying rates at three distinct thresholds-20% (low), 40% (medium), and 60% (high) highlighting that copying proportion is a primary determinant of classification accuracy. Similarly, Wollack (2006) and Sotaridona and Meijer (2002) used copying percentages in the test of 10%, 20%, 30%, and 40%. Karabatsos (2003) varied the number of aberrant responses between 5%, 10%, 25%, and 50%. Grounded in these prior findings, this study adopted copying levels of 10% (low), 40% (medium), and 60% (high) to comprehensively investigate the indices' performance across practical scenarios.

The ability level of individuals engaging in cheating behaviors is a crucial factor influencing both the Type I error rate and the detection power of answer copying and similarity indices. Research suggests that lower-ability examinees tend to have less confidence in their responses, thereby exhibiting a higher propensity to cheat (Voncken, 2014). For example, Sotaridona and Meijer (2002) considered source individuals with ability levels at the 90th and 60th percentiles. Similarly, Yormaz and Sunbul (2017) categorized source ability across five ranges 40-49%, 50-59%, 60-69%, 70-79%, and 80-90%and reported that the performance of copying detection indices improved as the source's ability increased. In line with these findings, Sunbul and Yormaz (2018a) examined various ability intervals for both sources and suspected copiers, including (-3, -1.51), (-1.5, 0), (0.01, 1.5), and (1.51, 3). Reflecting this, the present study varied the suspected copiers' ability levels at two distinct intervals: very low (-3, -1.51) and low (-1.50, 0).

Item difficulty is another key factor influencing the effectiveness of answer copying and similarity indices. For example, Zopluoglu (2016) employed average test difficulty parameters of 0.8 and 0.5 in his analyses, while Sunbul and Yormaz (2018a) categorized item difficulty into ranges of (-2.5, 0.00) for easier items and (0.01, 2.5) for more difficult items. Lower item difficulty has been linked to decreased detection power. To account for this variability, the present study incorporated item difficulty parameters at two levels, representing easier tests (-2.5, 0.00) and more difficult tests (0.01, 2.5).

Data Simulation

To evaluate the statistical power of the indexes, we designed 12 experimental conditions by combining three manipulated factors: copying percentage (3 levels), test difficulty (2 levels), and person ability (2 levels), resulting in a total of 12 unique scenarios. For each condition, data were simulated 100 times, generating 1,200 datasets in total for the power analysis. Similarly, for the analysis of Type I error rates, four conditions were created by crossing test difficulty (2 levels) and person ability (2 levels), with 100 replications per condition, yielding 400 datasets for this purpose. Altogether, the study utilized 1,600 datasets for both power and Type I error analyses. Each dataset comprised response data for 1,500 examinees whose ability parameters were drawn from a standard normal distribution. The response matrices followed a unidimensional Rasch model with binary (1-0) scoring for each of the 30 test items.

In designing the simulation class, special attention was given to increasing the likelihood that the suspected copier would choose a source with a higher ability level than themselves. To achieve this, the sample of 150 individuals was randomly selected with an intentional distribution favoring higher ability levels: 5% of the sample consisted of very low-ability individuals, 15% low-ability, 60% medium-ability, and 20% high-ability individuals. Following the sample selection, suspected copiers were identified by randomly choosing individuals from the very low and low ability groups as distinct cases. Finally, the seating positions of all individuals were assigned randomly to reflect realistic test conditions.

The steps described in this paragraph were conducted separately for the type I error and power analyses of the indexes. After randomly placing the individuals, a source was selected from those seated next to the suspected copier (in front, to the right, or to the left). It was assumed that the suspected copier would choose only one of these individuals as a source, with the source being determined randomly.

• For the type I error analysis, data were produced without any manipulation of the individuals' response patterns.

• For the power analysis, suspected copiers copied answers from 10%, 40%, or 60% of the source's responses. Thus, the responses of the copier were manipulated by replacing them with those of the source. Finally, the manipulated datasets were obtained and used for the power analysis of the indexes.

Analysis

Type I Error and Power Analysis

In the no-stage analysis, we calculated the probability values of the copy and similarity indexes using the "CopyDetect" package (Zopluoglu, 2018) to obtain a type I error rate. We compared the probability values of the answer copying and similarity indexes with $\alpha = 0.05$. We assigned 1 to those with probability values less than or equal to 0.05, and 0 to those with a greater probability. [A value of 1 represents "test fraud due to answer copying" and 0 represents "no test fraud due to answer copying" for the suspected copier and source pair]. Thus we obtained a frequency table. We summed up the "1" values in the frequency table, and divided the value we got from the total by the number of all selected pairs. Thus, we obtained the probability that the decision made for the selected pairs was wrong (type I error rate). We performed the same operations on the manipulated data. Based on these data, we obtained the probability that the decision made for the selected pairs was true (power ratio).

In the two-stage analysis, we initially computed person-fit statistics (l_z and l_z^* statistics) using the "PerFit" package (Tenderio, 2018) and calculated the Kullback-Leibler (KL) divergence index with the "LaplaceDemon" package (Singmann, 2020) in R. These calculations provided person-fit and KL values for each individual. To identify suspected copiers, we applied cutoff scores determined differently for each statistic: the cutoff for the person-fit statistics was generated by the “PerFit” package, while the cutoff for the KL index was established using the “OptimalCutpoints” package (Raton, 2014). Suspected copiers were then detected based on these thresholds. Subsequently, potential sources, defined as individuals seated directly in front of, to the right, or to the left of the suspected copier, were identified and included in the analysis. This approach limited the analysis to pairs consisting of suspected copiers and their potential sources, rather than involving all sample members.

The second stage followed the procedure of the no-stage analysis: using the "CopyDetect" package (Zopluoglu, 2018), we calculated the probability values of the copying and similarity indices to estimate the Type I error rate. We created a frequency table summing all instances where the test indicated copying (“1” values). The Type I error rate was calculated by dividing this sum by the total number of copiers, representing the proportion of false positive decisions. Likewise, for the power analysis, the same operations were conducted on the manipulated data representing actual copying behavior. The power was then calculated as the proportion of correct detections, obtained by dividing the summed “1” values by the number of true copiers. This two-stage procedure allowed for a more focused and rigorous evaluation of the indices’ performance.

RESULTS

Figure 1 demonstrates that the Type I error rate for the ω index increases with rising test difficulty among low-ability examinees when applied in a two-stage analysis following the l_z statistic. Importantly, under conditions of high test difficulty across all ability levels, the GBT index applied with the KL index consistently produced the lowest Type I error rates, followed by the GBT index used with l_z^*. These outcomes indicate that a two-stage approach based on the KL index offers improved control over false positives. Conversely, the ω index showed the highest Type I error rates in the no-stage analysis across every condition examined. Furthermore, both indices (ω and GBT) displayed markedly lower Type I error rates in two-stage analyses compared to the no-stage approach, underscoring the benefit of including person-fit indices to minimize incorrect classifications.

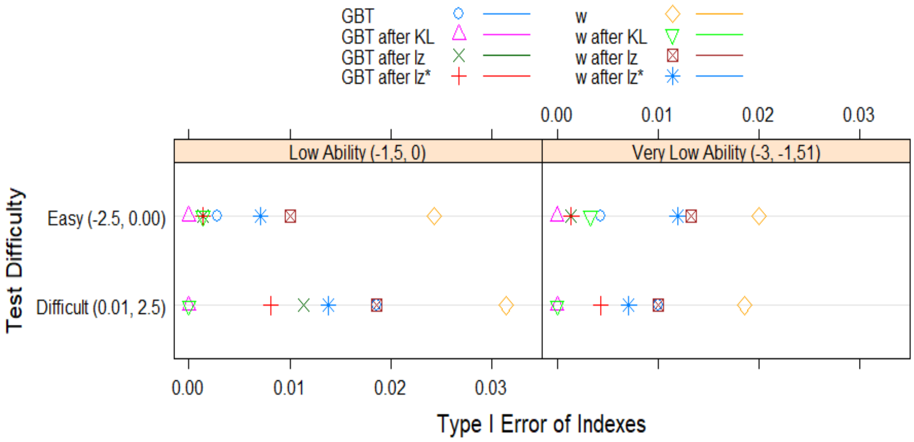

Figure 1. *The conditions' interaction effects for type I error of answer copying and similarity indexes.*

Figure 2 reveals that the GBT index, when employed after the KL index within a two-stage framework, consistently exhibits the highest detection power across all scenarios. This finding supports the effectiveness of combining KL and GBT in a sequential detection model for detecting copying behaviors. In comparison, the ω and GBT indices tend to have lower power when applied alongside person-fit statistics (l_z and l_z^*) relative to their performance with the KL index. However, despite this relative decrease, both indices demonstrate improved power when used following l_z and l_z^* compared to the no-stage analysis. These results suggest that initiating detection with either KL or person-fit statistics strengthens the capacity of both indices.

Collectively, the combination of KL and GBT delivers the most robust detection performance for identifying examinees involved in copying, thereby affirming the advantage of a two-stage detection system.

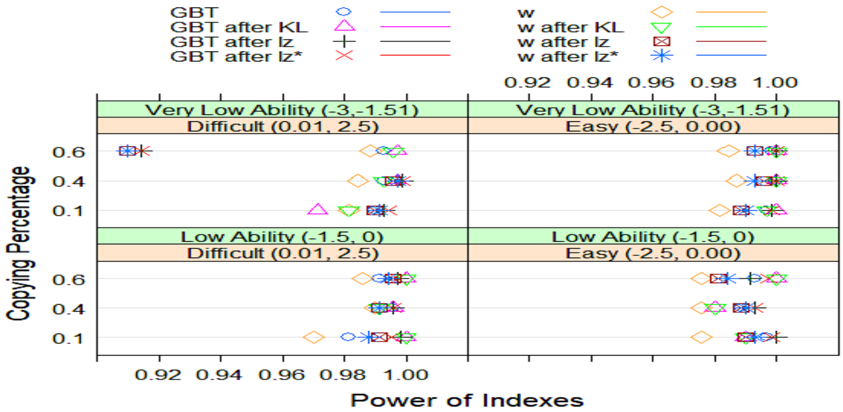

Figure 2. *The conditions' interaction effects for power of answer copying and similarity indexes.*

DISCUSSION and CONCLUSION

According to the results of the no-stage analysis, the Type I error rates of the ω index were higher than those of the GBT index across all conditions, a finding consistent with previous research (Sunbul & Yormaz, 2018a, 2018b; Zopluoglu & Davenport, 2012). This suggests that the ω index is more prone to falsely detecting non-cheating individuals as cheaters, which raises concerns regarding its specificity under single-stage conditions. Furthermore, both indices, ω and GBT, produced higher Type I error rates in the no-stage analysis than in the two-stage analysis, underscoring the value of integrating preliminary screening methods.

This aligns with Seol and Rubright (2013), who reported consistently lower Type I error rates when answer copying indices were used following initial screening via the KL index in a two-stage approach. Our findings corroborate this trend, revealing a notable reduction in false positives when the indices were applied after person-fit and KL-based filtering. In particular, the GBT index, when combined with the KL index, yielded the lowest Type I error rates under all conditions, emphasizing its potential robustness in conservative test security applications.

However, the effectiveness of these indices is sensitive to examinee ability levels. Tenderio and Meijer (2012) observed inflated Type I error rates of the l_z statistic among low-ability individuals. Similarly, Armstrong *et al.* (2007) noted that the l_z statistic can be unreliable in response patterns containing numerous incorrect answers, leading to over-flagging. Consistent with these findings, the current study observed elevated Type I error rates for both the ω and GBT indices when paired with the l_z statistic in two-stage analyses under low-ability conditions. Nevertheless, these rates remained lower than in the no-stage analysis, suggesting a partial mitigation of this issue through multi-stage modeling.

It is worth noting that the simulation design used in this study employed a relatively small sample size (n = 150), which may influence the generalizability and stability of the findings. Previous studies have shown that increasing sample size can reduce random errors and improve the accuracy of person-fit and KL-based detection methods (Tenderio & Meijer, 2014). As sample size grows, parameter estimates become more reliable and the underlying assumptions of the IRT model are better satisfied, which in turn reduces both Type I and Type II error rates. This consideration should be addressed in future research through larger-scale simulations or empirical validation studies.

The person-fit statistics and the Kullback-Leibler (KL) index detect aberrant test-taking behavior using different detection mechanisms. Person-fit statistics (e.g., l_z and l_z^*) detect individuals whose response patterns deviate from expectations based on the Item Response Theory (IRT) model. In contrast, the KL index detects unusual patterns by comparing the empirical ability distributions of individuals. Due to its sensitivity to differences in distributions, the KL index is effective in distinguishing genuine performance from potentially manipulated responses. In this study, the KL index was used to compare ability distributions under conditions with and without simulated test fraud. Individuals with significant differences in distributions were flagged as suspicious, supporting the use of KL as a useful tool for preliminary detection.

Beyond theoretical utility, the KL index also offers practical advantages. It can be applied in real-world testing scenarios by comparing prior (non-fraudulent) test data with current (potentially compromised) test results, especially in assessments where item pools include both common and unique items across test-takers. This flexibility broadens its applicability in operational settings where test security is a growing concern.

Consistent with prior findings by Sunbul and Yormaz (2018a, 2018b), this study found that both the ω and GBT indices demonstrated acceptable power levels in detecting answer copying across varying test difficulty conditions at the α = 0.05 significance level. Notably, the power of the ω index remained relatively stable across analyses conducted after both l_z and l_z^*, suggesting internal consistency in performance. However, under challenging conditions, such as when the average item difficulty was high, the suspected copier's ability was very low, and the copying rate was 60%, the power of the ω index was lower in the no-stage analysis compared to the two-stage analysis involving the KL index. This finding supports the use of a sequential approach, particularly when the copier's ability is substantially mismatched with the source.

Similarly, although the GBT index showed weaker performance in the two-stage analyses involving l_z and l_z^*than in no-stage and KL-enhanced two-stage analyses, its integration with the KL index resulted in superior power outcomes. These results align with Belov and Armstrong (2010), who emphasized the consistent performance of two-stage procedures involving the KL and K indices in detecting cheating behaviors. Our findings reinforce the potential value of the KL index as a reliable screening tool that enhances the detection capacity of both ω and GBT indices.

Furthermore, the increased power observed in two-stage analyses using person-fit statistics suggests that such procedures may effectively compensate for limitations inherent in individual indices. While prior research has shown variability in the sensitivity of these methods to low-ability examinees (Armstrong *et al*., 2007; Tenderio & Meijer, 2012), our results indicate that performance can be stabilized when sequential detection methods are applied.

Test security is a critical concern for educational institutions, examination bodies, and test developers, particularly in high-stakes contexts where individual performance is directly linked to educational or professional outcomes. The presence of cheating not only threatens the validity and fairness of assessments but may also result in the invalidation of test scores, with significant implications for test-takers and institutions alike (ÖSYM, 2010). Given the complexity of test fraud, relying on a single detection method is insufficient for making defensible judgments. This study reinforces the necessity of using multi-method detection strategies, particularly two-stage procedures, to more accurately identify aberrant response behaviors while minimizing false positives.

The comparative performance of detection indices revealed that the Type I error rates of both the ω and GBT indices were consistently higher in the no-stage analyses than in the two-stage analyses. Likewise, the statistical power of the ω index was generally lower than that of the GBT index, especially in the no-stage analysis. These results suggest that two-stage analysis not only improves detection accuracy but also provides more reliable evidence when suspecting test fraud. For instance, under the condition with the lowest Type I error rate for the ω index in the no-stage analysis, approximately three individuals (0.0185×150) with very low ability levels were falsely flagged as suspicious in a sample of 150 test-takers. In contrast, in the two-stage analysis the same condition yielded only one false positive with l_z (0.01 × 150), one with l_z^* (0.0071 × 150), and virtually none with the KL index (0.00001 × 150), illustrating the substantial reduction in false accusations when using a sequential approach. As sample sizes increase, the number of false positives rises proportionally when using detection indices with high Type I error rates, thereby exacerbating the potential adverse consequences. The methodological triangulation of multiple evidentiary sources (video-audio recordings, response time analyses) is fundamental for detecting aberrant behaviors. This approach minimizes measurement errors to ensure examination fairness while strengthening the evidentiary basis for high-stakes classification decisions.

Based on these findings, it is recommended that researchers and practitioners employ two-stage detection procedures in contexts where tests of higher difficulty are administered, and where examinees are unlikely to have had prior access to test content. In settings such as school-wide assessments or class-level assessments, if pretest integrity is assumed, applying the KL index as a first-stage screening tool followed by similarity indices like ω and GBT can enhance the credibility of detection outcomes. Moreover, the performance of these indices is influenced by various design factors including sample size, test length, item characteristics, and the dimensionality of the construct being measured.

In this study, the sample size and other factors such as test length and the copier ratio were not manipulated; however, future research should consider systematically varying these parameters to assess their effect on detection accuracy. Simulation studies could also incorporate different ability levels for both the source and the copier, alternative item structures, and multidimensional test models. In addition, the use of alternative divergence or distance metrics to quantify discrepancies in ability parameters between test-takers may provide further insight into potential fraud detection strategies.

The results obtained in this study offer practical contributions to the field of education, particularly regarding the development of reliable and secure assessment practices. In contexts where assessment outcomes are used to inform educational decisions, the detection of test fraud becomes a critical concern for maintaining the accuracy and fairness of measurement. The findings that the two-stage analysis reduced the Type I error rate and improved the detection power of the ω and GBT indices suggest that relying on a single method may be insufficient when identifying individuals suspected of cheating. The use of layered and complementary methods appears to be a more effective strategy for enhancing test security.

In educational practice, this points to the importance of integrating statistical detection methods into assessment processes, especially in settings where exams are conducted on a large scale or without supervision. Educators and institutions are encouraged to consider not only test scores but also test-taking behavior when interpreting results. For example, in classroom or institutional exams, applying a two-stage detection procedure that includes a preliminary screening step-such as the KL index-followed by a more targeted similarity analysis can help reduce the misidentification of non-cheating individuals as suspicious. In addition, adopting assessment designs that reduce the likelihood of copying, such as using different item orders or item variants, can contribute to the overall integrity of the testing process. From this perspective, the findings support the view that ensuring the validity of test results requires a combination of psychometric methods and instructional planning, rather than a reliance on scores alone.

While this study contributes to understanding the performance of similarity and person-fit indices, certain limitations must be considered. The simulations were conducted with fixed values for sample size, test length, and copier ratio, which may not reflect the diversity encountered in real testing environments. As a result, the findings might not generalize to conditions where these variables vary significantly. Moreover, the modeled cheating behavior followed structured patterns, whereas actual instances of test misconduct can involve more complex or unpredictable strategies. Additionally, the analyses were limited to response accuracy data; no auxiliary indicators such as response latency or examinee activity logs were incorporated, which could enhance detection in more comprehensive frameworks. Future studies are encouraged to expand the parameter range and explore richer data sources.

While this study contributes to understanding the performance of similarity and person-fit indices, certain limitations must be considered. The simulations were conducted with fixed values for sample size, test length, and copier ratio, which may not reflect the diversity encountered in real testing environments. As a result, the findings might not generalize to conditions where these variables vary significantly. Moreover, the modeled cheating behavior followed structured patterns, whereas actual instances of test misconduct can involve more complex or unpredictable strategies. Additionally, the analyses were limited to response accuracy data; no auxiliary indicators such as the time participants spend navigating and responding to each item, as well as the interaction-based digital records accumulated during the testing session were incorporated, which could enhance detection in more comprehensive frameworks. Future studies are encouraged to expand the parameter range and explore richer data sources.

Building on the findings of this study, several avenues warrant further exploration. Future research should consider varying key factors such as sample size, test length, and the proportion of suspected cheating behaviors to evaluate their influence on the detection indices' accuracy and robustness. Expanding simulation designs to include a wider range of cheating strategies beyond those predefined here-could provide deeper insight into the performance of detection methods under more complex and realistic conditions. Additionally, incorporating supplementary data sources, including item-level response times and detailed logs of test-taker interactions, may enhance the sensitivity and specificity of cheating detection algorithms. Exploring alternative distance or divergence measures within multi-stage detection frameworks also holds promise for improving accuracy. Finally, conducting empirical studies with real testing data would be valuable to validate simulation results and ensure practical applicability in operational settings.

Acknowledgments

This study was derived from the doctoral dissertation of the first author and was conducted under the supervision of the second author.

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. **Ethics Committee Number**: Ankara University, 30.10.2020-134.

Contribution of Authors

Arzu Uçar: Literature review, Resources, Methodology, Data analysis, Reporting, and Writing-original draft. **Celal Deha Doğan**: Supervision.

Orcid

Arzu Uçar https://orcid.org/0000-0002-0099-1348
Celal Deha Doğan https://orcid.org/0000-0003-0683-1334

REFERENCES

Armstrong, R., & Shi, M. (2009). Model-free CUSUM methods for person fit. *Journal of Educational Measurement, 46*, 408-428. https://doi.org/10.1111/j.1745-3984.2009.00090.x

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

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

Belov, D.I., Pashley, P.J., Lewis, C., & Armstrong, R.D. (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.

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

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

Belov, D. (2014a). *Detection of aberrant answer changes via Kullback Leibler divergence*. (Report No. RR-14-04). LSAC.

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

Belov, D. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40(2)*, 83-9. https://doi.org/10.1177/0146621615603327

Cizek, G., & Wollack, J. (2017). Exploring cheating on tests – the context, the concern, and the challenges. In G. Cizek & J. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp 3-19). Routledge.

Hauser, C., Kingsbury, G.G., & Houser, R.L. (2011, April 9-11). *Individual score validity: Using the wariness index to identify test performance to treat with caution*. [Paper presentation]. National Council on Measurement in Education Annual Meeting, New Orleans, LA, USA.

He, Q., Meadows, M., & Black, B. (2018). *Statistical techniques for studying anomaly in test results: A review of literature*. The Office of Qualifications and Examinations Regulation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/690007/Statistical_techniques_for_studying_anomaly_in_test_results-_a_review_of_literature.pdf

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

Krimpen-Stoop, E.M.L.A., & Meijer, R.R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26(2)*, 199-217. https://doi.org/10.3102/10769986026002199

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79-86. http://mathfaculty.fullerton.edu/sbehseta/Kullback.pdf

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

Magis, D., Raîche, G., & Be´land, S. (2012). A didactic presentation of Snijders's l_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37*(1), 57-81. https://doi.org/10.3102/1076998610396894

Measurement, Selection and Placement Center (ÖSYM). (2010, September 17). *Announcement of cancellation of the 2010 KPSS undergraduate educational sciences test*. https://www.osym.gov.tr/TR,3125/2010-kpss-lisans-egitim-bilimleri-testinin-iptali-17092010.html

Meijer, R.R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18(4)*, 311-314. https://doi.org/10.1177/014662169401800402

Meijer, R.R. (1996). Person-Fit research: An introduction. *Applied Measurement in Education, 9*(1), 3-8. https://doi.org/10.1207/s15324818ame0901_2

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., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing*. Law School Admission Council. https://ris.utwente.nl/ws/portalfiles/portal/5129730/LSAC_CT-03-03.pdf

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.

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

Schnipke, D.L., & Scrams, D.J. (1999). *Exploring issues of test taker behavior: Insights gained from response-time analyses*. Law School Admission Council.

Seol, J., & Rubright, J.D. (2013). *The impact of test characteristics on Kullback-Leibler divergence index to identify examinees with aberrant responses* [Paper presentation]. The 2nd Annual Conference on Statistical Detection of Potential Test Fraud, Madison, WI, USA.

Shu, Z. (2010*). Detecting test cheating using a deterministic, gated item response theory model* [Unpublished doctoral dissertation]. The University of North Carolina.

Sideridis, G.D., & Zopluoglu, C. (2018). Validation of response similarity analysis for the detection of academic cheating: An experimental study. *Journal of Applied Measurement*, *19*(1), 59-75.

Singmann, H. (2020). Complete environment for bayesian inference (LaplaceDemon) (Version 16.1.4) [Computer software manual]. https://cran.r-project.org/web/packages/LaplacesDemon/LaplacesDemon.pdf

Sinharay, S., & Johnson, M.S. (2020). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 397-419. https://doi.org/10.1111/bmsp.12187

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

Sotaridona, L.S., & Meijer, R.R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, *40*(1), 53-70. https://www.jstor.org/stable/1435054

Sunbul O., & Yormaz, S. (2018a). Effects of test level discrimination and difficulty on answer copying indices. *International Journal of Evaluation and Research in Education*, *7*(1), 32-38. http://doi.org/10.11591/ijere.v7i1.11488

Sunbul, O., & Yormaz, S. (2018b). 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

Sahin, A. (2012). *Madde Tepki Kuramı'nda test uzunluğuve örneklem büyüklüğünün model veri uyumu, madde parametreleri ve standart hata değerlerine etkisinin incelenmesi* [An investigation on the effects of test length and sample size in item response theory on model-data fit, item parameters and standard error values] [Unpublished doctoral dissertation]. Hacettepe University.

Raton, M. (2014). OptimalCutpoints: An R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, *61*(8), 1-36. https://doi.org/10.18637/jss.v061.i08

Tendeiro, J.N. (2018 Person Fit (PerFit) (Version 1.4.3) [Computer software manual]. https://cran.rproject.org/web/packages/PerFit/PerFit.pdf

Tendeiro, N.J., & Meijer, R.R. (2012). A CUSUM to detect person misfit: a discussion and some alternatives for existing procedures, *Applied Psychological Measurement 36*(5), 420-442. https://doi.org/10.1177/0146621612446305

Tendeiro, N.J., & 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

Thiessen, B. (2008). *Relationship between test security policies and test score manipulations*. [Unpublished doctoral dissertation]. University of Iowa.

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*, 287-308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W.J. (2009). A bivariate log-normal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, *34*(3), 378-394. https://doi.org/10.3102/1076998609332107

van der Linden, W.J., & Guo, F.M. (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., & 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

Voncken, L. (2014). *Comparison of the lz[*] person-fit index and ω copying*-*index in copying detection*. [First Year Paper]. Tilburg University. http://arno.uvt.nl/show.cgi?fid=135361

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

Wise, S.L., & DeMars, C.E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19-38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S.L., & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S.L., Ma, L., Kingsbury, G.G., & Hauser, C. (2010, May 1-3). An investigation of the relationship between time of testing and test-taking effort [Paper presentation]. *National Council on Measurement in Education Annual Meeting*, Denver, CO ,USA.

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. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement*, *40*(3), 189-205. https://doi.org/10.1111/j.1745-3984.2003.tb01104.x

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., & Cohen, A.S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement*, *22(2)*, 144-152. https://doi.org/10.1177/01466216980222004

Wollack, J.A., & Maynes, D. (2011, April 9-11). *Detection of test collusion using item response data* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New Orleans, LA, USA.

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

Yormaz, S. (2019). *Test güvenliği açısından bireyler arasındaki olası iş birliğinin incelenmesi* [Investigation of possible collusion between examinees in terms of test security] [Unpublished doctoral dissertation]. Mersin University.

Zhan, P., Jiao, H., Man, K., Wang, W.C., & He, K. (2021). Variable speed across dimensions of ability in the joint model for responses and response times. *Frontiers in Psychology*, *12*, Article 469196. https://doi.org/10.3389/fpsyg.2021.469196

Zopluoglu, 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

Zopluoglu, C. (2017). Similarity, answer copying, and aberrance: Understanding the status quo. In G. Cizek & J. Wollack (Eds.), *Handbook of Detecting Cheating on Tests* (pp. 44-72). Routledge.

Zopluoglu, C., & Davenport, E.C., Jr. (2012). The empirical power and type I error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, *72*(6), 975-1000. https://doi.org/10.1177/0013164412442941

Zopluoglu, C. (2018). Computing response similarity indices for multiple-choice tests (CopyDetect) (Version 1.3) [Computer software manual]. https://cran.r-project.org/web/packages/CopyDetect/index.html