Using Data Forensics to Detect Potential Test Fraud

Association of Test Publishers

Authored by the Test Content Infringement & Fraud Subcommittee of the ATP Security Committee

April 2025

Published by the

Association of Test Publishers

601 Pennsylvania Ave., N.W., Suite 900

Washington D.C. 20004

Suggested Citation:

Association of Test Publishers (2025). Using data forensics to detect potential test fraud.

https://www.testpublishers.org/white-papers

CONTENTS

PREFACE

This White Paper offers insights on how data forensics can help detect and mitigate potential fraud in testing programs. It is aimed at assessment professionals, leaders of organizations that develop and deliver tests, companies and other groups that hire testing professionals, and other test and exam stakeholders. This paper does not require statistical knowledge from its readers and does not contain any confidential or proprietary information.

Data forensics is a set of statistical techniques used to detect anomalies in test-taking behavior that may suggest test fraud. For example, data forensics can reveal instances in which test-takers have the same responses to test items, suggesting copying or collusion. While data forensics cannot directly prove prohibited behaviors or cheating, it can highlight anomalies to review for potential fraud.

The validity of test scores depends on test security. Breaches in test security undermine the integrity of testing programs and can have real-world impacts in areas ranging from classrooms to airline cockpits, to nuclear reactor control rooms, to operating rooms. As a result, data forensics complements other methods for protecting the integrity of testing programs and their stakeholders.

This paper is authored by the ATP Test Content Infringement and Fraud (TCIF) Subcommittee of the Association of Test Publishers (ATP) Security Committee, which aims to bolster the assessment community's efforts to enhance test security and integrity.

Overview of Sections:

- Section 1: Introduces data forensics and common forms of test fraud.
- Section 2: Explores data forensics, how it works, and its benefits.
- Section 3: Emphasizes need for policies and procedures.
- Section 4: Highlights potential future advances in data forensics.

There is also an Appendix with real-life examples and a comprehensive set of references.

Acknowledgements

The ATP wishes to acknowledge the following individuals who served on the initiative subcommittee and contributed to the white paper:

Authors (in alphabetical order by first name)

Amy Mann	(College Board)
Anna Rubin	(Cisco)
Jim Hussey	(ACT)
John Kleeman	(Learnosity)
John Weiner	(JW Advisers)
Kim Cohen	(ISACA)
Kirk Becker	(Pearson Vue)
Regi Mucino	(PSI)
Sarah Morrisey	(ISACA)
Sarah Toton	(Caveon)
Steve Addicott	(Caveon)

ATP Security Subcommittee on Test Content Infringement and Fraud (TCIF) Chairs:

Co-Chair:
Kim Cohen
ISACA

Co-Chair:
Jarret M. Dyer
College of DuPage

The Chairs would like to extend special thanks to John Kleeman for his leadership, organization and dedication to this project.

1. Introducing Data Forensics and Test Fraud

Data forensics uses statistical methods to detect and analyze abnormalities in test-taker patterns that may indicate different types of test fraud. Using data gathered during test administrations, statistical analyses can be conducted to model expected patterns and pinpoint outliers. Among the anomalies that can be detected through data forensics are unusual erasure patterns for paper-based testing, unexpectedly short response times for computer-based testing, unlikely similarities in item responses across test-takers or within groups, and significant changes in test scores between attempts. These and other anomalies are linked to different types of fraud, as described in the section below.

Combined with contextual information such as unusual data being observed at specific test centers or schools, trends may also be detected in certain locations, which can then prompt investigations into group collusion and other widespread fraudulent testing activities.

Test Fraud

Test fraud includes any actions that can unfairly advantage an individual test-taker or a group of test-takers, calling into question the validity of the scores those test-takers receive and, on a wider scale, potentially the ability of the exam to validate an individual's skills or knowledge.

When thinking about test fraud, it is important to consider the full range of potential threats. Test fraud is often thought of as "cheating" to earn a higher score but can also include illicit activities such as stealing test content.

This white paper focuses on the threats that can be detected through data forensics. These include:

- pre-knowledge of test content,
- proxy testing,
- copying,
- stealing test items,
- collusion with test center staff, and
- use of unauthorized resources, people, or tools (e.g., an AI model, another person, a smartphone) during the test.

Guidelines for Technology-Based Assessment (2022, Chapter 8), published jointly by ATP and the International Test Commission, provides an expanded discussion of score validity threats.

Test fraud threatens exam validity, the financial investments made in testing programs by credentialing and testing organizations, individual opportunities, institutional reputations, and the larger public if unqualified – and potentially unsafe – individuals are fraudulently certified.

It is important to note that, by itself, data forensics does not necessarily confirm "cheating," as "cheating" encompasses a variety of behaviors and implies malicious intent. While data

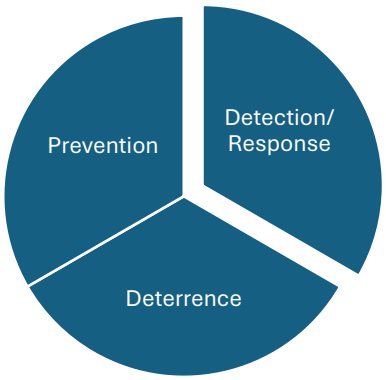
forensics can detect anomalous patterns, it must be supplemented with other information to draw conclusions about behaviors or intentions.

For example, test-takers may exhibit behaviors that suggest pre-knowledge of exam content, but data alone does not provide direct insight into the motivations or thought processes that occur during an exam. It is important to have additional resources, including both plans and investigators (e.g., contingency plans and investigative teams), to follow up on detected abnormalities. These investigators should be ready and equipped to conduct site audits, employ secret shoppers, review video taken during administrations and, in the case of live, remotely proctored administrations, to review background application logs.

Testing organizations must also inform test-takers of their test security policies as a form of individual deterrence and institutional safeguarding. At a minimum, policies should include rigorous terms of use, clearly stated ethical expectations for test-takers, and easy-to-access resources for whistleblowers.

How Data Forensics fits into a testing program

Guidelines for Technology-Based Assessment suggests that a testing program needs security solutions to prevent, deter and detect/respond. Data forensics provides *post hoc* detection of unusual patterns and current methods are not preventive. Data forensics is an effective tool for detection/response and can help with deterrence.



Data forensics should be integrated into testing programs alongside other security measures derived from risk analysis and test security concerns and metrics. Other security measures might include schedules for test item and form development and deployment, secure item banking, length/overlap[1] of administration windows, test-taker identification (e.g., traditional IDs or biometrics), delivery platforms (e.g., lockdown browsers), physical security (e.g., proctoring, onsite visits, securing the testing environment), as well as policies and procedures for responding to test security incidents and for deterrence.

Testing organizations should conduct regular test security audits to comprehensively evaluate their programs' security strengths and weaknesses, including how data forensics currently supports and could further enhance their test security measures.

[1] With overlapping or longer test event windows, some test-takers might be able to obtain copies of test items and share them with test-takers who have not yet taken the same test.

2. Data Forensics Methods

Data forensics methods are used at a variety of levels – for “people” at the individual and group levels, and for “programs” at the item, form, and exam levels. Data forensics methods may share foundational models with psychometrics, but their objectives, specific methodologies, and interpretations often differ.

The feasibility of a potential data forensics analysis depends on its goals, combined with factors such as test length and sample size, since the baseline data must be both sufficient and of high quality. Extremely short multiple-choice tests may not allow for statistically significant findings. Similarly, small sample sizes may not provide enough normative data to confidently interpret what appear to be unexpected or unusual results.

For example, if an analysis seeks to determine whether specific companies have employees engaged in prohibited behavior, but only a few employees from each of those companies have been tested, the sample size for each group (i.e., the number of employees from each company) may be insufficient to support making conclusions. In short, if the goals of the analysis are not aligned with the appropriate and available data, the use of the findings may be limited.

Many considerations go into the design of data forensics analyses, including the “window” of data to be analyzed, if historical data will be used in estimating the models, the frequency of analysis, and other specific methods to be used. Analyses should be customized based on test design, primary test security threats and vulnerabilities, the data that are available, and the needs of the testing programs.

Categories of data forensic methods include:

- Response similarities across test-takers.
- Item-level response times and total test times for computer-based examinations,
- Consistency of performance over item subsets,
- Consistency of performance over time,
- Number and nature of answer changes, and
- Unusual response patterns.

Combinations of these methods can also be useful. For example, a test-taker flagged for extremely fast responses may not be particularly interesting, but if that test-taker also had significantly higher performance on scored items than on unscored “field test” items and belonged to a group that is consistently flagged with similar anomalies, the concern may be great enough to merit review or investigation.

The strength of these methods differs by the category and characteristics of each statistic used. In general, the strongest statistics control for as many relevant factors possible and have no plausible alternative explanations outside a violation of test security.

Additionally, statistics should be considered in the context of performance. Unusual patterns are generally associated with lower performance (e.g., they may reflect test-taker fatigue,

illness, or guessing). When unusual patterns are associated with higher performance the potential risk associated with test fraud is elevated.

Response similarity
Response similarity can be conceptualized as matching responses across test-takers. These indices are designed to detect possible collusion or pre-knowledge of the test content and answers.

For single-select multiple-choice items, generally only one response is correct, and all the other responses are incorrect. Thus, two or more test-takers having identical incorrect responses, especially across multiple items, is unlikely -- barring consistently popular distractors or improbably similar patterns of guessing.

Test-taker response patterns from the same test centers, geographic locations, testing windows, or across the life of an exam can be used to determine the statistical probability of those patterns. Response similarity analyses focus on which items were answered correctly, which were answered incorrectly, and which incorrect responses were selected. The degree of similarity discovered for each of these factors can provide sufficient evidence to consider whether further investigation and potential disciplinary action may be warranted.

Item and test times
For online or computer-based testing, response times can be useful forensic data for detecting possible proxy testing, pre-knowledge, or test content exposure. How long does it take a test-taker to read and digest an item, consider the available responses, and select a response? In part, these answers depend on the test-takers, including their reading abilities and familiarity with the test content. The analysis also depends on whether the test-takers selected correct or incorrect responses. Some items may take less time to answer than others because they are shorter, simpler, or involve less higher-order thinking. In a typical testing session, the time spent on items, individually or in total, varies greatly.

So, when are response times unusual enough to warrant additional review?

Extremely fast responses, combined with high scores, may warrant additional attention and evaluation. For example, if a test-taker answers 90 percent or more of items correctly in less than 10 minutes for a test that typically takes an hour to complete, a flag should likely be raised. Or, if a test-taker spends less than 5 seconds on 30 percent or more of the test items, this could indicate purposeful skipping through test content. The thresholds for acting vary based on the factors being analyzed, and the risk levels the testing organization is willing to tolerate.

Not only are extremely fast responses concerning, but so are response times that do not follow expected patterns – for example, spending consistent times on every item, regardless of difficulty or complexity. Response times that do not align with typical patterns can indicate pre-knowledge, the use of AI or web search tools, and other threats to test security.

Consistency of performance over item subsets
Some items may be more likely to be compromised than others. For example, older items may be more likely to be compromised than newer items because of increased exposure.

New test forms are published with either 100 percent new content, or some percentages of new and current (or “resting”) content. The percentage of items a test-taker correctly answers with new content, compared to the percentage answered correctly with old content, can potentially identify test-takers who are familiar with previously used content. For example, if a test-taker answers 95 percent of the old content correctly but only 10 percent of the new content, it warrants a closer look.

Similar analyses can be performed for any relevant item subset (new vs old, scored vs unscored, items with higher vs lower exposure rates for CAT (Computer Adaptive Testing) / LOFT (Linear on the Fly Testing), equating items vs not, etc.). It is best practice to account for the number or percentage of items in each item subset and the distribution of that content across the exam when determining thresholds for action.

Consistency of performance over time
Performance over time can be useful information to detect potential item harvesting or general exposure of content. Large swings in performance may suggest test security concerns, especially if concentrated within a particular group. Both score gains and losses may be of interest, since current or previous scores could be untrustworthy.

Answer changes
For paper-based tests, erasure statistics can assess if there are unusual numbers of wrong-to-right answer changes, or more wrong-to-right answer changes than other types of answer changes – i.e., right-to-wrong or wrong-to-wrong. For computer-based tests, these are referred to as “answer change statistics.” It is important to note the number of answer changes generally differs by delivery mode, with answer changes rarer for paper-based than computer-based testing. There are several test fraud scenarios associated with abnormally high frequencies of answer changes that, in turn, may relate to test-takers or other motivated stakeholders having access to the test key.

Unusual response patterns
Person-fit or “aberrance” statistics measure responses that do not fit the selected psychometric model or expected pattern of test responses. For example, when test-takers with high total scores get the answers wrong on easy questions, it may indicate potential pre-knowledge or access to a purported or quasi scoring key. Typically, unusual responses are associated with lower performance. Potential causes include uneven content knowledge, guessing, and fatigue, but also pre-knowledge with respect to higher performance. At the individual level, these statistics are not strong enough to recommend action, but they can provide contextual information and can also be useful at the group level.

Other data
There are other useful data that do not fit into the categories above. These include, but are not limited to:

- Test-taker characteristics (e.g., how many times the test-taker previously tested, their employer, where they trained)
- Identification (e.g., ID type, keystroke logs, evaluation of photo IDs or videos for similar backgrounds)
- Mismatched locations (e.g., inconsistent countries of residence and test site)

- Policy violations (e.g., violations of the retake policy or testing after normal site hours)
- Registration anomalies (e.g., same email address used for many test-taker accounts, or an older adult taking a test intended for high school students)
- Test session monitoring including that of computer processes.

Some other information like biometrics, facial comparison or monitoring test-taker behavior may be possible for some testing organizations, but there are significant privacy and in some cases regulatory issues to consider if capturing/processing such data.

Power of multiple factors
While these methods often stand alone in providing valuable forensic data, combining multiple data forensics methods in an investigation can provide powerful support for action. Layers of data provide a strong foundation for inferring the likely presence of test security violations . Still, decisions to use any data forensic methods, alone or in concert, to invalidate, challenge or suspend a test-taker's score should be made with legal counsel and psychometric support.

Group analyses
Data forensics performed at the individual level can be aggregated to conduct group-level analyses. The methods that are strongest at the individual level are also the strongest at the group level. Moreover, methods that may not be strong enough to recommend individual actions may be strong enough to recommend group-level actions. For example, if a group of test-takers that studied at a particular location consistently performs better on older items used in previous tests, that may merit investigation for possible collusion or a breach of test content.

Groups of interest
Groups of interest will vary by testing program. For K-12 education, groups of interest often include school, classroom, teacher, district, accommodations, and grade-subject. For certification programs, groups may include company of employment, first-time attempts vs retakes, training location, geographic location, IP address, mode of administration, test site, and proctor. Test performance and unusual patterns, as measured using the methods described above, provide opportunities for robust analyses of potential test security issues.

Analyses of exams and items for disclosed content
In addition to assessing individuals and groups of individuals, data forensics can assess the security of exams, forms, and even individual items.

Item compromise, whether sharing within groups or through other disclosures, can cause items to become easier and potentially less discriminating over time. Changes in pass rates, especially for identifiable subgroups, may also indicate disclosed content or exposure.

Each of the methods above can be combined with performance information to assess if anomalies are associated with better performance. Findings that indicate potentially disclosed items can be used to retire items or to guide other analyses.

3. Operational Policies and Procedures

When developing a data forensics program, testing organizations would be well-advised to first consider the types of security threats and test fraud the program is intended to mitigate. From there, appropriate data forensic methods can be designed, and their uses specified, as a part of the test security program.

Cizek (1999) recommends that testing organizations establish goals for their data forensics programs and that they evaluate whether their policies and procedures support those goals. These policies and procedures should reflect:

- The types of analyses to be conducted.
- The thresholds or levels of statistical results that might trigger actions.
- What those triggered actions would include.

The uses of data forensics results can vary considerably. In most cases the ultimate uses are driven by a testing program's policies and procedures – for mature programs, as those policies have already been adopted and implemented, and for newer programs, as policies and procedures are under consideration.

Broad vs. Deep

In general, data forensics analyses may be broken into two categories – broad scans and deep investigative analyses.

A broad scan is like an annual health checkup. Perhaps no malady has been identified, but there is enough concern to have an image created and reviewed by an expert to identify undetected threats. Like patients visiting medical professionals, this is also the perspective of testing organizations in scanning their data at least annually to identify and measure potential problems so they can better protect their testing programs and results.

A deep investigative analysis may be prescribed when a medical professional suspects a specific illness or ailment – or when a testing organization detects a specific anomaly. In the medical context, an MRI may identify whether a condition exists, its potential impact, and help inform an appropriate remedy. The same concepts apply with data forensics: analysts sift through the data and focus on worrisome situations, locations, or test-takers.

Having identified focuses for their analyses, testing organizations need to consider what actions they should take based on the results and their Test-Taker Agreement. Potential actions to be included in a written threat policy and procedures document include, but are not limited to:

- Formal Warnings – In many cases, these also require some form of explanation.
- Score Invalidations – With score invalidations, it is recommended that testing organizations focus communications on test results, not test-taker behaviors. Rather than identifying a test-taker as a "cheater," simply indicate the test result is "indeterminate" and requires a retest. Some programs choose to have an appeal process in place.

- Revocation and/or Suspensions – If a result is received by the test-taker before an analysis is conducted and/or an investigation is resolved, some programs will revoke or suspend the result and anything associated with it (e.g., certificate or credential). As with score invalidations, there is often an appeal process available.
- Formal Interviews – Usually involving a group (e.g., several individuals from a school or test center), a program may formally interview test-takers or test center staff to learn more about unusual test instances.
- Security Investigations – In more severe cases, a program may move to an actual investigation. It is recommended that investigation protocols be implemented in concert with the data forensics program.
- Program Responses – Appropriate responses may include adjusting the exam or its processes by refurbishing, retiring, or replacing test items, republishing test forms, eliminating identified security vulnerabilities, updating policies, revising training, and initiating site monitoring.
- Additional Recommendations – Some situations may require the involvement of a program board or similar oversight body to levy other types of sanctions.

Deterrence is a useful aspect of test security programs. Testing organizations should communicate potential actions and sanctions to test-takers in advance, including in the Test-Taker Agreement, as this awareness can reduce the amount of cheating by making test-takers consider its consequences.

It is critical to maintain procedures that deliver equitable test-taker experiences, while also efficiently identifying test-taking concerns. A two-tiered approach might start with a broad scan to assess the integrity of the overall administration, which may then move to deeper investigative analyses of test-taking data that suggest specific concerns.

Whatever policies and sanctions a testing organization adopts and implements, it is crucial that its data forensics and test security programs are communicated clearly, and administered and executed consistently, for all test-takers.

There are legal examples where a test-taker has claimed in court that the sanction they received, based on data forensics results, was “arbitrary and capricious.” However, because the testing organization had communicated, administered, and executed its policies consistently and uniformly for years, the court ruled in favor of the testing organization.

To help ensure consistency, program managers are advised to develop and use decision-criteria matrices to support the actions they take. Maintaining matrices allows for equitable flagging across all test-takers and supports consistency with the resulting responses. Criteria matrices should be regularly reviewed and refined to ensure organizations capture new threats as bad actors develop new means to gain an unfair advantages.

As cases are resolved, it is also important for programs to track historical outcomes and their metrics. To extend the earlier medical metaphor, key data can help inform decisions and enhance the health of future administrations. Metrics are critical in measuring program performance and integrity, the effectiveness of security measures, and the health and validity of items and forms. Metrics help program managers identify trends, anomalies, and areas of concern. Strong metrics inform business decisions that provide for better customer experiences and sustained testing integrity.

A Data Forensics Case Study: Response Similarity Reveals Exposure

Consider a short example of a real-life use of data forensics. There are brief descriptions of other real-life examples in the Appendix from different program types.

During a “health check” of one of its exams, an IT organization identified a cluster of test-takers with similar response patterns. A Response Similarity Index (RSI) analysis revealed many of the test-takers in this cluster had highly matching responses even though they only answered 70 percent of the exam questions correctly.

Through data forensics, this cluster of 100 test-takers was flagged as having paired examinations with statistically improbable numbers of matching responses. The likelihood of these pairs providing such similar responses independently or without pre-knowledge of the test items was 10^{-9}, or less than one in a billion.

These results prompted the organization’s security team to investigate. They discovered this cluster of test-takers worked in the same location and had organized a large study group to help prepare for the exam. The group studied using a harvested copy of the test questions and answers, and many of the answers had been marked incorrectly.

Data forensics provided the evidence needed to not only identify the test-takers who had pre-knowledge, but to connect the test-taker who initially distributed the illicit material within the study group with the person who had previously harvested the test content in a different country.

4. The Future

Technology is rapidly reshaping assessment, resulting in new challenges and the need to advance security analytics and data forensics. As assessment evolves, so too will the issues addressed by forensic analyses.

For example:

- Artificial Intelligence and Machine Learning-based algorithms will leverage expanded, integrated databases from multiple sources ("Big Data"), enabling more robust indices of anomalies in testing, as well as new forensic methods and metrics for detecting cheating and content exposure, providing that compliance is maintained with AI and privacy regulation.

  Test-takers' digital footprints are expanding beyond basic response data (e.g., time, similarity) to include data such as test event records flagged in proctoring systems (both human and AI), system activity reports indicating users sought outside help, and richer response pattern data modeled to detect anomalies and potential cheating.

- Generative AI (GenAI) will be a two-edged sword. On one side, GenAI is already being built into educational programs and the workplace. As its use becomes increasingly accepted, we can expect GenAI to be incorporated into assessments, with forensic methods needing to adapt to new definitions of "cheating."

  Conversely, certain fraudulent or disallowed uses of GenAI will require new definitions of "anomalies" that forensic methods must then detect.

- In the longer term, high-stakes assessments will change as written multiple-choice tests are supplemented or replaced by increased use of technology in performance-based assessments using constructed responses and simulations (which may include the use of GenAI).

  These changes, in turn, will spur the development of new forensic methods and metrics.

While it is always difficult to predict the future, we can expect the continued transformation of testing, accompanied by a corresponding evolution in our forensic methods. The Test Content Infringement & Fraud Subcommittee of the ATP Security Committee plans to produce further documents as technology and practice develops.

Appendix: Real-Life Examples

The following chart describes how several testing organizations have discovered or investigated test security issues through data forensics programs and used the results to take action. Some rows represent generalizations created using multiple real-life examples. While these examples represent results from specific types of programs, the problems identified may be found in all types of assessment environments.

Program Type(s)	**Impetus and Data Forensics Findings**	**Outcomes and Actions Taken**
IT Certification, Professional Certification, and Medical Board	Routine data forensics analysis identified several very large clusters of test instances with extremely similar responses, matching even on several rare, incorrect answers. There were also large clusters of identical test instances with particular responses selected. These did not appear to be concentrated within any group (e.g., company, geographic region, test site). In some cases, proctors have reported test-takers accessing cell phones, smart watches, or other devices during the exam.	Test-taker scores were invalidated and continue to be based on routine data forensics monitoring. Details of the large-scale invalidations were published in a press release. Further investigation identified several brain dumps (i.e., documents containing live test items resulting from “item harvesting” by test-takers) available on the internet in private groups. The brain dump items were later also found in documents from authorized test preparation providers. The program retired compromised items, republished affected forms, repurposed the compromised items for practice exams, and implemented increased messaging around the ethical and behavioral expectations of test-takers (e.g., clarifying authorized vs. prohibited test preparation materials).
IT Certification and Professional Certification	Data forensics flags identified clusters of test instances with extremely similar responses, fast response times, shared email addresses, and/or with individual test-takers showing multiple levels of ability on the same skillset.	Review of videos showed the same individual testing for multiple test-takers with the same background location. The scores were invalidated, and the proxy test-taker was banned from future testing. In some cases, proxy testers have submitted tips to the testing organization or employer that a test-taker used the proxy service but did not pay and did not deserve the certification. In others, test-takers have been blackmailed by proxy testers who have evidence the test-taker of record was not truly taking the test.

IT Certification and Professional Certification	A web crawling service discovered a brain dump of an exam (i.e., document containing live test items). The brain dump contained the items in a particular order. Additionally, pretest items were administered from a pool for this program, so the brain dump contained a particular subset of pretest items. Data forensic analyses estimated the probability of the items being administered in that order with that particular subset of pretest items. It was exceedingly rare and only one test-taker in the data group had that item order and those pretest items.	The item harvester was identified, his/hers core was invalidated, the test-taker was banned from future testing, and he/she was charged with copyright infringement. A DMCA notice was sent to the website, and the brain dump was taken down. The testing program was able to compare test-takers' responses to the known brain dump to assess access to it and focus item refurbishment efforts on disclosed items.
IT Certification	A testing program piloted remote proctoring. Within a few hours there were many flags for extremely similar, or even identical, responses and very fast response times on both fixed form and computer-adaptive testing (CAT) exams. These flags did not appear to be concentrated within any group (e.g., company, geographic region, test site).	Scores from the remote proctoring pilot were cancelled. The testing program decided to increase the size and health of its forms and item pools, explore new item designs, and administer through brick-and-mortar sites. Later, the program implemented remote proctoring with deeper and more diversified item pools for some geographic regions but not others. It regularly uses data forensics results to select items for refurbishment or retirement.
Medical Board	A test site was flagged with a high rate of test-takers performing better on older than on newer items. This site primarily served one medical residency program and was detected consistently.	An investigation discovered that former medical residents started a review course and encouraged students to memorize items and share them with the review course founders. The review course founders had their certifications revoked, scores for test-takers who used the review course were invalidated (across a few test sites), and a new test was published.

Medical Board	A pair of test-takers was flagged with extremely similar responses. Both test-takers were from the same medical residency program and tested on the first and last day of the testing window, respectively. The test-taker who tested later had much better performance on items administered to the first test-taker than on other items.	An investigation into the test-takers' social media discovered that the test-takers had a close personal relationship. Despite their claims that they studied together, the statistical findings were strong enough that the test-takers' scores were invalidated, and the case was submitted to the ethics board for additional investigation.
Medical Board	A proctor at a test site discovered notes in a bathroom. Data forensics analyses assessed answer changes before and after breaks for test-takers in that test site with access to that bathroom. One test-taker took many breaks and changed many answers, with a large positive score gain resulting from those changes.	The test-taker's score was invalidated. Proctor training was supplemented to include more thorough monitoring of unscheduled breaks and bathroom environments. Keystroke analyses were implemented for test-takers with many unscheduled breaks and/or who were observed accessing personal belongings.
State Department of Education	A tip was received that students appeared to have suspiciously high scores given their reading levels. Answer sheets and test booklets were analyzed. There were high rates of similarity in their responses. There were large score gains for these students compared to the year before. There were high rates of wrong-to-right erasures in the test booklets, but not the answer sheets. These flags were associated with a few teachers and dates.	The teachers, school, and principal were investigated for coaching and/or answer-tampering. Three educators (one teacher, one vice principal, and one principal) had their teaching certificates revoked for 1-20 years.
State Department of Education	There were high rates of similarity in the responses for students within a school. There appeared to be many extremely similar pairs and a few larger clusters within classrooms.	Student scores were invalidated. An investigation found that proctors were ineffectively maintaining a secure testing environment – for example, instead of active monitoring, they were reading books or leaving the testing room in some cases. The test administration manual and proctor trainings were revised, and site monitors were deployed in those schools the next year.

State Department of Education	There were high rates of fast response times, and low scores observed widely in a particular district.	Engagement interventions and messaging on the importance of doing your best on the test were implemented in this district for both students and educators.

References

References: General

Cizek, G. J. (1999) *Cheating on tests: how to do it, detect it, and prevent it.* Mahwah, New Jersey: Routledge.

Cizek, G. J., & Wollack, J. A. (Eds.). (2017). *Handbook of Quantitative Methods for Detecting Cheating on Tests.* New York, NY: Routledge.

Kingston, N. & Clark, A. (Eds.) (2014). *Test Fraud: Statistical Detection and Methodology.* Mahwah, New Jersey: Routledge.

Wollack, J. A. & Fremer, J. J. (Eds.) (2013). *Handbook of Test Security.* Mahwah, New Jersey: Routledge.

References: Response similarity

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. Journal of the American Statistical Association, 69 (345), pp. 44-49.

Becker, K. A., & Meng, H. (2022). Identifying statistically actionable collusion in remote proctored exams. Journal of Applied Testing Technology, 22(2), 1-8.

Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error similarity analysis. Teaching of Psychology, 16(3), 151-155. https://doi.org/10.1207/ s15328023top1603_15

Belov, D. I. (2013). Detection of test collusion via Kullbach-Leibler divergence. Journal of Educational Measurement, 50(2), 141-163. http://www.jstor.org/stable/24018104

Bird, C. (1927). The detection of cheating in objective examinations. School and Society, 25 (635), 261-262

Dickenson, H. F. (1945). Identical errors and deception. The Journal of Educational Research, 38(7), 534-542.

Hanson, B., A., Harris, D. J., & Brennan, R. L. (1987) A comparison of several statistical methods for examining allegations of copying. ACT Research Report Series 87-15.

Holland, P. W. (1996). Assessing unusual agreement between the incorrect answers of two examinees using the K-index: statistical theory and empirical support (ETS Technical Rep. No. 96-4). Princeton, NJ: Educational Testing Service.

Hurtz, G.M, & Weiner, J.A. (2019). Analysis of Test-Taker Profiles Across a Suite of Statistical Indices for Detecting the Presence and Impact of Cheating. Journal of Applied Testing Technology, 20(1) 1-15.

Hurtz, G.M., & Weiner, J.A. (2022). Comparability and Integrity of Online Remote Vs. Onsite Proctored Credentialing Exams. Journal of Applied Testing Technology, 23, 36-45.

Maynes, D. (2014). Detection of non-independent test taking by similarity analysis. In N. M. Kingston and A. K. Clark (Eds.) Test fraud: Statistical detection and methodology (pp. 53-82). Routledge.

Maynes, D. (2017). Detecting Potential Collusion Among Individual Examinees Using Similarity Analysis. In G. J. Cizek & J. A. Wollack (Eds.) Handbook of Quantitative Methods for Detecting Cheating on Tests (pp. 47-69). Routledge.

Meng, H., & Ma, Y. (2023). Machine Learning–Based Profiling in Test Cheating Detection. Educational Measurement: Issues and Practice, 42(1), 59-75.

Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. Journal of Educational Measurement, 40(1), 53-69.

Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2006). Detecting answer copying using the kappa statistic. Applied Psychological Measurement, 30(5), 412-431.

Weiner, J.A., & Hurtz, G.M (2017). A Comparative Study of Online Remote Proctored versus Onsite Proctored High-Stakes Exams. Journal of Applied Testing Technology, Vol 18(1), 13-20.

Wollack, J. A. (1997). A nominal response model approach to detect answer copying. Applied Psychological Measurement, 21(4), 307-320. https://doi.org/10.1177%2F01466216970214002

Zopluoglu, C. (2017). Similarity, answer copying, and aberrance-understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), Handbook of quantitative methods for detecting cheating on tests (pp. 25-46). Routledge.

Zopluoglu, C. (2019). Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (XGBoost). Educational and Psychological Measurement, 79(5), 931-961.

References: Item and test times

Fox, J. P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. Journal of Educational Measurement, 54(2), 243-262.

Hurtz, G. M. & Mucino, R. (2024). Expanding the Lognormal Response Time Model Using Profile Similarity Metrics to Improve the Detection of Anomalous Testing Behavior. Journal of Educational Measurement. DOI: 10.1111/jedm.12395

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.

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.

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

Sinharay, S. (2018). A new person-fit statistic for the lognormal model for response times. Journal of Educational Measurement, 55(4), 457-476.

Sinharay, S. (2020). Detection of item preknowledge using response times. Applied Psychological Measurement, 44(5), 376-392.

Toton, S., & Maynes, D. (2019). Detecting Examinees With Pre-knowledge in Experimental Data Using Conditional Scaling of Response Times. *Frontiers in Education, 4,* 1-18.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. Journal of Educational Measurement, 46(3), 247-272.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. Psychometrika, 73(3), 365-384.

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.

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.

References: Consistency of performance over item subsets

Belov, D. (2019, October). Detection of Item Preknowledge by Exploiting Uncompromised Subset of Items. Paper presented at the annual meeting of the Conference on Test Security, Miami, FL.

References: Consistency of performance over time

Bishop, S., & Egan, K. (2017). Detecting Erasures and Unusual Gain Scores: Understanding the Status Quo. In G. J. Cizek & J. A. Wollack (Eds.) Handbook of Quantitative Methods for Detecting Cheating on Tests (pp. 193-213), Routledge.

References: Answer changes

Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting test tampering using item response theory. Educational and Psychological Measurement, 75(6), 931-953.

Sinharay S., & Johnson, M.S. (2017). Three New Methods for Analysis of Answer Changes. Educational and Psychological Measurement, 77(1), 54-81. doi: 10.1177/0013164416632287. PMC5965521.

References: Unusual response patterns

Egberink, I. J., Meijer, R. R., Veldkamp, B. P., Schakel, L., & Smid, N. G. (2010). Detection of aberrant item score patterns in computerized adaptive testing: An empirical example using the CUSUM. Personality and Individual Differences, 48(8), 921-925.

Lewis, C., Lee, Y-H, & von Davier, A. A. (2014). Test Security for Multistage Tests: A Quality Control Perspective. In N. M. Kingston & A. Clark (Eds.) Test Fraud (pp. 230-238). Routledge.

Man, K., & Harring, J. R. (2021). Assessing preknowledge cheating via innovative measures: A multiple-group analysis of jointly modeling item responses, response times, and visual fixation counts. Educational and Psychological Measurement, 81(3), 441-465.

Meijer R. R., van Krimpen-Stoop E. M. L. A. (2010). Detecting person misfit in adaptive testing. In van der Linden W. J., Glas C. A. (Eds.), Elements of adaptive testing (pp. 315–329). Springer.

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

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.

van Krimpen-Stoop, E. M., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In Computerized adaptive testing: Theory and practice (pp. 201-219). Dordrecht: Springer Netherlands.

References: Analyses of exams and items for disclosed content

DeMars, C.E. (2004). Detection of item parameter drift over multiple test administrations. Applied Measurement in Education, 17(3), 265-300.

Han, N. (2003). Using moving averages to assess test and item security in computer-based testing. (Research Report No. 468). University of Massachusetts, School of Education, Center for Educational Assessment.

Han, N., & Hambleton, R. (2004, April). Detecting exposed test items in computer-based testing. Paper presented at the National Council on Measurement in Education, San Diego, CA. http://iacat.org/content/detecting-exposed-test-items-computer-based-testing

O'Leary, L. S., & Smith, R. W. (2017). Detecting candidate preknowledge and compromised content using differential person and item functioning. In G. J. Cizek & J. A. Wollack (Eds.) Handbook of Quantitative Methods for Detecting Cheating on Tests (pp. 151-163). Routledge.