LSAC RESEARCH REPORT SERIES

A New Approach to Detecting Cluster Aberrancy

Dmitry I. Belov

Law School Admission Council
Research Report 16-05
November 2016

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation that provides unique, state-of-the-art products and services to ease the admission process for law schools and their applicants worldwide. Currently, 222 law schools in the United States, Canada, and Australia are members of the Council and benefit from LSAC's services. All law schools approved by the American Bar Association are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also members. Accredited law schools outside of the United States and Canada are eligible for membership at the discretion of the LSAC Board of Trustees; Melbourne Law School, the University of Melbourne is the first LSAC-member law school outside of North America. Many nonmember schools also take advantage of LSAC's services. For all users, LSAC strives to provide the highest quality of products, services, and customer service.

Founded in 1947, the Council is best known for administering the Law School Admission Test (LSAT®), with about 100,000 tests administered annually at testing centers worldwide. LSAC also processes academic credentials for an average of 60,000 law school applicants annually, provides essential software and information for admission offices and applicants, conducts educational conferences for law school professionals and prelaw advisors, sponsors and publishes research, funds diversity and other outreach grant programs, and publishes LSAT preparation books and law school guides, among many other services. LSAC electronic applications account for nearly all applications to ABA-approved law schools.

Table of Contents

Executive Summary

This report addresses a general type of cluster aberrancy in which a subgroup of test takers has an unfair advantage on some subset of administered items. Examples of cluster aberrancy include item preknowledge and test collusion. In general, cluster aberrancy is hard to detect due to the multiple unknowns involved: Unknown subgroups of test takers have an unfair advantage on unknown subsets of items. The issue of multiple unknowns makes the detection of cluster aberrancy a challenging problem from the standpoint of applied mathematics. This report presents a novel algorithm to detect cluster aberrancy. The algorithm is general and applicable to all types of testing programs: paper-and-pencil testing, computer-based testing, multistage testing, and computerized adaptive testing; it can also be applied in areas outside of psychometrics, such as finance (e.g., detecting financial fraud). Both simulated and real data were used to study the performance of this algorithm.

Introduction

Cluster aberrancy occurs when a subgroup of test takers (called *aberrant test takers*) has an unfair advantage on some subset of administered items (called *compromised items*). Thus, all test takers and items are partitioned into pairs of clusters connected by the aberrancy (see Figure 1). Assuming that an aberrant test taker does not possess a high ability, each aberrant test taker performs better on compromised items than on uncompromised items, and on each compromised item the aberrant test takers perform better than the nonaberrant population. When the number of aberrant test takers and compromised items is large, the corresponding testing program and its stakeholders (universities, companies, government organizations, etc.) are negatively affected because they are given invalid scores.

Examples of cluster aberrancy include item preknowledge and test collusion. *Item preknowledge* describes a situation in which a subgroup of test takers had access to a subset of items from an administered test prior to the exam. *Test collusion* is the sharing of test materials or answers to test questions, where the source of the shared information could be a teacher, a test-preparation company, the Internet, or test takers communicating on the day of the exam (Wollack & Maynes, 2011).

Assuming there is only one subgroup of aberrant test takers with a corresponding subset of compromised items, it is easy to show that if the compromised items are known, one can detect[1]

[1] A collection of detectors was analyzed recently by Belov (2016); all of these detectors can be modified to detect compromised items.

the aberrant test takers and vice versa. For example, the final log odds ratio (FLOR) by McLeod, Lewis, and Thissen (1999) can detect aberrant test takers given a probability of compromise for each item; conversely FLOR can be transformed to detect compromised items as well (McLeod & Schnipke, 2006). However, the assumption of only one subgroup of aberrant test takers and only one subset of compromised items is not realistic. Three approaches to addressing this issue can be distinguished:

- The first approach is to use a statistic that does not explicitly take into account information about compromised items, such as the lz statistic (Drasgow, Levine, & Williams, 1985). Multiple parametric and nonparametric statistics (Karabatsos, 2003) can be applied; however, their power is usually low. The CUSUM method (van Krimpen-Stoop & Meijer, 2001) performs well only when compromised items are positioned sequentially in the test (Tendeiro & Meijer, 2012). Cluster analysis (Wollack & Maynes, 2011) and factor analysis (Zhang, Searcy, & Horn, 2011) were applied to detect item preknowledge; however, both methods rely on the number of response matches, which is not applicable to multistage testing (MST) and computerized adaptive testing (CAT), where the actual test varies across test takers (e.g., the items administered and possibly the test length).
- The second approach is to exploit new data recently available during the test. For example, response time (used in computer-based testing [CBT], MST, and CAT) is currently considered to be the "holy grail" for detecting item preknowledge. It is commonly assumed that if a test taker has preknowledge of an item, he or she will respond unusually quickly to this item. However, as pointed out by Stone (2016), if a test taker with item preknowledge is aware of this type of monitoring, he or she may deliberately take longer to answer the item.
- The third approach is to solve the actual problem by simultaneously detecting subgroups of aberrant test takers and corresponding subsets of compromised items. Belov (2014) demonstrated that this is feasible under certain assumptions. His method (the 3D algorithm), which exploits information theory and combinatorial optimization, is applicable to all testing programs, and in CAT simulations it performed substantially better than lz. However, this algorithm is based on the assumption that a group of test takers (e.g., a test center) cannot have more than one subgroup of aberrant test takers. Another major limitation of the 3D algorithm is its dependence on a special subset Q that should contain mostly compromised items. However, many concepts behind the 3D algorithm are very useful in developing a more general and more effective framework, and that is what will be presented in this report.

Cluster aberrancy is hard to detect due to the multiple unknowns involved: Unknown subgroups of test takers have an unfair advantage on unknown subsets of items. In practice, however, a subset of uncompromised items (here it will be called an *uncompromised subset*) is often given, which makes the detection of cluster aberrancy feasible. The existence of an uncompromised subset is crucial for the presented approach to correctly detect compromised items and aberrant test takers. Let us consider two major cases of cluster aberrancy:

Item preknowledge: In CAT, items that are administered for the first time form an uncompromised subset. An uncompromised subset is represented by a *variable section* in paper-and-pencil (P&P) and CBT, and by a *variable testlet* in MST; in MST and CBT, new items may also be distributed throughout the test as opposed to being grouped all in one section or testlet.

Test collusion: If test takers work together on a common part of the test during the exam (P&P, CBT, and MST), then the variable section can be considered an uncompromised subset. In CAT, if test takers work together on common items, items with a low exposure rate can be considered uncompromised. If test takers get help during the break between sections, then they will try to change answers to items from previous sections; therefore, items without answer changes can be considered uncompromised. Similarly, when a teacher changes answers from incorrect to correct for a group of students, the items without answer changes can be considered uncompromised.

Thus, the central assumption of this report is that for each test taker it is possible to identify a subset of uncompromised items. The approach presented in this report relies on this assumption and may fail when the assumption is violated. Computer simulation demonstrated that the approach is robust to the violation of this assumption, but only up to a point: It is robust for a violation of 15% but not for a violation of 30%. More specifically, it can identify aberrant test takers considerably better than the lz statistic, even when 15% of the items in an uncompromised subset are actually compromised (see the results of computer simulations below).

Conceptually, the presented algorithm has three stages. Stage 1 computes an initial estimate of all aberrant test takers via a Monte Carlo method. Stage 2 performs the following: Fix initial aberrant test takers and detect compromised items, then fix detected compromised items and detect aberrant test takers, then fix detected aberrant test takers and detect compromised items, and so on until convergence. Stage 2 relates to a general process called *alternating minimization.* Psychometricians are familiar with an instance of this process called the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), which is commonly used to estimate latent traits and item parameters. Stage 2 converges to a local extremum containing a

subgroup of aberrant test takers with corresponding compromised items. Stage 3 detects multiple subgroups of aberrant test takers with corresponding subsets of compromised items: Test takers detected at Stage 2 are removed from the initial estimate of all aberrant test takers, and Stage 2 is called if the initial estimate is still not empty; otherwise the algorithm stops.

The next sections formalize the problem, present an analysis of the problem, develop the detection algorithm, perform its analysis using simulated and real data, and conclude with a summary.

Formalization of the Problem

Consider a group of test takers that includes aberrant test takers. The aberrant test takers can be represented by disjoint *aberrant subgroups*, where each aberrant subgroup has an unfair advantage on some *compromised subset* of items (Figure 1).

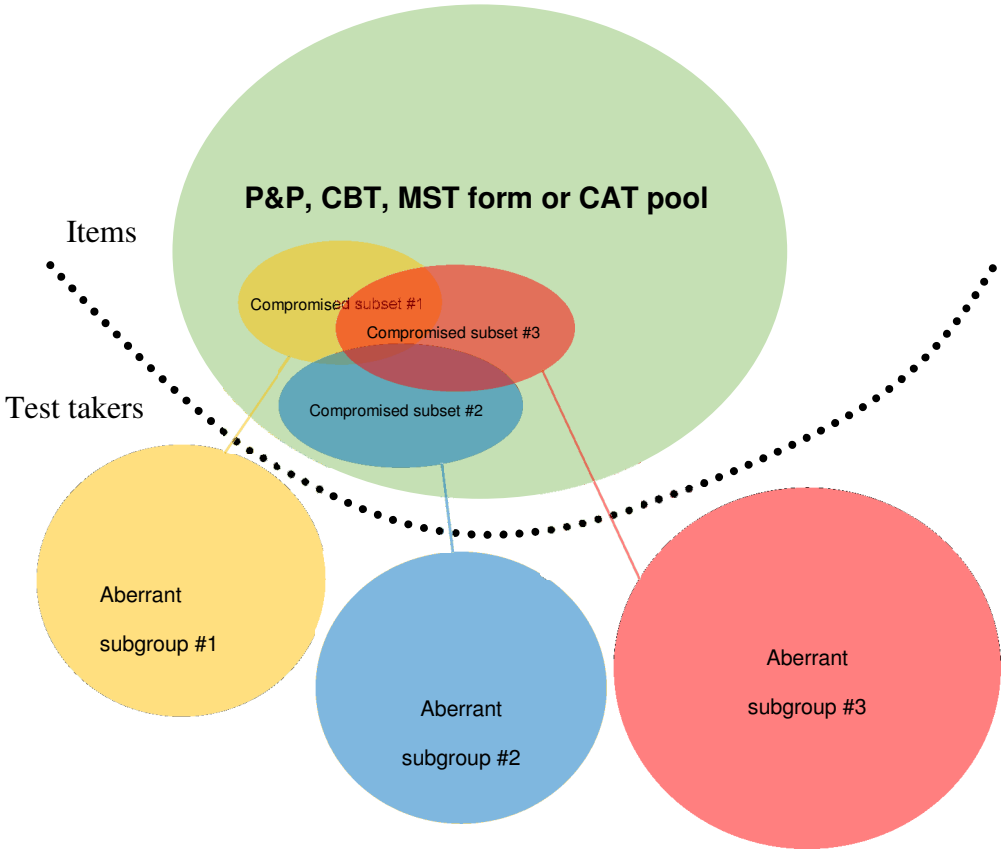

FIGURE 1. *Terms used to describe cluster aberrancy*

Note that while aberrant subgroups are disjoint, compromised subsets are not, suggesting that multiple aberrant subgroups may take unfair advantage of the same items (see Figure 1). The above terms are chosen in order to avoid confusion in the text: *groups* and *subgroups* refer to test takers; *sets* and *subsets* refer to items. Thus, the detection problem addressed in this report is now stated as follows: How does one detect each aberrant subgroup and its corresponding compromised subset?

Alternating Detector of Cluster Aberrancy

The main problem is divided into five subproblems, which are stated and solved in the next subsections. Throughout the report the following notation is used:

- Lowercase letters $a, b, c, \ldots$ denote scalars.
- Lowercase Greek letters $\alpha, \beta, \gamma, \ldots$ denote random variables.
- Capital letters $A, B, C, \ldots$ denote general sets, subsets of items, groups (or subgroups) of test takers, and sequences of scalars; the number of elements in S is denoted by $|S|$.
- Bold capital letters $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$ denote functions.

A test taker j is defined by two random variables: an unobservable latent trait (ability) θ_j and an observable response χ_{ij} to item i. Denote $\mathbf{P}_i(\chi_{ij}|z)$ as the probability of response χ_{ij} to item i conditioned on $\theta_j = z$, where ability level z belongs to a fixed finite sequence of ability levels Z. Consider an arbitrary subset of items I administered to test taker j. Subset I may vary among test takers (e.g., CAT, MST) and should be denoted as I_j; however, without loss of generality, this index can be skipped. Then, Bayes' theorem can be applied to compute the discrete posterior distribution of θ_j with a uniform prior distribution:

$$\mathbf{F}_{I,j}(z) = \frac{\prod_{i \in I} \mathbf{P}_i(\chi_{ij}|z)}{\sum_{y \in Z} \prod_{i \in I} \mathbf{P}_i(\chi_{ij}|y)}, \; z \in Z \, . \tag{1}$$

The posterior $\mathbf{F}_{I,j}(z)$ characterizes how well test taker j performs on the subset of items I. The more $\mathbf{F}_{I,j}(z)$ is shifted to the higher ability, the better the performance. Similarly, one can

estimate how well a subgroup of test takers J performs on item i by computing the following posterior distribution:

$$\mathbf{F}_{i,J}(z) = \frac{\prod_{j \in J} \mathbf{P}_i(\chi_{ij} \mid z)}{\sum_{y \in Z} \prod_{j \in J} \mathbf{P}_i(\chi_{ij} \mid y)}, \; z \in Z . \tag{2}$$

Subgroup J may vary among items (e.g., CAT, MST) and should be denoted as J_i; however, without loss of generality, this index can be skipped.

Measure of Test-Taker Aberrancy

From an aberrant test-taker standpoint, the administered test is partitioned into two disjoint subsets: the first subset with compromised items and the second subset with uncompromised items. Naturally, the aberrant test taker will perform better on the first subset than on the second subset. Therefore, detection of an aberrant test taker may be based on a measurement of performance gain in the first subset relative to the second subset: the larger the gain, the higher the probability that the test taker is aberrant.

Given a compromised subset C, consider a test taker j taking a test I, which can be partitioned into two disjoint subsets $I \cap C$ (compromised items, intersection of the test and the compromised subset) and $I \setminus C$ (uncompromised items, test items without items from the compromised subset).

The following measure, called the posterior shift (PS), is based on the assumption that, for an aberrant test taker, the posterior computed from the responses to $I \cap C$ should be shifted more toward higher ability than the posterior computed from the responses to $I \setminus C$. The PS measures the difference between the two posteriors in the largest right boundary region, where the first posterior is higher than the second posterior (Figure 2). Clearly, the value of the PS is always between 0 and 1; for a given test taker j it is denoted as $\mathbf{S}(\mathbf{F}_{I \cap C, j} \| \mathbf{F}_{I \setminus C, j})$, where posteriors $\mathbf{F}_{I \cap C, j}$ and $\mathbf{F}_{I \setminus C, j}$ are computed by Equation (1). A formal definition of the PS and an analysis of its properties are described in the Appendix.

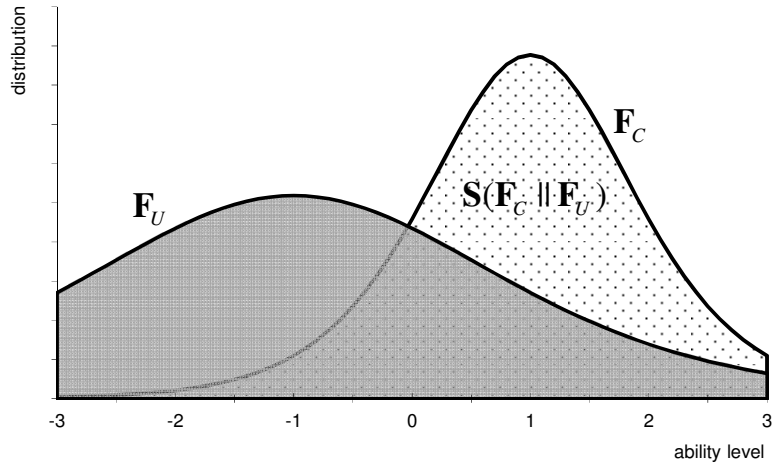

FIGURE 2. *An illustration for the definition of the PS statistic. The dotted, unshaded area corresponds to the posterior shift* $\mathbf{S}(\mathbf{F}_C \| \mathbf{F}_U)$ *between probability mass functions* $\mathbf{F}_C$ *and* $\mathbf{F}_U$*, where the largest right boundary region of ability level on which* $\mathbf{F}_C$ *is higher than* $\mathbf{F}_U$ *is approximately [0,3].*

Measure of Item Compromise

From a compromised item standpoint, all test takers who took that item are partitioned into two disjoint subgroups: the first subgroup with aberrant test takers and the second subgroup with nonaberrant test takers. Naturally, the test takers from the first subgroup will perform better on this item than the test takers from the second subgroup. Therefore, detection of compromised items may be based on some measurement of performance gain by the first subgroup relative to the second subgroup: The larger the gain, the higher the probability that the given item is compromised.

Given a group of test takers J taking item i and aberrant subgroup A, test takers in J can be partitioned into two disjoint subgroups $J \bigcap A$ (aberrant test takers) and $J \setminus A$ (nonaberrant test takers). By analogy with the previous subsection, the measure of item compromise for item i is given by the statistic $\mathbf{S}(\mathbf{F}_{i,J\cap A} \| \mathbf{F}_{i,J\setminus A})$, where posteriors $\mathbf{F}_{i,J\cap A}$ and $\mathbf{F}_{i,J\setminus A}$ are computed by Equation (2).

Subproblem 1

In this subproblem, a compromised subset C is already detected. One then needs to detect a corresponding aberrant subgroup.

One can follow a statistical approach by selecting test takers with unusual performance gains on items from subset C. The measure of test-taker aberrancy (see above) can be interpreted as a statistic computed for each test taker. Then, given a significance level, one can compute the critical value as, for example, a corresponding percentile of the simulated null distribution. Then each test taker for whom the value of the statistic is larger than the critical value is detected as aberrant. Formally stated, given a group of test takers J and a compromised subset C, the statistic $\mathbf{S}(\mathbf{F}_{C\cap I,j} \,\|\, \mathbf{F}_{I\setminus C,j})$ is computed for each test taker $j \in J$. The statistical approach works well when the null distribution fits the real data.

A more general approach to solving Subproblem 1 is based on combinatorial optimization. According to the definition of the measure of item compromise (see above), items from subset C should maximize this measure on the corresponding aberrant subgroup. In other words, Subproblem 1 is solved by searching through all possible subgroups of test takers maximizing an expectation of the measure of item compromise, where the expectation is taken over the compromised items. Formally stated, the following combinatorial optimization problem is to be solved:

$$\arg\max_{A\subset J} \frac{1}{|C|}\sum_{i\in C}\mathbf{S}(\mathbf{F}_{i,J\cap A} \,\|\, \mathbf{F}_{i,J\setminus A})\,. \tag{3}$$

Problem (3) can be solved, for example, by simulated annealing similar to that found in Belov (2014). The convergence can be improved by removing from J all test takers with a measure of aberrancy below a certain critical value, which can be computed from the null distribution given a larger significance level in order to adjust for possible data misfit. Thus, from a combinatorial optimization standpoint, the above two approaches are connected: The statistical approach provides a feasible solution to Problem (3), while an optimal solution to Problem (3) is a feasible solution A that maximizes $\frac{1}{|C|}\sum_{i\in C}\mathbf{S}(\mathbf{F}_{i,J\cap A} \,\|\, \mathbf{F}_{i,J\setminus A})$.

Subproblem 2

In this subproblem, an aberrant subgroup A is already detected. One then needs to detect a corresponding compromised subset.

One can follow a statistical approach by selecting items on which aberrant subgroup A has an unusual performance gain. The measure of item compromise (see above) can be interpreted as a statistic computed for each item. Then, given a significance level, one can compute a critical value as, for example, a corresponding percentile of the simulated null distribution. Then each

item for which the statistic is larger than the critical value is detected as compromised for the corresponding aberrant subgroup. Formally stated, given aberrant subgroup A from a group J, the statistic $\mathbf{S}(\mathbf{F}_{i,J\cap A} \,\|\, \mathbf{F}_{i,J\setminus A})$ is computed for each item $i \in I$. The statistical approach works well when the null distribution fits the real data.

A more general approach to solving Subproblem 2 is based on combinatorial optimization. Due to the definition of the measure of test-taker aberrancy (see above), test takers from subgroup A should maximize this measure on the corresponding compromised subset. In other words, Subproblem 2 is solved by searching through all possible subsets of items maximizing an expectation of the measure of test-taker aberrancy, where the expectation is taken over the aberrant test takers. Formally stated, the following combinatorial optimization problem is to be solved:

$$\arg\max_{C} \frac{1}{|A|} \sum_{j \in A} \mathbf{S}(\mathbf{F}_{C\cap I,j} \,\|\, \mathbf{F}_{I\setminus C,j}) . \tag{4}$$

Belov (2014) successfully solved Problem (4) using simulated annealing. The convergence can be improved by excluding from the search all items with a measure of item compromise below a certain critical value, which can be computed from the null distribution given a larger significance level in order to adjust for possible data misfit. Thus, from a combinatorial optimization standpoint, the above two approaches are connected: The statistical approach provides a feasible solution to Problem (4), while an optimal solution to Problem (4) is a feasible solution C that maximizes $\frac{1}{|A|} \sum_{j \in A} \mathbf{S}(\mathbf{F}_{C\cap I,j} \,\|\, \mathbf{F}_{I\setminus C,j})$.

Subproblem 3

In this subproblem, assuming there is only one aberrant subgroup with a corresponding compromised subset, one has to detect them both.

This report exploits the idea of alternating minimization (Csiszár & Tusnády, 1984), which has been a foundation of multiple computational schemas (e.g., the EM algorithm; Dempster et al., 1977). Consider a 2D plane (Figure 3). Given two convex regions, one has to find a pair of points (one from the first region, another from the second region) that are closest to each other. Start with an arbitrary point A_0 in the first region. Then point C_0 closest to A_0 is found in the second region. Then point A_1 closest to C_0 is found in the first region. This pattern continues until a convergence to a desired pair of closest points (A_2, C_2) is established. This schema

always converges to a global minimum, and it can be generalized to any two convex sets with a convex function as the distance between elements of the sets (Csiszár & Tusnády, 1984).

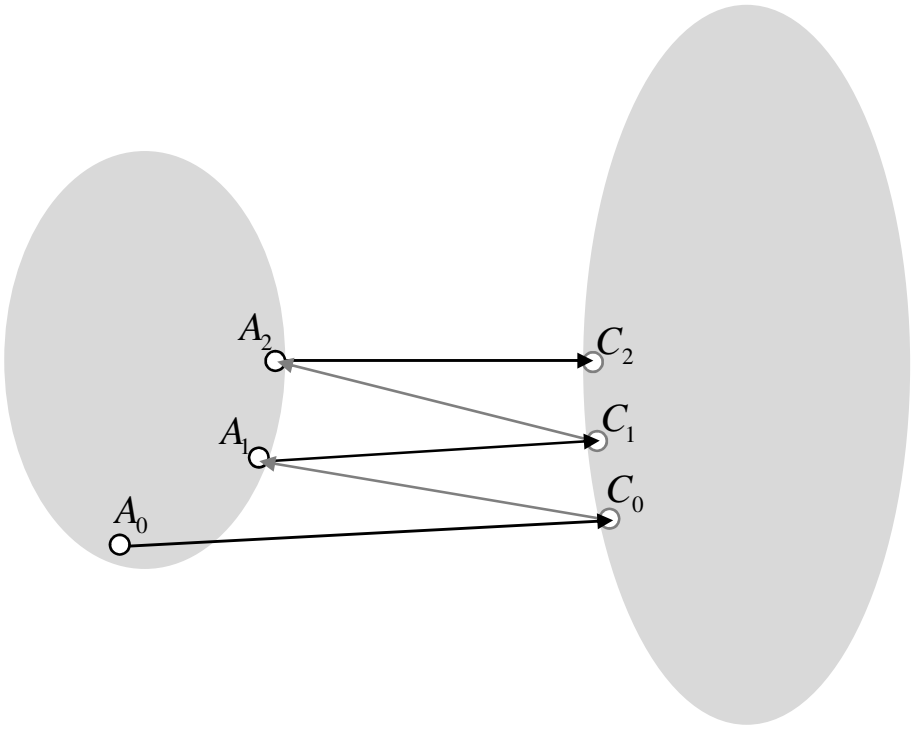

FIGURE 3. *Illustration of alternating minimization*

In the context of detecting cluster aberrancy, the above process can be applied as follows. An initial estimate of the aberrant subgroup is identified (see the next subsections), denoted as A_0. After A_0 is fixed, Subproblem 2 is solved, resulting in compromised subset C_0. After C_0 is fixed, Subproblem 1 is solved, resulting in aberrant subgroup A_1. After A_1 is fixed, Subproblem 2 is solved, resulting in compromised subset C_1. This pattern continues until a stopping condition is achieved. This corresponds to alternating between optimization Problems (4) and (3), i.e., an alternating maximization. If the statistical approach—providing feasible solutions to optimization Problems (4) and (3)—is used for solving Subproblems 1 and 2, respectively, then the stopping condition will correspond to a significance level $\alpha_{1,2}$ reaching below a certain threshold. For example, start with $\alpha_{1,2} = 0.1$, and at each iteration multiply $\alpha_{1,2}$ by 0.9; finally, stop when $\alpha_{1,2} \leq 0.01$. If the optimization approach is used, then the alternating schema should stop when an increase in objective functions from (3) and (4) is below a certain threshold.

The above schema will not converge to a global extremum because it alternates between a collection of all subgroups of test takers and a collection of all subsets of items, where both

collections are, obviously, not convex. Thus, the choice of the initial solution is crucial for the resultant local extremum to be close to the global extremum (see the next subsections).

Another important aspect of the convergence is that, at an early iteration k, the estimates A_k and C_k may have an error, meaning that A_k may have nonaberrant test takers and C_k may have uncompromised items. Because of this error, the measure of test-taker aberrancy $\mathbf{S}(\mathbf{F}_{I\cap C_k,j} \,\|\, \mathbf{F}_{I\setminus C_k,j})$ and the measure of item compromise $\mathbf{S}(\mathbf{F}_{i,J\cap A_k} \,\|\, \mathbf{F}_{i,J\setminus A_k})$ (used while solving Subproblems 1 and 2, respectively) are negatively affected, because subgroup $J \setminus A_k$ may have aberrant test takers and subset $I \setminus C_k$ may have compromised items. For example, if half of the compromised items are in C_k then the other half will be in subset $I \setminus C_k$ which may cause the value of $\mathbf{S}(\mathbf{F}_{I\cap C_k,j} \,\|\, \mathbf{F}_{I\setminus C_k,j})$ for an aberrant test taker j to be closer to zero, so test taker j will not be included in subgroup A_{k+1} (corresponding to the next iteration). Thus, a small error at early iterations may lead to a larger error in later iterations, resulting in an incorrect solution to the cluster aberrancy detection problem. To resolve this issue, a subgroup of nonaberrant test takers J_0 and a subset of uncompromised items I_0 are introduced to purify two subgroups $J \bigcap A_k$ and $J \setminus A_k$ and two subsets $I \bigcap C_k$ and $I \setminus C_k$. Then the measure of test-taker aberrancy and the measure of item compromise are computed as $\mathbf{S}(\mathbf{F}_{[I\cap C_k]\setminus I_0,j} \,\|\, \mathbf{F}_{[I\setminus C_k]\cap I_0,j})$ and $\mathbf{S}(\mathbf{F}_{i,[J\cap A_k]\setminus J_0} \,\|\, \mathbf{F}_{i,[J\setminus A_k]\cap J_0})$, respectively. Nonaberrant subgroup J_0 can be formed either by using simulated responses (as in this report) or by using real responses from previous administrations, when corresponding items were administered for the first time. The uncompromised subset I_0 is often given in practice (see Introduction).

Monte Carlo Estimation of Aberrancy Level

Monte Carlo methods provide solutions to multiple practical problems by estimating means of corresponding random variables. In the context of this report, the desired value to be estimated is an aberrancy level for each test taker. The aberrancy level is defined here as a mean of random variable $\mathbf{S}(\mathbf{F}_{[I\cap R]\setminus I_0,j} \,\|\, \mathbf{F}_{[I\setminus R]\cap I_0,j})$ of test taker j given random subset R, and it can be estimated by averaging the measure of test-taker aberrancy computed over multiple random subsets of items that do not intersect with uncompromised subset I_0. According to the definition of the measure of test-taker aberrancy (see above), the estimated mean should be larger for aberrant test takers because of multiple intersections of random subsets with the actual compromised subset. Figure 4 illustrates this concept; obviously, the larger the intersection of a random subset R with

the gray blob, the larger the value of $\mathbf{S}(\mathbf{F}_{[I\cap R]\setminus I_0,j} \,\|\, \mathbf{F}_{[I\setminus R]\cap I_0,j})$. Thus, the aberrancy level is estimated as follows:

$$\mathbf{D}(j) = \frac{1}{m}\sum_{k=1}^{m} \mathbf{S}(\mathbf{F}_{[I\cap R_k]\setminus I_0,j} \,\|\, \mathbf{F}_{[I\setminus R_k]\cap I_0,j}), \tag{5}$$

where m is the number of random subsets $R_1, R_2, ..., R_m$, where each R_k does not intersect with I_0 (see Figure 4). The aberrancy level $\mathbf{D}(j)$ is obviously between 0 and 1.

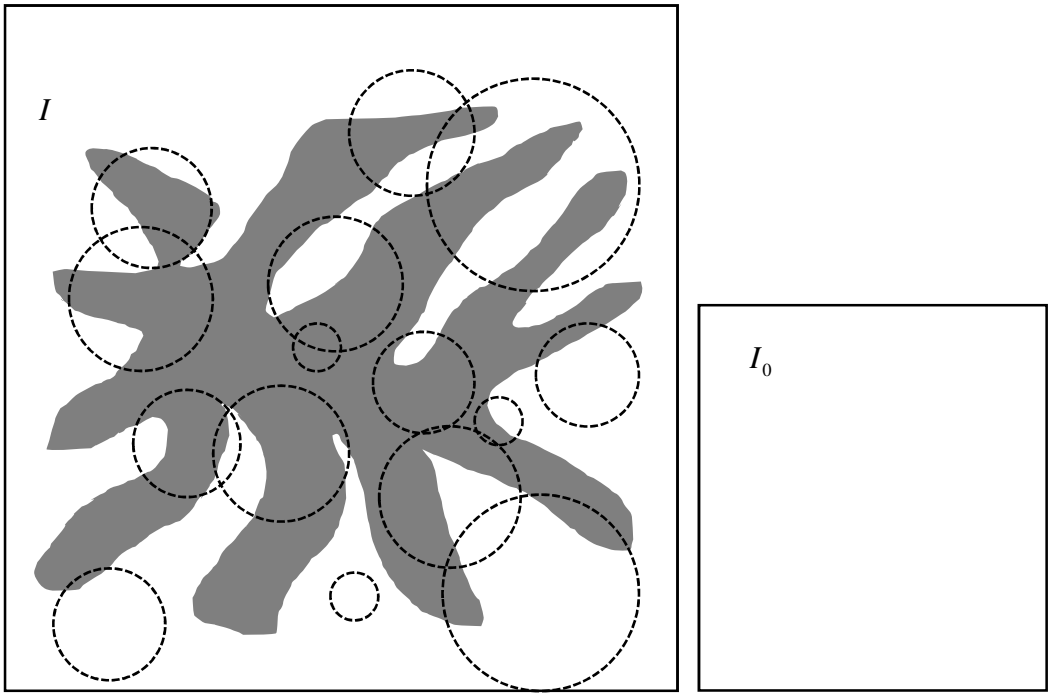

FIGURE 4. *Illustration of the Monte Carlo estimation of test-taker aberrancy level. The left-hand rectangle represents items from test* I, *and the right-hand rectangle represents items from uncompromised subset* I_0. *Inside the left-hand rectangle, the gray blob represents the compromised subset and the circles represent random subsets.*

Subproblem 4

In this subproblem, one needs to identify the initial estimate of aberrant subgroup A_0. Estimate (5) can be interpreted as a statistic computed for each test taker. Then, given a significance level α_4, one can compute the critical value as, for example, the corresponding percentile of the simulated null distribution. Then each test taker for whom the value of the

statistic is larger than the critical value is added to subgroup A_0. Formally stated, the statistic $\mathbf{D}(j)$ is computed for each test taker j from group J.

Subproblem 5

In this subproblem, one needs to detect multiple aberrant subgroups, each with a corresponding compromised subset.

If there is only one aberrant subgroup with a corresponding compromised subset, the main problem can be trivially solved as follows. First, solve Subproblem 4 to find the initial estimate of the aberrant subgroup and then solve Subproblem 3, resulting in the solution to the main problem. This is practical when test takers are partitioned into disjoint groups that are homogeneous with respect to their potential for item sharing, thereby allowing separate investigations to be conducted for each group. Each test taker has a profile, which includes variables potentially useful for test security purposes, such as test center where the test is taken, former high school, former undergraduate college, test-prep center, or current group in a social network. Such variables can be used to partition all test takers into disjoint groups. For example, the following variables partition test takers by the geographic location where test takers take a test: room, college, state, region, and country. Variables related to geographic location are most common; however, as Belov (2013) pointed out, other profile variables mentioned above could potentially help detect aberrant test takers, even if they take the exam at different geographic locations.

This report addresses a general case where within a group of test takers there might be multiple aberrant subgroups, each with a corresponding unique compromised subset (see Figure 1). In this case, the solution to Subproblem 4 (the initial estimate of aberrant subgroup A_0) contains aberrant test takers from different subgroups. Obviously, the alternating process used to solve Subproblem 3 will converge to some aberrant subgroup with a corresponding compromised subset. Then, remove the detected subgroup from the initial estimate A_0 and repeat the alternating process again; this should provide another aberrant subgroup with a corresponding compromised subset. Do this until the initial estimate A_0 becomes empty. These iterations form a heuristic that does not guarantee that each aberrant subgroup with a corresponding compromised subset will be detected. The success of this heuristic depends on how good the initial estimate A_0 is—"good" meaning that A_0 contains many aberrant test takers from all aberrant subgroups. The performance analysis of this heuristic will be given via computer simulations in the next section.

Alternating Detector

It is at this point that an algorithm to detect cluster aberrancy can be formulated. The Alternating Detector (AD) algorithm solves a sequence of subproblems analyzed previously by performing the following steps:

Step 1: Detect an initial estimate of all aberrant test takers (i.e., solve Subproblem 4).

Step 2: Detect an aberrant subgroup with a corresponding compromised subset (i.e., solve Subproblem 3 with embedded Subproblems 1 and 2 above).

Step 3: Remove the detected subgroup from the initial estimate of all aberrant test takers, and if the estimate is still not empty, go to Step 2 (i.e., solve Subproblem 5); otherwise stop.

Final Check

The output of the AD contains disjoint subgroups of test takers, each with a corresponding subset of items (see Figure 1). Consider a detected subgroup A and its detected subset C, where each test taker j from subgroup A has a high value for the statistic $\mathbf{S}(\mathbf{F}_{[I\cap C]\setminus I_0,j} \,\|\, \mathbf{F}_{[I\setminus C]\cap I_0,j})$. Clearly, if subset C is indeed correctly detected, then that test taker should also have a high value for the statistic $\mathbf{S}(\mathbf{F}_{C\cap I,j} \,\|\, \mathbf{F}_{I\setminus C,j})$. Thus, an additional hypothesis test can be applied here for statistic $\mathbf{S}(\mathbf{F}_{C\cap I,j} \,\|\, \mathbf{F}_{I\setminus C,j})$ to check if the detected test takers are indeed aberrant. This hypothesis test, called the *final check*, has the potential to improve the Type I error with a possible loss of power. The implications of the final check are studied via computer simulations below. Moreover, after the final check is applied, one can solve Subproblem 2 in order to purify the detected compromised subsets; this additional purification, however, is not considered here.

Experiments With Simulated Data

This section presents the results of simulated experiments to study properties of the AD, where the statistical approach to solving Subproblems 1 and 2 was applied. The source code (written in C++ by the author) for these experiments can be easily adapted to arbitrary item response theory (IRT) models, types of tests, or types of distribution for nonaberrant and aberrant test takers.

The lz Statistic

A baseline detector in several experiments was provided by the lz statistic (Drasgow et al., 1985). Critical values for lz were computed from simulated data with 1,000 nonaberrant test takers drawn from an $N(0,1)$ distribution, where the significance level was always chosen to be equal to the Type I error of the AD. This way both detectors had similar Type I errors.

Simulated Test

A previously administered form from the Law School Admission Test (LSAT) was used with 100 items in the operational part (consisting of four sections, about 25 items each) and about 25 items in the variable section. In this high-stakes P&P test, the operational part is the same for all test takers and the variable section varies across test takers (in this simulation, there were 10 different variable sections, and each test taker was randomly assigned a variable section). The variable section is assumed to have uncompromised items only.

Simulated Design

In each experiment, 10 test centers were simulated, each with 100 test takers, for a total of 1,000 test takers. Nonaberrant test takers were drawn from an $N(0,1)$ population, where the probability of a correct response to an item was modeled by the three-parameter logistic model (3PLM; Lord, 1980).

Simulation of Item Preknowledge

The size of each aberrant subgroup and the size of each compromised subset varied across experiments. Each aberrant subgroup was created by drawing a given number of test takers from $U(-3, 0)$. For each aberrant subgroup, a compromised subset was created as follows: Randomly select a section from the operational part; randomly select a given number of items from the selected section. Responses of aberrant test takers were simulated, where the probability of a correct answer to an item from the corresponding compromised subset was 1 and the probability of a correct answer to any other item was modeled by 3PLM. A given number of aberrant subgroups was created and randomly assigned to test centers such that each test center could not contain more than one aberrant subgroup. Finally, nonaberrant test takers drawn from $N(0,1)$ were added to all test centers such that each test center contained exactly 100 test takers.

Parameters of the Alternating Detector

The AD was run four times, each time with a different operational section as a search region for compromised items. Therefore, each test taker could be detected up to four times; however, there was no correction for significance levels.

When solving Subproblem 3 (with embedded Subproblems 1 and 2), the significance level $\alpha_{1,2}$ was changed as follows: Start with $\alpha_{1,2} = 0.01$, and at each iteration multiply $\alpha_{1,2}$ by 0.9; finally, stop when $\alpha_{1,2} \leq 0.001$. Given the current estimate of the compromised subset, the null distribution for Subproblem 1 was generated from the simulated responses of 1,000 nonaberrant test takers drawn from $N(0,1)$. Given the current estimate of the aberrant subgroup, the null distribution for Subproblem 2 was generated from the simulated responses of 1,000 nonaberrant test takers drawn from $N(0,1)$ to 1,000 items randomly drawn from the operational part.

When solving Subproblem 4, the number of generated random subsets $R_1, R_2, ..., R_m$ was $m = 100$, where the size of each subset was randomly selected from 5 to 20 (because each operational section has about 25 items) and the significance level was $\alpha_4 = 0.1$.

Finally, in Equations (1) and (2), the sequence of ability levels was as follows: $Z: -5 < -4.9 < ... < 5$.

Performance Measures for Detecting Aberrant Test Takers

The following performance measures were applied in order to compare the performance of the AD and lz. Because lz does not distinguish between aberrant subgroups, these measures did not take into account the existence of multiple aberrant subgroups and, therefore, are only suitable for a crude analysis of the AD's performance.

The Type I error (i.e., the empirical probability of falsely detecting a test taker) was computed as follows:

$$\frac{[\text{number of detected test takers}] - [\text{number of correctly detected test takers}]}{[\text{number of nonaberrant test takers}]}. \qquad (6)$$

The detection rate was computed as follows:

$$\frac{[\text{number of correctly detected test takers}]}{[\text{number of aberrant test takers}]}. \tag{7}$$

The precision was computed as follows:

$$\frac{[\text{number of correctly detected test takers}]}{[\text{number of detected test takers}]}. \tag{8}$$

Performance Measures for Detecting Compromised Items

The following performance measures were designed to be analogous to previous performance measures. These measures do not take into account the existence of multiple compromised subsets and, therefore, are only suitable for a crude analysis of the AD's performance. The Type I error was computed as follows:

$$\frac{[\text{number of detected items}]-[\text{number of correctly detected items}]}{[\text{number of uncompromised items}]}. \tag{9}$$

The detection rate was computed as follows:

$$\frac{[\text{number of correctly detected items}]}{[\text{number of compromised items}]}. \tag{10}$$

The precision was computed as follows:

$$\frac{[\text{number of correctly detected items}]}{[\text{number of detected items}]}. \tag{11}$$

Experiment 1

Experiment 1 is a comparison study between lz and AD, where the size of each compromised subset $|C|$ and the size of each aberrant subgroup $|A|$ varied as 5, 10, and 20; meanwhile, the number of aberrant subgroups was fixed at 4. Thus, there were $3 \times 3 \times 1 = 9$ scenarios, where each scenario was replicated 10 times in order to get averaged estimates of the performance measures. The resultant estimates are provided in Figures 5 and 6.

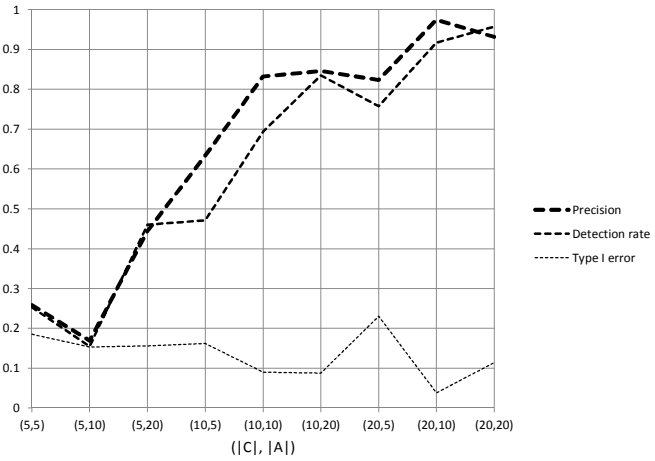

FIGURE 5. *Performance of the AD in detecting compromised items, where* $|C|$ *denotes the size of each compromised subset and* $|A|$ *denotes the size of each aberrant subgroup*

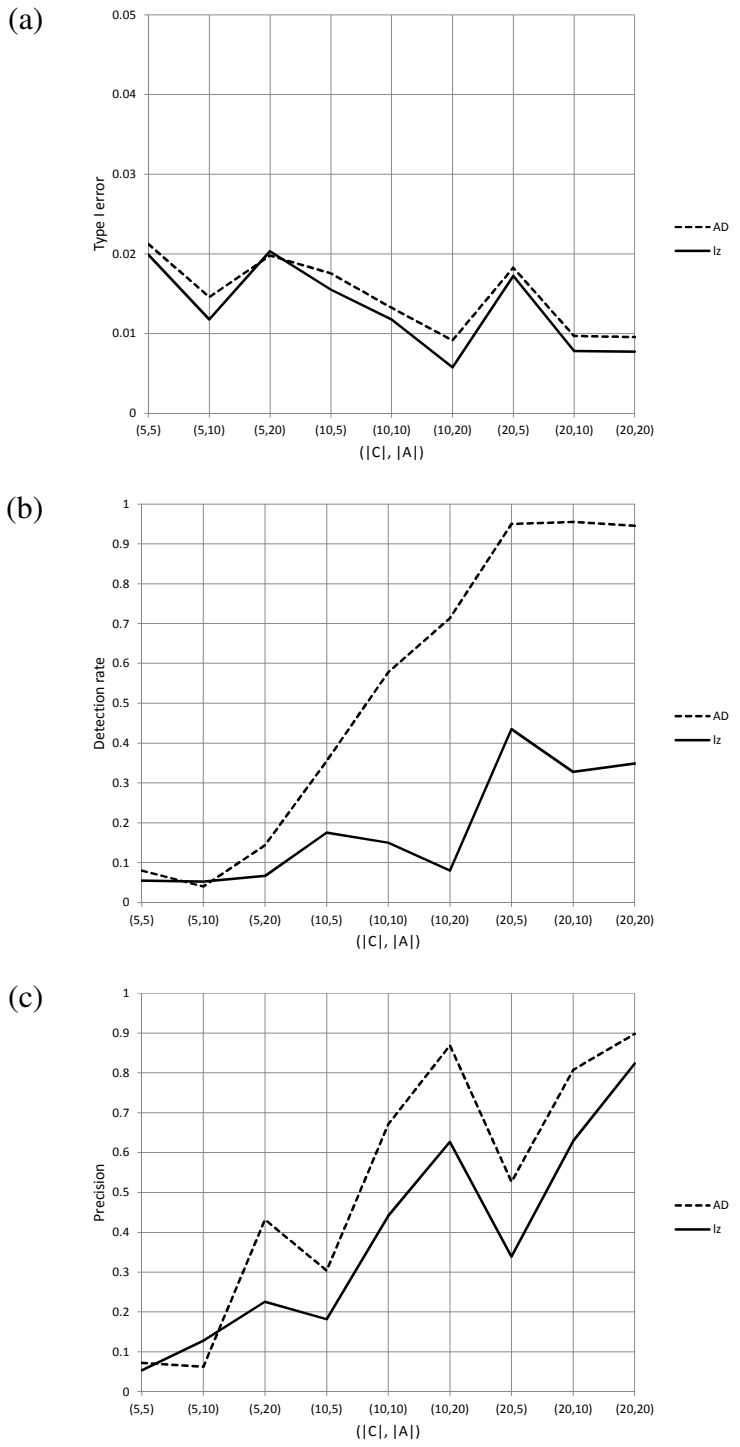

FIGURE 6. *Comparison between lz and the AD in detecting aberrant test takers: (a) Type I error; (b) detection rate; (c) precision.* Note: $|C|$ *denotes the size of each compromised subset and* $|A|$ *denotes the size of each aberrant subgroup.*

In detecting aberrant test takers, the AD resulted in a higher detection rate and a higher precision than the lz across almost all conditions while having a similar Type I error (see Figure 6). In detecting compromised items, the AD performed above 0.5 for the detection rate and the precision when the size of the compromised subset was at least 10 (see Figure 5), which is 10% of the operational part of the test. The same can be said about the AD when detecting aberrant test takers (see Figure 6). Thus, the size of the compromised subset exerts the largest effect on the performance of the AD (see Figures 5 and 6), and a threshold value of 10 causes a threshold performance of the AD. *Threshold performance* here means that when a parameter (the size of the compromised subset) passes a certain threshold value (in this case, 10), the performance of the AD improves dramatically (see Figures 5 and 6).

Experiment 2

Experiment 2 begins a series of five experiments to carefully study the threshold performance of the AD within its parametric space. The size of each compromised subset was fixed at 10; the number of aberrant subgroups and their size varied such that the total number of aberrant test takers was always 40. In particular, there were 6 scenarios, such that the number of aberrant subgroups and the size of each aberrant subgroup were changing: 0, 1, 2, 4, 5, 8 and 0, 40, 20, 10, 8, 5, respectively. The first scenario with 0 aberrant subgroups was intended to check the resultant Type I error for nonaberrant data. Each scenario was replicated 10 times in order to get averaged estimates of the performance measures. The resultant estimates are provided in Figure 7. One can observe a common pattern of a gradual decrease in detection rate and precision as the number of aberrant subgroups increases and their sizes decrease. Such behavior is compatible with the results from Figures 5 and 6. An exception to this pattern can be observed in Figure 7(b) for precision. Since each aberrant subgroup has a unique compromised subset with exactly 10 items (see above), increasing the number of aberrant subgroups increases the number of compromised items.

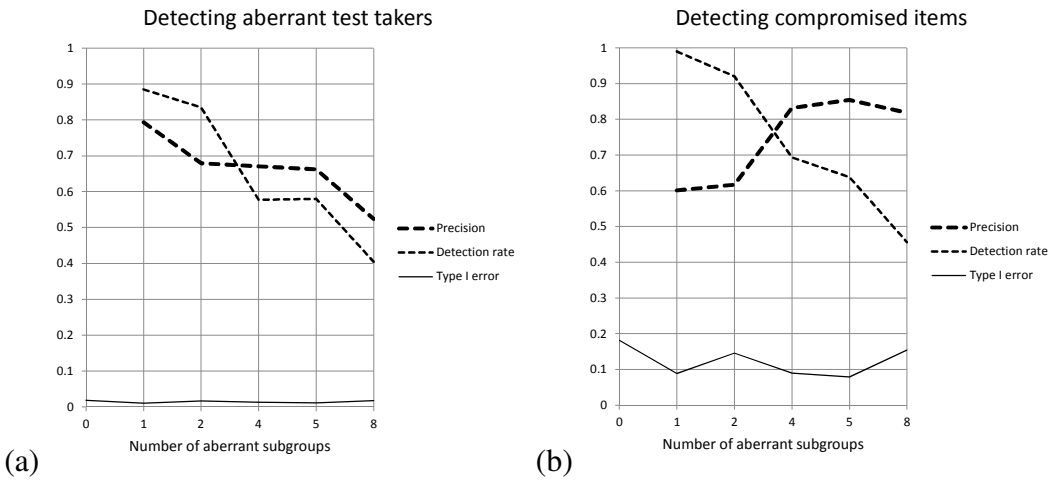

FIGURE 7. *Performance measures of the AD: (a) for detecting aberrant test takers; (b) for detecting compromised items*

Experiment 3

From the definition of the PS above, it follows that the AD should perform better when compromised items have a higher difficulty. Experiment 3 has the same setup as Experiment 2, the only difference being that all compromised items had a difficulty higher than 0 (on the ability level scale). The results are presented in Figure 8, where one can observe a better performance of the AD in comparison with results from Figure 7, where there was no control over the selection of compromised items.

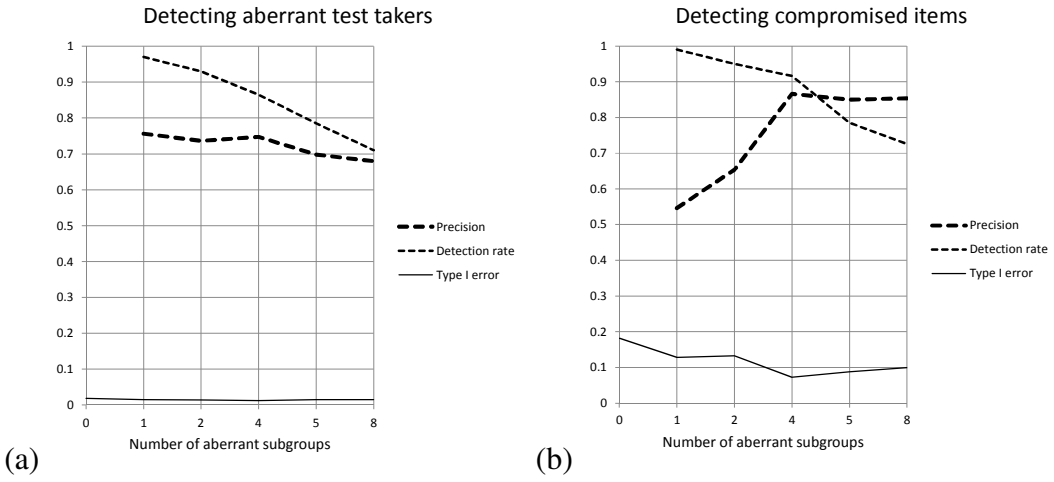

FIGURE 8. *Performance measures of the AD, where compromised items had a difficulty higher than 0: (a) for detecting aberrant test takers; (b) for detecting compromised items*

Experiment 4

Experiment 4 has the same setup as Experiment 2, the only difference being that the final check was applied. As pointed out above, it is expected that in this case the Type I error should be lower and the precision should be higher when detecting aberrant test takers. The results in Figure 9(a) confirm this expectation: one can observe a lower Type I error and a higher precision when detecting aberrant test takers in comparison with results from Figure 7(a), where there was no final check.

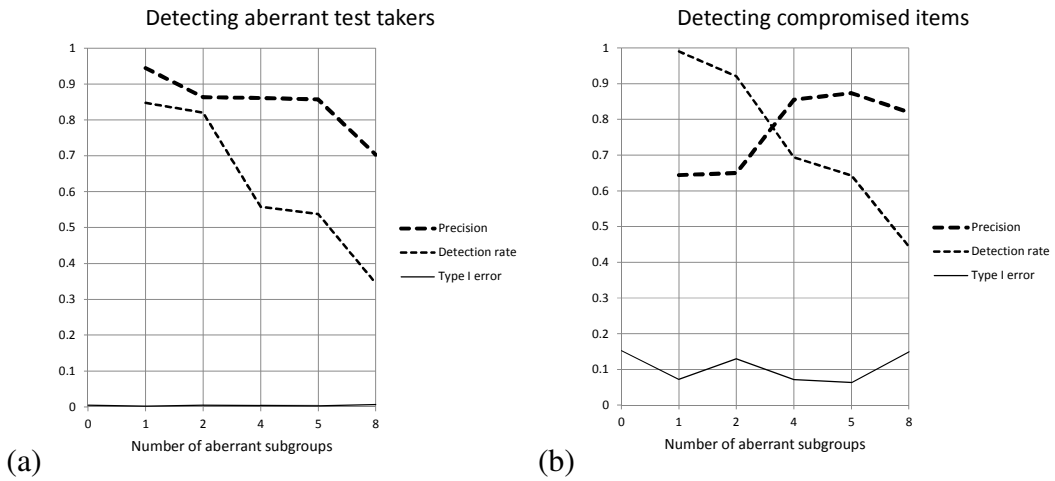

FIGURE 9. *Performance measures of the AD, where the final check was applied: (a) for detecting aberrant test takers; (b) for detecting compromised items*

Experiment 5

Experiment 5 has the same setup as Experiment 2, the only difference being that the AD was run separately for each test center. Since each aberrant subgroup was located within a single test center, it is expected that the performance of the AD should be higher. The results in Figure 10 confirm this expectation: One can observe a much better performance of AD in comparison with results from Figure 7, where the AD was applied to all test centers simultaneously (this confirms statements made in Subproblem 5, above). Also, notice the drop of the Type I error to almost zero (see Figure 10).

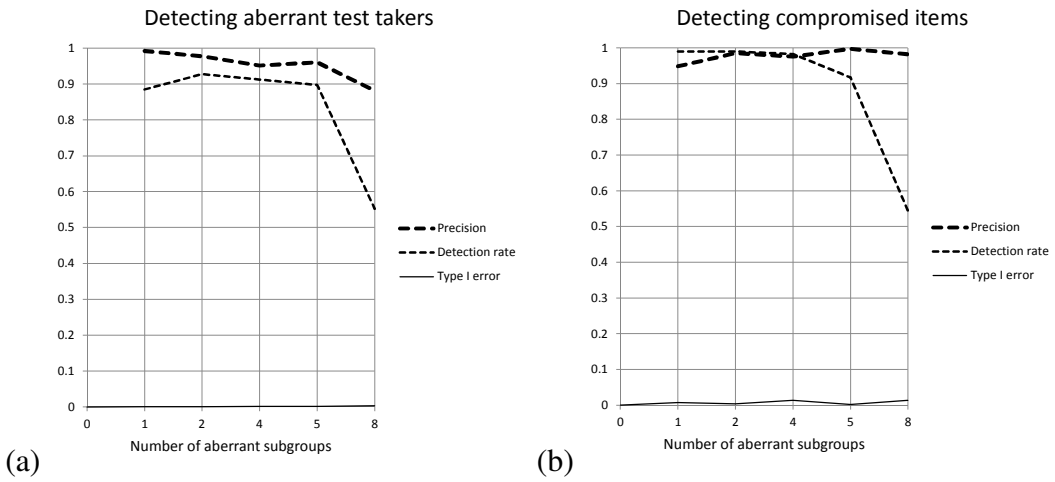

FIGURE 10. *Performance measures of the AD, where the AD was applied separately for each test center: (a) for detecting aberrant test takers; (b) for detecting compromised items*

Experiment 6

Experiment 6 has the same setup as Experiment 2, the only difference being that for each aberrant test taker, 15% or 30% of items in a variable section were compromised. Such a violation of the major assumption is expected to have a negative effect on the performance of the AD, and the amount of this effect should characterize the robustness of the AD. The major task of this experiment is to determine the percentage of violation under which the AD can still provide a higher detection rate than the lz (recall that all simulations were designed such that both detectors had similar Type I errors). The results are shown in Figure 11. One can observe that the AD performs better than the lz when the violation is 15%, although the AD's detection rate does drop by 0.1–0.3 with respect to its detection rate in Figure 7(a). However, when the violation is 30%, the AD's performance drops dramatically and lands below the performance of the lz.

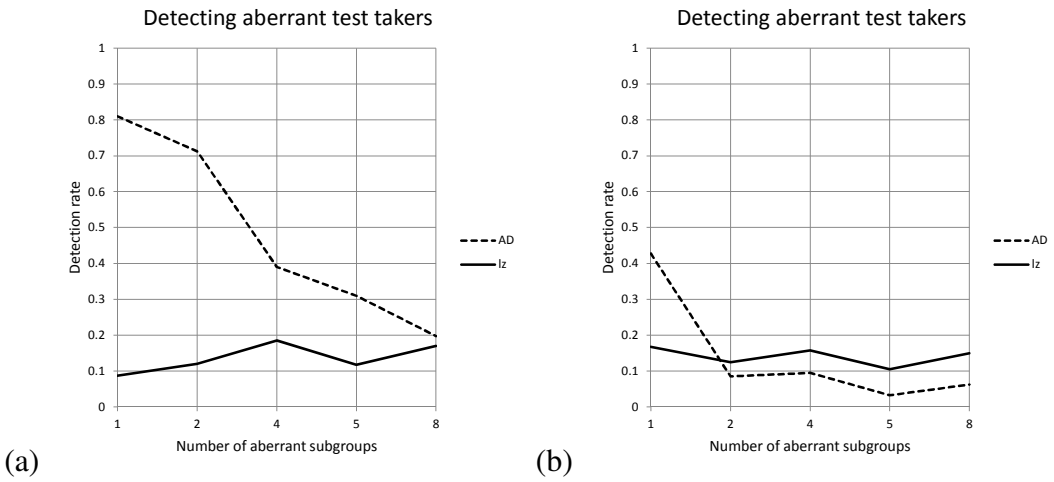

FIGURE 11. *Detection rate of the AD and lz, where for each aberrant test taker some percentage of items in a variable section were compromised: (a) 15% of the items in a variable section; (b) 30% of items in a variable section*

Experiments With Real Data

This section presents results of a comparison study between the AD and the lz using real data.

Description of Real Data

Responses to the LSAT (its description is given in the subsection titled Simulated Test) were used from three administrations given in 3 consecutive years of the same month (Table 1).

Parameters of the Detectors

The lz was applied the same way as in the previous section. The parameters of the AD were described above except for the following changes: Significance level $\alpha_4 = 0.01$, the final check was applied, and the AD was run separately for each test center.

TABLE 1
Characteristics of real datasets

Dataset Name	No. of Different Variable Sections	No. of Test Takers	No. of Test Centers	Mean (*SD*) of No. of Test Takers per Test Center
Data #1	14	20,634	556	37 (36)
Data #2	10	18,961	557	34 (33)
Data #3	10	22,088	556	40 (39)

Detecting Item Preknowledge Embedded in Real Data

Test takers from each real dataset were considered nonaberrant. Then the following process embedded simulated aberrant test takers into each real dataset. Each aberrant subgroup was created by drawing a given number of test takers from U(−3, 0). For each aberrant subgroup, a compromised subset was created as follows: Randomly select a section from the operational part; randomly select a given number of items from the selected section. The responses of aberrant test takers were simulated, where the probability of a correct answer to an item from the corresponding compromised subset was 1 and the probability of a correct answer to any other item was modeled by the 3PLM. A given number of aberrant subgroups was created and randomly assigned to real test centers such that each test center could not contain more than one aberrant subgroup and the original number of test takers remained the same. For the latter, randomly chosen real test takers were substituted with simulated aberrant test takers.

Similarly to simulated experiments from the previous section, the size of each compromised subset was fixed at 10. The number of aberrant subgroups and their size varied, such that the total number of aberrant test takers was always 40. In particular, there were 5 scenarios, such that the number of aberrant subgroups and the size of each aberrant subgroup were changing: 1, 2, 4, 5, 8 and 40, 20, 10, 8, 5, respectively. Each scenario was simulated 10 times in order to get averaged estimates of the performance measures. The resultant estimates are provided in Figure 12. One can observe a common pattern of a gradual decrease in the AD's detection rate as the number of aberrant subgroups increased and their sizes decreased. Such behavior is compatible with the results from Figures 6–11. However, the AD performs considerably better than the lz across all performance measures, datasets, and conditions (see Figure 12).

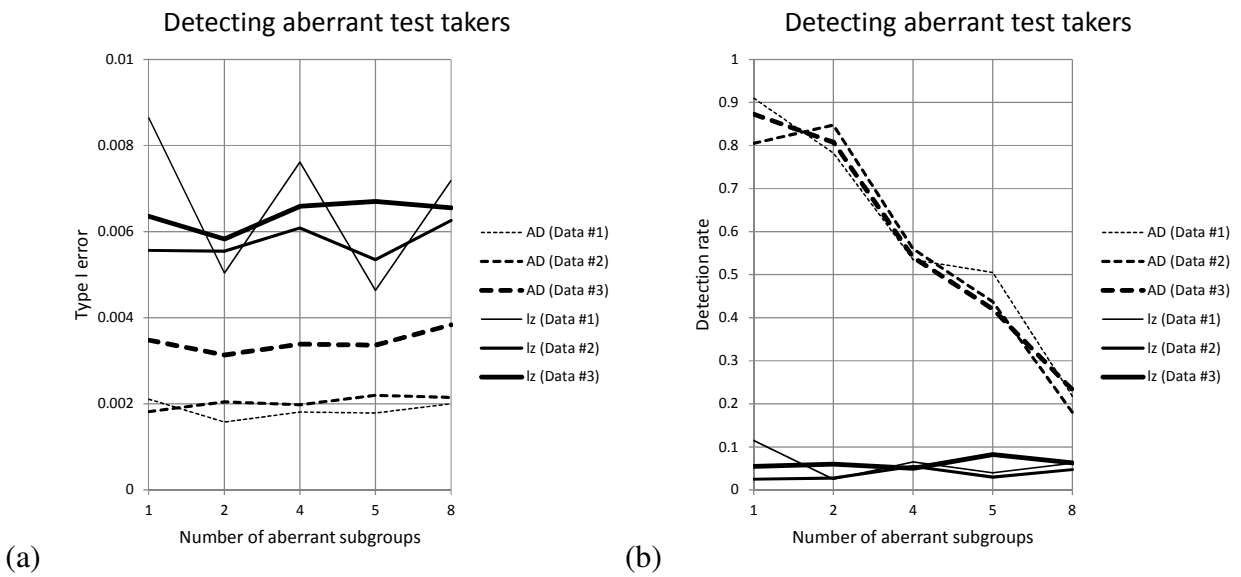

FIGURE 12. *Performance measures of the AD and lz where three real datasets with embedded simulated item preknowledge were used: (a) Type I errors; (b) detection rates. Notice that the detection rate of the lz fluctuated only slightly; that is because the total number of aberrant test takers was fixed at 40.*

Summary

In general, cluster aberrancy is hard to detect due to the multiple unknowns involved: Unknown subgroups of test takers have an unfair advantage on unknown subsets of items (see Figure 1). The objective of this report was to identify all multiple unknowns by employing elements of statistics and combinatorial optimization.

The major result of this report is formulated as the Alternating Detector (AD) algorithm. This algorithm employs the alternating minimization process and Monte Carlo methods in order to identify aberrant subgroups and their compromised subsets. In comparison with the lz statistic, the AD performed better on simulated (see Figure 6) and real (see Figure 12) data when detecting aberrant test takers. To the best of the author's knowledge, the AD is the first algorithm that can simultaneously identify aberrant subgroups of test takers (see Figure 6) and compromised subsets of items (see Figure 5) solely on the basis of responses and information about uncompromised items.

A comprehensive simulation study demonstrated the advantages and limitations of the AD. The most crucial assumption—the existence of an uncompromised subset—is realistic, as pointed out in the Introduction. Moreover, even if this assumption is violated up to 15%, the AD still outperforms the lz (see Figure 11). The performance of the AD decreases as the size of the compromised subsets and aberrant subgroups decreases and as the number of aberrant subgroups increases (see Figures 5–7). It was shown that when the number of aberrant subgroups is fixed,

the decrease in the size of the compromised subsets has the highest negative impact on the performance of the AD (see Figures 5 and 6). The results of the simulation studies showed that if this size equals at least 10% of the operational test length, then the AD performs well (see Figures 5 and 6). When compromised items have a higher difficulty, the performance of the AD increases (see Figure 8). When test takers are partitioned into disjoint groups that are homogeneous with respect to their potential for item sharing, thereby allowing separate investigations to be conducted for each group, the AD applied separately for each group demonstrates higher performance (see Figure 10). In general, adding the final check decreases the Type I error and increases the precision of the AD (see Figures 9 and 12) when detecting aberrant test takers.

The AD is able to detect item preknowledge and test collusion. According to the general definition of the measures for test-taker aberrancy and item compromise, the algorithm is applicable to all types of testing programs: P&P, CBT, MST, and CAT.

The AD is an algorithmic framework where embedded subroutines and statistics can be modified in order to adjust for a specific testing program. As a future work, the following modifications are possible:

- Different measures of aberrancy and compromise can be used (for possible candidates, see Belov, 2016). Also, the AD can be applied for CAT, MST, and CBT with posteriors of speed (for details on response time modeling, see van der Linden, 2011]. In this case, the following measures of aberrancy and compromise can be used:

$$\frac{1}{2}\left[\mathbf{S}(\mathbf{F}_{I\cap C,j} \,\|\, \mathbf{F}_{I\setminus C,j}) + \mathbf{S}(\mathbf{V}_{I\cap C,j} \,\|\, \mathbf{V}_{I\setminus C,j})\right], \tag{12}$$

$$\frac{1}{2}\left[\mathbf{S}(\mathbf{F}_{i,J\cap A} \,\|\, \mathbf{F}_{i,J\setminus A}) + \mathbf{S}(\mathbf{V}_{i,J\cap A} \,\|\, \mathbf{V}_{i,J\setminus A})\right], \tag{13}$$

  where $\mathbf{V}$ is the posterior of speed computed from response times. Equations (12) and (13) assume that each aberrant test taker performs faster on compromised items than on uncompromised items and on each compromised item the aberrant test takers perform faster than the nonaberrant population (in general, this assumption is incorrect; see Introduction). The measure of test-taker aberrancy (12) and the measure of item compromise (13) use additional information from response times, which may improve the overall performance of the AD.

- Instead of using a statistical approach for solving Subproblems 1 and 2, one can rely on a more subtle combinatorial optimization approach; see Problems (3) and (4). In this case possible solvers include simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983); greedy heuristic (Papadimitriou & Steiglitz, 1982); genetic algorithm (Mitchell, 1996); or tabu search (Glover & Laguna, 1997).

Cluster aberrancy goes beyond psychometrics and educational measurement. Theoretically, the AD can be applied to detect a subgroup of individuals with illegal access to some information, giving them an unfair advantage in a certain activity. For example, in financial markets, a subgroup of individuals with illegal access to corporate information may perform unusually well on some stocks in contrast to other stocks and other players. This is becoming a serious issue because all corporate information is currently kept in private or public clouds; meanwhile, tools for hacking into such systems are constantly improving. The Financial Industry Regulatory Authority (FINRA) currently monitors 3,940 securities firms with approximately 641,155 brokers. It reported more than 800 fraud cases referred for prosecution in 2015 and $191.7 million in fines and restitution levied in 2015 (www.finra.org/about, 2015). For such applications, the appropriate measures of aberrancy and compromise should be developed.

References

Belov, D. I. (2013). Detection of test collusion via Kullback-Leibler divergence. *Journal of Educational Measurement*, *50*, 141–163.

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

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*, 83–97.

Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52.

Csiszár, I., & Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions: Supplement Issue 1*, 205–237.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1–38.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67–86.

Glover, F., & Laguna, M. (1997). *Tabu search.* Boston: Kluwer Academic Publishers.

Lehmann, E. L. (1999). *Elements of large-sample theory*. New York, NY: Springer.

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277–298.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). *Optimization by simulated annealing. Science*, *220*, 671–680.

McLeod, L. D., Lewis, C., & Thissen, D. (1999). *A Bayesian method for the detection of item preknowledge in CAT* (LSAT Computerized Testing Report, CT 98-07). Newtown, PA: Law School Admission Council.

McLeod, L. D, & Schnipke, D. L. (2006). *Detecting items that have been memorized in the CAT environment* (LSAT Computerized Testing Report, CT 99-05). Newtown, PA: Law School Admission Council.

Mitchell, M. (1996). *An introduction to genetic algorithms*. Cambridge, MA: The MIT Press.

Papadimitriou, C. H., & Steiglitz, K. (1982). *Combinatorial optimization: Algorithms and complexity*. Englewood Cliffs, NJ: Prentice-Hall.

Stone, E. (2016, April). *Integrating digital assessment meta-data for psychometric and validity analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Tendeiro, J. N., & 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.

van der Linden, W. J. (2011). Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling*, *53*, 334–358.

van 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*, 199–217.

Wollack, J. A., & Maynes, D. (2011, April). *Detection of test collusion using item response data.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Zhang, Y., Searcy, C. A., & Horn, L. (2011, April). *Mapping clusters of aberrant patterns in item responses.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Appendix: Definition and Properties of the Posterior Shift

Let us consider two probability mass functions $\mathbf{F}_C$ and $\mathbf{F}_U$ defined on a fixed finite sequence of ability levels Z. Consider Z as an increasing sequence of ability levels $Z: z_1 < z_2 < ... < z_n$ (e.g., $Z: -3 < -2.9 < ... < 3$). Find indices l and r, $1 \le l < r \le n$, such that $\mathbf{F}_C(z_j) > \mathbf{F}_U(z_j)$ for each $j = 1, 2, ..., l$ and $\mathbf{F}_C(z_j) > \mathbf{F}_U(z_j)$ for each $j = r, r+1, ..., n$; otherwise, assume $l = 0$ or $r = n+1$. Then the following equation, called the posterior shift (PS), measures how far the posterior $\mathbf{F}_C$ is shifted toward the higher ability with respect to the posterior $\mathbf{F}_U$:

$$\mathbf{S}(\mathbf{F}_C \| \mathbf{F}_U) = \begin{cases} \sum_{r \le j \le n} \left[\mathbf{F}_C(z_j) - \mathbf{F}_U(z_j)\right], & \sum_{r \le j \le n} \left[\mathbf{F}_C(z_j) - \mathbf{F}_U(z_j)\right] > \sum_{1 \le j \le l} \left[\mathbf{F}_C(z_j) - \mathbf{F}_U(z_j)\right] \\ 0, & \sum_{r \le j \le n} \left[\mathbf{F}_C(z_j) - \mathbf{F}_U(z_j)\right] \le \sum_{1 \le j \le l} \left[\mathbf{F}_C(z_j) - \mathbf{F}_U(z_j)\right] \end{cases} \quad \text{(A-1)}$$

Note: if $l = 0$ then $\sum_{1 \le j \le l} \left[\mathbf{F}_C(z_j) - \mathbf{F}_U(z_j)\right] = 0,$ and if $r = n+1$ then $\sum_{r \le j \le n} \left[\mathbf{F}_C(z_j) - \mathbf{F}_U(z_j)\right] = 0$.

By definition, $\mathbf{S}(\mathbf{F}_C \| \mathbf{F}_U)$ is asymmetric and non-negative, located from 0 to 1. For example, in Figure 2, $\mathbf{S}(\mathbf{F}_C \| \mathbf{F}_U) > 0,$ but $\mathbf{S}(\mathbf{F}_U \| \mathbf{F}_C) = 0.$

To establish the asymptotic distribution and power of the PS statistic, the following two lemmas and one theorem will be proved.

Lemma 1: Given two probability densities $\mathbf{H}(x)$ and $\mathbf{G}(x),$ the following holds for arbitrary x_0:

$$\int_{-\infty}^{x_0} \left(\mathbf{H}(x) - \mathbf{G}(x)\right) dx = \int_{x_0}^{+\infty} \left(\mathbf{G}(x) - \mathbf{H}(x)\right) dx. \quad \text{(A-2)}$$

Proof: This immediately follows from the definition of a probability density.

Lemma 2: Given two probability densities $\mathbf{H}(x)$ and $\mathbf{G}(x),$ consider the following function:

$$\mathbf{K}(y) = \int_{-\infty}^{y} \mathbf{H}(x)dx - \int_{-\infty}^{y} \mathbf{G}(x)dx. \quad \text{(A-3)}$$

Then all its local maxima and minima are located at points where $\mathbf{H}(x)$ and $\mathbf{G}(x)$ are equal.

Proof: Function $\mathbf{K}(y)$ is continuous by its definition; therefore, its local maxima and minima are located at each point where the derivative of $\mathbf{K}(y)$ is zero. The derivative of function $\mathbf{K}(y)$ is $\mathbf{H}(y)-\mathbf{G}(y)$.

Theorem: Let us consider two continuous probability densities $\mathbf{H}(x)$ and $\mathbf{G}(x)$, where there are no more than two points where $\mathbf{H}(x)$ and $\mathbf{G}(x)$ are equal (e.g., two normal distributions with different means). Consider a point y_0 such that:

$$\begin{array}{l}\mathbf{H}(x)=\mathbf{G}(x), x=y_0\\ \mathbf{H}(x)<\mathbf{G}(x), x>y_0\end{array}. \tag{A-4}$$

Also, if there is point $z_0 < y_0$ where $\mathbf{H}(z_0)=\mathbf{G}(z_0)$, then the following inequality holds:

$$\int_{-\infty}^{z_0}\left[\mathbf{G}(x)-\mathbf{H}(x)\right]dx<\int_{y_0}^{+\infty}\left[\mathbf{G}(x)-\mathbf{H}(x)\right]dx. \tag{A-5}$$

Then the following equality holds:

$$\sup_y\left|\mathbf{K}(y)\right|=\sup_{y<y_0}\left|\mathbf{K}(y)\right|=\mathbf{K}(y_0)=\int_{y_0}^{+\infty}\left[\mathbf{G}(x)-\mathbf{H}(x)\right]dx. \tag{A-6}$$

Proof: From Lemma 1, it follows that $\mathbf{K}(y_0)=\int_{y_0}^{+\infty}\left[\mathbf{G}(x)-\mathbf{H}(x)\right]dx.$ There are only two possible cases:

Case 1: Point y_0 is the only point where $\mathbf{H}(x)$ and $\mathbf{G}(x)$ are equal (e.g., two normal distributions with different means and equal variances).

Case 2: There is point $z_0 < y_0$ where $\mathbf{H}(z_0)=\mathbf{G}(z_0)$ and inequality (A-5) holds. Since $\mathbf{H}(x)$ and $\mathbf{G}(x)$ are continuous, the inequality $\mathbf{H}(x)<\mathbf{G}(x)$ holds for $x<z_0$. Then the following inequality follows:

$$\left|\mathbf{K}(y_0)\right| - \left|\mathbf{K}(z_0)\right| = \left|\int_{-\infty}^{y_0} \left[\mathbf{H}(x) - \mathbf{G}(x)\right] dx\right| - \left|\int_{-\infty}^{z_0} \left[\mathbf{H}(x) - \mathbf{G}(x)\right] dx\right| =$$
$$= \int_{y_0}^{+\infty} \left[\mathbf{G}(x) - \mathbf{H}(x)\right] dx - \int_{-\infty}^{z_0} \left[\mathbf{G}(x) - \mathbf{H}(x)\right] dx > 0 \qquad \text{(A-7)}$$

From inequality (A-7) and Lemma 2, it follows that $\sup_{y<y_0} \left|\mathbf{K}(y)\right| = \mathbf{K}(y_0)$. Since function $\mathbf{L}(y) = \int_{y}^{+\infty} \left[\mathbf{G}(x) - \mathbf{H}(x)\right] dx$ is positive and monotonically decreasing on $y > y_0$, then due to Lemma 1 the equality $\sup_{y} \left|\mathbf{K}(y)\right| = \sup_{y<y_0} \left|\mathbf{K}(y)\right|$ holds. ■

According to Chang and Stout (1993), the posteriors of ability $\mathbf{F}_C(x) = \mathbf{G}(x)dx$ and $\mathbf{F}_U(x) = \mathbf{H}(x)dx$ are asymptotically normal. Then due to Equation (A-1), the positive value of $\mathbf{S}(\mathbf{F}_C \,\|\, \mathbf{F}_U)$ corresponds to a situation described in Theorem 1, where $\mathbf{S}(\mathbf{F}_C \,\|\, \mathbf{F}_U)$ is a discrete approximation of $\mathbf{K}(y_0) = \int_{y_0}^{+\infty} \left[\mathbf{G}(x) - \mathbf{H}(x)\right] dx$. Then from Theorem 1 it follows that a hypothesis test to detect an aberrant test taker or a compromised item based on a positive PS statistic is equivalent to the Kolmogorov–Smirnov test (Lehmann, 1999) that analyzes statistic $\sup_{y} \left|\mathbf{K}(y)\right|$. It is known (Lehmann, 1999) that the asymptotic power of the Kolmogorov–Smirnov test is 1.