Journal of Computerized Adaptive Testing

Volume 2 Number 3

October 2014

Detecting Item Preknowledge in Computerized Adaptive Testing Using Information Theory and Combinatorial Optimization

Dmitry I. Belov

DOI 10.7333/1410-0203037

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing
Volume 2, Number 3, October 2014
DOI 10.7333/1410-0203037
ISSN 2165-6592

Detecting Item Preknowledge in Computerized Adaptive Testing Using Information Theory and Combinatorial Optimization

Dmitry I. Belov
Law School Admission Council

Item preknowledge occurs when some examinees (called *aberrant examinees*) have had access to a subset of items (called a *compromised subset*) from an administered test prior to an exam. As a result, aberrant examinees might perform better on compromised items as compared to uncompromised items. When the number of aberrant examinees is large, the corresponding testing program and its users might be negatively affected because the aberrant examinees might be given invalid scores. There are numerous item preknowledge detection methods exploiting the difference in an examinee's performance between compromised items and uncompromised items. These methods are based on an incorrect assumption that the compromised subset is known and that it does not vary across subgroups of aberrant examinees. Computer simulations demonstrated that when this assumption is slightly violated the detection rate drops dramatically. This paper introduces a new algorithm—the *3D algorithm*—merging information theory and combinatorial optimization for detecting subgroups of aberrant examinees and their corresponding compromised subsets of items.

Keywords: *test security, item preknowledge, hypothesis testing, Kullback-Leibler divergence, combinatorial optimization, simulated annealing.*

Item preknowledge occurs when some examinees (called *aberrant examinees*) have access to a subset of items (called the *compromised subset*) from an administered test prior to an exam. As a result, aberrant examinees will likely perform better on compromised items as compared to uncompromised items. When the number of aberrant examinees is large, the corresponding testing program—whether using paper-and-pencil tests (P&P), computer-based tests (CBT), multistage tests (MST), or computerized adaptive tests (CAT)—and its users (e.g., universities, companies, government organizations) might be negatively affected because the aberrant examinees might be given invalid scores.

There are numerous person-fit statistics that exploit differences in performance between compromised items and uncompromised items in order to detect aberrant examinees (Karabatsos, 2003; McLeod, Lewis, & Thissen, 2003; Belov, Pashley, Lewis, & Armstrong, 2007; Shu, Henson, & Luecht, 2013). These statistics assume that the compromised subset is known and does not change from one subgroup of aberrant examinees to another, which is not realistic. For example, the compromised subset can be defined by assigning to each item a probability of preknowledge (McLeod et al. 2003); however, Hui (2010) demonstrated through computer simulations that when the number of items with a high probability of preknowledge increases (e.g., due to a large number of aberrant subgroups each with a unique compromised subset), the detection rate drops dramatically.

There are alternative approaches that operate without the above assumption. The CUSUM method (van Krimpen-Stoop & Meijer, 2001; Armstrong & Shi, 2009) is only applicable when compromised items are positioned sequentially in the test (Tendeiro & Meijer, 2012). Response time modeling (van der Linden & Guo, 2008) has great promise to detect item preknowledge. However, the actual response times are only available in CAT, where examinees cannot return to previously seen items; otherwise, as is the case in CBT or MST, it is not clear how to compute the time that an examinee actually dedicated to each item (i.e., examinees might still think about previously seen items in order to get back to them if there are any doubts about chosen answers); and additionally, response times can be realistically faked. Cluster analysis (Wollack & Maynes, 2011) and factor analysis (Zhang, Searcy, & Horn, 2011) have been applied to detect item preknowledge; however, both methods rely on number of response matches, which is not applicable in MST and CAT where the actual test varies across examinees.

Thus, item preknowledge is difficult to detect due to multiple unknowns involved—unknown subgroups of examinees (at unknown schools or test centers) accessing unknown compromised subsets of items prior to taking the test. This paper demonstrates that disentangling the problem becomes feasible via combinatorial optimization.

This study is based on research by Belov (2013). In particular, he studied the problem of item preknowledge (with particular emphasis on CAT) in three distinct cases: Case 1, when the compromised subset is known; Case 2, when the compromised subset is covered by a known collection of subsets; and Case 3, when the compromised subset is unknown. In order to make Case 3 tractable, Belov made three assumptions about the compromised subset. Then, he applied random search criteria to identify each compromised subset, which in computer simulations resulted in good detection rates. However, his approach for Case 3 has two major practical weaknesses:

1. The second assumption by Belov (2013) about known bounds on size of compromised subsets might be violated in practice. If these bounds are too narrow, then the random search will miss the actual compromised subset. On the other hand, if these bounds are too wide then it will make the search space too large for the random search to converge to a compromised subset in a reasonable time.
2. Random search for the compromised subset was biased by a probability for each item to be potentially compromised, where the probability was computed as normalized item exposure without item exposure control (Belov, 2013). However, when item exposure is controlled, many items will have the same probability (right on the boundary of the exposure constraint), which will also make the search space too large for the random search to converge in a reasonable time. In addition, in P&P and CBT all administered items have the same level of exposure.

This paper suggests a new approach to address these two weaknesses. This approach addresses Case 3 from Belov (2013) where bounds on the size of compromised subsets are unknown and all items have equal probability of potential compromise.

Throughout the paper the following notation is used:

- lowercase letters $a,b,c,...$ denote scalars;
- capital letters $A,B,C,...$ denote sets; $|S|$ denotes the number of elements in a set S; and $\varnothing$ denotes an empty set;

- bold capital letters $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$ denote functions (including discrete distributions defined by probability mass functions);
- lowercase Greek letters $\alpha, \beta, \gamma, \ldots$ denote random variables; and
- capital Greek letters $\Omega, \Psi, \Theta, \ldots$ denote collections of subsets; $|\Omega|$ denotes the number of subsets in a collection Ω.

Problem Statement

It is assumed that there is a relation between examinees that partitions them into non-intersecting groups. For example, the following relation partitions examinees into test centers—the same geographic location where examinees take a test (e.g., room, college, state, region, country). The same geographic location is the most common relation. However, as Belov (2013) pointed out, there are other relations highly practical for test security: same high school, same undergraduate college, same test-prep center, or same group in a social network. Using these relations could potentially help to detect aberrant examinees, even if they take an exam at different geographic locations.

This study assumes that each group cannot have more than one subgroup of aberrant examinees. Considering how small a group can be (e.g., class) or how specific a corresponding relation can be (e.g., same group in a social network), this assumption is realistic. The detector of item preknowledge presented here is based on this assumption. When a group has multiple subgroups of aberrant examinees and their compromised subsets have small intersection, the detection rate might drop. The methodology developed here can be generalized to be effective even when multiple subgroups of aberrant examinees are present within a group. Such generalization is briefly described.

Large-Scale Item Preknowledge

Aberrant examinees and compromised items can be partitioned into *aberrant subgroups* and *compromised subsets*, respectively, where each aberrant subgroup has preknowledge of a unique compromised subset (Figure 1). A group (e.g., a test center) with aberrant examinees is called an *affected group*. Note that in CAT aberrant examinees are administered items drawn from the whole CAT item bank. Therefore, it is possible that an aberrant examinee might be administered none or just a few items from the corresponding compromised subset. G*roup* and *subgroup* always refer to examinees; *set* and *subset* always refer to items.

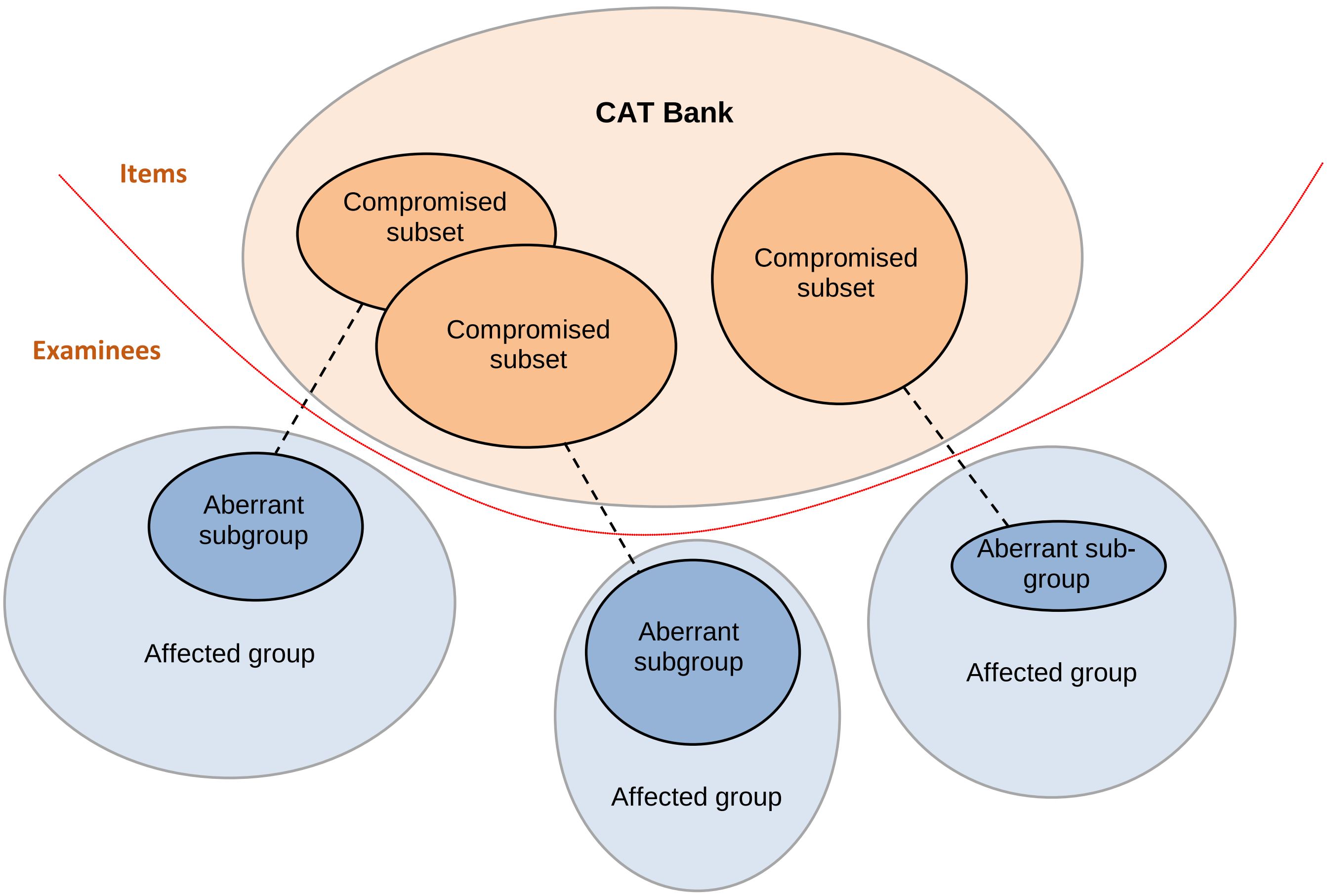
The problem addressed in this paper is now stated as follows: how to detect all triplets, each consisting of an *affected group*, its *aberrant subgroup* of examinees, and the corresponding *compromised subset* of items?

Analysis of the Problem

Due to item preknowledge, if a group X^* is affected (see Figure 1) and the corresponding compromised subset S^* (see Figure 1) is known, then the distribution of a person-fit statistic that is sensitive to item preknowledge, computed for examinees from affected group X^*, should be unusual among distributions of the person-fit statistic computed for examinees from unaffected groups.

Denote a compromised subset of items as S^* and consider a corresponding aberrant examinee $j^* \in X^*$ who is administered a test T_{j^*}. There are multiple person-fit statistics sensitive to a large difference in performance of examinee j^* on administered items from T_{j^*} that belong to

Figure 1. Terminology of Large-Scale Item Preknowledge



$S*$ versus administered items that do not belong to $S*$. Without loss of generality, denote the corresponding person-fit statistic as $d_{S*,j*}$, such that aberrant examinees should be located at the right tail of the corresponding null distribution. Note that the case when $S*\bigcap T_{j*}=\varnothing$ for each aberrant examinee $j*$ eliminates any danger for test security. Therefore, it is assumed that $S*\bigcap T_{j*}\neq\varnothing$ for each aberrant examinee $j*$.

All examinees form group J. Given an arbitrary subset of items S, the empirical distribution of the statistic $d_{S,j}$ computed for examinees from group $X\subset J$ is denoted as $\mathbf{H}_{S,X}$. Given an affected group $X*\subset J$ with corresponding compromised subset $S*$, the item preknowledge causes the empirical distribution $\mathbf{H}_{S*,X*}$ to be dissimilar from the empirical distribution $\mathbf{H}_{S*,Y}$, where Y is an unaffected group. The unaffected group can be created by random sampling of examinees from given data or by simulation. This paper uses unaffected groups with simulated examinees (called *simulated groups*).

Given an analyzed subset of items S, one can estimate how unusual is the distribution of $d_{S,j}$, $j\in X$ at analyzed group $X\subset J$ by computing the following statistic, which is a modification of the statistic g_c by Belov (2013):

$$\mathbf{G}(S,X)=\sum_{i=1}^{n}\left[\mathbf{D}(\mathbf{H}_{S,X}\parallel\mathbf{H}_{S,Y_i})+\mathbf{D}(\mathbf{H}_{S,Y_i}\parallel\mathbf{H}_{S,X})\right], \qquad (1)$$

where $\{Y_1,Y_2,...,Y_n\}$ is a fixed random sample of simulated groups. The function $\mathbf{D}(\mathbf{H}_1\parallel\mathbf{H}_2)$ is computed according to the following equation (Cover & Thomas, 1991; Kullback & Leibler, 1951):

$$\mathbf{D}(\mathbf{H}_1\parallel\mathbf{H}_2)=\sum_{l\in L}\mathbf{H}_1(l)\ln\frac{\mathbf{H}_1(l)}{\mathbf{H}_2(l)}, \qquad (2)$$

where $\mathbf{H}_1$ and $\mathbf{H}_2$ are discrete distributions defined on a finite set L. Due to the definition of Kullback–Leibler divergence (KLD) in Equation 2, the larger the divergence $\mathbf{D}(\mathbf{H}_1\parallel\mathbf{H}_2)$, the higher the dissimilarity between distributions $\mathbf{H}_1$ and $\mathbf{H}_2$. The value of $\mathbf{D}(\mathbf{H}_1\parallel\mathbf{H}_2)$ is always non-negative and equals zero only if the two distributions are identical. KLD is asymmetric; that is, in general, $\mathbf{D}(\mathbf{H}_1\parallel\mathbf{H}_2)\neq\mathbf{D}(\mathbf{H}_2\parallel\mathbf{H}_1)$. The sum in Equation 1 is used to balance the asymmetry of the KLD.

If a group of examinees $X^*\subset J$ is affected with a corresponding compromised subset of items S^*, then the value of $\mathbf{G}(S^*,X^*)$ should be the largest among all subsets of set Q; in other words, $S^*=\arg\max_{S\subset Q}\mathbf{G}(S,X^*)$, where Q contains items with a high risk to be compromised. Practical examples of Q include a previously pretested section in P&P or CBT form, a previously pretested testlet in MST form, or a subset of items with high exposure in previous CAT administrations.

Thus, for a given affected group. Equation 1 can be used as an objective in a combinatorial search for a corresponding compromised subset. But how to find out which groups are affected? To detect affected groups the following statistic is introduced:

$$c_X=\sum_{i=1}^{m}\mathbf{G}(S_i,X), \qquad (3)$$

where subsets $S_1,S_2,...,S_m$ are randomly generated such that $S_i\subset Q$, $i=1,2,...,m$. Since each compromised subset can intersect with multiple probes S_i, the statistic in Equation 3 should have larger values for affected groups. From the above analysis it follows that each affected group X^* can be detected using the statistic c_X, then $\mathbf{G}(S,X^*)$ can be used as an objective in a combinatorial search for a compromised subset S^*.

Detection Algorithm

A new algorithm for detecting large-scale item preknowledge, called the *3D algorithm*, is named as such because it performs the following sequence of detections: (1) **D**etect affected groups; given each affected group, (2) **D**etect a corresponding compromised subset; given each affected group and corresponding compromised subset, (3) **D**etect the aberrant subgroup.

3D Algorithm (Conceptual Overview)

Step A: The set of relations R is built based on data automatically acquired from examinees, which might include the following relations: same geographic location, same high school, same undergraduate college, same test-preparation center, or same group in a social network.

Step B: For each relation $r \in R$ repeat the following steps:

Step 1: Partition all examinees J into groups using the relation r.

Step 2: Each group $X^* \subset J$ having value of c_{X^*} above the corresponding critical value (computed for significance level α_1) is detected as affected.

Step 3: For each affected group X^*, a combinatorial search identifies the corresponding compromised subset S^* maximizing $\mathbf{G}(S, X^*)$, $S \subset Q$. Each examinee j^* from the affected group X^* having value of d_{S^*, j^*} above the corresponding critical value (computed for significance level $\alpha_2 / |R|$) is detected as aberrant.

Having many relations might cause a problem of multiple comparisons. Following the Bonferroni correction (Abdi, 2007), the significance level α_2 is divided by the number of relations in set R (see Step 3 above). The Type I error of the 3D algorithm can be estimated theoretically. If $|R| = 1$ (only one relation is considered) and the number of examinees is 10,000 (100 groups with 100 examinees per each group), then the number of incorrectly detected examinees can be bounded from above by $100 \times \alpha_1 \times 100 \times \alpha_2$ (this is possible if critical values are computed from the empirical distributions, see details below). In general, the number of incorrectly detected examinees can be bounded from above by

$$\sum_{r \in R} |\Omega_r| \alpha_1 \max_{X \in \Omega_r} |X| \alpha_2 / |R|, \tag{4}$$

where Ω_r is a collection of groups partitioning J for a given relation $r \in R$.

Using Simulated Annealing to Detect a Compromised Subset

For each group X^* detected as affected, the 3D algorithm runs a simulated annealing as proposed by Kirkpatrick, Gelatt, & Vecchi (1983) in order to detect subset S^* that would maximize Equation 1. Simulated annealing is a heuristic for finding an optimal solution to a given unconstrained optimization problem. The name comes from annealing in metallurgy, a technique involving heating and controlled cooling of a material to reduce its defects (van Laarhoven & Aarts, 1987). The convergence of simulated annealing can be analyzed by its reduction to a Markov chain (Bertsimas & Tsitsiklis, 1993). Simulated annealing was successfully applied to solve a number of practical problems in psychometrics (Veldkamp, 1999; van der Linden, Veldkamp, & Carlson, 2004; Brusco, Koehn, & Steinley, 2013). To the author's best knowledge, this is the first application of simulated annealing (as well as combinatorial optimization in general) to test security.

For each affected group X^*, the simulated annealing starts with an initial solution that provides the largest value of $\mathbf{G}(S_i, X^*)$ (see random probes $S_1, S_2, ..., S_m$ introduced for Equation 3). The initial solution is assigned to a global solution S^* and a local solution S_0. Then multiple trials are performed to improve $S*$ by exploring the following random modifications of S_0:

$$\begin{array}{llll} \text{Modification:} & 1 & 2 & 3 \\ \text{Probability:} & \mathbf{P}(1) & \mathbf{P}(2) & \mathbf{P}(3) \end{array}, \qquad (5)$$

where

$\mathbf{P}(1)+\mathbf{P}(2)+\mathbf{P}(3)=1$ (actual values are given in the section on computer simulations),

Modification 1 adds a random item from $Q \setminus S_0$ to S_0,

Modification 2 swaps a random item from $Q \setminus S_0$ with the random item from S_0, and

Modification 3 removes the random item from S_0.

If a current random modification results in an improvement of the global solution, then this modification is accepted for both $S*$ and S_0; otherwise, this modification is accepted only for S_0 with a probability, which is gradually decreasing according to a “cooling” schedule. Accepting non-optimal modifications prevents getting stuck in a local maximum.

Simulated annealing has the following “cooling” schedule. Probability of accepting a non-optimal modification depends on the parameter t (called temperature). The initial value of t is $t_0 = 10 \max\limits_{X \subset J} \max\limits_{i=1}^{m} G(S_i, X)$. After each trial the temperature t is reduced as $t = t \times d$, where $0 < d < 1$ (called the *cooling parameter*). The simulated annealing terminates when $t \le t_0 / h$. Parameters t_0, h, and d control the convergence of the simulated annealing to a local (possibly global) maximum: the larger their values the slower the convergence but larger the value of Equation 1.

Detailed Description of the 3D Algorithm

The following describes detailed steps of the 3D algorithm for a currently selected relation $r \in R$. Step B selects each group with the value of Equation 3 from a critical region identified at Step A. For each selected group $X*$, Steps 1–5 implement simulated annealing for a combinatorial search of a corresponding compromised subset $S*$. Step 7 detects each examinee $j* \in X*$ with value of the statistic $d_{S*,j*}$ from a critical region identified at Step 6.

Step A: For each group $X \subset J$ compute the statistic c_X according to Equation 3. Given the significance level α_1 compute the critical value v_1 as the $(1-\alpha_1)$ percentile of c_X, $X \subset J$.

Step B: For each group $X* \subset J$, such that $c_{X*} > v_1$, repeat the following steps:

Step 1: Set the global solution $S* = \arg\max\limits_{i=1}^{m} \mathbf{G}(S_i, X*)$, the local solution $S_0 = S*$, and the temperature $t = t_0$.

Step 2: Set $S = S_0$. Simulate random variable $\delta \in \{1,2,3\}$ according to the discrete distribution of Expression 5 and modify S, respectively.

Step 3: If $\mathbf{G}(S, X*) > \mathbf{G}(S*, X*)$, indicating that an improvement to the global solution has been found, then set $S_0 = S$ and $S* = S$ (update the global solution) and go to Step 5; otherwise continue to Step 4.

Step 4: Simulate uniformly distributed $\gamma \in [0,1)$. If $\gamma < \exp\left(\left[\mathbf{G}(S, X^*) - \mathbf{G}(S_0, X^*)\right]/t\right)$ (probability of accepting a modification to the local solution S_0 that did not improve the global solution S^*) then set $S_0 = S$ (update the local solution).

Step 5: If $t > t_0 / h$ then $t = t \times d$ and go to Step 2 (perform more trials to improve the global solution); otherwise continue to Step 6 (no more trials).

Step 6: Given significance level α_2, compute the critical value v_2 as $(1-\alpha_2)$ percentile of $d_{S^*,j}$, $j \in X^*$ (i.e., the statistic $d_{S^*,j}$ is computed only for examinees from X^*). This will guarantee that Expression 4 is an upper bound for the Type I error.

Step 7: Each examinee $j^* \in X^*$ with $d_{S^*,j^*} > v_2$ is detected as aberrant.

Generalized 3D Algorithm

The 3D algorithm can be generalized to detect multiple aberrant subgroups within each affected group. This generalization is described below, where for each affected group X^* at Step 2, all potentially compromised subsets are detected and added to the corresponding collection Δ_{X^*}. For each affected group X^* at Step 3, aberrant examinees with different compromised subsets from the collection Δ_{X^*} are detected, where the Bonferroni correction (Abdi, 2007) takes into account the size of the collection Δ_{X^*}. A computational study analyzing performance of this generalization goes beyond this paper.

Steps in the Generalized 3D Algorithm

Step A: A set of relations R is built based on data automatically acquired from examinees, which might include the following relations: same geographic location, same high school, same undergraduate college, same test-preparation center, or same group in a social network.

Step B: For each relation $r \in R$ repeat the following steps:

Step 1. Partition all examinees J into groups using the relation r. For each group $X \subset J$ compute the statistic c_X. Given the significance level α_1, compute the critical value v_1 of c_X, $X \subset J$. Each group $X^* \subset J$ with $c_{X^*} > v_1$ is added to a collection of affected groups Ψ and the corresponding collection of compromised subsets is initialized $\Delta_{X^*} := \{\varnothing\}$; also, a clone of X^* denoted as $\tilde{X}^*$ is added to a collection $\tilde{\Psi}$.

Step 2. Select the first group $\tilde{X}^*$ from the collection $\tilde{\Psi}$. A combinatorial search identifies the corresponding compromised subset S^* maximizing $\mathbf{G}(S, \tilde{X}^*)$. Add S^* to the collection Δ_{X^*} of potentially compromised subsets for group X^*. Given the significance level $\tilde{\alpha}$ compute the critical value $\tilde{v}$ of $d_{S^*,j}$, $j \in \tilde{X}^*$. Each examinee $j \in \tilde{X}^*$ with $d_{S^*,j} > \tilde{v}$ is substituted by a simulated non-aberrant examinee. Recompute $c_{\tilde{X}^*}$. If $c_{\tilde{X}^*} \le v_1$ then remove $\tilde{X}^*$ from the collection $\tilde{\Psi}$ (no more aberrant subgroups left in $\tilde{X}^*$). Repeat Step 2 until the collection $\tilde{\Psi}$ becomes empty.

Step 3. At each aberrant group $X^* \in \Psi$ perform the following two steps for each compromised subset $S^* \in \Delta_{X^*}$:

Step 3.1. For significance level $\alpha_2 / |R| / |\Delta_{X^*}|$ compute the critical value v_2 of $d_{S^*,j}$.

Step 3.2. Each examinee $j \in X^*$ with $d_{S^*,j} > v_2$ is detected as aberrant.

Computer Simulations

This section presents two experiments. The first experiment demonstrates that the use of information about compromised items does help detect aberrant examinees with high power. The second experiment demonstrates that when this information has relatively small noise the power drops dramatically but the 3D algorithm is able to prevent this drop.

Three detectors of item preknowledge were compared. The first detector applied the l_z statistic (Drasgow, Levine, & Williams, 1985) and it was used as a baseline in all studies. The second detector applied the KLD statistic—a person-fit statistic developed by Belov et al. (2007). The third detector was the 3D algorithm incorporating the KLD statistic. All detectors were implemented in C++ by the author.

Detector Based on the l_z Statistic

l_z (Drasgow et al., 1985) is a normalization of the statistic by Levine & Rubin (1979) and it is computed as follows:

$$d_j = -\frac{\lambda - e_\lambda}{v_\lambda}, \tag{6}$$

$$\lambda = \sum_{i \in T_j} \chi_i \ln \mathbf{P}_i(1|\theta) + (1-\chi_i) \ln \mathbf{P}_i(0|\theta), \tag{7}$$

$$e_\lambda = \sum_{i \in T_j} \mathbf{P}_i(1|\theta) \ln \mathbf{P}_i(1|\theta) + \mathbf{P}_i(0|\theta) \ln \mathbf{P}_i(0|\theta), \tag{8}$$

$$v_\lambda = \sum_{i \in T_j} \mathbf{P}_i(1|\theta) \ln \mathbf{P}_i(1|\theta) \left(\ln \frac{\mathbf{P}_i(1|\theta)}{\mathbf{P}_i(0|\theta)} \right)^2, \tag{9}$$

where $\chi_i \in \{0,1\}$ is the observed response to item $i \in T_j$, $\mathbf{P}_i(1|\theta)$ is the probability of a correct response to item i for latent trait θ, and $\mathbf{P}_i(0|\theta) = 1 - \mathbf{P}_i(1|\theta)$. However, the true value of θ is unknown in practice and an estimate $\hat{\theta}$ is commonly used instead. This study used the expected a posteriori (EAP) estimator with a uniform prior.

The statistic l_z is often used as a baseline in computational studies of item preknowledge (Levine & Drasgow, 1988; Karabatsos, 2003; Shu et al., 2013). However, l_z does not explicitly take into account information about compromised items. l_z is operationalized in three steps:

1. For each examinee $j \in J$ compute d_j (Equations 6–9).
2. Given a significance level α, compute the critical value v as the $(1-\alpha)$ percentile of $d_j,\ j \in J$.
3. Each examinee $j \in J$ with $d_j > v$ is detected as aberrant.

Detector Based on the KLD Statistic

The KLD between posteriors of θ computed from responses to uncompromised and compromised items, respectively, was chosen as a person-fit statistic (Belov et al., 2007). This person-fit statistic was demonstrated to be effective for detecting item preknowledge in CAT (Belov, 2011, 2013; Chao, Chen & Chen, 2011) and P&P tests (Belov et al., 2007).

More precisely, given a subset of items S and examinee j who is administered a set of items T_j, the person-fit statistic $d_{S,j} = \mathbf{D}(\mathbf{P}_{T_j \setminus S} \,\|\, \mathbf{P}_{S \cap T_j})$ (Equation 2) is computed, where $\mathbf{P}_{T_j \setminus S}$ is the posterior of θ computed from responses of examinee j to administered items that do not belong to S, and $\mathbf{P}_{S \cap T_j}$ is the posterior of θ computed from responses of examinee j to administered items that belong to S. If the examinee j has preknowledge of items in subset S, then the distribution $\mathbf{P}_{S \cap T_j}$ will be shifted toward higher levels of θ more than the distribution $\mathbf{P}_{T_j \setminus S}$ will. Clearly, this shift will be even larger for lower ability examinees. Since low-ability examinees involved in item preknowledge have the largest negative impact on the scoring, the statistic $d_{S,j}$ is highly practical. Note that the use of the word *shift* is only for illustrative purposes; the actual dissimilarity might not be so easily described. However, this dissimilarity can be measured by the KLD $\mathbf{D}(\mathbf{P}_{T_j \setminus S} \,\|\, \mathbf{P}_{S \cap T_j})$ computed by Equation 2. The KLD detector is, therefore, implemented with the following three steps:

1. Assume a subset of compromised items $S \subset Q$. For each examinee $j \in J$ compute $d_{S,j} = \mathbf{D}(\mathbf{P}_{T_j \setminus S} \,\|\, \mathbf{P}_{S \cap T_j})$.
2. Given a significance level α ,compute the critical value v as the $(1-\alpha)$ percentile of $d_{S,j},\ j \in J$.
3. Each examinee $j \in J$ with $d_{S,j} > v$ is detected as aberrant.

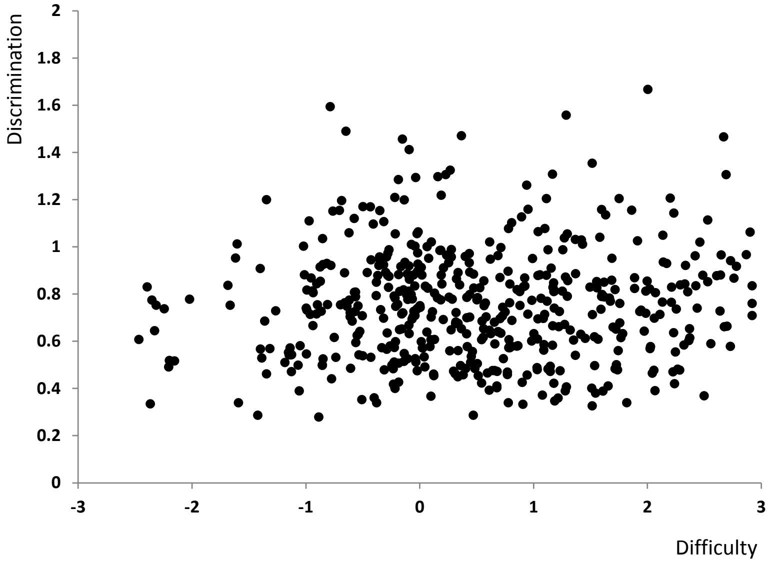
Simulation Design

Multiple simulation studies were conducted using disclosed real-life Logical Reasoning (LR) items of the Law School Admission Test (LSAT). The response probability for each item was modeled by the three-parameter logistic (3PL) model with D = 1.7 (Lord, 1980). The CAT item bank consisted of 500 LR items. The distribution of discrimination (a_i), difficulty (b_i), and guessing (c_i) parameters of the items in the CAT bank had the following minimums, maximums, means, and standard deviations, respectively: (a_i), minimum 0.28, maximum 1.67, mean 0.75, standard deviation 0.24; (b_i) minimum −2.47, maximum 2.92, mean 0.49, standard deviation 1.13; and (c_i) minimum 0.00, maximum 0.52, mean 0.17, standard deviation 0.1. The distribution of discrimination and difficulty in the CAT bank is shown in Figure 2.

The item selection criterion for CAT was the maximization of Fisher information at the current estimate of ability $\hat{\theta}$. The estimator of θ was the EAP estimator with a uniform prior. The ability estimate was initialized at $\hat{\theta} = 0$. The test length was fixed at 50 items for each examinee. The item exposure constraint was set to 0.4.

The Type I error rate was computed as follows (this was an empirical probability for an examinee to be falsely detected):

Figure 2. Distribution of Discrimination and Difficulty in the CAT Item Bank



$$\frac{[\text{number of detected examinees}]-[\text{number of correctly detected examinees}]}{[\text{number of all nonaberrant examinees}]}. \qquad (10)$$

The detection rate was computed by

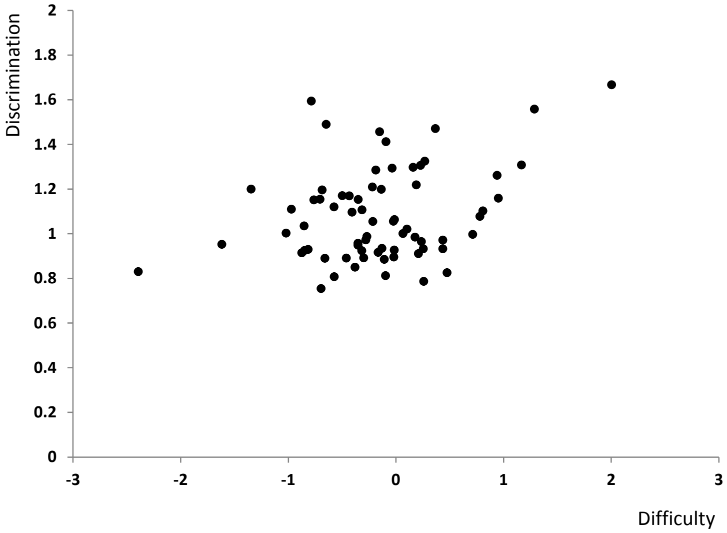
$$\frac{[\text{number of correctly detected examinees}]}{[\text{number of all aberrant examinees}]}. \qquad (11)$$

Each simulation study was performed as follows:

1. All 10,000 examinees were randomly partitioned into 100 test centers (groups) with approximately 100 examinees per test center. Non-aberrant examinees were simulated with abilities drawn from N(0, 1).
2. CAT was simulated without item preknowledge (no aberrant examinees), where item exposure was bounded from above by 0.4.
3. A set, Q, of items with the highest exposure (equal to 0.4 because the item exposure constraint was 0.4) was created. This step resulted in Q with 69 items (Figure 3). Set Q represented potentially compromised items.
4. Aberrant examinees were simulated with abilities drawn from U(–3,–2), U(–2,–1), or U(–1,0), where the probability of correct response to a compromised item equals 1 (since memorizing a correct answer for a multiple-choice item is trivial). The percentage of affected test centers was 1% and 5%. The percentage of aberrant examinees in each affected test center was 10% and 20%. At each affected test center, aberrant examinees substituted random non-aberrant examinees such that the total number of examinees per each affected test center fluctuated around 100 in order to size affected test centers similarly to unaffected test centers.

5. Each affected test center X_k^* was assigned a unique random compromised subset $S_k^* \subset Q$, such that $\left|S_k^*\right| = 45$ and compromised items were drawn uniformly from Q. Steps 1–5 simulate a realistic scenario of item preknowledge. Although more exposed items have higher probability of being compromised, it does not guarantee that each aberrant subgroup of examinees will have access to all of these items prior to the exam. Such a realistic design makes the detection problem much more difficult than in the current literature. Often, researchers assume that the compromised items have high difficulty (Karabatsos, 2003; McLeod et al., 2003), which is not true in real CAT where highly exposed items (with highest risk to be compromised) have difficulties fluctuating around 0 (see Figure 3). Also, researchers assume a large number of compromised items administered to each aberrant examinee; for example, Shu et al., (2013) assumed lower bounds of 30%, 50%, and even 70%. The present study assumed a lower bound of only 20%, similar to Karabatsos (2003).
6. Simulate CAT with item preknowledge, where item exposure was bounded from above by 0.4.
7. Run each item preknowledge detector and compute its Type I error rate (Equation 10) and detection rate (Equation 11).

Figure 3. Distribution of Discrimination and Difficulty in Q



Parameters for the Detectors

Each posterior of θ was computed for the following θ levels: −5, −4.9, − 4.8, …, 5. Detectors l_z and KLD ran for significance level α of 0.005, 0.01, 0.02, 0.03, 0.04, and 0.05. The 3D algorithm ran under the following conditions:

1. The KLD statistic was used as the person fit statistic $d_{S,j}$.
2. Parameters for detecting affected test centers:
    a. Given test center X and subset of items S, the empirical distribution $\mathbf{H}_{S,X}$ of the KLD statistic $d_{S,j}$ was computed using five bins [0, 10], [10, 20], …, [40, 50].

   b. Number of simulated test centers $Y_1, Y_2, ..., Y_n$ used in Equation 1 was $n = 10$.
   c. Number of random probes used in Equation 3 was $m = 100$, where for each random probe S_i, $i = 1, 2, ..., m$, $S_i \subset Q$ and $0.2|Q| \le |S_i| \le 0.8|Q|$.
   d. Significance level $\alpha_1 = 0.1$; thus, it was assumed that the number of affected test centers could not exceed 10% of all test centers.
3. Parameters for searching each compromised subset by the simulated annealing:
   a. Every possible modification to a local solution had equal probability to be selected, i.e., $\mathbf{P}(1) = \mathbf{P}(2) = \mathbf{P}(3) = 1/3$.
   b. Parameters defining cooling schedule: $h = 10000$, $d = 0.95$.
4. Parameters for detecting aberrant examinees: The significance level $\alpha_2 = \alpha / \alpha_1$ was chosen from {0.05, 0.1, 0.2, 0.3, 0.4, 0.5} to guarantee that the 3D algorithm had the same Type I error rate as the other detectors.

Experiment 1

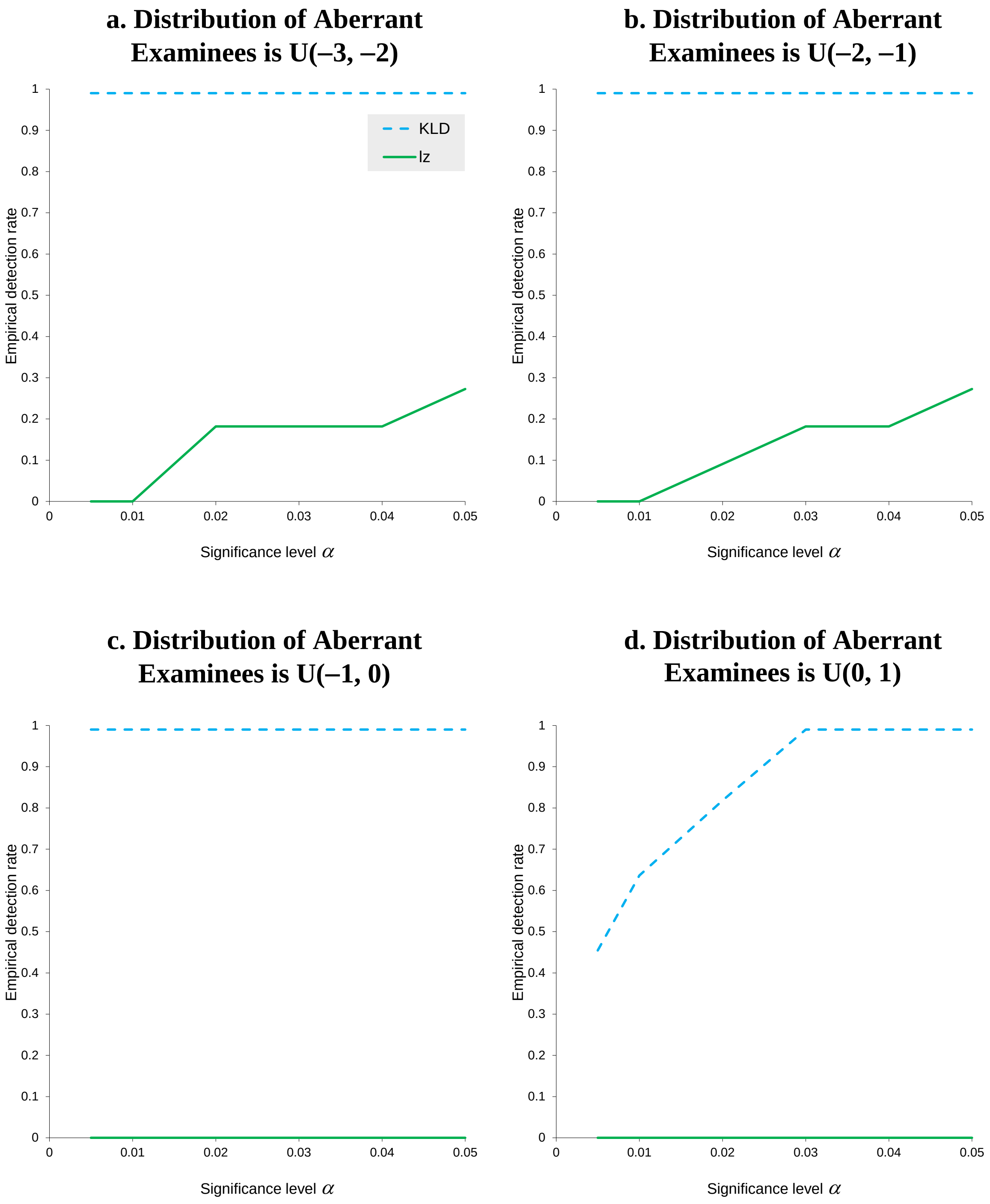
This experiment was designed to determine if the use of information about compromised items allowed detecting aberrant examinees with high power. Two detectors were compared (l_z and KLD), where only one test center was affected (with 10 aberrant examinees) and the corresponding compromised subset was known. There were four scenarios for four different distributions of aberrant examinees: U(–3,–2), U(–2,–1), U(–1,0), and U(0,1). In all scenarios, both detectors resulted in similar Type I error rates, which were equal or just under nominal significance level $\alpha \in \{0.005, 0.01, 0.02, 0.03, 0.04, 0.05\}$. This is expected behavior because both detectors used critical values computed as percentiles of corresponding empirical distributions (see above). The empirical detection rates are presented in Figure 4, where it can be seen that KLD was sensitive to the θ distribution of aberrant examinees (compare its detection rates for distribution U(0,1) with other distributions). Overall, KLD clearly outperformed l_z. This is due to the fact that the compromised subset was known. But what will happen when the compromised subset is unknown and it is assumed that the compromised items are the whole set Q? Intuitively, a large drop in detection rates would not be expected for KLD (or for any other statistic that explicitly uses information about compromised items) since each compromised subset was 45 items long and all of these items were chosen randomly from set Q with 69 items. In other words, the noise in the assumption seems relatively small. A counterintuitive answer to this question is given by Experiment 2.

Experiment 2

This experiment considered 3 (distribution of aberrant examinees) × 2 (number of affected test centers) × 2 (number of aberrant examinees at each affected test center) + 1 = 13 scenarios simulating item preknowledge in CAT. Thus, 13 × 10,000 = 130,000 response vectors were analyzed for item preknowledge by three different detectors. One additional scenario simulated CAT without item preknowledge. Total running time on a personal computer with Intel® Core™ i7 CPU 860 2.8 GHz was about one hour.

In each scenario, all detectors resulted in similar Type I error rates, which were equal or just under the nominal significance level $\alpha \in \{0.005, 0.01, 0.02, 0.03, 0.04, 0.05\}$. This was expected behavior because each detector used critical values computed as percentiles of corresponding empirical distributions (see Step 2 of the l_z and KLD detectors; and Steps A and 6 of the detailed description of the 3D algorithm).

Figure 4. Detection Rates With One Affected Test Center and 10 Aberrant Examinees Drawn From Different Distributions
(Solid Green Line Corresponds to Detector l_z; Short Dashed Blue Line Corresponds to Detector KLD With Known Compromised Subset)



a. Distribution of Aberrant Examinees is U(–3, –2)

b. Distribution of Aberrant Examinees is U(–2, –1)

c. Distribution of Aberrant Examinees is U(–1, 0)

d. Distribution of Aberrant Examinees is U(0, 1)

The detection rates for the 12 scenarios modeling item preknowledge are presented in Figures 5–7. The detection rate of KLD dropped to zero (compare Figures 5–7 with Figure 4). This drop agrees with the results of simulation studies by Hui (2010) using a person-fit statistic proposed by McLeod et al., (2003). This drop indicates that all detectors explicitly relying on information about compromised items are not stable in the presence of noise. At the same time, the use of the 3D algorithm prevented this drop (Figures 5–7), which can be explained by the following:

1. Simulated annealing provides a good approximation of each compromised subset, which decreases the amount of noise in the information about compromised items.
2. The nested structure of the 3D algorithm (first, detect affected test centers and then, for each affected test center detect its aberrant examinees) allows using a larger (than targeted) significance level for detecting aberrant examinees, which results in higher power but still with the targeted Type I error rate (see Equation 4).
3. For the simulation design described above, the statistic in Equation 3 for detecting affected test centers performed well. In each scenario modeling item preknowledge, the 3D algorithm detected all affected test centers.

Discussion and Conclusions

Item preknowledge is difficult to detect due to multiple unknowns involved—unknown subgroups of examinees from unknown affected groups (e.g., affected test centers) accessing unknown compromised subsets of items prior to taking the test. The major objective of this research was to explore the possibility of disentangling this problem using combinatorial optimization. This is an important practical problem because if compromised items can be identified then it is possible to detect aberrant examinees with high power (see the results of Experiment 1 in Figure 4).

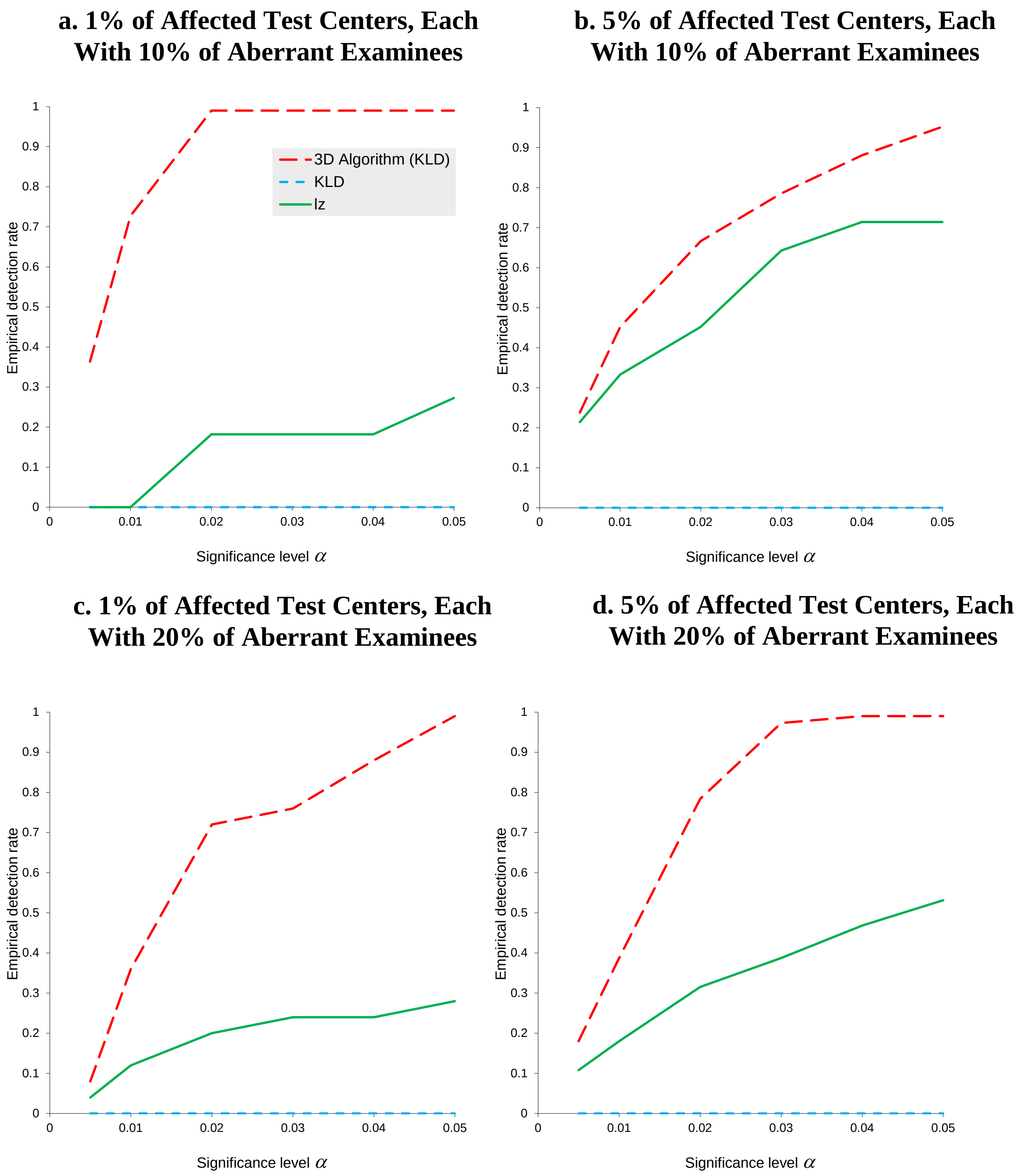
The major result of this study is formulated as the 3D algorithm that performs the following sequence of nested steps:

1. Detect *affected groups* of examinees.
2. For each affected group, detect the corresponding *compromised subset* of items using combinatorial optimization.
3. For each affected group and corresponding compromised subset, detect an *aberrant subgroup* of examinees.

In computer simulations, the 3D algorithm outperformed two modern detectors (l_z and KLD) and demonstrated great promise to meet the objective (see Figures 5–7). In particular, multiple aberrant subgroups were simulated, each having access to a unique random subset of set Q, where Q contained items that had high exposure in previous CAT administrations. When the KLD detector was applied alone (under the incorrect assumption that the compromised subset was the whole Q), the resultant detection rates dropped to zero (compare detection rates of KLD from Figure 4 to Figures 5–7). At the same time, when the KLD statistic was incorporated into the 3D algorithm the resultant detection rates were higher than for the detector l_z (see Figures 5–7).

Performance of the 3D algorithm depends on a good choice of the set Q. Ideally, this set should include only compromised items. In other words, for aberrant subgroups of examinees $E_1, E_2, \ldots$ with corresponding compromised item subsets $Z_1, Z_2, \ldots$ the ideal choice of Q is $Q = Z_1 \cup Z_2 \cup \ldots$. This study used a simple method of identifying Q as a set of items with expo-

Figure 5. Detection Rates When Aberrant Examinees Were Drawn From U(–3, –2)



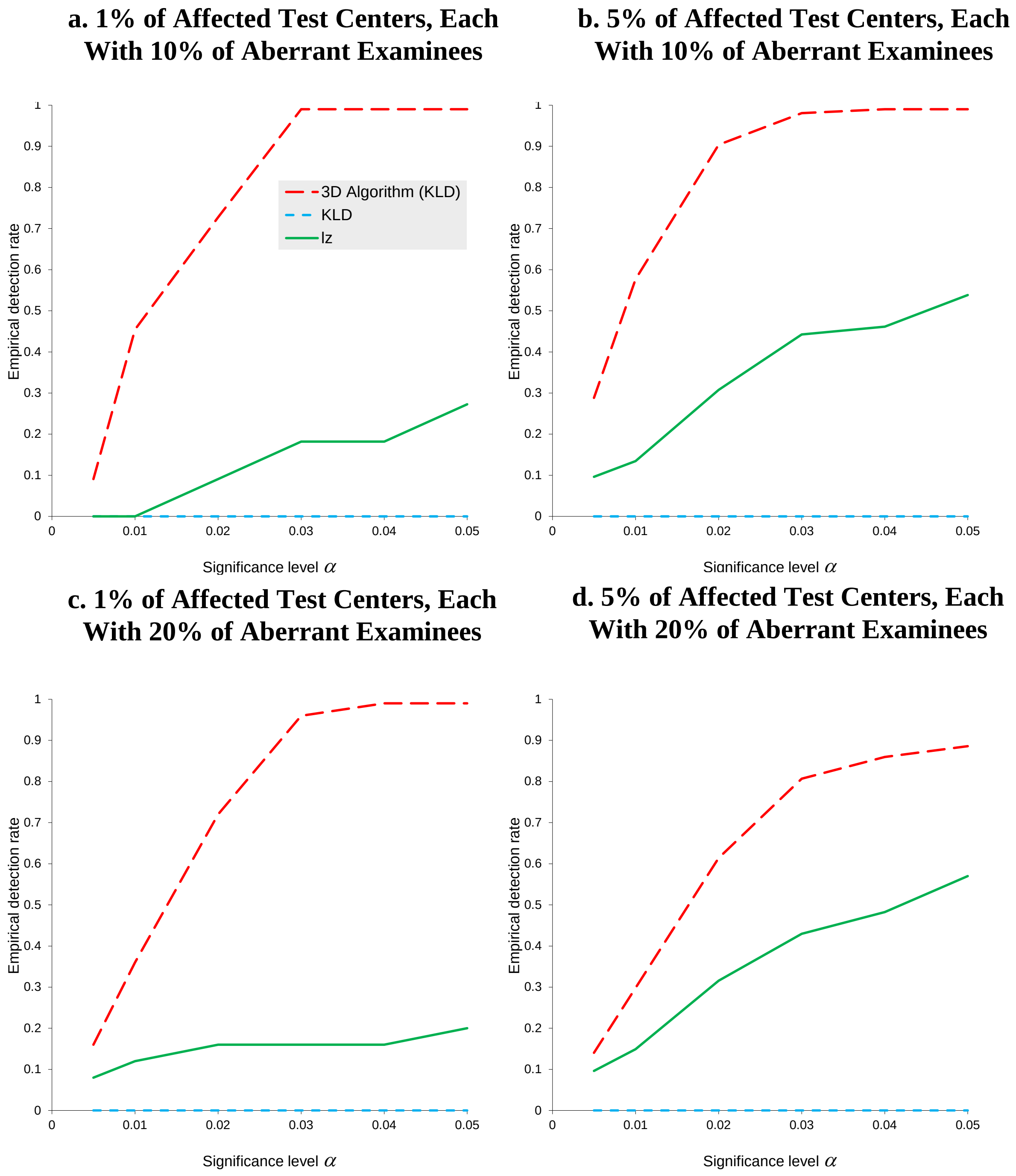
a. 1% of Affected Test Centers, Each With 10% of Aberrant Examinees

b. 5% of Affected Test Centers, Each With 10% of Aberrant Examinees

c. 1% of Affected Test Centers, Each With 20% of Aberrant Examinees

d. 5% of Affected Test Centers, Each With 20% of Aberrant Examinees

Figure 6. Detection Rates When Aberrant Examinees Were Drawn From U(−2, −1)



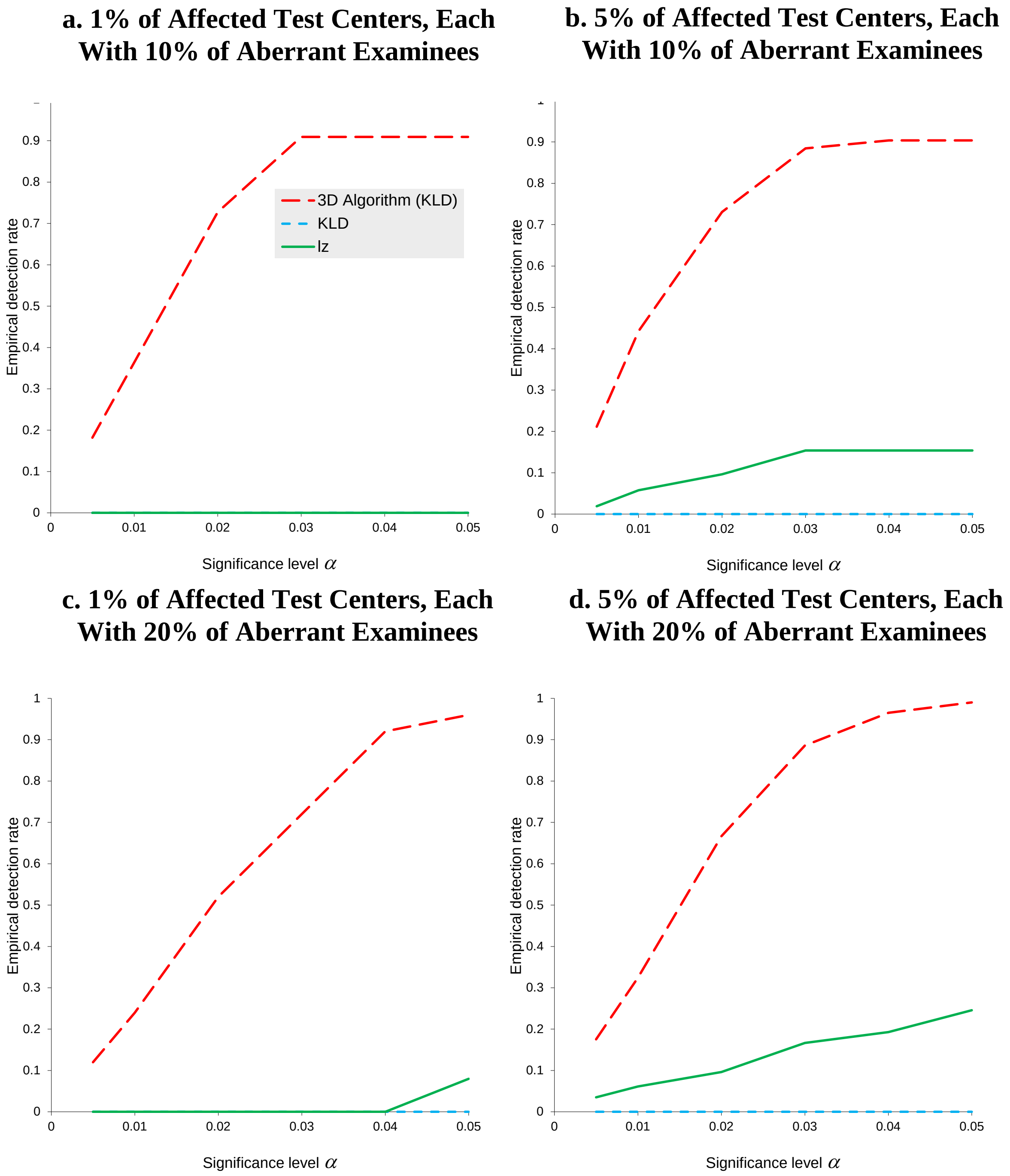
a. 1% of Affected Test Centers, Each With 10% of Aberrant Examinees

b. 5% of Affected Test Centers, Each With 10% of Aberrant Examinees

c. 1% of Affected Test Centers, Each With 20% of Aberrant Examinees

d. 5% of Affected Test Centers, Each With 20% of Aberrant Examinees

Figure 7. Detection Rates When Aberrant Examinees Were Drawn From U(–1, 0)



a. 1% of Affected Test Centers, Each With 10% of Aberrant Examinees

b. 5% of Affected Test Centers, Each With 10% of Aberrant Examinees

c. 1% of Affected Test Centers, Each With 20% of Aberrant Examinees

d. 5% of Affected Test Centers, Each With 20% of Aberrant Examinees

sure (in previous CAT administration) higher than a fixed threshold. Clearly, the detection rate of the 3D algorithm might drop if this threshold is too high (Q does not include enough compromised items) or too low (Q includes too many uncompromised items). Recently, multiple statistical methods were developed to detect compromised items (Choe, 2014; Obregon, 2013; van der Linden & Guo, 2008). These methods might potentially improve the identification of Q.

Given the subset Q, the performance of the 3D algorithm mainly depends on the following parameters: m, the number of random probes (see Equation 3), and two parameters of the cooling schedule h and d. With a higher value of the m parameter, more affected groups can be detected and better initial solutions for the simulated annealing can be found. Higher values of h and d allow simulated annealing to get closer to each compromised subset. Thus, higher values of these parameters increase the resultant detection rate but decrease speed. In order to increase speed, the 3D algorithm can be parallelized for modern multicore CPUs by running simulated annealing for each affected group in a separate thread.

Large-scale item preknowledge is a type of test collusion. Test collusion can be described as large-scale sharing of test materials or answers to test questions. The source of the shared information could be a teacher, a test-preparation company, the Internet, or examinees communicating on the day of the exam (Wollack & Maynes, 2011). Any type of test collusion causing an unusual change of examinees' performance from one subset of items to another can be detected by the 3D algorithm incorporating the corresponding person-fit statistic. Practical examples of test collusion detectable by the 3D algorithm include: a teacher correcting answers to difficult items for a subgroup of students in a class, and a subgroup of examinees working together on a subset of items at some test center. Clearly, the 3D algorithm is applicable to all major test delivery methods: P&P, CBT, MST, and CAT.

This research and its results should be interpreted as a work in progress because of the following:

1. The assumption about one aberrant subgroup per affected group (Figure 1), though certainly realistic, is a limitation. This limitation is addressed in this paper by formulation of the generalized 3D algorithm (see above), but its study goes beyond this paper.
2. The first stage of the 3D algorithm (detecting affected groups) is crucial for its overall performance because if an affected group is undetected then all its examinees will skip further stages of detection. The 3D algorithm uses a high value of $\alpha_1 = 0.1$ and employs random search (see Equation 3), which might converge too slowly when subset Q is larger and/or compromised subsets are smaller. Therefore, the first stage needs a separate study and, perhaps, an improvement.
3. Computational studies were rather limited since their purpose was only to demonstrate a proof of concept. Comparison study using real data, where all aberrant examinees are tagged, is needed.
4. The 3D algorithm (as well as the generalized 3D algorithm) is an algorithmic framework where embedded subroutines and statistics can be modified in order to improve overall performance for a specific testing program and/or type of test collusion. The following modifications are possible:
   a. The 3D algorithm can be applied for CAT, MST, and CBT with posteriors of speed [see van der Linden (2011) for details on response time modeling]. In this case, the following person-fit statistics can be used:

(12)

$$d_{S,j} = \mathbf{D}(\mathbf{P}_{T_j \setminus S} \,\|\, \mathbf{P}_{S \cap T_j}) + \mathbf{D}(\mathbf{V}_{T_j \setminus S} \,\|\, \mathbf{V}_{S \cap T_j})$$

$$\begin{aligned} d_{S,j} = {} & \mathbf{D}(\mathbf{P}_{T_j \setminus S} \,\|\, \mathbf{P}_{S \cap T_j}) + \mathbf{D}(\mathbf{P}_{S \cap T_j} \,\|\, \mathbf{P}_{T_j \setminus S}) + \\ & \mathbf{D}(\mathbf{V}_{T_j \setminus S} \,\|\, \mathbf{V}_{S \cap T_j}) + \mathbf{D}(\mathbf{V}_{S \cap T_j} \,\|\, \mathbf{V}_{T_j \setminus S}), \end{aligned} \tag{13}$$

where $\mathbf{V}_{S \cap T_j}$ is the posterior of speed computed from response times of examinee j to the administered items T_j that belong to S, and $\mathbf{V}_{T_j \setminus S}$ is the posterior of speed computed from response times of examinee j to the administered items T_j that do not belong to S. The statistics in Equations 12 and 13 use additional information from examinees (response times) which should improve the overall performance of the 3D algorithm.

b. The following alternatives for the combinatorial search of the compromised subsets can substitute the simulated annealing within the 3D algorithm: greedy heuristic (Papadimitriou & Steiglitz, 1982); genetic algorithm (Mitchell, 1996); or tabu search (Glover & Laguna, 1997).

c. Critical values for the 3D algorithm can be estimated via asymptotic distributions or empirical distributions computed from simulated data.

References

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics*. Thousand Oaks, CA: Sage.

Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, *33*, 391–410. CrossRef

Belov, D. I. (2011, October). *Detection of test collusion in computerized adaptive testing*. Paper presented at the IACAT Conference, Pacific Grove, CA.

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

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

Bertsimas, D., & Tsitsiklis, J. (1993). Simulated annealing. *Statistical Science*, *8*(1), 10–15. CrossRef

Brusco, M. J., Koehn, H.-F., & Steinley, D. (2013). Exact and approximate methods for a one-dimensional minimax bin-packing problem. *Annals of Operations Research, 206, 611-626.* CrossRef

Chao, H. Y., Chen, J. H., & Chen, S. Y. (2011, July). *Applying Kullback-Leibler divergence to detect examinees with item pre-knowledge in computerized adaptive testing*. Paper presented at the international meeting of the Psychometric Society, Hong Kong.

Choe, E. (2014, April). *Utilizing response time in sequential detection of compromised items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley & Sons, Inc. CrossRef

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. CrossRef

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

Hui, H.-fai. (2010). *Stability and sensitivity of a model-based person-fit index in detecting item pre-knowledge in computerized adaptive test*. Dissertation Abstracts International Section A: Humanities and Social Sciences. University of Hong Kong.

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

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

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86. CrossRef

Levine, M. V. & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53, 161–176. CrossRef

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269–290. CrossRef

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

McLeod, L. D., Lewis, C., & Thissen, D. (2003). A Bayesian method for detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*, 121–137. CrossRef

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

Obregon, P. (2013, April). *A Bayesian approach to detecting compromised items*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, *78*(3), 481–497. CrossRef

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. CrossRef

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

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

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement*, *28*, 317–331. CrossRef

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. CrossRef

van Laarhoven, P. J. M., & Aarts, E. H. L. (1987). *Simulated annealing: Theory and applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers. CrossRef

Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement*, *36*, 253–266. CrossRef

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.

Acknowledgments

The author thanks the editor, David J. Weiss, the associate editor, Bernard P. Veldkamp, and two anonymous reviewers for their valuable comments and suggestions on previous versions of the manuscript.

Author Address

Dmitry I. Belov, Psychometric Research, Law School Admission Council, 662 Penn Street, Newtown, PA 18940, U.S.A. Emails: dbelov@lsac.org; belovd@mail.ru