LSAC RESEARCH REPORT SERIES

- Comparison Study of Item Preknowledge Detectors
  Dmitry I. Belov

- Law School Admission Council
  Research Report 14-01
  March 2014
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, 218 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 98 percent of all applications to ABA-approved law schools.

© 2014 by Law School Admission Council, Inc.

All rights reserved. No part of this work, including information, data, or other portions of the work published in electronic form, may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, 662 Penn Street, PO Box 40, Newtown, PA, 18940-0040.

This study is published and distributed by LSAC.
# Table of Contents

**Executive Summary** ................................................................................................................. 1

**Introduction** ................................................................................................................................. 1

**Item Preknowledge Detection Statistics** .................................................................................. 2
  - Statistic \( I_z \) ................................................................................................................................. 3
  - Statistic Based on Score Ratio ........................................................................................................ 4
  - Statistic Based on Ability Difference ............................................................................................. 4
  - Statistic Based on Kullback–Leibler Divergence ........................................................................... 4
  - Statistic Based on Posterior Shift ................................................................................................... 4
  - Statistic Based on the Approach by Shu, Henson, and Luecht ..................................................... 5
  - Modified \( I_z \) .................................................................................................................................. 5
  - Statistic Based on Neyman–Pearson Lemma .................................................................................. 6

**Computer Simulations** ................................................................................................................. 7
  - Setup ............................................................................................................................................. 7
  - Results ........................................................................................................................................... 9

**Summary** ....................................................................................................................................... 15

**References** .................................................................................................................................... 18
Executive Summary

When a test taker has prior knowledge about an administered test question (item), then this event is called item preknowledge, the test taker is called aberrant, and the item is called compromised. Item preknowledge negatively affects the corresponding testing program and its test score users (universities, companies, government organizations) because the scores produced for aberrant test takers will be invalid. The performance of eight statistics for detection of item preknowledge (five existing, two modified, and one new) was studied via computer simulations. Three major factors that could potentially influence the performance of the statistics were considered: (a) the type of test (adaptive, in which the next administered item is selected based on the test taker’s responses to previously administered items; or nonadaptive, in which all test takers are administered the same group of items); (b) distribution of the aberrant population (normal or uniform); and (c) noise in the information about compromised items (since different groups of aberrant test takers may have prior knowledge of different groups of items). The last factor demonstrated the highest negative impact on the performance of all of the statistics: the greater the noise, the lower the detection of item preknowledge. Several methods to address this problem are discussed.

Introduction

Item preknowledge occurs when some test takers (aberrant test takers) have access to some items (compromised items) from an administered test prior to an exam. As a result, aberrant test takers perform better on compromised items as compared to uncompromised items. When the number of aberrant test takers is large, the corresponding testing program (paper-and-pencil testing [P&P], computer-based testing [CBT], multiple-stage testing [MST], and computerized adaptive testing [CAT]) and its users (universities, companies, government organizations, etc.) are negatively affected because the scores produced for aberrant test takers will be invalid.

Item preknowledge is a special case of test collusion that has recently received much attention in test security research and practice (Maynes, 2013). Test collusion may 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 test takers communicating on the day of the exam (Wollack & Maynes, 2011). Often, the detection of different types of test collusion can be reduced to the detection of item preknowledge. For example, both (a) a teacher correcting answers to hard items for a group of students and (b) students working together on a subset of items are events that can be detected as item preknowledge.

There are numerous statistics that can be applied to detect item preknowledge (Karabatsos, 2003). Eight statistics (five existing, two modified, and one new) were selected for the comparison study based on their ability to be sensitive to differences in test takers’ responses to compromised versus uncompromised items. This report follows Karabatsos (2003) and uses the area under the receiver operating characteristic (ROC) curve as the measure of performance. Multiple factors potentially influencing the performance of the statistics are considered. From a practical standpoint, one needs to
know answers to the following questions. Which statistic has the best performance? Which factor has the most negative influence on the performance? Which statistic is least affected by the most negative factor? How can we alleviate the most negative factor? This report provides answers to these questions.

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; an estimator of a random variable \( \theta \) is denoted as \( \hat{\theta} \).

Capital letters \( A, B, C, \ldots \) denote sets and sequences; \( |S| \) denotes the size of \( S \).

Bold capital letters \( A, B, C, \ldots \) denote functions.

**Item Preknowledge Detection Statistics**

A test taker is defined by two random variables: unobservable latent trait (ability) \( \theta \) and observable response \( \chi_i \in \{0, 1\} \) to item \( i \). Consider a subset of dichotomous items \( I \) administered to the test taker, which may vary among test takers (e.g., CAT) or be fixed (e.g., P&P). Then several characteristics can be computed. The following random variable is the resultant score:

\[
\varphi_i = \sum_{i \in I} \chi_i. \quad (1)
\]

Bayes’ theorem is applied to compute the discrete posterior distribution of \( \theta \) with uniform prior:

\[
F_i(y) = \frac{\prod_{i \in I} P_i(\chi_i | y)}{\sum_{z \in Y} \prod_{i \in I} P_i(\chi_i | z)}, \quad y \in Y, \quad (2)
\]

where \( F_i(y) \) is the probability of \( \theta = y \), \( P_i(\chi_i | y) \) is the probability of response \( \chi_i \) to item \( i \) conditioned on \( \theta = y \), and set \( Y \) contains ability levels (this report used \( Y = \{-5, -4.9, \ldots, 5\} \)). Then the expected a posteriori (EAP) estimator of \( \theta \) is computed as follows:

\[
\hat{\theta}_i = \sum_{y \in Y} yF_i(y). \quad (3)
\]
Let us assume that the subset of compromised items $S$ is known and does not vary among test takers. Consider a test taker taking a test $T$, which can be partitioned into two nonintersecting subtests $C = T \cap S$ (compromised items) and $U = T \setminus S$ (uncompromised items). Then the following characteristics can be computed for $C$, $U$, and $T$: scores $\varphi_C$, $\varphi_U$, $\varphi_T = \varphi_C + \varphi_U$; posterior distributions of ability $F_C$, $F_U$, $F_T$; and ability estimates $\hat{\theta}_C$, $\hat{\theta}_U$, $\hat{\theta}_T$. These characteristics will be used to describe the following eight statistics for detection of item preknowledge. For each test taker, these statistics measure the difference in responding between compromised items $C$ and uncompromised items $U$.

**Statistic $I_z$**

This statistic by Drasgow, Levine, and Williams (1985) is a normalization of the statistic by Levine and Rubin (1979), and it is computed as follows:

$$-\frac{\hat{\lambda} - e_{\lambda}}{v_{\lambda}},$$

where

$$\hat{\lambda} = \sum_{i \in T} \chi_i \ln P_i(1 | \hat{\theta}_T) + (1 - \chi_i) \ln P_i(0 | \hat{\theta}_T)$$

$$e_{\lambda} = \sum_{i \in T} P_i(1 | \hat{\theta}_T) \ln P_i(1 | \hat{\theta}_T) + P_i(0 | \hat{\theta}_T) \ln P_i(0 | \hat{\theta}_T)$$

$$v_{\lambda} = \sum_{i \in T} P_i(1 | \hat{\theta}_T) \ln P_i(1 | \hat{\theta}_T) \left( \ln \frac{P_i(1 | \hat{\theta}_T)}{P_i(0 | \hat{\theta}_T)} \right)^2$$

$$P_i(0 | \hat{\theta}_T) = 1 - P_i(1 | \hat{\theta}_T)$$

Drasgow, Levine, and Williams (1985) introduced $I_z$ based on the true value of $\theta$, which, however, is unknown in practice; an estimate $\hat{\theta}_T$ is commonly used instead [see Equation (5)]. Statistic (4) is often used as a baseline in computational studies on detecting item preknowledge (Levine & Drasgow, 1988; Shu, Henson, & Luecht, 2013).
Statistic Based on Score Ratio

This simple statistic has an immense practical advantage in that it is computable without any assumption about how observable variable $X_i$ depends on latent trait $\theta$:

$$\frac{\varphi_C}{\varphi_U + 1}$$

Statistic Based on Ability Difference

Another simple statistic that takes advantage of the information about compromised items is:

$$\hat{\theta}_C - \hat{\theta}_U$$

Statistic Based on Kullback–Leibler Divergence

The use of Kullback–Leibler divergence for detecting aberrant responding was proposed by Belov, Pashley, Lewis, & Armstrong (2007). Kullback–Leibler divergence for discrete distributions is computed as follows (Cover & Thomas, 1991; Kullback & Leibler, 1951):

$$D(F_C \| F_U) = \sum_{y \in Y} F_C(y) \ln \frac{F_C(y)}{F_U(y)}$$

According to the definition of Kullback–Leibler divergence (8), the larger the divergence $D(F_C \| F_U)$, the higher the dissimilarity between posterior distributions of ability $F_C$ and $F_U$. The value of $D(F_C \| F_U)$ is always non-negative and equals zero only if the two posterior distributions are identical. Kullback–Leibler divergence is asymmetric; that is, in general, $D(F_C \| F_U) \neq D(F_U \| F_C)$.

Statistic Based on Posterior Shift

Consider the ordered sequence of ability levels $y_1 > y_2 > \ldots > y_n$, $y_j \in Y$, $j = 1, 2, ..., n$ with subsequence $Z: y_1 > y_2 > \ldots > y_k$, such that $F_C(y_j) \geq F_U(y_j)$, $j = 1, 2, ..., k$, $k \leq n$. Then the following statistic measures how far the posterior distribution $F_C$ is shifted toward the higher ability with respect to the posterior distribution $F_U$:

$$\sum_{z \in Z} F_C(z) - F_U(z).$$

This is a nontrivial statistic, and to the best of my knowledge, it has never been used before.
Statistic Based on the Approach by Shu, Henson, and Luecht

Given two random variables \( \theta_C \) and \( \theta_U \) drawn from posterior distributions \( F_C \) and \( F_U \), respectively, the following statistic estimates the probability of \( \theta_C > \theta_U \) (Shu, Henson, & Luecht, 2013):

\[
\sum_{j=1}^{m} \left\{ \begin{array}{ll}
1, & \theta_C > \theta_U \\
0, & \theta_C \leq \theta_U \\
m & \\
\end{array} \right.
\]

\[
\hat{\theta}_C \sim F_C, \\
\hat{\theta}_U \sim F_U
\]

where at each iteration out of \( m \) (this report used \( m = 1,000 \)), variables \( \theta_C \) and \( \theta_U \) are drawn from posterior distributions \( F_C \) and \( F_U \), respectively. This statistic also measures how far the posterior distribution \( F_C \) is shifted toward the higher ability with respect to the posterior distribution \( F_U \). In computer simulations, statistics (9) and (10) demonstrated similar performances (see below).

Modified Iz

Equation (5) can be modified to take advantage of the information about compromised items. The idea is to use ability estimate \( \hat{\theta}_C \) when analyzing responses to uncompromised items and ability estimate \( \hat{\theta}_U \) when analyzing responses to compromised items. According to the definition of statistic Iz, such modification will be more sensitive to a difference in responding between subsets \( C \) and \( U \). Thus, the modified statistic is computed as follows:

\[
\frac{\gamma - e_\gamma}{v_\gamma}
\]

(11)
where

\[ \gamma = \sum_{i \in \mathcal{C}} \chi_i \ln P_i(1|\hat{\theta}_u) + (1 - \chi_i) \ln P_i(0|\hat{\theta}_u) + \sum_{i \in \mathcal{U}} \chi_i \ln P_i(1|\hat{\theta}_c) + (1 - \chi_i) \ln P_i(0|\hat{\theta}_c) \]
\[ e_\gamma = \sum_{i \in \mathcal{C}} P_i(1|\hat{\theta}_u) \ln P_i(1|\hat{\theta}_u) + P_i(0|\hat{\theta}_u) \ln P_i(0|\hat{\theta}_u) + \]
\[ \sum_{i \in \mathcal{U}} P_i(1|\hat{\theta}_c) \ln P_i(1|\hat{\theta}_c) + P_i(0|\hat{\theta}_c) \ln P_i(0|\hat{\theta}_c) \]  

(12)

\[ v_\gamma = \sum_{i \in \mathcal{C}} P_i(1|\hat{\theta}_u) \ln P_i(1|\hat{\theta}_u) \left( \frac{P_i(1|\hat{\theta}_u)}{P_i(0|\hat{\theta}_u)} \right)^2 + \]
\[ \sum_{i \in \mathcal{U}} P_i(1|\hat{\theta}_c) \ln P_i(1|\hat{\theta}_c) \left( \frac{P_i(1|\hat{\theta}_c)}{P_i(0|\hat{\theta}_c)} \right)^2 \]

**Statistic Based on Neyman–Pearson Lemma**

Applying the Neyman–Pearson lemma for detecting item preknowledge when the percentage of compromised items is known was proposed by Levine and Drasgow (1988). Theoretically, this approach should demonstrate the highest power among all item preknowledge detection statistics with the same Type I error (Lehman, 1999). This report considers a more specific case, that is, when compromised items are known, which greatly simplifies all computations. Precisely, the likelihood ratio between the probability of a response vector assuming item preknowledge and the probability of a response vector assuming no item preknowledge can be approximated as follows:

\[ \sum_{y \in \mathcal{Y}} H(y) \prod_{i \in \mathcal{T}} P_i(\mathcal{X}_i | y) \prod_{i \in \mathcal{C}} p^{x_i} (1 - p)^{1 - x_i} \]
\[ \sum_{y \in \mathcal{Y}} G(y) \prod_{i \in \mathcal{T}} P_i(\mathcal{X}_i | y) , \]  

(13)

where \( H \) is the probability density function (PDF) of aberrant test takers, \( p \) is the probability of the correct response to a compromised item for an aberrant test taker, and \( G \) is the PDF of nonaberrant test takers. This report assumes \( p = 0.95 \) and \( H, G \) to be standard normal PDFs:

\[ H(y) = G(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} . \]  

(14)

Obviously, statistic (13) is much easier to compute than the more general statistic by Levine and Drasgow (1988). Computer simulations demonstrated that statistic (13) has the highest performance among the previous statistics when the assumption about distribution \( H \) holds (see below).
Computer Simulations

The performance of the above eight statistics was analyzed via computer simulations. Multiple factors potentially influencing the performance were studied. The area under the ROC curve (ROC area) is the primary measure of performance. The secondary measure of performance is stability, which is computed for each statistic as follows:

$$1 - (u - l),$$

(15)

where $u$ is the maximal ROC area across factors and $l$ is the minimal ROC area across factors.

The above statistics and the simulation environment were implemented in standard C++. Therefore, the resultant software is more scalable and runs much faster than existing tools written in script languages such as R. The source code of this software can be easily adapted to arbitrary IRT models, types of tests, or types of distribution for nonaberrant and aberrant test takers.

Setup

Nonaberrant test takers were drawn from N(0,1) distribution. In each simulated scenario, the number of nonaberrant test takers was 1,000.

Type of Distribution of Aberrant Test Takers

Aberrant test takers were drawn from N(0,1) or U(-3,0) distributions.

Amount of Aberrancy

The number of aberrant test takers was 5%, 10%, or 20% of the nonaberrant population, resulting in 50, 100, or 200 aberrant test takers. Thus, the total number of test takers (aberrant and nonaberrant) was 1,050, 1,100, or 1,200 test takers.

Amount of Noise in the Information About Compromised Items

The information about compromised items is represented by a subset of compromised items $S$. In reality, however, this information has noise because different groups of aberrant test takers may have access to different (and unknown to us) subsets of items. In order to make simulations more realistic, the aberrant test takers were partitioned into 10 groups, where each group had preknowledge of a unique random subset of items. These compromised subsets $S_1, S_2, ..., S_{10}$ were the same size as $S$, but were formed in such a way that their intersection $S^* = S_1 \cap S_2 \cap ... \cap S_{10}$ varied as follows:
• All subsets were equal to the assumed subset of compromised items \( S \) (i.e., \( S = S^* \)). This is an ideal situation where the amount of noise is 0%.
• Each subset included 75% of the assumed subset of compromised items \( S \) (i.e., \( |S \cap S^*| = 0.75|S| \)). This is a situation with a small amount of noise (25%).
• Each subset included 25% of the assumed subset of compromised items \( S \) (i.e., \( |S \cap S^*| = 0.25|S| \)). This is a situation with a large amount of noise (75%).

**Type of Test**

Two types of tests were studied: adaptive (CAT) and nonadaptive (P&P). Simulations were conducted using disclosed items of the Law School Admission Test (LSAT). Each aberrant test taker had a 0.9 probability of responding correctly to each compromised item; otherwise, the response probability was modeled by the three-parameter logistic model (Lord, 1980).

As in Levine and Drasgow (1988), the parameters of items were known. This simulated a realistic situation of using pretest parameters. When items are new (and their pretest parameters are estimated from a subpopulation) then the item preknowledge cannot exist, unless there is a leak in the testing organization.

In both tests, the ratio of compromised items to test length was about 0.26. Results of computer simulations for two other different ratios (0.13 and 0.51) demonstrated stability of answers to the questions stated above in the introduction. Therefore, this factor was not considered.

**CAT**

Multiple simulation studies were conducted using disclosed Logical Reasoning (LR) items of the LSAT. The CAT pool consisted of 500 LR items. The distribution of (a) discrimination, (b) difficulty, and (c) guessing parameters of the items in the CAT pool have the following minimums, maximums, means, and variances, respectively: (a) minimum 0.28, maximum 1.67, mean 0.75, variance 0.06; (b) minimum -2.47, maximum 2.92, mean 0.49, variance 1.27; and (c) minimum 0.00, maximum 0.52, mean 0.17, variance 0.01.

The item-selection criterion for CAT was the maximization of Fisher information at the current estimate of ability \( \hat{\theta} \). The test length was fixed at 50 items for each test taker. The estimator of \( \theta \) was the EAP estimator with a uniform prior. The ability estimate was initialized at \( \hat{\theta} = 0 \). There was no item-exposure control.
The following simulation study was performed:

1. Simulate CAT without item preknowledge.
2. Compute item exposure.
3. Form a subset of potentially compromised items \( S \) with exposure higher than 0.6. This step resulted in \( S \) with 13 items.
4. To each group of aberrant test takers, assign a unique random compromised subset formed around \( S \) (see previous section about noise). Steps 1–4 simulate a realistic scenario of item preknowledge. Although items with greater exposure have a higher probability of being compromised, this does not guarantee that each aberrant subgroup of test takers will have access to all of these items.
5. Simulate CAT with item preknowledge.

**P&P**

Multiple simulation studies were conducted using a disclosed form of the LSAT comprising the following item types: Analytical Reasoning (AR), LR, and Reading Comprehension (RC). The test consists of an operational part (one AR section, two LR sections, and one RC section) and a pretest part (one AR, LR, or RC section). The operational part consists of approximately 100 items, and the pretest part consists of approximately 25 items. When the LSAT is administered, adjacent test takers usually have different pretest parts. A test taker cannot distinguish which section is operational and which is pretest. More information on the LSAT can be found at [www.LSAC.org](http://www.LSAC.org).

The following simulation study was performed:

1. One operational LR section was assumed to be memorized during a previous administration of the LSAT and then later partially distributed to various groups of aberrant test takers. A subset of potentially compromised items \( S \) contains all items from this section (26 items).
2. To each group of aberrant test takers, assign a unique random compromised subset formed around \( S \) (see previous section about noise).
3. Simulate P&P with item preknowledge.

**Results**

The above setup resulted in \( 2 \) (type of distribution for aberrant test takers) \( \times 3 \) (amount of aberrancy) \( \times 3 \) (noise) \( \times 2 \) (type of test) = 36 simulation conditions. Detailed results are presented in Figures 1–4. The ROC areas of statistics averaged over 3 (amount of aberrancy) \( \times 3 \) (noise) = 9 simulation conditions are shown in Figure 5. The stabilities of statistics computed over 3 (amount of aberrancy) \( \times 3 \) (noise) = 9 simulation conditions are presented in Figure 6. A detailed analysis of the results follows in the next section.
FIGURE 1. ROC areas for CAT, where aberrant test takers’ proficiency was drawn from N(0, 1)
FIGURE 2. ROC areas for CAT, where aberrant test takers’ proficiency was drawn from $U(-3.0)$. When aberrancy and noise were 5% and 75% (see chart in the upper right corner), respectively, the ROC area for the statistic based on score ratio was 0.446.
FIGURE 3. ROC areas for P&P, where aberrant test takers were drawn from N(0,1)
FIGURE 4. ROC areas for P&P, where aberrant test takers were drawn from $U(-3,0)$.
FIGURE 5. ROC areas averaged over aberrancy and noise
FIGURE 6. Stability computed over aberrancy and noise

Summary

This section is structured as a list of questions with answers acquired from the analysis of Figures 1–6. Taking into account practical considerations, the following questions are formulated:

How are these results compatible with results from existing publications?

The most well-known comparison study that used ROC analysis is by Karabatsos (2003), who estimated item parameters from analyzed data. Therefore, the contamination of data was the most negative factor; see Figure 2 in
Karabatsos (2003). However, when the contamination was low, the results for IZ in Karabatsos (2003) are similar to results for IZ in this report: Compare Figure 2 in Karabatsos (2003) with Figures 2 and 4 in this report.

Which statistic has the best performance on average?

Figure 5 demonstrates that on average, when the distribution of aberrant test takers’ proficiency is N(0,1), the statistic based on the Neyman–Pearson lemma has the best performance. However, when the distribution of aberrant test takers’ proficiency is U(−3,0), assumption (14) is violated and this statistic consequently loses its power. When distribution of aberrant test takers’ proficiency is U(−3,0), the best average performer for CAT is the statistic based on posterior shift and the best average performer for P&P is the modified IZ.

By looking at detailed performance in Figures 1–4, one can see that when the amount of noise is 0% or 25%, the statistic based on the Neyman–Pearson lemma has the best performance in most cases, even when the distribution of aberrant test takers is not N(0,1).

How did the simplest statistic perform?

The simplest statistic based on score ratio (6) performed surprisingly well. It outperformed IZ in many cases. Taking into account its immense practical advantage in that it is computable without any assumption about how observable variable \( \chi_i \) depends on latent trait \( \theta \), the results shown in Figures 1–4 are quite encouraging in terms of the practical use of this statistic.

How do factors affect averaged performance of the statistics?

Figure 5 shows that all statistics performed better for P&P. This can be explained by fluctuations in statistics caused by variability of the test among test takers in CAT.

When distribution of aberrant test takers is U(−3,0), all statistics demonstrated better performance (see Figure 5). This follows from a larger difference compared to N(0,1) in aberrant test-taker responses between compromised and uncompromised items.

Which factor has the most negative influence on the performance of the statistics?

Figures 1–4 show that noise has the most negative effect on the performance of all statistics: the larger the noise, the smaller the ROC area. One can see that without noise, the statistic based on the Neyman–Pearson lemma has the best performance.
Which statistic is the most stable to the most negative factor?

Figure 6 demonstrates that in most cases statistic \( l_z \) has the highest stability. This follows from the definition of \( l_z \); see Equations (4) and (5). In particular, \( l_z \) does not explicitly take into account the information about compromised items (in contrast with, for example, the modified \( l_z \)). Therefore, when this information has noise, the performance of \( l_z \) does not change as dramatically as the performance of statistics defined based on subset \( S \) (the subset representing the information about compromised items).

How can we alleviate the most negative factor?

As shown above, the most negative factor affecting the performance of all statistics is noise in the assumption about which items are compromised (Figures 1–4). There are two approaches to address this problem.

The first approach is to operate without the above assumption (like statistic \( l_z \)). CUSUM (Armstrong & Shi, 2009; van Krimpen-Stoop & Meijer, 2001) demonstrates excellent performance when compromised items are positioned sequentially in the test; however, in general, its performance is questionable. Response time modeling (van der Linden & Guo, 2008) holds great promise for detecting item preknowledge. However, there are two caveats: (a) the actual response times are only available in CAT, where the test taker cannot return back to a previously seen item; it is not clear (in CBT or MST) how to compute the time a test taker actually dedicated to each item; and (b) response times can be faked. 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 MST and CAT, where the actual test varies across test takers.

The second approach (Belov, 2013a, 2013b) is to identify the actual compromised subsets and then apply an item preknowledge detection statistic (IPDS) for each found subset. Each group of aberrant test takers exists within a test center (called an affected test center). It is realistic to assume that each affected test center has only one group of aberrant test takers with a corresponding compromised subset because the definition of test center is not limited by the geographic location (e.g., room, class, college, state) and can be extended to support various relations among test takers (e.g., from the same high school, undergraduate college, test preparation center, or group in a social network). Due to item preknowledge, if test center \( R \) is affected and the corresponding compromised subset \( S \) is known, then the distribution of a given IPDS computed at \( R \) should be unusual among distributions of IPDS computed at nonaberrant test centers. This can be formally represented by a statistic \( G(R,S) \) that accumulates differences (e.g., measured by Kullback–Leibler divergence) between the distribution of IPDS computed at \( R \) and distributions of IPDS computed at nonaberrant test centers. (*Note:* there are multiple alternatives for generating nonaberrant test centers, such as by simulation). If \( R \) is
affected, then a combinatorial optimization algorithm (e.g., simulated annealing) can be applied to search for the corresponding compromised subset \( S \) by selecting items maximizing the value of \( G(R, S) \). To identify affected test centers, multiple random subsets \( S_1, S_2, \ldots \) of items are generated and the statistic \( C(R_j) = \sum G(R_j, S_i) \) is computed for each test center \( R_j \). Due to intersections of random subsets \( S_1, S_2, \ldots \) with the actual compromised subsets, statistic \( C(R_j) \) will have larger values for affected test centers. This approach has great potential considering the results from Figures 1–4. One can see that even when noise is 25%, the performance of statistics taking into account information about compromised items is higher than the performance of \( lz \). Thus, even an approximation of an actual aberrant subset can be sufficient for IPDS to have high power.

**References**


