LSAC RESEARCH REPORT SERIES

Statistical Algorithms for Detection of Test Collusion

Dmitry I. Belov

Law School Admission Council
Research Report 11-03
October 2011

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 admission products and services to ease the admission process for law schools and their applicants worldwide. More than 200 law schools in the United States, Canada, and Australia are members of the Council and benefit from LSAC's services.

LSAC fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, Credential Assembly Service (CAS), and other matters may change without notice at any time. Up-to-date LSAC policies and procedures are available at LSAC.org.

Table of Contents

Executive Summary

The development of statistical methods for detecting test collusion is a new research direction in the area of test security. Test collusion may be described as large-scale sharing of test materials or answers to test questions. The source of the test materials could be a teacher, a test-preparation company, the Internet, or test takers communicating on the day of the exam. The danger of test collusion for high-stakes testing programs is that it can seriously affect the scoring of the exam because of the potentially large number of test takers involved. This is a serious concern for the test users (law schools, universities, companies, government organizations, etc.), who will be given invalid scores for test takers involved in test collusion. Therefore, identifying such test takers in order to remove their responses from the data is an important task.

Test collusion often influences the test performance or speed of involved test takers for the affected portion of the test. However, the portion of the test where the collusion took place is often unknown, and searching through all possible portions will decrease the detection power or increase the false-positive rate of a statistical test. This report introduces an algorithm that resolves this problem by working in two stages. In the first stage, test centers with potential collusion are identified as aberrant. In the second stage, potentially aberrant test takers from the aberrant test centers are identified. A simulation study demonstrates the advantages of using this algorithm for the Law School Admission Test (a case of a partially stolen section of the LSAT was simulated). However, the algorithm is general and can be applied to other testing formats, including computer-based testing (CBT), multiple-stage testing (MST), and computerized adaptive testing (CAT).

Introduction

Current research on test security is moving beyond the classical problem of detecting answer copying between two test takers (Angoff, 1974; Belov & Armstrong, 2010; Frary, 1993; Harpp & Hogan, 1993, 1998; Harpp, Hogan, & Jennings, 1996; Holland, 1996; Sotaridona, van der Linden, & Meijer, 2006; Wesolowsky, 2000; Wollack, 1997; Wollack & Cohen, 1998) and is focusing on methods of detecting larger groups involved in test collusion (Jacob & Levitt, 2003; Wollack & Maynes, 2011; Zhang, Searcy, & Horn, 2011).

Wollack and Maynes (2011) outline the following types of test collusion:

> [I]llegal coaching by a teacher or test-prep school, examinees accessing stolen test content posted on the world wide web, examinees communicating about test answers during an exam, examinees harvesting and sharing exam content using e-mail or the Internet, and teachers or administrators changing answers after tests have been administered. (p. 2)

They point out that the “last activity is actually tampering, but when the tampering is performed on several answer sheets consistently, it can be detected using collusion analysis” (p. 2).

In order to identify classrooms where teachers have changed answers, Jacob and Levitt (2003) analyzed the joint distribution of two summary statistics: answer strings summary and unexpected score fluctuations summary. Both cluster analysis (Wollack & Maynes, 2011) and factor analysis (Zhang et al., 2011) were demonstrated to be applicable for detecting various types of test collusion. However, all the above methods rely on statistics also used in detecting answer copying, and therefore they lose power when there is a lack of matching between incorrect responses within a group of test takers involved in test collusion. Furthermore, statistics based on matching responses are generally useless for MST and CAT, where the actual exam varies among test takers. This report presents a new approach that does not rely on response matching statistics but instead uses the difference between posteriors measured by the Kullback–Leibler divergence (Kullback & Leibler, 1951).

Throughout the report the following notation is used:

- Lowercase letters $a,b,c,...$; $\alpha,\beta,\gamma,...$ denote scalars (including random variables).
- Capital letters $A,B,C,...$ denote sets; $|S|$ denotes the number of elements in a set S.
- Bold capital letters $\mathbf{A},\mathbf{B},\mathbf{C},...$ denote functions (including discrete distributions defined by probability mass functions).
- Capital Greek letters $\Omega,\Psi,\Theta,...$ denote collections of subsets.

Test collusion often influences the posterior $\mathbf{P}_S$ of the ability or speed of involved test takers, estimated on a subset S of items in test T. Practical examples of S include difficult items, stolen items, items with high exposure during previous CAT administration(s), and items with a large number of erasures. For example, if answers to stolen items S are known to a group of test takers, then there should be a large difference between the posteriors of ability $\mathbf{P}_{T\setminus S}$ and $\mathbf{P}_S$ for each low-ability test taker from the group. Therefore, the Kullback–Leibler divergence between corresponding posteriors $\mathbf{D}(\mathbf{P}_{T\setminus S} \,||\, \mathbf{P}_S)$ can be used as an instrumental statistic. Its properties (Cover & Thomas, 1991; Kullback & Leibler, 1951) include $\mathbf{D}(\mathbf{P}_{T\setminus S} \,||\, \mathbf{P}_S) \geq 0$, $\mathbf{D}(\mathbf{P}_{T\setminus S} \,||\, \mathbf{P}_S) = 0 \Leftrightarrow \mathbf{P}_{T\setminus S} \equiv \mathbf{P}_S$, and in general, $\mathbf{D}(\mathbf{P}_{T\setminus S} \,||\, \mathbf{P}_S) \neq \mathbf{D}(\mathbf{P}_S \,||\, \mathbf{P}_{T\setminus S})$. When S is known, the use of Kullback–Leibler divergence is effective for detecting aberrant response patterns (Belov & Armstrong, 2011; Belov, Pashley, Lewis, & Armstrong, 2007), including detection of pairs of test takers involved in answer copying (Belov & Armstrong, 2010). When S is unknown but belongs, for example, to a collection $\Omega = (S_1, S_2, ..., S_m)$ of test sections (other examples can be found in Table 1 below), then the multiple comparison problem will arise, because each empirical distribution of $\mathbf{D}(\mathbf{P}_{T\setminus S_i} \,||\, \mathbf{P}_{S_i})$, $i = 1,2,...,m$ has to be analyzed. A standard approach to resolving this problem is to use Bonferroni or Šidák corrections for significance level (Abdi, 2007), which will decrease the statistical power.

The objective of this report is to apply Kullback–Leibler divergence in order to detect test takers involved in test collusion resulting in an unusual performance gain for the items in $S \subset T$. The following assumptions are made:

1. Subset S is unknown.
2. Subset S is partially covered by a known collection $\Omega = (S_1, S_2, ..., S_m)$, $S_i \subset T$, $i = 1, 2, ..., m$ (i.e., there exist $1 \le j \le m$ such that $S \cap S_j \neq \varnothing$).

Test takers involved in test collusion will be called *aberrant test takers*. They form *aberrant groups* or just *groups*. Test centers with aberrant groups will be called *aberrant test centers*.

Detection of Test Collusion

Given a test T, its subsets $\Omega = (S_1, S_2, ..., S_m)$, and a set of test takers E, a straightforward algorithm (Algorithm 1) searching for aberrant test takers is introduced. For each test taker it computes statistic h: Kullback–Leibler divergence between corresponding posteriors. Then given a significance level, the algorithm computes a critical value of the statistic using its empirical distribution.

Algorithm 1

Step 1: For each test taker $e \in E$, the statistics $h_{e,i} = \mathbf{D}(\mathbf{P}_{T \setminus S_i} \,||\, \mathbf{P}_{S_i})$, $i = 1, 2, ..., m$ are computed.
Step 2: For each $i = 1, 2, ..., m$, the empirical distribution $\mathbf{H}_i$ of $h_{e,i}$, $e \in E$ is computed.
Step 3: Given a significance level $\alpha_{\mathbf{H}} / m$, the critical value v_i for each $\mathbf{H}_i$ is computed.
Step 4: A test taker e is reported as aberrant if $h_{e,i} > v_i$ for some $i \in \{1, 2, ..., m\}$.

Note the Bonferroni correction (Abdi, 2007) at Step 3, since at Step 4 each test taker has to pass m comparisons. Algorithm 1 loses its power when m is large and/or in the presence of large aberrant groups. One can address the latter data contamination problem by computing the critical value via simulations. However, the use of a simulated distribution may lead to a large increase in the Type I error rate. The next algorithm addresses the data contamination problem by the following process: (1) potentially aberrant test centers are identified via a new statistic g indicating how the distribution of the statistic h within a particular test center is different from other test centers; (2) posteriors of test takers from the identified test centers are ignored when computing empirical distribution of the statistic h.

Given a test T, its subsets $\Omega = (S_1, S_2, ..., S_m)$, a set of test takers E, and a set of test centers C, Algorithm 2 searches for aberrant test takers (detailed explanations of major steps follow the description of the algorithm). The subset $E_c \subseteq E$ contains test takers taking a test at test center $c \in C$.

Algorithm 2

Step 1: For each test taker $e \in E$, the statistics $h_{e,i} = \mathbf{D}(\mathbf{P}_{T \setminus S_i} \,||\, \mathbf{P}_{S_i})$, $i = 1, 2, ..., m$ are computed.

The following is the first iteration of the search for aberrant test centers (test centers in resultant subsets A and B will be considered potentially aberrant). Subset A will include test centers with high values of statistic $g_{c,i}$, $i = 1, 2, ..., m$. Subset B will include test centers with high values of statistic g_c.

Step 2: A subset of test centers is initialized as $A = \varnothing$.
Step 3: For each $i = 1, 2, ..., m$ the following steps are performed:
Step 3.1: For each test center $c \in C$, the empirical distribution $\mathbf{H}_{c,i}$ of $h_{e,i}$, $e \in E_c$ is computed.
Step 3.2: For each test center $c \in C$, the statistic $g_{c,i} = \sum_{x \in C} \left(\mathbf{D}(\mathbf{H}_{c,i} \,||\, \mathbf{H}_{x,i}) + \mathbf{D}(\mathbf{H}_{x,i} \,||\, \mathbf{H}_{c,i}) \right)$ is computed.
Step 3.3: The empirical distribution $\mathbf{G}_i$ of $g_{c,i}$, $c \in C$ is computed.
Step 3.4: Given significance level $\alpha_{\mathbf{G}}$, the critical value v_i for $\mathbf{G}_i$ is computed.
Step 3.5: For each test center $c \in C$, if $g_{c,i} > v_i$ then let $A = A \bigcup \{c\}$.

Step 4: A subset of test centers is initialized as $B = \varnothing$.
Step 5: For each test center $c \in C$, the empirical distribution $\mathbf{H}_c$ of $\sum_{i=1}^{m} h_{e,i}$, $e \in E_c$ is computed.
Step 6: For each test center $c \in C$, the statistic $g_c = \sum_{x \in C} \left(\mathbf{D}(\mathbf{H}_c \,||\, \mathbf{H}_x) + \mathbf{D}(\mathbf{H}_x \,||\, \mathbf{H}_c) \right)$ is computed.
Step 7: The empirical distribution $\mathbf{G}$ of g_c, $c \in C$ is computed.
Step 8: Given significance level $\alpha_{\mathbf{G}}$, the critical value v for $\mathbf{G}$ is computed.
Step 9: For each test center $c \in C$, if $g_c > v$ then let $B = B \bigcup \{c\}$.

The following is the second iteration of the search for aberrant test centers (test centers with a resultant nonempty subset I_c will be considered potentially aberrant). Subsets A and B (identified at the first iteration) are used to remove data from potentially aberrant test centers. Then for each test center $c \in C$, its index set I_c will

include indices of elements from collection Ω *inducing high values of statistic* g_z *computed for each* $c \in C$ *and* $i = 1,2,...,m$.

Step 10: For each test center $c \in C$, the following steps are performed:

Step 10.1: A subset of test centers $Z = \left(C \setminus \left(A \cup B\right)\right) \cup \{c\}$ is built.

Step 10.2: The index set is initialized as $I_c = \varnothing$.

Step 10.3: For each $i = 1,2,...,m$ the following steps are performed:

Step 10.3.1: For each test center $z \in Z$, the empirical distribution $\mathbf{Q}_z$ of $h_{e,i}$, $e \in E_z$ is computed.

Step 10.3.2: For each test center $z \in Z$, the statistic $g_z = \sum_{x \in Z} \left(\mathbf{D}(\mathbf{Q}_z \,||\, \mathbf{Q}_x) + \mathbf{D}(\mathbf{Q}_x \,||\, \mathbf{Q}_z)\right)$ is computed.

Step 10.3.3: The empirical distribution $\mathbf{G}$ of g_z, $z \in Z$ is computed.

Step 10.3.4: Given significance level $\alpha_{\mathbf{G}}$, the critical value v for $\mathbf{G}$ is computed.

Step 10.3.5: If $g_c > v$, then let $I_c = I_c \cup \{i\}$.

Search for aberrant test takers.

Step 11: For each $i = 1,2,...,m$ the empirical distribution $\mathbf{H}_i$ of $h_{e,i}$, $e \in \bigcup_{c \in C \setminus (A \cup B)} E_c$ is computed.

Step 12: Given significance level $\alpha_{\mathbf{H}} / m$, the critical value v_i for each $\mathbf{H}_i$ is computed.

Step 13: For each test center $c \in C$ with $I_c \neq \varnothing$, a test taker $e \in E_c$ is reported as aberrant if $h_{e,i} > v_i$ for some $i \in I_c$.

Informally, Algorithm 2 works in two stages. In Stage 1, potentially aberrant test centers are identified; in Stage 2, potentially aberrant test takers at potentially aberrant test centers are reported. If one assumes independence between Stage 1 and Stage 2, then the Type I error rate should have an upper bound $u = m\alpha_{\mathbf{G}}\alpha_{\mathbf{H}} / m \ll \min(\alpha_{\mathbf{G}}, \alpha_{\mathbf{H}})$ (see results of computer simulations below). Note the Bonferroni correction (Abdi, 2007) at Step 12, since at Step 13 each test taker has to pass $|I_c| \leq m$ comparisons. However, Steps 3 and 10 do not apply the Bonferroni correction in the search for aberrant test centers, because the ultimate goal of Algorithm 2 is to identify aberrant test takers. Steps 2–9 build two subsets of aberrant test centers A and B, which are used at Step 11 to compute empirical distributions $\mathbf{H}_i$ from data without the responses of too many test takers involved in test collusion. Step 3 is sensitive to a large group of aberrant test takers within a test center. Steps 5–9 are sensitive to multiple small groups each gaining performance at different subsets of the test. Step 10 looks for aberrant test

centers using subsets A and B. Step 13 identifies aberrant test takers at potentially aberrant test centers.

Often in practice, no more than one group of aberrant test takers per test center is expected. In that case, Algorithm 2 can be modified in order to avoid multiple comparisons at Step 13. For this, at Step 10.3.5 the index set I_c should have only one index that provides the maximum value of corresponding g_c, as follows:

Step 10.3.5: If $g_c > v$ and ($I_c = \varnothing$ or $g_c > g_{c,\max}$), then $I_c = \{i\}$, $g_{c,\max} = g_c$.

Then at Step 12 the significance level should be α_H, because at Step 13 no more than one comparison for each test taker can be made. Such simple modification should improve the power of Algorithm 2 when all groups of aberrant test takers from a test center gain performance on the same unknown subset S of items.

A particular set of test centers C influences the performance of Algorithm 2. Commonly, the subset $E_c \subseteq E$ contains test takers taking a test at the corresponding geographic location $c \in C$ (room, college, state, region, country, and so on). However, each element $c \in C$ can be a logical condition defining a relation between test takers. Examples of such relations include same high school (or group of high schools geographically close to one another), same undergraduate college (or group of colleges geographically close to one another), and relationships identified from Facebook or other social networks. This allows the detection of groups of aberrant test takers even if they take the actual exam at different geographic locations.

Algorithms 1 and 2 are general and should be implemented for a particular Ω, which depends on the testing format: paper-and-pencil testing (P&P), computer-based testing (CBT), multiple-stage testing (MST), or computerized adaptive testing (CAT). Table 1 shows examples of different elements from collection Ω (subsets of test T) and corresponding testing formats. Note that in MST and CAT, test T varies among test takers. This, however, is not an obstacle because Algorithms 1 and 2 employ the difference between posteriors.

TABLE 1
Examples of different elements from collection Ω (subsets of test T) and corresponding testing formats

Subset of Test T	Applicable Testing Format
A scored section or a group of scored sections	P&P, CBT
Items with a large number of erasures or corrections	P&P, CBT
Items with difficulty above a certain level	P&P, CBT, MST, CAT
Stolen items	P&P, CBT, MST, CAT
High-exposure items from previous administration(s)	MST, CAT
A testlet or a group of testlets	MST

CAT = computerized adaptive testing; CBT = computer-based testing; MST = multiple-stage testing; P&P = paper-and-pencil testing.

Computer Simulations

Multiple simulation studies were conducted using a disclosed form of the Law School Admission Test (LSAT). The LSAT comprises the following item types: Analytical Reasoning (AR), Logical Reasoning (LR), and Reading Comprehension (RC). The scored part of the test consists of one AR section, two LR sections, and one RC section. Each section has approximately 25 items. More information on the LSAT can be found at http://www.LSAC.org.

Nonaberrant test takers were simulated with abilities drawn from N(0, 1); aberrant test takers (i.e., test takers involved in test collusion) were simulated with abilities drawn from U(−3, 0). The response probability for each item was modeled by the three-parameter logistic (3PL) model (Lord, 1980). There were 100 test centers with 100 test takers per test center. The LSAT is $T = S_{\text{AR}} \cup S_{\text{LR1}} \cup S_{\text{LR2}} \cup S_{\text{RC}}$, where S_{AR} is the AR section, S_{LR1} is the first LR section, S_{LR2} is the second LR section, and S_{RC} is the RC section. It was assumed that answers to a certain percentage δ of items from a section were known to a group of aberrant test takers, where $\delta \in \{50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$. For each group, the position of the first known item was selected randomly. The number of groups per aberrant test center was $n_g \in \{1, 2, 4, 5\}$. The number of aberrant test centers was $n_c \in \{5, 10, 15, 20, 30\}$. The number of aberrant test takers (from all groups) in each aberrant test center was $n_e \in \{5, 10, 15, 20, 30\}$. Each group was randomly assigned a section. Each group was also randomly assigned to a test center such that the total number of test takers in each test center would be 100. Algorithm 1 was implemented in C++ by the author for $\Omega = \{S_{\text{AR}}, S_{\text{LR1}}, S_{\text{LR2}}, S_{\text{RC}}\}$.

The Type I error rate was computed as follows (this is an empirical probability for a test taker to be falsely reported):

$$\frac{[\text{number of reported examinees}] - [\text{number of correctly reported examinees}]}{[\text{number of all nonaberrant examinees}]} \quad (1)$$

The detection rate was computed according to the following:

$$\frac{[\text{number of correctly reported examinees}]}{[\text{number of all aberrant examinees}]} \quad (2)$$

The results for $\alpha_{\text{H}} = \{0.0025, 0.005, 0.0075, 0.01, 0.025, 0.05\}$, $\alpha_{\text{G}} = 0.1$, $\delta \in \{50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$, $n_g = 1$, $n_c = 20$, and $n_e = 20$ are presented in Figures 1 and 2. Both algorithms performed with Type I error rates that were less than the nominal level (Figure 1), and the error rate for Algorithm 2 demonstrated insensitivity to the change of δ. Algorithm 2 had a much higher detection rate than Algorithm 1 (Figure 2); the difference was highest for lower values of significance level α_{H}. Algorithm 1 demonstrated a low sensitivity of the detection rate to the change of δ (Figure 2).

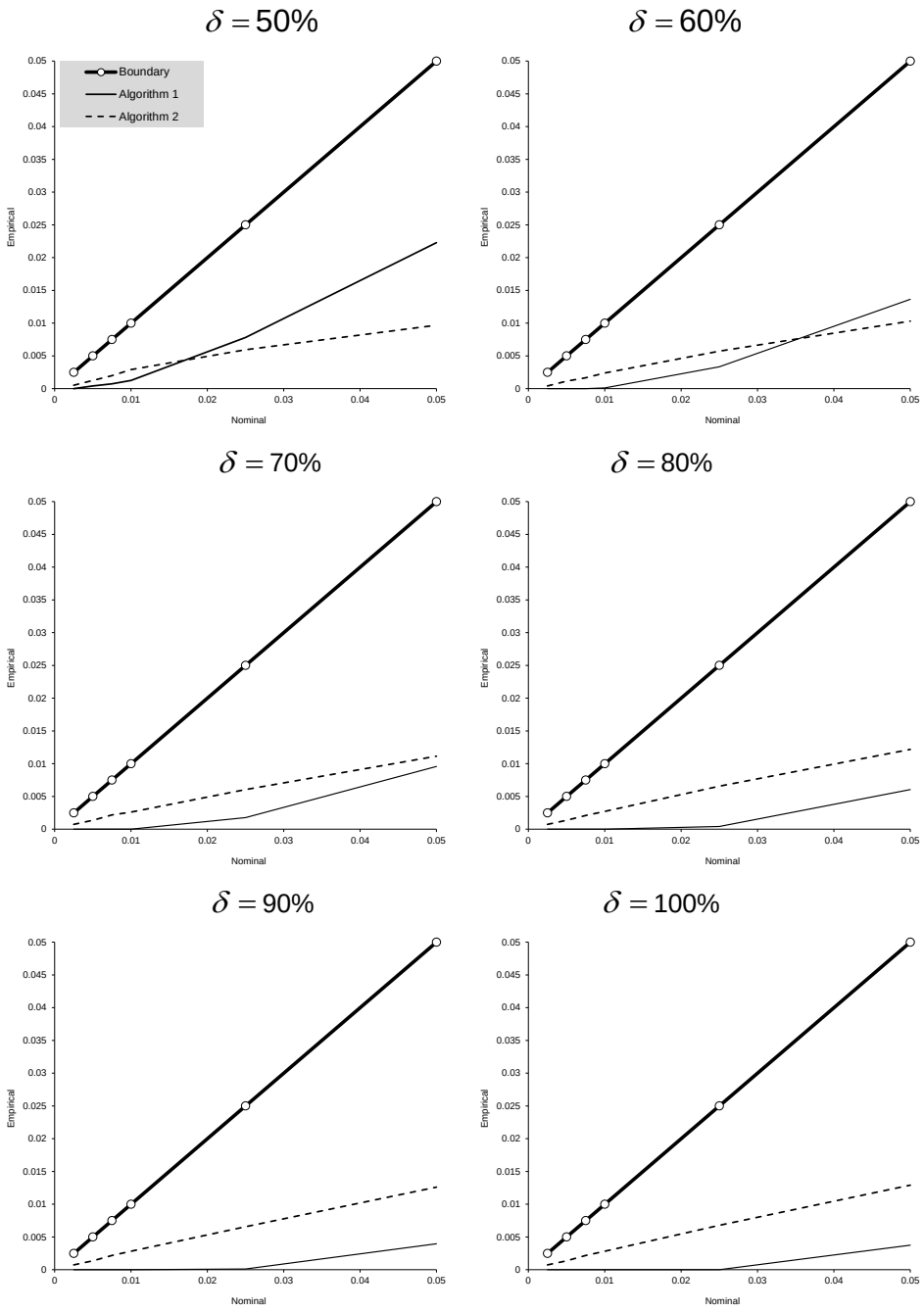

FIGURE 1. *Empirical Type I error rates ($\alpha_{\mathbf{G}} = 0.1$, $n_g = 1$, $n_c = 20$, and $n_e = 20$), where the abscissa is* $\alpha_{\mathbf{H}}$

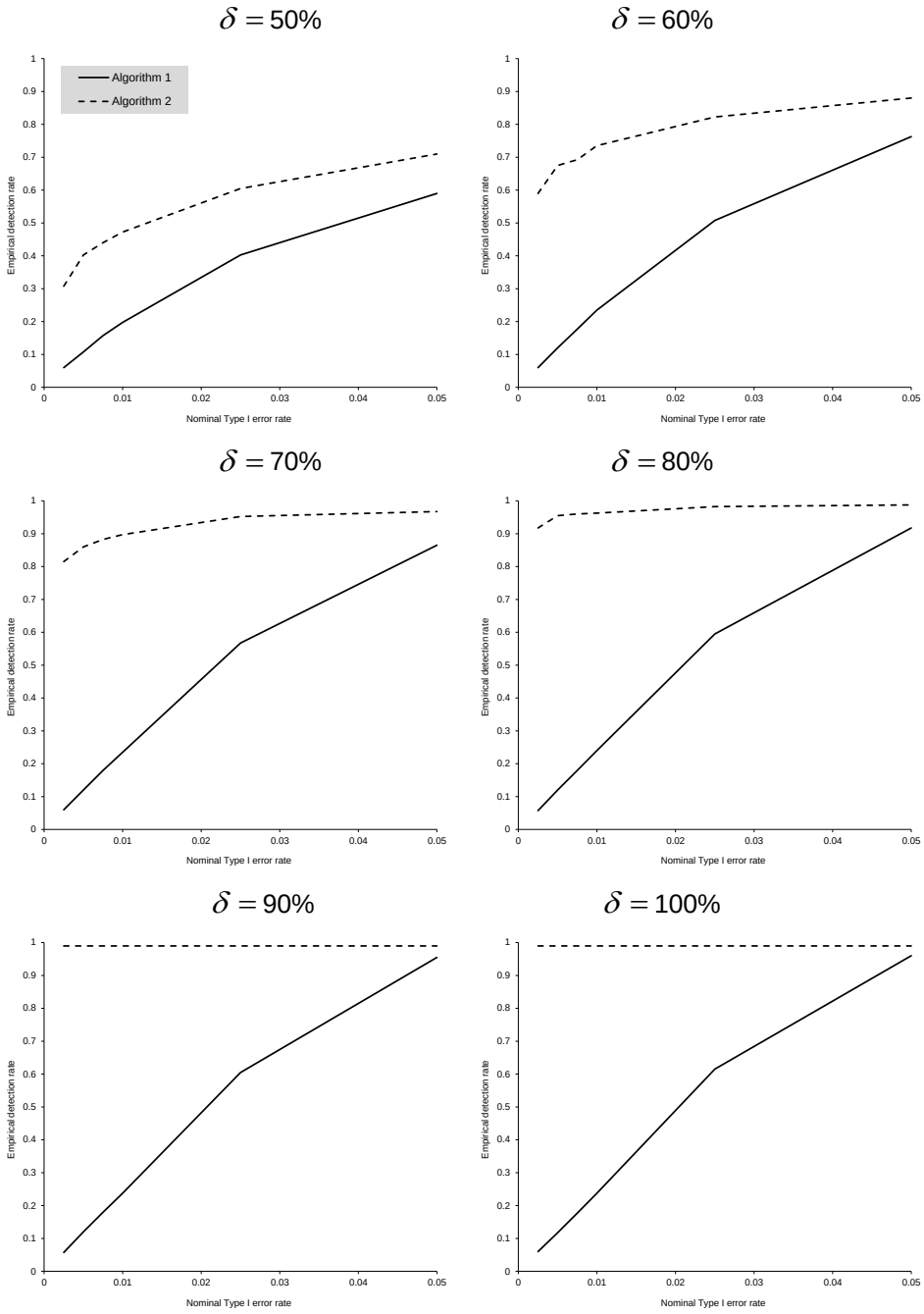

FIGURE 2. *Empirical detection rates ($\alpha_{\mathbf{G}} = 0.1$, $n_g = 1$, $n_c = 20$, and $n_e = 20$), where the abscissa is $\alpha_{\mathbf{H}}$*

Two additional computer simulations studied the influence of n_c and n_e on the performance of Algorithms 1 and 2. In both studies $\alpha_{\mathbf{H}} = 0.01$, $\alpha_{\mathbf{G}} = 0.1$, $\delta \in \{50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$, and $n_g = 1$. In the first study, $n_c = \{5, 10, 15\}$ and $n_e = \{5, 10, 15\}$, which resulted in test collusions on a smaller scale. In the second study, $n_c = \{10, 20, 30\}$ and $n_e = \{10, 20, 30\}$, which resulted in test collusions on a larger scale. The empirical Type I error rates were distributed with the following characteristics: for Algorithm 1 (first study: mean ≈ 0.003, standard deviation ≈ 0.002; second study: mean ≈ 0.001, standard deviation ≈ 0.001); for Algorithm 2 (first study: mean ≈ 0.003, standard deviation ≈ 0.000; second study: mean ≈ 0.002, standard deviation ≈ 0.001). The empirical detection rates are presented in Figures 3 and 4. One can see that in test collusions on a smaller scale (Figure 3), the power of Algorithm 1 drops when n_c is fixed and n_e increases. For the most part, in both studies Algorithm 1 loses power but Algorithm 2 remains stable when n_c and n_e increase (Figures 3 and 4). When $n_c = 30$, Algorithm 2 becomes unstable and loses power, which indicates an upper bound (30% of aberrant test centers) of applicability for Algorithm 2 (Figure 4). However, in practice, one would expect a smaller value of n_c.

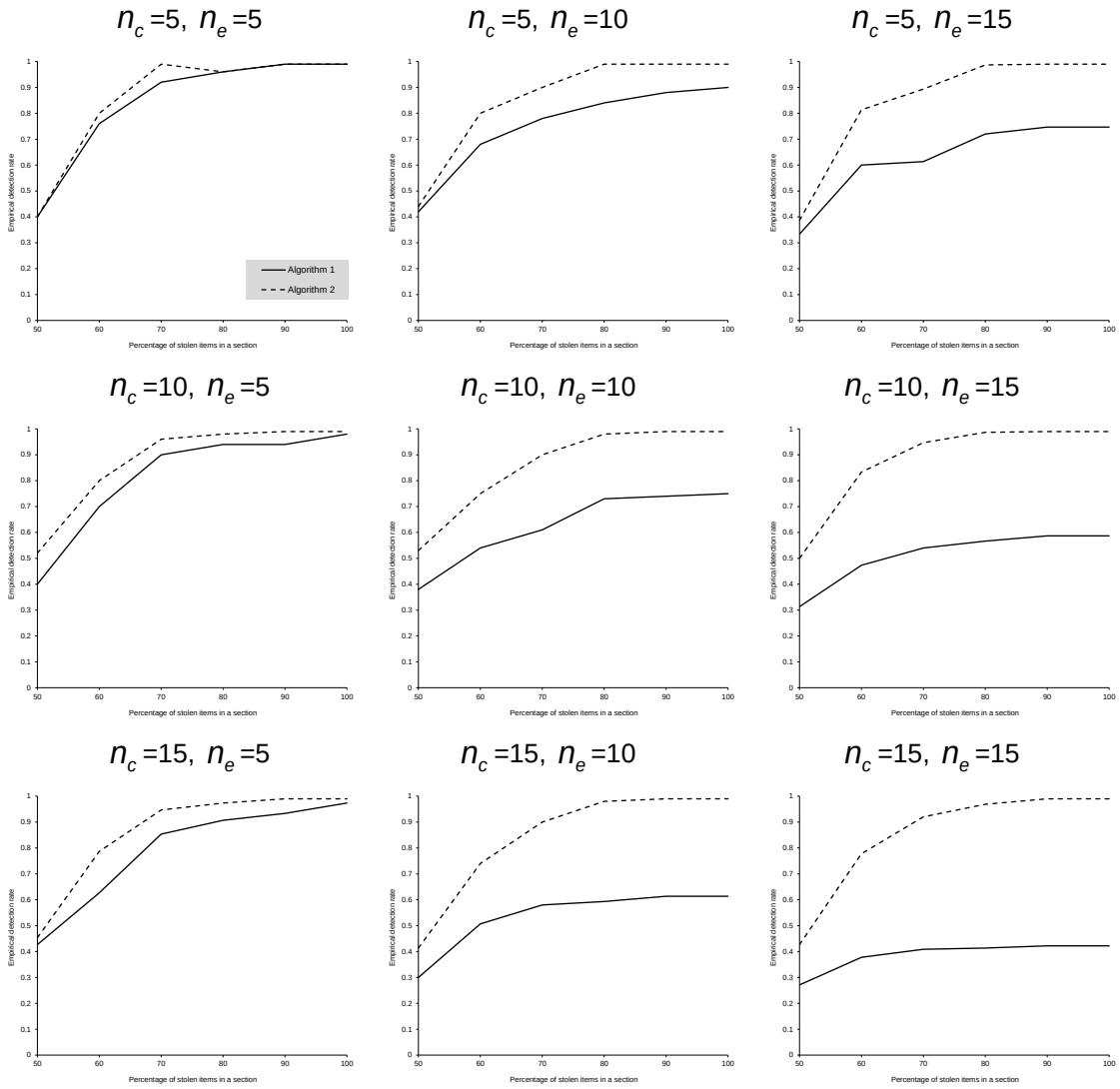

FIGURE 3. *Empirical detection rates (test collusions on a smaller scale,* $\alpha_{\mathbf{H}} = 0.01$, $\alpha_{\mathbf{G}} = 0.1$, *and* $n_g = 1$*), where the abscissa is* δ

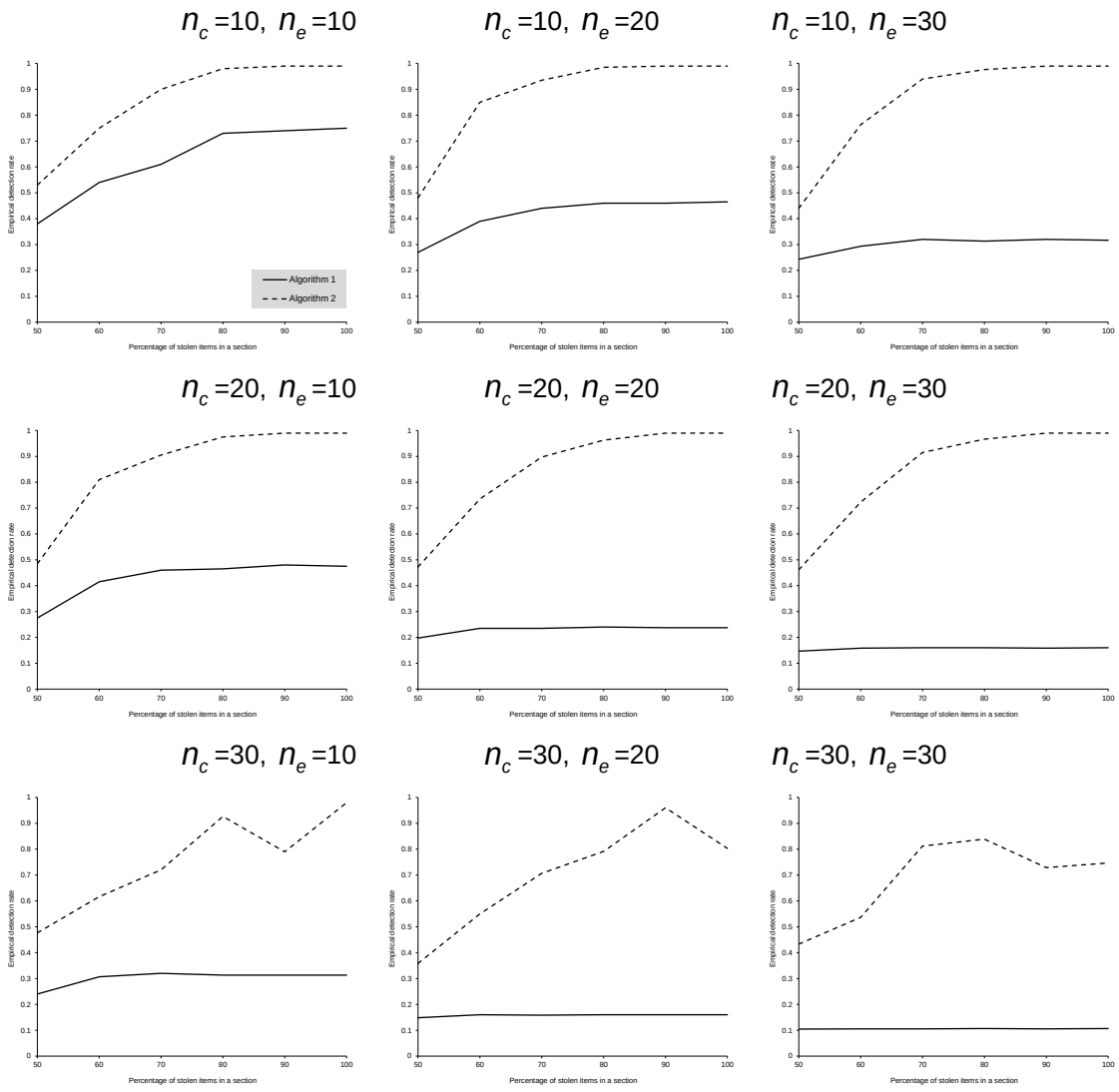

FIGURE 4. *Empirical detection rates (test collusions on a larger scale,* $\alpha_{\mathbf{H}} = 0.01$, $\alpha_{\mathbf{G}} = 0.1$, *and* $n_g = 1$*), where the abscissa is* δ

A final computer simulation studied the influence of n_g on the performance of both algorithms, where for each aberrant test center the corresponding n_g was randomly chosen from {1, 2, 4, 5}. The results for $\alpha_{\mathbf{H}} = \{0.0025, 0.005, 0.0075, 0.01, 0.025, 0.05\}$, $\alpha_{\mathbf{G}} = 0.1$, $\delta \in \{50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$, $n_c = 20$, and $n_e = 20$ are presented in Figures 5 and 6. One can see that these results are mostly identical to the results shown in Figures 1 and 2, which suggests an insensitivity of both algorithms to change of n_g.

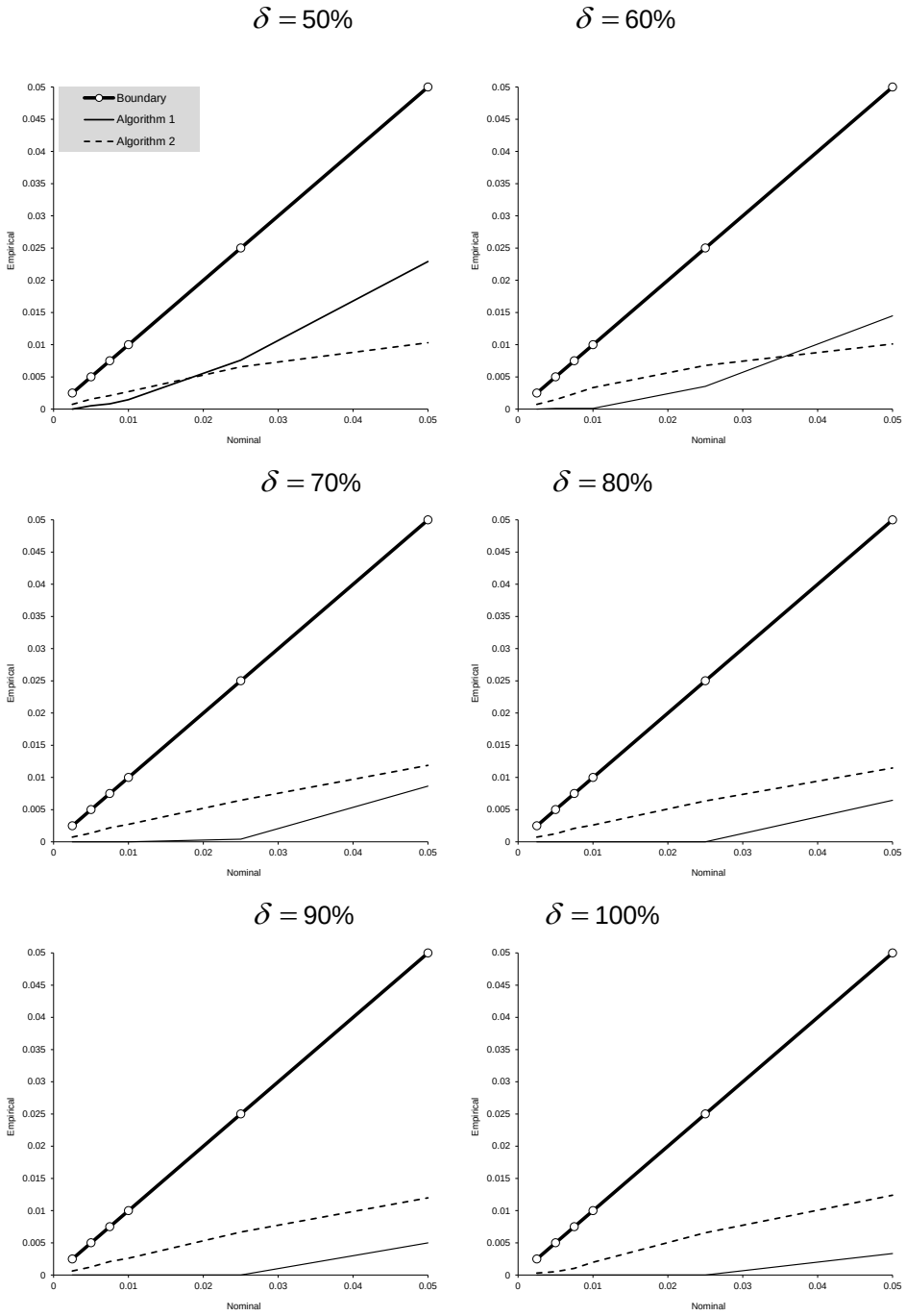

FIGURE 5. *Empirical Type I error rates ($\alpha_{\mathbf{G}} = 0.1$, $n_g \sim U(\{1,2,4,5\})$, $n_c = 20$, and $n_e = 20$), where the abscissa is $\alpha_{\mathbf{H}}$*

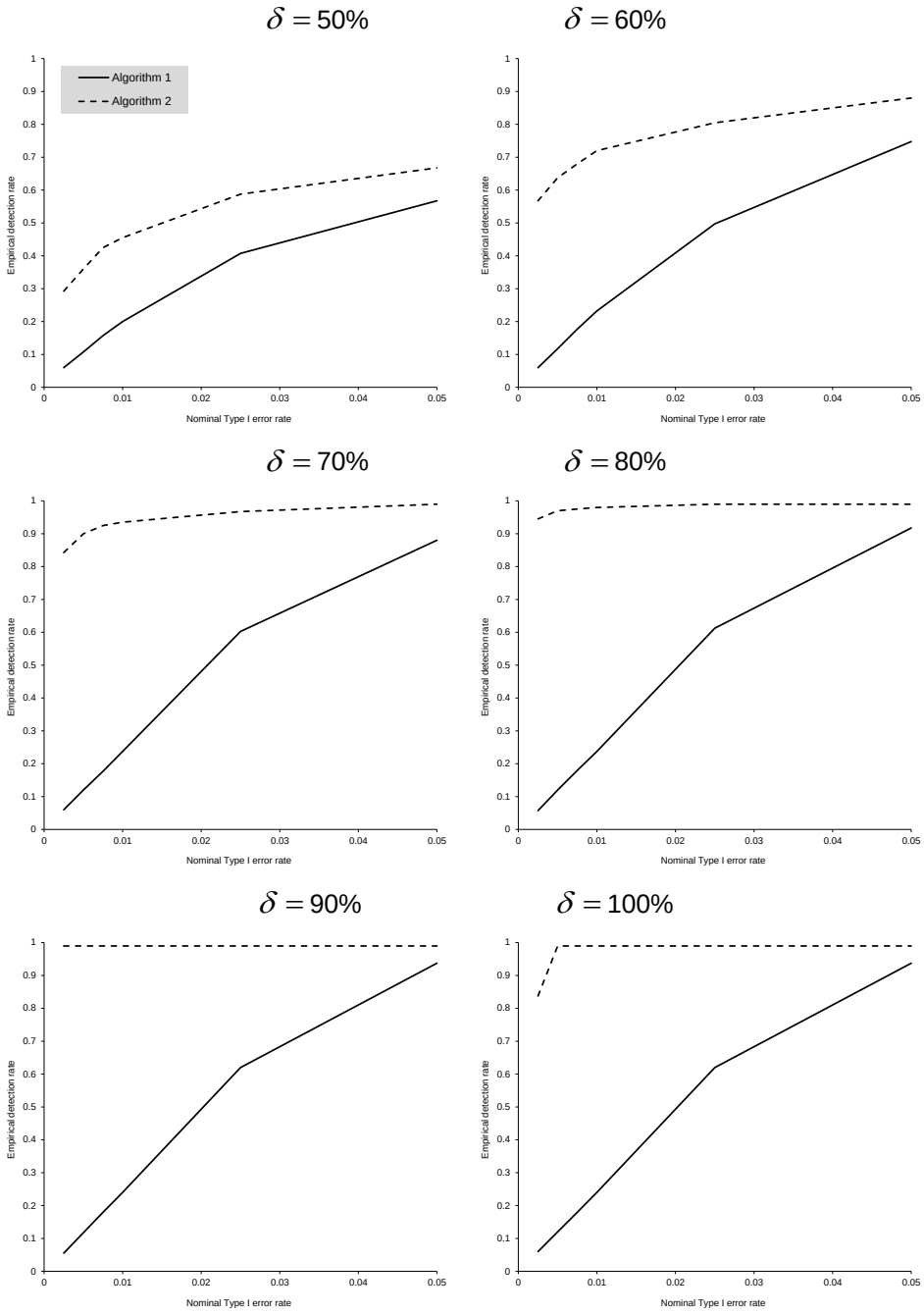

FIGURE 6. *Empirical detection rates ($\alpha_{\mathbf{G}} = 0.1$, $n_g \sim U(\{1,2,4,5\})$, $n_c = 20$, and $n_e = 20$), where the abscissa is $\alpha_{\mathbf{H}}$*

Summary

This report addresses an important practical problem of test security: detection of test takers involved in test collusion resulting in an unusual performance gain in a part of the test. Two algorithms based on Kullback–Leibler divergence are introduced, and they are applicable for testing formats such as P&P, CBT, MST, and CAT.

A simulation study demonstrated the advantages of using Algorithm 2 for a high-stakes P&P test (e.g., the LSAT). When Ω covers more than 50% of an unknown subset S, then Algorithm 2 has a high detection rate (Figures 2–4 and 6). Otherwise, a cluster analysis of response patterns of test takers can be applied to identify more subsets for inclusion in Ω, which has the potential to lead to a higher than 50% coverage.

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.

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

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback–Leibler divergence and K-index. *Applied Psychological Measurement, 34,* 379–392.

Belov, D. I., & Armstrong, R. D. (2011). Distributions of the Kullback–Leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology, 64,* 291–309.

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.

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

Frary, R. B. (1993). Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education, 6,* 153–165.

Harpp, D. N., & Hogan, J. J. (1993). Crime in the classroom: Detection and prevention of cheating on multiple-choice exams. *Journal of Chemical Education, 70,* 306–311.

Harpp, D. N., & Hogan, J. J. (1998). Crime in the classroom part III: The case of the ultimate identical twin. *Journal of Chemical Education, 75,* 482–483.

Harpp, D. N., Hogan, J. J., & Jennings, J. S. (1996). Crime in the classroom part II: An update. *Journal of Chemical Education, 73,* 349–351.

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-Index: Statistical theory and empirical support* (ETS Technical Report 96-4). Princeton, NJ: Educational Testing Service.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics, 118,* 843–877.

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

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

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

Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics, 27,* 909–921.

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement, 21,* 307–320.

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22,* 144–152.

Wollack, J. A., & Maynes, D. (2011). *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). *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

I would like to thank Charles Lewis and Bernard Veldkamp for valuable comments and suggestions on previous versions of the report.