On the Optimality of the Detection of Examinees With Aberrant Answer Changes

Applied Psychological Measurement
2017, Vol. 41(5) 338–352
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621617692077
journals.sagepub.com/home/apm

Dmitry I. Belov¹

Abstract

In standardized multiple-choice testing, examinees may change their answers for various reasons. The statistical analysis of answer changes (ACs) has uncovered multiple testing irregularities on large-scale assessments and is now routinely performed at many testing organizations. This article exploits a recent approach where the information about all previous answers is used only to partition administered items into two disjoint subtests: items where an AC occurred and items where an AC did not occur. Two optimal statistics are described, each measuring a difference in performance between these subtests, where the performance is estimated from the final responses. Answer-changing behavior was simulated, where realistic distributions of wrong-to-right, wrong-to-wrong, and right-to-wrong ACs were achieved under various conditions controlled by the following independent variables: type of test, amount of aberrancy, and amount of uncertainty. Results of computer simulations confirmed the theoretical constructs on the optimal power of both statistics and provided several recommendations for practitioners.

Keywords

test security, answer changes, erasure analysis, hypothesis testing, Neyman–Pearson lemma, Kullback–Leibler divergence, Kolmogorov–Smirnov test

Introduction

Statistical analyses of answer changes (ACs) are routinely performed at most testing organizations for large-scale assessments (Maynes, 2013), as changes to answers may indicate testing irregularities (i.e., aberrant ACs). Examples of aberrant ACs may include a teacher changing the answers of low-ability examinees from incorrect to correct, or an examinee changing answers after getting some help from a peer during the break between sections or from a friend via smartphone.

Two popular statistics that are used to detect examinees with aberrant ACs are as follows: (a) the number of wrong-to-right ACs (denoted here as WTR) and (b) the WTR divided by the total number of ACs (denoted here as WTR/NAC). Both the WTR and the WTR/NAC are based on measuring the differences between initial and final responses, where a higher WTR indicates a higher probability of aberrancy.

¹Law School Admission Council, Newtown, PA, USA

Corresponding Author:

Dmitry I. Belov, Psychometric Research, Law School Admission Council, 662 Penn Street, Newtown, PA 18940, USA.
Email: dbelov@lsac.org

Although a higher WTR may be indicative of aberrant ACs, there are also many other nonaberrant reasons why a higher WTR may be observed, thus causing an uncertainty in AC data. This uncertainty may be due to technological factors as well as human factors. Human factors that can attribute more uncertainty to AC data include warm-up effect (Qualls, 2001), fatigue, misunderstanding of the passage in a set-based item, or certain answering strategies (Maynes, 2013). Technological factors include scanning errors in paper-and-pencil (P&P) testing and the presence of multiple ACs for a single item in computer-based testing (CBT), multiple-stage testing (MST), and computerized adaptive testing (CAT). To illustrate the latter issue in a CBT scenario, assume that an examinee has multiple ACs for a single item and the sequence of ACs for this item was ABCDBA (where A is the correct choice). Based on this sequence, if only the initial and final answers are considered (both of which are correct), then a new type of AC emerges: a right-to-right (RTR) AC, where in this case $RTR = 1$. However, if all ACs in the sequence are analyzed sequentially, the following results occur: $RTR = 0$, right-to-wrong (RTW) = 1, wrong-to-wrong (WTW) = 3, and $WTR = 1$.

As a result of a combination of the above factors, common statistics that rely on the difference between previous and final responses (e.g., WTR, WTR/NAC) may result in larger Type I errors. For example, Maynes (2013) found that a certain answering strategy causes the WTR to result in a large Type I error. In this instance, a group of examinees in the detected class had quickly selected preferred answers in the first stage of testing but later performed a thorough review of their selected answers. This two-stage answering strategy led to a significantly large WTR, and hence a large Type I error.

Belov (2015) addressed the issue of the uncertainty in AC data by using only final responses and ignoring the difference between previous and final responses. Using his approach, items that are administered to an examinee are partitioned into two disjoint subtests: first subtest with items where ACs occurred and second subtest with items where ACs did not occur. Once these subtests are defined, a statistic is calculated that measures the difference in performance between these subtests using only final responses. The rationale for this method is that a *nonaberrant examinee* (e.g., an examinee who has no aberrant ACs) should perform similarly across these subtests, whereas an *aberrant examinee* (e.g., an examinee who has aberrant ACs) should perform much better on the first subtest than on the second subtest. For example, if a teacher corrected some of the answers of a low-ability examinee, then this examinee would have an unusually large number of correct final responses in the first subtest compared with the second subtest. Belov pointed out that the choice of measure of the difference in performance between the two subtests is crucial. A difference between point estimates (e.g., score, ratio of score to number of items, expected ability) may be unstable if there is an insufficient number of items in one of the subtests. Therefore, the difference between posteriors of ability computed from responses to corresponding subtests was chosen as a measure (Belov, 2015). Due to the Bayesian approach and smoothness of the item characteristic curve, the posterior changes continuously with each additional item in a subtest. Belov applied Kullback–Leibler divergence (KLD; Kullback & Leibler, 1951) between the posteriors, and demonstrated via computer simulations its advantages over WTR and WTR/NAC in terms of Type I, Type II errors and robustness to the uncertainty in AC data.

The rest of this section will discuss two recently published statistics to detect examinees with aberrant ACs. The next section will describe two optimal statistics for detecting examinees with aberrant ACs, followed by comparison studies and an analysis of their asymptotic distributions. Finally, a short summary completes the article.

Throughout the article, the following notation is used: Lowercase letters $a, b, c, \dots$ denote scalars, lowercase Greek letters $\alpha, \beta, \gamma, \dots$ denote random variables, capital letters $A, B, C, \dots$ denote sets and sequences, and bold capital letters $\mathbf{A}, \mathbf{B}, \mathbf{C}, \dots$ denote functions.

An examinee is defined by two random variables: unobservable latent trait (ability) θ and observable final response $\chi_i \in \{0, 1\}$ to item i . Let $\mathbf{P}_i(\chi_i|y)$ be the probability of response χ_i to item i conditioned on $\theta=y$, where $y \in Y$ and set Y contains ability levels. Consider an examinee taking a test T (where T may vary among examinees), which can be partitioned into two disjoint subtests: C (items with ACs) and U (items without ACs).

Bayes' theorem is applied to compute the discrete posterior distribution of θ with a uniform prior from the responses to items without ACs (subtest U):

$$\mathbf{F}_U(y) = \frac{\prod_{i \in U} \mathbf{P}_i(\chi_i|y)}{\sum_{z \in Y} \prod_{i \in U} \mathbf{P}_i(\chi_i|z)}, y \in Y, \quad (1)$$

where $\mathbf{F}_U(y)$ is the probability of $\theta=y$. Similarly, the posterior $\mathbf{F}_C(y)$ is computed from the final responses to items with ACs (subtest C) as follows:

$$\mathbf{F}_C(y) = \frac{\prod_{i \in C} \mathbf{P}_i(\chi_i|y)}{\sum_{z \in Y} \prod_{i \in C} \mathbf{P}_i(\chi_i|z)}, y \in Y. \quad (2)$$

KLD Statistic

Formally, KLD for discrete distributions $\mathbf{F}_U$ and $\mathbf{F}_C$ is computed as follows (Kullback & Leibler, 1951):

$$\mathbf{D}(\mathbf{F}_U||\mathbf{F}_C) = \sum_{y \in Y} \mathbf{F}_U(y) \ln \frac{\mathbf{F}_U(y)}{\mathbf{F}_C(y)}. \quad (3)$$

According to the definition of KLD, the larger the divergence $\mathbf{D}(\mathbf{F}_U||\mathbf{F}_C)$, the higher the dissimilarity between $\mathbf{F}_U$ and $\mathbf{F}_C$. The value of $\mathbf{D}(\mathbf{F}_U||\mathbf{F}_C)$ is always nonnegative and equals 0 only if the two distributions are identical. KLD is asymmetric; that is, in general, $\mathbf{D}(\mathbf{F}_U||\mathbf{F}_C) \neq \mathbf{D}(\mathbf{F}_C||\mathbf{F}_U)$.

Belov (2015) established conditions under which the KLD is asymptotically chi-square with 1 degree of freedom. He developed three algorithms to satisfy the conditions, where each algorithm filters a subtest that is longer than the other one, such that the posteriors of ability computed from the responses to these two subtests have similar variances. Alternatively, similar variances can also be achieved by bootstrapping the responses to the shorter subtest (in this case, information from responses to longer subtest is not lost).

One issue with the application of KLD for detecting examinees with aberrant ACs is illustrated in Figure 1, where posterior distributions of ability are shown for two examinees: a non-aberrant examinee performs better on subtest U than on subtest C , see Figure 1a; an aberrant examinee performs better on subtest C than on subtest U , see Figure 1b. In both cases, the KLD has the same value of 2, which creates the potential for both examinees to be either detected or undetected simultaneously, thus increasing Type I and Type II errors. Belov (2017) addresses this issue by suggesting a modified KLD.

Erasure Detection Index (EDI) Statistic

Another statistic studied in this article is the EDI by Wollack, Cohen, and Eckerly (2015), which can be defined as follows:

Figure 1. Posterior distributions of ability for (a) a nonaberrant examinee and (b) an aberrant examinee, where $\mathbf{F}_C$ was computed from the final responses to subtest C , and $\mathbf{F}_U$ was computed from responses to subtest U .

$$\frac{\sum_{i \in C} \chi_i - \sum_{i \in C} \mathbf{P}_i(1|\theta_U) + c}{\sqrt{\sum_{i \in C} \mathbf{P}_i(1|\theta_U) \mathbf{P}_i(0|\theta_U)}}, \quad (4)$$

where θ_U is a maximum-likelihood estimate of ability computed from responses to subtest U . Wollack et al. (2015) showed that if the correction for continuity $c = -0.5$, then the EDI demonstrated the best performance; therefore, this article will assume $c = -0.5$ in Equation 4.

From its definition, it follows that the EDI is asymptotically standard normal only when $c = 0$. In this case, however, the Type I error of the EDI is inflated (Wollack et al., 2015).

Optimal Statistics to Detect Examinees With Aberrant ACs

Two statistics presented in this section, similarly to KLD, exploit the idea of measuring the difference in performance between C and U , relying on the assumption that, for an aberrant examinee, the posterior $\mathbf{F}_C$ should be shifted more toward the higher ability than the posterior $\mathbf{F}_U$. Thus, these two statistics are expected to have a similar robustness to, but a higher power than, the KLD.

Statistic Based on Posterior Shift (PS)

The PS statistic is based on the assumption that, for an aberrant examinee, the posterior computed from the final responses to items with ACs should be shifted more toward the higher ability than the posterior computed from the responses to items without ACs. Its formal definition presented next corrects the definition of PS statistic by Belov (2016) to analyze its asymptotic behavior.

Let us consider set Y as an increasing sequence of ability levels $Y : y_1 < y_2 < \dots < y_n$ (e.g., $Y : -3 < -2.9 < \dots < 3$). Find indices l and r , $1 \leq l < r \leq n$, such that $\mathbf{F}_C(y_j) > \mathbf{F}_U(y_j)$ for each $j = 1, 2, \dots, l$ and $\mathbf{F}_C(y_j) < \mathbf{F}_U(y_j)$ for each $j = r, r+1, \dots, n$; otherwise, assume $l = 0$ and/or $r = n+1$. Then, the following statistic measures how far the posterior $\mathbf{F}_C$ is shifted toward the higher ability with respect to the posterior $\mathbf{F}_U$ (see Figure 2):

Figure 2. Probability mass functions F_C and F_U are computed for ability levels $-3, -2.9, \dots, 3$.

Note. The area of dotted, unshaded domain equals to the posterior shift $S(F_C || F_U)$.

$$S(F_C || F_U) = \begin{cases} \sum_{r \leq j \leq n} [F_C(y_j) - F_U(y_j)], & \sum_{r \leq j \leq n} [F_C(y_j) - F_U(y_j)] > \sum_{1 \leq j \leq l} [F_C(y_j) - F_U(y_j)] \\ 0, & \sum_{r \leq j \leq n} [F_C(y_j) - F_U(y_j)] \leq \sum_{1 \leq j \leq l} [F_C(y_j) - F_U(y_j)] \end{cases} \quad (5)$$

Note that if $l=0$ then $\sum_{1 \leq j \leq l} [F_C(y_j) - F_U(y_j)] = 0$, and if $r=n+1$ then $\sum_{r \leq j \leq n} [F_C(y_j) - F_U(y_j)] = 0$. By definition, the divergence measure $S(F_C || F_U)$ is asymmetric and nonnegative, located from 0 to 1. For example in Figure 2, $S(F_C || F_U) > 0$, but $S(F_U || F_C) = 0$; at the same time, both KLDs $D(F_C || F_U)$ and $D(F_U || F_C)$ are positive. One can prove that the PS statistic is asymptotically optimal (see below).

Statistic Based on the Neyman–Pearson Lemma (NPL)

Levine and Drasgow (1988) calculated an appropriateness index based on the NPL (Lehmann, 1986) to detect item preknowledge under the assumption that the percentage of compromised items was known. Drasgow, Levine, and Zickar (1996) considered a more specific case, where the actual compromised items were known, which greatly simplified all computations. Their approach is applied here; the NPL statistic is simply the ratio between the probability of a response vector assuming aberrant ACs and the probability of a response vector assuming no aberrant ACs:

$$\frac{\prod_{i \in C} p^{x_i} (1-p)^{(1-x_i)} \sum_{y \in Y} \mathbf{H}(y) \prod_{i \in U} \mathbf{P}_i(x_i | y)}{\sum_{y \in Y} \mathbf{G}(y) \prod_{i \in T} \mathbf{P}_i(x_i | y)}, \quad (6)$$

where $\mathbf{H}$ is the probability density function (PDF) of aberrant examinees, p is the probability of a correct final response to an item with ACs for an aberrant examinee, and $\mathbf{G}$ is the PDF of non-aberrant examinees. Similarly to Belov (2016), this article assumes that $p = .95$ (see details on how to choose this parameter in Drasgow et al., 1996), and that $\mathbf{H}$ and $\mathbf{G}$ are standard normal PDFs. It is a common assumption that distributions of aberrant and nonaberrant examinees are standard normal. If these assumptions hold, then due to the NPL (Lehmann, 1986), the statistic (6) should demonstrate a power higher than any other statistic to detect response vectors with aberrant ACs performing with the same Type I error (see results in the next section).

Computer Simulations

This section evaluates the performance of the above statistics (KLD, EDI, PS, and NPL) using simulated ACs according to the model defined by the constraints gleaned from the real data (Belov, 2015): number of items in the test: $C[1] = [83, 119]$; percentage of examinees with at least one AC: $C[2] = [50\%, 84\%]$; percentage of WTR ACs: $C[3] = [54\%, 65\%]$; percentage of RTW ACs: $C[4] = [12\%, 24\%]$; percentage of WTW ACs: $C[5] = [20\%, 28\%]$; average number of ACs: $C[6] = [3.61, 5.43]$; average number of WTR ACs: $C[7] = [1.95, 3.53]$; $C[8]$, percentage of WTR ACs should increase as score increases; $C[9]$, percentage of RTW ACs should decrease as score increases; $C[10]$, percentage of WTW ACs should decrease as score increases. Then, the following procedure (Belov, 2015) for modeling nonaberrant ACs was used:

Procedure 1 (Simulate Response Data With Nonaberrant ACs)

Step 1: Simulate *final responses* (corresponding to final answer choices) of multiple examinees with their θ drawn from $N(0, 1)$, where each item in a disclosed form of a large-scale admission test with 100 items (constraint $C[1]$ is satisfied) is modeled by the three-parameter logistic (3PL) model (Lord, 1980).

Step 2: For each examinee, compute the number of ACs x as follows:

$$x = \min(90, \max(0, \lambda)), \lambda \sim N(3, 36). \quad (7)$$

Parameters ($M = 3$, variance = 36) for the distribution of λ were chosen empirically to satisfy the above constraints $C[2]$ – $C[7]$.

Step 3: For each examinee, form a subset with x randomly selected items $S = \{i_1, i_2, \dots, i_x\}$ where the AC took place. For each item $i \in S$ simulate the *previous response* r_i (corresponding to the previous answer choice) with the following probability of a correct response (computed according to the 3PL model; Lord, 1980):

$$\mathbf{P}_i(r_i = 1 | \theta). \quad (8)$$

Due to Equation 8, higher ability examinees (with true $\theta > 0$) will have a greater number of WTR changes, and lower ability examinees (with true $\theta < 0$) will have a greater number of RTW changes, which corresponds to constraints $C[8]$ – $C[10]$.

All statistics and the simulation environment were implemented by the author in standard C++. Therefore, the resultant software is more scalable and runs much faster than existing tools written in script languages like R. The source code for this software can be easily adapted to

arbitrary item response theory (IRT) models, types of tests, or types of distribution for nonaberrant and aberrant examinees. The source code is available upon e-mail request.

The measure of performance is based on the receiver operating characteristic (ROC) curve. The ROC curve allows estimating the quality of a binary classifier (Green & Swets, 1966) by showing the relation between empirical detection rates and Type I errors under changing significance levels. The ROC allows one to integrate Type I and Type II errors in just one chart. This is obviously more informative for a practitioner than separate reports of Type I and Type II errors. The quantitative interpretation of the ROC is given by the area under the ROC curve (called the *ROC area*); the higher the area, the better the classifier. In the context of test security, one is only interested in the ROC curve for the region of small Type I errors. Therefore, the measure of performance is computed here as the ROC area for the region of Type I errors below 0.1 divided by 0.1. This measure is called the *truncated ROC area* (Belov, 2016), and it is always between 0 and 1.

KLD, EDI, PS, and NPL statistics were compared via computer simulations under various conditions controlled by the following independent variables: type of test, amount of aberrancy, and amount of uncertainty (this uncertainty is different from the uncertainty in AC data discussed in the Introduction). Each statistic is computed, such that aberrant examinees should be located at the right tail of the corresponding null distribution. Given a significance level and the type of test, the critical value of each statistic was estimated from a null distribution computed from the responses of 10,000 nonaberrant examinees drawn from $N(0,1)$. To control each simulation scenario, the following independent variables were considered:

- Primoli and Liassou (2013) showed that the type of test (P&P or CBT) directly influences the number of ACs. In particular, the average number of ACs in CBT is much larger than that in P&P (Primoli & Liassou, 2013). In this article, the number of ACs is controlled by random variable λ from Equation 7. Belov (2015) demonstrated that Procedure 1 (see above) simulates realistic ACs in P&P when $\lambda \sim N(3, 36)$ and realistic ACs in CBT when $\lambda \sim N(30, 36)$.
- Amount of aberrancy is defined by the percentage of aberrant examinees in the population. The following values were considered: 1%, 5%, 10%, and 20%.
- C-uncertainty affects responses to subtest *C* (items with ACs). It is defined by the percentage of ACs that are not aberrant for each aberrant examinee. This allows one to simulate a real situation where not every AC is aberrant for an aberrant examinee. The following values were considered: 0% (no uncertainty), 25% (moderate uncertainty), and 50% (high uncertainty).
- U-uncertainty affects responses to subtest *U* (items without ACs). It is defined by the percentage of responses to subtest *U* that are aberrant for each aberrant examinee. This allows one to simulate a real situation where some responses to items from subtest *U* are aberrant for an aberrant examinee (e.g., due to item preknowledge or test collusion). The following values were considered: 0% (no uncertainty), 25% (moderate uncertainty), and 50% (high uncertainty).

The above variables resulted in 2 (type of test) $\times$ 3 (level of C-uncertainty) $\times$ 3 (level of U-uncertainty) = 18 simulation scenarios, where the amount of aberrancy was embedded within each scenario. Such embedding was done because Belov (2015, 2016) demonstrated that a separate treatment of the amount of aberrancy did not have any noticeable effect on the performance of studied statistics. For each scenario, the following steps were repeated 10 times for each amount of aberrancy (1%, 5%, 10%, and 20%) to compute 40 estimates of the truncated ROC area:

Step 1: Given the type of test, data were simulated with nonaberrant ACs using Procedure 1 (see above) with 1,000 nonaberrant examinees drawn from an $N(0,1)$ distribution.

Step 2: Given the amount of aberrancy, the corresponding number of examinees with more than a certain number of ACs (3 for P&P, 30 for CBT) was randomly selected. These examinees were considered aberrant.

Step 3: For each of the aberrant examinees, $100\% - [C\text{-uncertainty}]\%$ of items from subtest C were randomly selected, and the final response for each of these items was set to the correct response.

Step 4: For each of the aberrant examinees, $[U\text{-uncertainty}]\%$ of items from subtest U were randomly selected, and the final response for each of these items was set to the correct response.

Step 5: Given the simulated data, the truncated ROC area for each statistic was computed.

The results are presented in Figures 3 and 4, where mean, mean minus standard deviation, and mean plus standard deviation of the truncated ROC areas are shown for each statistic.

Let us first discuss the case of zero uncertainty (see Figures 3 and 4, upper left-hand chart). As expected, the NPL statistic performed the best (with the PS statistic following closely behind). Also, as expected, the KLD statistic did not perform as well as the PS and NPL statistics. All statistics performed better in CBT than in P&P (i.e., the means were higher, and the standard deviations were lower). This can be explained by a larger number of ACs in CBT than in P&P. In particular, Belov (2015) reported the average number of ACs in P&P to be 3.87 and the average number of ACs in CBT to be 29.51.

Under zero C-uncertainty and positive U-uncertainty, the performance of each of the statistics decreased. However, PS and NPL demonstrated more stability under this condition than KLD and EDI (see Figures 3 and 4, upper charts).

Under zero U-uncertainty and positive C-uncertainty, the performance of each of the statistics also decreased. However, C-uncertainty had a larger negative effect on performance than U-uncertainty (Figures 3 and 4; compare left charts with upper charts). The EDI statistic was the most stable to the C-uncertainty due to its construction (see Equation 4).

When both uncertainties were positive, the performance of all statistics drastically decreased (Figures 3 and 4), though the NPL performed the best under the highest levels of uncertainty (see Figures 3 and 4; lower right-hand chart).

When both C- and U-uncertainties were 25% or lower, the EDI statistic had a mean of the truncated ROC area above 0.4 for P&P (Figure 3), and the NPL statistic had a mean of the truncated ROC area above 0.8 for CBT (Figure 4).

When one of the uncertainties was 0, the second best-performing statistic was always PS; the best-performing statistic alternated between NPL and EDI under this condition (see Figures 3 and 4; left and upper charts).

Based on the results of computer simulations, one can formulate several recommendations for a practitioner discussed in the Summary.

Asymptotic Distributions of PS and NPL Statistics

In this section, the PS and NPL statistics will be analyzed to identify corresponding asymptotic distributions. This section is optional, as it is straightforward to compute critical values for each statistic via simulations. However, simulated data may misrepresent real data (Belov, 2015). Also, the asymptotic analysis can shed some light on the performance of a particular statistic. Therefore, it is practical to have some information about the asymptotic behavior of PS and NPL statistics.

Figure 3. Estimations of truncated ROC areas (mean, mean minus standard deviation, and mean plus standard deviation) for P&P given different levels of uncertainty.
Note. ROC = receiver operating characteristic; P&P = paper-and-pencil; KLD = Kullback–Leibler divergence; EDI = erasure detection index; PS = posterior shift; NPL = Neyman–Pearson lemma.

PS Statistic

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

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

Figure 4. Estimations of truncated ROC areas (mean, mean minus standard deviation, and mean plus standard deviation) for CBT given different levels of uncertainty.
Note. ROC = receiver operating characteristic; CBT = computer-based testing; KLD = Kullback–Leibler divergence; EDI = erasure detection index; PS = posterior shift; NPL = Neyman–Pearson lemma.

$$\int_{-\infty}^{x_0} (\mathbf{H}(x) - \mathbf{G}(x))dx = \int_{x_0}^{+\infty} (\mathbf{G}(x) - \mathbf{H}(x))dx. \tag{9}$$

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

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

$$\mathbf{K}(y) = \int_{-\infty}^y \mathbf{H}(x)dx - \int_{-\infty}^y \mathbf{G}(x)dx. \quad (10)$$

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

Proof. Function $\mathbf{K}(y)$ is continuous by its definition; therefore, its local maxima and minima are located where the derivative of $\mathbf{K}(y)$ is 0. The derivative of function $\mathbf{K}(y)$ is $\mathbf{H}(y) - \mathbf{G}(y)$.

Theorem 1. Let us consider two continuous probability densities $\mathbf{H}(x)$ and $\mathbf{G}(x)$ where there are no more than two points where $\mathbf{H}(x)$ and $\mathbf{G}(x)$ are equal (e.g., two normal distributions with different means). Consider a point y_0 such that

$$\begin{aligned} \mathbf{H}(x) &= \mathbf{G}(x), x = y_0, \\ \mathbf{H}(x) &< \mathbf{G}(x), x > y_0. \end{aligned} \quad (11)$$

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

$$\int_{-\infty}^{z_0} [\mathbf{G}(x) - \mathbf{H}(x)]dx < \int_{y_0}^{+\infty} [\mathbf{G}(x) - \mathbf{H}(x)]dx. \quad (12)$$

Then, the following equality holds:

$$\sup_y |\mathbf{K}(y)| = \sup_{y < y_0} |\mathbf{K}(y)| = \mathbf{K}(y_0) = \int_{y_0}^{+\infty} [\mathbf{G}(x) - \mathbf{H}(x)]dx. \quad (13)$$

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

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

Case 2: There is a point $z_0 < y_0$ where $\mathbf{H}(z_0) = \mathbf{G}(z_0)$ and inequality (12) holds. As $\mathbf{H}(x)$ and $\mathbf{G}(x)$ are continuous, the inequality $\mathbf{H}(x) < \mathbf{G}(x)$ holds for $x < z_0$. Then, the following inequality follows:

$$\begin{aligned} |\mathbf{K}(y_0)| - |\mathbf{K}(z_0)| &= \left| \int_{-\infty}^{y_0} [\mathbf{H}(x) - \mathbf{G}(x)]dx \right| - \left| \int_{-\infty}^{z_0} [\mathbf{H}(x) - \mathbf{G}(x)]dx \right| = \\ &= \int_{y_0}^{+\infty} [\mathbf{G}(x) - \mathbf{H}(x)]dx - \int_{-\infty}^{z_0} [\mathbf{G}(x) - \mathbf{H}(x)]dx > 0. \end{aligned} \quad (14)$$

Then from Lemma 2, it follows that $\sup_{y < y_0} |\mathbf{K}(y)| = \mathbf{K}(y_0)$. As function

$\mathbf{L}(y) = \int_y^{+\infty} [\mathbf{G}(x) - \mathbf{H}(x)]dx$ is positive and monotonically decreasing on $y > y_0$, then due to Lemma 1 the following holds: $\sup_y |\mathbf{K}(y)| = \sup_{y < y_0} |\mathbf{K}(y)|$.

According to Chang and Stout (1993), the posteriors of ability $\mathbf{F}_C(x) = \mathbf{G}(x)dx$ and $\mathbf{F}_U(x) = \mathbf{H}(x)dx$ are asymptotically normal. Then, the positive value of $\mathbf{S}(\mathbf{F}_C || \mathbf{F}_U)$ corresponds

to a situation described in Theorem 1, where $\mathbf{S}(\mathbf{F}_C||\mathbf{F}_U)$ is a discrete approximation of $\mathbf{K}(y_0) = \int_{y_0}^{+\infty} [\mathbf{G}(x) - \mathbf{H}(x)]dx$. Then from Theorem 1, it follows that a hypothesis test to detect aberrant examinees based on a positive PS statistic is equivalent to the Kolmogorov–Smirnov test (Lehmann, 1999) that analyzes statistic $\sup_y |\mathbf{K}(y)|$, denoted here as KS. It is known (Lehmann, 1999) that the asymptotic power of the KS test is 1. Thus, the discovered connection between PS and KS explains the high performance of the PS statistic in the case of zero uncertainty. Interestingly, the performance of PS increases when $\mathbf{F}_U$ is used as a prior to compute $\mathbf{F}_C$ in Equation 2. However, then the stability of the PS to the C-uncertainty drops further down.

Figure 5 illustrates how Theorem 1 holds for two types of tests—P&P and CBT—where the PS and KS statistics were computed for 10,000 nonaberrant examinees drawn from $N(0,1)$. Dots on each chart characterize a match between positive values of $\mathbf{S}(\mathbf{F}_C||\mathbf{F}_U)$ (vertical axis, PS) and $\sup_y |\mathbf{K}(y)|$ (horizontal axis, KS), where a discrete approximation of $\sup_y |\mathbf{K}(y)|$ was computed as follows:

$$\max_{1 \leq i \leq n} \left| \sum_{1 \leq j < i} \mathbf{F}_C(y_j) - \sum_{1 \leq j < i} \mathbf{F}_U(y_j) \right|. \quad (15)$$

If the conditions of Theorem 1 are satisfied for each examinee with a positive value of the PS statistic, then all dots should be located on the identity line PS = KS (Figure 5). One can observe an almost perfect match between PS and KS; for CBT, eight dots were off the identity line because corresponding densities did not satisfy the conditions of Theorem 1. For one of those outliers, Figure 6 shows two corresponding densities with three points where these two densities are equal; this violates the conditions of Theorem 1.

NPL Statistic

The numerator and denominator in the definition of the NPL statistic (6) are not likelihoods. There is a famous result reported by Wilks (1938) on the asymptotic distribution of the likelihood ratio, but this is only for the case of likelihoods with parameters maximizing these likelihoods. Thus, the asymptotic distribution of the NPL statistic (6) is unknown. One can use a parametric version of the NPL statistic based on maximum-likelihood estimates of ability θ_U , θ_C , and θ_T computed from responses to U , C , and $T = U \cup C$, respectively:

$$2 \ln \left[\frac{\prod_{i \in U} \mathbf{P}_i(\chi_i|\theta_U) \prod_{i \in C} \mathbf{P}_i(\chi_i|\theta_C)}{\prod_{i \in T} \mathbf{P}_i(\chi_i|\theta_T)} \right]. \quad (16)$$

The asymptotic distribution of statistic (16) is chi-square with 1 degree of freedom (Wilks, 1938). However, the performance of this parametric version is always lower than the performance of the NPL statistic (6) because a hypothesis test with statistic (16) is a composite hypothesis test; meanwhile Neyman–Pearson lemma is only applicable for a simple hypothesis test, like the hypothesis test with statistic (6). In computer simulations for both types of tests and zero uncertainty, the performance of statistic (16) was lower than EDI’s performance.

Figure 5. Distribution of dots (KS, PS) under different conditions.

Note. Different conditions caused a different number of dots (KS, PS) corresponding to positive values of the PS statistic. KS = Kolmogorov–Smirnov test; PS = posterior shift; P&P = paper-and-pencil; CBT = computer-based testing.

Summary

This article exploits a robust approach by Belov (2015) to detect examinees with aberrant ACs by measuring performance difference between items with ACs and items without ACs, where the performance is estimated from the final responses. His approach is equivalent to detecting item preknowledge (Belov, 2016), where a corresponding statistic measures a difference in performance between compromised and uncompromised items. Two statistics were applied to detect examinees with aberrant ACs, with the objective being to maximize the power of the corresponding hypothesis test. The PS statistic (similar to the KLD) is based on the idea of testing the goodness of fit of one probability distribution to another. The NPL statistic by Drasgow et al. (1996) provides the uniform most powerful test (Lehmann, 1986) for detecting a response vector with aberrant ACs, given that all assumptions for the NPL statistic are true.

Four statistics (KLD, EDI, PS, and NPL) were compared via computer simulations under various conditions controlled by the following independent variables: type of test (P&P or CBT), amount of aberrancy, C-uncertainty, and U-uncertainty. The measure of performance for each statistic was based on the ROC curve. The results demonstrated that all statistics operate well under zero uncertainty, with NPL and PS performing the best. Under moderate uncertainty, some acceptable bounds on average performance were achieved: for P&P, the EDI statistic performed with a lower bound of 0.4 (Figure 3); for CBT, the NPL statistic performed with a lower bound of 0.8 (Figure 4). However, under high uncertainty, the performance of all statistics dropped drastically, with the C-uncertainty having a larger negative effect on the performance than the U-uncertainty (Figures 3 and 4).

The asymptotic behavior of the four statistics was analyzed. This analysis provides critical values without using computer simulations. It also sheds some light on the performance of each statistic.

Based on the presented research, several recommendations for a practitioner can be provided. If one expects a low C-uncertainty in the data, then the best-performing statistics should be PS and NPL (Figures 3 and 4). In CBT, the NPL statistic should keep performing well under moderate C-uncertainty (Figure 4). For both test types, however, in the case of a higher C-uncertainty one should also use EDI. As usual, it is a good idea to apply several statistics simultaneously (e.g., EDI, PS, and NPL) and report examinees detected by all or several statistics; a

Figure 6. Example of two densities violating the conditions of Theorem 1, where three dots indicate points where these densities are equal.

Bonferroni or other correction (Abdi, 2007) should be applied to the significance level to control the Type I error.

Classification of the amount of C- and U-uncertainty in given data is an interesting topic for future research. Let us discuss a possible way to address this. From the definition of C- and U-uncertainty, it follows that larger C- and U-uncertainty should cause a larger number of Guttman errors (Meijer, 1994) at subtests *C* and *U*, respectively. Therefore, statistics such as C^* or H^T (Tendeiro & Meijer, 2014) can be used to classify data into the following C- and U-uncertainty categories: low, moderate, or high. Corresponding confidence intervals can be computed from simulated data.

The four statistics analyzed in this article detect examinees with aberrant ACs. However, these statistics can easily be incorporated into the approach by Belov (2013, 2014, 2017) to detect groups (e.g., classrooms, schools, etc.) with aberrant ACs on a common subset of items.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 103-107). Thousand Oaks, CA: SAGE.
- Belov, D. I. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement*, 50, 141-163.
- Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, 2(3), 37-58.
- Belov, D. I. (2015). Robust detection of examinees with aberrant answer changes. *Journal of Educational Measurement*, 52, 437-456.
- Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, 40, 83-97.

- Belov, D. I. (2017). Identification of item preknowledge by the methods of information theory and combinatorial optimization. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 164-176). New York, NY: Routledge.
- Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, 58, 37-52.
- Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, 9, 47-64.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79-86.
- Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York, NY: Wiley.
- Lehmann, E. L. (1999). *Elements of large-sample theory*. New York, NY: Springer.
- Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53, 161-176.
- Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Maynes, D. (2013). Educator cheating and the statistical detection of group-based test security threats. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 173-199). New York, NY: Routledge.
- Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, 18, 311-314.
- Primoli, V., & Liassou, D. (2013, October). *Comparing answer-changing behavior by test mode*. Paper presented at the 2nd Annual Statistical Detection of Potential Test Fraud Conference, Madison, WI.
- Qualls, A. L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practice*, 20(1), 9-16.
- Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, 51, 239-259.
- Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9, 60-62.
- Wollack, J. A., Cohen, A. S., & Eckert, C. A. (2015). Detecting test tampering using item response theory. *Educational and Psychological Measurement*, 75, 931-953.