**Gregory M. Hurtz**
PSI Services LLC and
California State University

**John A. Weiner**
PSI Services LLC

# Analysis of Test-Taker Profiles Across a Suite of Statistical Indices for Detecting the Presence and Impact of Cheating

## Abstract

Preventing cheating on tests is crucial, but in the event of failure to prevent it, detecting its presence and impact on scores is equally important. Many statistical indices have been developed for this purpose in nearly 100 years of research, often in a quest for the "best" index. The premise of this study is that: (a) Cheating is manifested in different patterns of aberrant responding, (b) No single index can effectively detect all such patterns, and thus (c) A suite of different indices working together should be more effective at identifying anomalous responding indicative of potential cheating. Two methods for simultaneous consideration of a suite of 7 indices were investigated for detecting three patterns of cheating that were manipulated in random subsamples of actual data from two high stakes credentialing tests. Discriminant function analysis distinguished presence and impact of cheating through different linear combinations of the indices. An alternative approach based on testing the fit of individuals' profiles across the 7 indices to heuristic model profiles for each cheating pattern led to similarly accurate classifications of test-taker's into their manipulated cheating groups. This profile similarity method provides a means for theory development and model testing, in aggregating any number of individual statistical indices for detecting fit to any modeled pattern of interest.

**Keywords:** Collusion Detection, Data Forensics, Test Security, Test Cheating

# 1. Introduction

Preventing cheating on tests in occupational and educational settings is crucial, but in the event of failure to prevent it, detecting its presence and impact on scores through forensic data analysis is equally important. Investigations into the statistical detection of likely cheating and other abnormalities in test-taker response behavior have been documented dating back nearly a century to Bird (1927, 1929). In recent years, attention paid to these methods has increased dramatically with the more widespread use of computer-based testing practices in high-stakes settings where test-taker's may be motivated to cheat. The increasing availability of data stored by computer-based testing systems has provided greater capability for building forensic analysis models from item response data as well as associated collateral data (e.g., response times) and meta-data (e.g., date, time, and location of tests).With such data more readily and immediately available for analysis, research in data analytics for forensic investigations into fraudulent test-taker behavior has increased in recent years.

In much of the research in statistical detection of aberrant response patterns over the years, cheating has implicitly been defined as one test-taker colluding with another while taking the test (e.g., Sotaridona et al. 2006; Wollack, 2003), and much of the research has the feel of a search for the single best detection index for that scenario. However, there is a range of scenarios in which cheating may occur either before the test (e.g., gaining advance access to pirated test content), during the test administration (e.g., copying answers directly from another test-taker) or after the test (e.g., answers being changed after-the-fact). Moreover, different indices may be better suited for detecting some patterns in the data than others so that identifying the best index depends on which pattern one wants to detect. Further still, a carefully selected set of multiple indices might, on the whole, allow for distinguishing between the different patterns in a way that any single index is incapable of. The current study explores detection of pre-knowledge and answer-copying patterns using a set of indices that are expected to detect different patterns.

# 2. Literature Review

## 2.1 Response Similarity

One traditional analysis strategy for detecting the presence of cheating effects in item responses is to look at the similarities in response patterns by counting the number of matching responses between test-taker's and comparing the number to an expected value such as the number expected by chance. Many indices are of this form, including those by Bird (1929), Saupe (1960), Angoff (1974), Frary et al. (1977), Bellezza and Bellezza (1989), and Wollack (1997). Some indices focus only on matching incorrect answers, while others focus on all matching answers. Criteria for detecting anomalies are based on comparisons to expected values for the number matching based on theoretical distributions in some cases and derived from empirical data in others.

Research comparing various indices for detecting cheating, and developing new strategies in this realm, has continued into the 2000s. Examples include development of refined expected values of matching responses based on assumptions regarding the typical response behavior of answer-copiers (van der Linden & Sotaridona, 2004), adaptations of Cohen's (1960) kappa for detecting inflated agreement in response choices beyond chance (Sotaridona et al., 2006), and use of regression models to identify outliers in observed vs. predicted matching response counts, taking into account total score (Weiner, et al., 2013; Weiner, et al., 2014). Each of these various indices has its strengths and weaknesses, possibly under different sets of circumstances, making it difficult to identify a single best index for all scenarios.

## 2.2 Score Similarity

Another strategy for detecting cheating is to analyze patterns of item scores, after the raw responses (e.g., A, B, C, D) are converted to correct/incorrect (i.e., 1, 0) status. Person-fit statistics (e.g., Karabatsos, 2003) may be used to flag individuals that produce score patterns substantially contradicting model predictions. These indices will, by their design, "miss" aberrant patterns that occur strictly among the incorrect options, such as an individual with a wrong response copying another test-taker's different wrong response, because all incorrect options are collapsed to a value of "0" before these statistics are computed. This suggests that such indices may only be effective when the behavior directly impacts item scores but will not detect all forms of answer-copying. On the other hand, they are likely to be effective at detecting pre-knowledge because prior access to the content or the answer key is more likely to directly impact scores. Again, there are many options for this type of index, including 36 that Karabatsos (2003) compared; while some appear generally stronger than others it is difficult to identify a single best index for all situations.

## 2.3 Presence vs. Impact of Cheating

The distinction between sensitivity to raw *item responses* and *item scores* is not often recognized in comparisons of alternative indices and does not in itself suggest one type of index is better than others. Instead, the distinction helps

to highlight a potentially important distinction between measures of the presence of cheating versus measures of the impact of cheating. One needs to look carefully at how different indices are computed in order to evaluate what patterns they might be capable or incapable of detecting, so that multiple indices can be selected to detect multiple potential patterns. Detecting the *presence* of cheating, despite its *impact* on scores, could be an important part of a bigger-picture test security monitoring system designed to look for security weaknesses at a system-level with a prevention and quality control focus. Detecting presence and impact together may be used in cases of detecting individual cases to investigate further for score validation. In any case, a suite of both presence and impact measures of aberrant response patterns would seem to have value for a test security analysis system, with multiple measures of each type used to replicate and corroborate detection of patterns across multiple indices. This may allow some indices to compensate for others if the different indices are differentially sensitive to elements of a particular situation (e.g., on easier exams, on shorter exams, when fewer items are copied, etc.).

### 2.4 Combining Detection Indices

While the use of multiple indices to corroborate findings or detect multiple patterns is not necessarily a novel idea, strategies for combining the multiple indices have not been extensively investigated in the extant literature. An exception is a study by Wollack (2006) who investigated different pair and triplet combinations of a set of indices to search for the combination with the best power and Type I error performance. Each index was computed and converted into a null hypothesis significance test, and the rejection decisions from each test were submitted to decision rules where, for example, the null hypothesis was rejected if at least one index in the pair was statistically significant at an adjusted alpha level of $\alpha/2$. A particular pair of indices that were each sensitive to different patterns (answer copying that occurs randomly vs. in runs or strings of adjacent items) emerged as most effective for the manipulated conditions in the study. The study supports the notion of combining information from multiple indices. However, the specific methodology was limited to indices with associated null hypothesis tests. Moreover, it relied on the dichotomous rejection decision from each individual test which could result in a loss of information on magnitude or effect size. This means it would likely suffer a loss in power if more than two or three tests were combined, given the increasing adjustments to the alpha level.

## 3. Present Study

The current study was carried out to evaluate an alternative approach toward combining multiple (possibly many) indices to detect different aberrant patterns reflecting different cheating scenarios. The general approach was to analyze test-taker profiles across multiple indices, drawing conclusions at the level of a test-taker's overall profile rather than on each of the individual tests within the profile. The study explored the development of a model using optimally-weighted composite scores across a set of alternative indices through discriminant function analysis. An alternative profile analysis approach was also developed to address potential limitations in the first approach. In both cases, theoretical meaningfulness and classification accuracy were considered.

## 4. Method

### 4.1 Data

The primary data were from a high-stakes credentialing exam with test-taker's ($N$ = 1551) who completed a 140-item form of an exam from a Rasch-calibrated item bank. A secondary cross-validation dataset ($N$ = 1169) from a different exam of the same length was also developed from a Rasch-calibrated item bank. In each case, the dataset was viewed as a "haystack", with randomly sampled cases drawn out, partially manipulated, and planted back into the haystack and conceptualized as "needles." Model building ensued by exploring detection of the needles in the haystack. The haystacks in this scenario were made up entirely of real examinees' responses to the items, and the needles were likewise based on a subset of examinee responses that were partially manipulated. Given prior screening of items for sufficient fit to the Rasch model, these were expected to provide suitably well-behaved datasets, with greater ecological validity than purely simulated data.

### 4.2 Design

The study included three experimental (needle manipulation) conditions and one control (natural response) condition. The needle manipulations were carried out following two paradigms, leading to three groups, with 5% of test-taker's falling into each group (for a total of 15% cheating). The first was a *response similarity* paradigm, where for each cheater another test-taker within the data set was sampled as their source, and some of

**TABLE 1. MANIPULATIONS OF "CHEATING" RELATED DATA PATTERNS**

| Data Manipulations | Copy Random | Copy Upward | Pre-Knowledge |
|---|---|---|---|
| % of Examinees Cheating | 5% | 5% | 5% |
| % of Items Cheated | 30-40% | 30-40% | 35% |
| Difficulty of Items Cheated | All | $p < 0.75$ | $p < 0.75$ |
| Ability Level of Cheaters | All | Low: $Z < -1$ | Low: $Z < -1$ |
| Ability Level of Source | All | High: $Z > 1$ | N/A |

the cheater's answers were overwritten by their source's answers. The second paradigm was an item *pre-knowledge* pattern (conceptualized similarly to Belov, 2016) where samples of items were selected, and each sample treated as a compromised set that a group of cheaters had access to, and each cheater's original answers were replaced with the correct answers on the majority of these items. The first paradigm was used to create two groups (Copy Random and Copy Upward) while the second paradigm was used to create one group (Pre-Knowledge). Table 1 outlines the conditions that were manipulated to create the three experimental groups.

### 4.2.1 Copy Random Group

For the first group, a random sample of 5% of test-taker's was selected from throughout the full ability range to serve as cheaters, and each was paired with a source that was likewise sampled from throughout the ability range. Each cheater "copied" their source's answers on a different random sample of 30-40% (average 35%) of test items which Wollack (2006) described as "major" copying. No constraints were placed on how difficult the items were that they copied. The purpose of this group was to evaluate a condition with no assumptions about who cheats, who they cheat from, or on which items they are most likely to cheat. For the purpose of the current study it creates a group for whom cheating is present, yet scores are not always impacted.

### 4.2.2 Copy Upward Group

The second group was constructed through a similar process to the first group except that constraints were placed on the ability levels of the cheater and source, and the difficulty levels of the items. The random sampling of cheaters was constrained such that their standard score on the test fell below -1, while the random sampling of sources was constrained to standard scores above +1. The random sampling of items was constrained such that they each copied a different (but likely overlapping) 30-40% of items that were generally among the more challenging (*p*-values below .75). This pattern is one where lower ability test-taker's improve their chances of obtaining correct item scores on difficult items, by copying off a higher ability test-taker. It should be noted that while the chances of improving item scores are increased by this manipulation overall, in individual cases some of the answer copying may be ineffective due to copying wrong answers.

### 4.2.3 Pre-knowledge Group

The third group was similar to the second group except that it was created with a pre-knowledge paradigm similar in concept to Belov (2016). Lower-ability test-taker's (standard scores below -1) were again sampled, and they were again manipulated to have cheated on the relatively more difficult items (*p*-values below .75). However, their cheating manipulation was not carried out by copying answers directly from another test-taker but instead through manipulation of the keyed response as if they had prior knowledge of the items. Four "compromised" item sets were created by taking separate but possibly overlapping, samples of 40% of the test items. These four item samples represented four sets that were hypothetically pirated and made available to test-taker's before the test. Each cheater was linked to one of the four item sets, and their original, natural responses to a randomly selected 90% of items in their set were replaced by the correct answer. Selecting a different 90% of the items for each test-taker served two purposes. First, it represented either fallibility in test-taker's' memories of the content they were exposed to or fallibility in what they understood to be the correct answers on their pirated content (stolen content does not necessarily have correct answers marked). Second, selecting 90% of the items results in a proportion of items with manipulated results that is similar in size to the average number of manipulated responses in each of the previously-defined groups.

### 4.2.4 Natural Responses Group

This group is the remaining 85% of test-taker's for whom no manipulation of item responses was introduced. The test-taker's' original responses were left untouched, making it a realistic comparison baseline. It should be noted that the answer-copiers' "sources" were selected from among this group – which implicitly assumes that sources are innocent victims of copying, that they allow others to copy from them, or that they help coach others beforehand without

themselves being influenced by the other test-taker's (making their own responses their "natural" responses).

### 4.3 Statistical Indices

Several indices were selected that, by their nature, were expected to differentially detect the presence versus impact of manipulated response patterns. Our goal was to select indices that were expected to detect the different patterns based on how they are computed. General descriptions are summarized in Table 2 and additional discussion follows.

#### 4.3.1 Measures of Impact

Two indices were selected for detecting cheating-related response behaviors that directly (and solely) impact scores: $H_i^T$ and $C_i$*. Both of these measures are based only on item scores, so they do not "see" the original responses test-taker's gave to the items. Our attention to these indices came from a review of Karabatsos' (2003) study comparing 36-person fit statistics, where he found $H_i^T$ (Sijtsma & Meijer, 1992) to be the strongest or among the strongest for detecting five manipulated score patterns with later studies by Dimitrov and Smith (2006) and Tendeiro and Meijer (2014) replicating the utility of $H_i^T$. Karabatsos also found Sato's (1975) caution index $C_i$ and Harnisch and Linn's (1981) modified caution index ($C_i$*) to be nearly as effective.

An advantage of $C_i$ and $C_i$* is that they involve only the focal test-taker's observed response pattern, comparing it against a Guttman pattern defined in accordance with rank ordered item difficulties, which can come from prior administrations of the items through pretesting. $H_i^T$ on the other hand involves covariances between a test-taker's observed item score patterns and the patterns of the other test-taker's in the sample under investigation, making it potentially more vulnerable to the effects of sample characteristics. Thus, while Karabatsos found $H_i^T$ to work slightly better in simulated data, it would seem more potentially vulnerable to the effects of sample characteristics since $C_i$ and $C_i$*, if computed from archived *p*-values not involving the sample under investigation, would not be influenced by sample characteristics. As our goal was to build a suite of indices that would allow for balance among their individual strengths and potential limitations, we included both $H_i^T$ and $C_i$* rather than picking just one.

#### 4.3.2 Measures of Presence

For detecting the presence of cheating regardless of impact on scores, we included several measures that are based on similarities in raw item responses. First, we elected to use an adaptation of Cohen's (1960) kappa following the logic of Sotaridona et al.'s (2006) recoding scheme, where correct options are coded as 1 and incorrect options are coded as 2, 3, and 4 in descending order of popularity (option *p*-values), conditioned on a test-taker's score level. We saw this measure as an incremental

**TABLE 2. INDICES OF ABERRANT RESPONSE PATTERNS SELECTED FOR DEFINING THE GROUP PROFILES**

| Index | Pattern detected | Brief Description |
|---|---|---|
| $H_i^T$ (Sijtsma & Meijer, 1992) | Impact | Covariances of an individual's item scores with others' item score patterns, relative to their maximum possible covariances. |
| $C_i^*$ (Harnisch & Linn, 1981) | Impact | Covariance of an individual's item scores with item difficulties, relative to the expected covariance if scores followed the Guttman pattern. |
| $k_r$ (Sotaridona et al., 2006) | Presence | Index of agreement between two test-taker's in their selection of item responses relative to chance, after recoding raw responses to align correct answer categories and popularity-ordered incorrect categories. |
| *B* (Angoff, 1974) | Presence | Number of matching errors with another test-taker, standardized against an expected mean and standard deviation of matching errors computed within strata of the product of their total numbers of errors. |
| $J_2$ (Weiner et al., 2013) | Presence | Standardized residual from regressing examinees' maximum observed number of matching responses with another test-taker onto their own number-correct test score. |
| $J_3$ | Presence | Standardized residual from regressing examinees' maximum observed number of matching responses with another test-taker onto their own mean of matching responses with others. |
| $Z_i$ | Presence | Number of matching responses with another test-taker, standardized against one's own mean and standard deviation of matching responses with others. |

step away from the "impact" measures that were sensitive only to the correct/incorrect status of an individual's responses, in that the recoded kappa ($\kappa_r$) index retains the distinction among the incorrect categories and uses this information instead of collapsing all incorrect categories into a score of zero. Another potential advantage of $\kappa_r$ is that it involves only the response patterns of two test-taker's, along with response option *p*-values which can be computed from prior administrations, while all other indices described below require a comparison against a pool of other test-taker's. This should make $\kappa_r$ less sensitive to sample characteristics.

Second, to move further away from the "impact" measures that are sensitive only to the correct/incorrect status of a response, we used Angoff's (1974) Index *B* which only includes analysis of similarities among incorrect answers—precisely the information that is lost in the impact measures by their collapsing of all incorrect answers into a score of zero. For *B*, the number of matching errors in each pairing of test-taker's is standardized against a mean and standard deviation of matching errors that are computed from the data and are conditioned on the stratum that each pairing of test-taker's falls into. The strata are defined along a continuum created from the products of the numbers of errors made by each member of each pairing of test-taker's.

Third, we sought to include measures that did not distinguish between correct and incorrect answers at all but instead simply looked at matches among the raw responses. A number of such indices can be found in the literature, including some widely-cited indices involving conditional probability computations from option-level probabilities derived from item response models (Frary et al., 1979; Wollack, 1997). For the current investigation we opted for simpler indices including, and derived from, the $J_2$ index (Weiner et al., 2013) that starts by simply counting the number of matching responses between each pairing of test-taker's and saving the maximum observed match count for each test-taker. For $J_2$, these maximum match count values are regressed onto the test-taker's number correct test scores to obtain predicted maximum values for each test score. $J_2$ for each test-taker is their standardized residual from this regression model. Large positive standardized residuals are taken to indicate that a test-taker had a maximum match count that exceeded the expected maximum value for people with the same number-correct score and may indicate collusion.

Two other indices presented here were derived as alternatives to $J_2$. First, an index we refer to as $J_3$ is computed identically to $J_2$ except that each test-taker's average number of matches across the pool of test-taker's is substituted into the regression equation in place of their number-correct scores as predictors of the maximum value. Positive standardized residuals from this model indicate that one's maximum observed value exceeds the expected maximum for people with their average match count. Since test scores (the sum of item scores) are not used in the index at all, it is even further away from the "impact" measures in how it is computed—the model is derived entirely from match count information. A related index we refer to as $Z_i$ is an individually-standardized value for each match count, where one's match count with each other test-taker is standardized against his or her own average and standard deviation of match counts across all pairings with other test-taker's.

### 4.3.3 Index Scaling

Some indices are computed as a single value for each test-taker ($C_i^*$, $H_i^T$, $J_2$, $J_3$) while others ($Z_i$, $\kappa_r$, and *B*) are computed separately for each pairing of all test-taker's. For the latter set, individual test-taker's' values for the current analysis were set to their maximum observed values, as these would be the most likely cases of collusion if any occurred at all. Next, all indices were placed onto the same metric to improve interpretability of subsequent analyses, by standardizing each with respect to its mean and standard deviation in the "natural response" group. In addition, the standardized $H_i^T$ scores were reflected so that, like all other measures in this study, higher values are expected to represent more aberrant responding.

## 4.4 Analysis Strategy

Discriminant function analysis was carried out to form a weighted composite of each test-taker's profile across the standardized index values, with weights that maximally differentiated between groups. With four groups to differentiate (including the baseline "natural response" group), the analysis produced three weighted composite functions and assigned cases to groups using the associated classification coefficients. Each successive function had prior functions partialled. out so that each provided unique separation of groups beyond the previously-defined functions. The "loadings" of each index onto each function were examined to understand which indices were most strongly associated with each function, then average function scores (group centroids) for each group were evaluated to investigate which groups tended to be differentiated by each function. Finally, a comparison of predicted group membership to actual group membership gave a sense of how accurately the functions, taken together as a set, were able to place individuals into their groups. Follow-up analyses involved comparison to classification accuracy based on each test-taker's degree of fit to model profiles across the standardized indices, as measured by the similarity index *D*.

# 5. Results

## 5.1 Descriptive Statistics

Table 3 provides descriptive statistics across groups, and Table 4 provides correlations among the raw values of the statistical detection indices. Not surprisingly, very high correlations were observed between $C_i$* and $H_i^T$ and between $J_3$ and $Z_i$, and other indices were moderately to strongly correlated as well.

## 5.2 Discriminant Functions from Standardized Indices

Table 5 summarizes the results of the discriminant function analysis. The squared canonical correlation coefficients ($R^2$) in Table 5 indicate strong effect sizes for the first two functions (.73 and .45, respectively), with a considerably smaller effect size for the third function (.12). The first two functions accounted for over 96% of the cumulative variance in the solution.

The structure coefficients for the first function reveal that it is most strongly associated with the $C_i$* and $H_i^T$ indices, and the group centroids reveal that this function distinguishes the groups whose scores were consistently impacted by their cheating strategies (Pre-Knowledge centroid = 6.22; Copy-Upward centroid = 3.25) from those whose cheating strategies did not systematically lead to a change in scores (Copy-Random, centroid = -0.02) and those whose responses were untouched (Natural-Response, centroid = -0.56). It is reasonable that $C_i$* and $H_i^T$ define this function, given that they are based on already-scored items and do not "see" any patterns among the incorrect answers; they are therefore sensitive only to effects on item scores specifically.

Structure coefficients for the second function reveal that it is most strongly associated with the $Z_i$, $J_2$, and $J_3$ indices. The group centroids suggest that after systematic impact on scores are accounted for by the first function, the second function provides further differentiation with respect to the overall presence of response similarities, in that the Copy-Upward group (centroid = 3.39) and to a much lesser degree the Copy-Random group (centroid = 0.62) are contrasted with the Pre-Knowledge group (centroid = -1.95), while the Natural-Response group is

**TABLE 3. DESCRIPTIVE STATISTICS FOR ALTERNATIVE INDICES**

| | Natural-Response ($N$ = 1317) | Copy-Random ($N$ = 78) | Copy-Upward ($N$ = 78) | Pre-Knowledge ($N$ = 78) | Total ($N$ = 1551) |
|---|---|---|---|---|---|
| Index | *M(SD)* | *M(SD)* | *M(SD)* | *M(SD)* | *M(SD)* |
| $H_i^T$ | 0.24 (0.04) | 0.23 (0.03) | 0.11 (0.05) | -0.02 (0.05) | 0.22 (0.08) |
| $C_i^*$ | 0.22 (0.05) | 0.23 (0.04) | 0.37 (0.06) | 0.54 (0.06) | 0.24 (0.09) |
| Max-$k_r$ | 0.39 (0.06) | 0.47 (0.07) | 0.41 (0.09) | 0.27 (0.04) | 0.39 (0.07) |
| Max-B | 3.26 (0.73) | 4.32 (1.07) | 4.02 (1.16) | 1.53 (0.51) | 3.27 (0.91) |
| $J_2$ | 0.01 (0.75) | 0.52 (0.88) | 1.54 (1.26) | -2.23 (0.62) | 0.00 (1.00) |
| $J_3$ | -0.21 (0.74) | 0.36 (0.85) | 2.66 (1.11) | 0.54 (0.57) | 0.00 (1.00) |
| Max-$Z_i$ | 2.26 (0.31) | 2.47 (0.35) | 3.66 (0.54) | 2.43 (0.30) | 2.35 (0.45) |

**TABLE 4. CORRELATIONS AMONG STATISTICAL INDICES**

| Index | $H_i^T$ | $C_i^*$ | $k_r$ | B | $J_2$ | $J_3$ | $Z_i$ |
|---|---|---|---|---|---|---|---|
| $H_i^T$ | 1.00 | | | | | | |
| $C_i^*$ | -0.97 | 1.00 | | | | | |
| $k_r$ | 0.53 | -0.55 | 1.00 | | | | |
| B | 0.43 | -0.42 | 0.74 | 1.00 | | | |
| $J_2$ | 0.46 | -0.47 | 0.74 | 0.69 | 1.00 | | |
| $J_3$ | -0.31 | 0.36 | 0.36 | 0.29 | 0.58 | 1.00 | |
| $Z_i$ | -0.32 | 0.36 | 0.16 | 0.36 | 0.55 | 0.84 | 1.00 |

*Note*: $H_i^T$ in this table is in its original form where lower values indicate more aberrant responding.

**TABLE 5. DISCRIMINANT FUNCTIONS ANALYSIS ON MULTIPLE STATISTICAL INDICES**

| | | Function 1 | Function 2 | Function 3 |
|---|---|---|---|---|
| Canonical R2: | | 0.73 | 0.45 | 0.12 |
| Eigenvalue: | | 2.75 | 0.80 | 0.13 |
| Cum. % of Var.: | | 74.6 | 96.4 | 100.0 |
| Index | Tolerance | Loadings and Coefficients | | |
| | | Structure | Structure | Structure |
| $C_i^*$ | 0.02 | ***.91*** | -.12 | -.17 |
| $H_i^T$ | 0.03 | ***.85*** | -.11 | -.08 |
| $Z_i$ | 0.13 | .32 | ***.88*** | -.10 |
| $J_2$ | 0.07 | -.21 | ***.79*** | .11 |
| $J_3$ | 0.08 | .34 | ***.71*** | .02 |
| $k_r$ | 0.16 | -.21 | .36 | ***.67*** |
| $B$ | 0.26 | -.20 | .54 | ***.66*** |
| | | Centroids | | |
| Natural Response: | | -0.56 | -0.12 | -0.08 |
| Pre-Knowledge: | | ***6.22*** | ***-1.95*** | 0.05 |
| Copy-Upward: | | ***3.25*** | ***3.39*** | -0.30 |
| Copy-Random: | | -0.02 | 0.62 | ***1.56*** |
| | | Interpretations | | |
| | | Identifies patterns that primarily impact item scores | Detects residual patterns of overall response similarity | Detects residual patterns of error similarity |

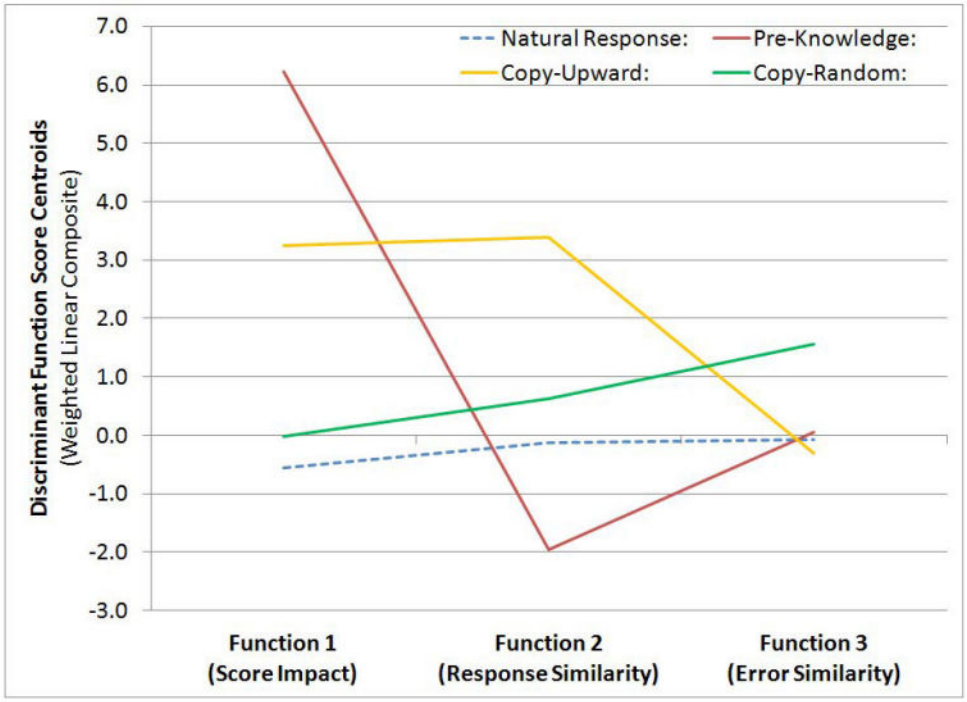
*Note*: All indices were entered in standardized form; $H_i^T$ was further reflected so high values indicate more aberrance.

relatively neutral (centroid = -0.12). It makes sense that $Z_i$, $J_2$, and $J_3$ define this function since all three of these indices use the overall number of matching responses between pairs of test-taker's as their basis and "see" the presence of patterns among the incorrect options that are missed by the first function.

The structure coefficients for the third function reveal that it is most strongly associated with the B and $\kappa_r$ indices, and group centroids suggest that it differentiates a residual pattern in the Copy Random group (centroid = 1.56) from the other three groups (centroids = -0.08, 0.05, and -0.30). Angoff's *B* index specifically analyzes matching errors, and the $\kappa_r$ index involves distinguishing correct responses from incorrect responses and sets up an error term that especially contrasts model-predicted patterns among the errors with chance responding. While this function was not as strong as the first and second and was not anticipated in advance as a distinct pattern from function 2, it suggests that error similarity indices may in fact provide somewhat distinct information for detecting the presence of fairly random patterns of cheating.

Figure 1 displays the group centroids of the discriminant function scores in a profile plot in order to more easily visualize how the functions differentiate groups. This plot clearly shows the different profile patterns for the Pre-Knowledge and Copy-Upward groups, and to a much lesser degree the Copy-Random group. The ability of these functions to accurately differentiate groups is summarized in Table 6 as classification accuracy results. The diagonal of the percent section reveals that 97.0% of the Natural Responses group was correctly classified as such, and most of the false positives for this group were classified as random copiers. The Pre-Knowledge group was 100% correctly classified. The Copy-Upward group had a 7.7% false negative rate with 91.0% correctly classified and 1.3% classified as aberrant but placed in the Pre-Knowledge group. The much less systematic Copy Random group had a much lower rate of correct classification with 32.1% correctly classified and a 67.9% false negative rate. Totals along the bottom of the table further show that relative to the size of the haystack (85% of sample) and needles (5% each), the Copy-Random group is 1.5% under-detected, with most of those cases being misclassified as natural responders.

**FIGURE 1.** GROUP PROFILES BASED ON DISCRIMINANT FUNCTION SCORES.




## 5.3 Discriminant Function Analysis Assumptions

While the discriminant function analysis has elucidated meaningful patterns of results in the data that support the distinction between presence and impact of aberrant response patterns, practical application as a method of computing composite scores across multiple indices would involve derivation of a stable set of classification coefficients to weight and score individuals' index profiles and categorize them into groups. At that level of specificity in deriving a set of stable parameters that will cross-validate across samples, it would become paramount that the parameters be derived within a context where model assumptions are sufficiently met.

To this end, homogeneity of variances was evaluated by taking the ratio of largest to smallest group variances for each of the seven indices, and ratios of 2.7 to 5.2 were found. These ratios indicate a modest degree of heterogeneous variances, and consistently, the "Copy-Up" group had the largest variance. In terms of normality, the $C_i$*, $H_i^T$, and $\kappa_r$ indices had relatively normal skewness and kurtosis values (well within ±1) across all groups but the other indices tended to have larger degrees of

**TABLE 6.** CLASSIFICATION ACCURACY OF DISCRIMINANT FUNCTION SCORES

| Type of Cheating | | | Predicted Group | | | | Total |
|---|---|---|---|---|---|---|---|
| | | | Nat-Resp | Pre-Know | Copy-Up | Copy-Rand | |
| Actual Group | *n* | Nat-Resp | 1278 | 0 | 9 | 30 | 1317 |
| | | Pre-Know | 0 | 78 | 0 | 0 | 78 |
| | | Copy-Up | 6 | 1 | 71 | 0 | 78 |
| | | Copy-Rand | 53 | 0 | 0 | 25 | 78 |
| | % | Nat-Resp | 97.0 | 0.0 | 0.7 | 2.3 | 100.0 |
| | | Pre-Know | 0.0 | 100.0 | 0.0 | 0.0 | 100.0 |
| | | Copy-Up | 7.7 | 1.3 | 91.0 | 0.0 | 100.0 |
| | | Copy-Rand | 67.9 | 0.0 | 0.0 | 32.1 | 100.0 |
| | | **Total %** | **86.2** | **5.1** | **5.2** | **3.5** | |

positive skew and leptokurtosis in the Natural Response and Pre-Knowledge groups. In terms of the independence assumption we did not run a statistical test but note that in the Pre-Knowledge condition there is notable potential for within-group non-independence resulting from the manipulated cheating patterns relating back to a common source of correct answers for each of the four item sets. There is also non-independence between each of the Copy groups and the Natural Response group where their sources resided. While these assumptions are typically most problematic for error rates associated with statistical significance tests which were not the primary focus in the study, violations can impact the estimation of model parameters as well.

While the analysis of standard assumptions above raised potential concerns, of likely more consequence in the practical application of a discriminant function model for this purpose is the high degree of collinearity among some of the alternative statistical indices of aberrant responding, indicated by high correlations in Table 4 and especially the low tolerance values in Table 5. Multicollinearity is known to produce instability in parameter estimation for discriminant analysis (e.g., Naes & Mevik, 2001), suggesting that deriving stable coefficients for practical application may be problematic. One strategy for eliminating multicollinearity is to remove highly correlated predictors, which we explored as a strategy in the current analysis by re-running the analysis with only three predictors, selecting the one from each function in Table 5 that had the highest structure coefficient (i.e., $C_i$*, $Z_i$, $\kappa_r$). A check on multicollinearity for these three predictors revealed a much-improved situation with tolerance values ranging from .49 to .69. The results were extremely similar to those in Table 5 in terms of the patterns of group centroids, with the most notable difference being a drop in the centroid value of the Copy-Random group on Function 3 from 1.56 to 1.22. Accordingly, comparison of classification accuracy to Table 6 revealed that the percents were identical for the first three groups, but the Copy-Random percentage dropped from 32.1 to 20.5.

For the first two patterns at least – namely *score impact* and *response similarity* – this analysis suggests that if the goal is to identify the most effective subset of indices that differentiates the groups in this sample, using the discriminant function analysis and selecting from the highest structure coefficients appears effective for this use. This strategy, however, leaves open the possibility that if the rank-orders of structure coefficients within each function vary across different Monte Carlo conditions or across different real-world scenarios, the results from the single selected index might not cross-validate. One of the initial objectives of the current study was to devise a method that allows for combining multiple indices in a way that might allow their strengths and weaknesses in sensitivities to varied scenarios in practice (e.g., test lengths, fit to item response models, etc.) to balance out. Thus, in line with the initial goal of the current study, we sought to explore an alternative methodology that would allow for retention of multiple correlated indices within each pattern (pre-knowledge, answer copying, etc.), on the premise that they might collectively prove to be more stable in the long run across different situations than any of them would be alone.

### 5.4 An Alternative Strategy: Profile Similarity Analysis

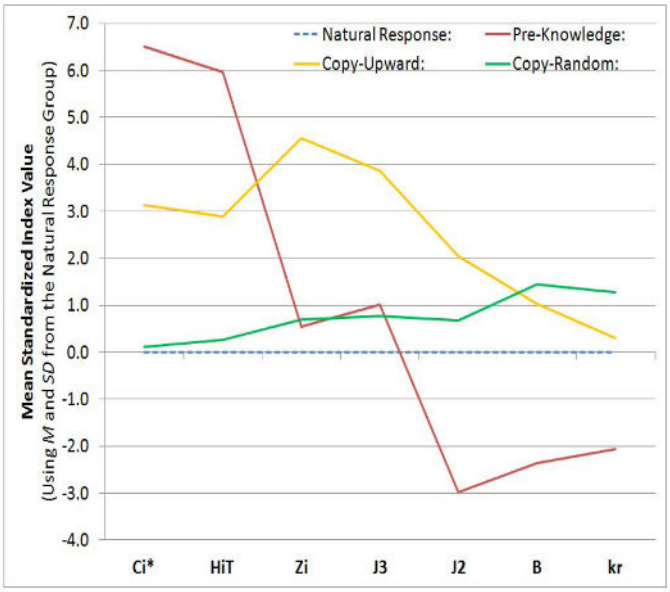
An alternative approach was explored to analyze similarities (i.e., fit) to model profiles across the observed measures, where the model profiles were derived to represent expected patterns resulting from the different cheating strategies. Test-taker's were then classified into the group they displayed the closest fit to.

The first step in this approach was to define the model profiles. The left panel of Figure 2 shows the average standardized index values across the seven indices for each group. Note that the graph reveals similar patterns of differentiations among the groups as the function profiles did in Figure 1. While these average profiles could be used directly, they are subject to sampling error and we sought to establish a heuristic set of values that might be applicable across multiple exams; we therefore used the averages to guide development of a set of model profiles shown in the pattern coefficients on right panel of Figure 2.

In doing this we essentially rounded the averages to their nearest integer; for example, the Pre-Knowledge averages were rounded to a pattern of 6, 6, 1, 1, -3, -2, -2. This method, however, resulted in a fairly flat line for the Copy-Random group with coefficients of 0, 0, 1, 1, 1, 1, 1 which was not very distinguishing from the Natural Response pattern of all 0's. In order to force a stronger distinction between this model profile and the Natural-Response profile, we increased the coefficients as shown in the right panel of Figure 2, to 0, 0, 2, 2, 2, 3, 3, which maintained the shape of the corresponding average profile in the left panel of the Figure 2 but slightly exaggerated the model pattern.

Next, fit of each test-taker's individual observed profile to each of the heuristic model profiles was assessed with the similarity index D, which is the square root of the sum of the squared deviations of a test-taker's index values from the model values. Smaller D values indicate stronger fit to a particular model profile. Table 7 shows average D values for each group on each profile. Each profile fit best to its expected group, and all groups except Copy-Random fit best to their expected profile. The Copy-Random group's fit to its own profile could be increased by adjusting the model coefficients that we exaggerated, but this has the effect of increasing false positive classifications among

**FIGURE 2.** GROUP PROFILES BASED ON AVERAGE (LEFT) AND HEURISTIC MODEL (RIGHT) STANDARDIZED INDEX VALUES.



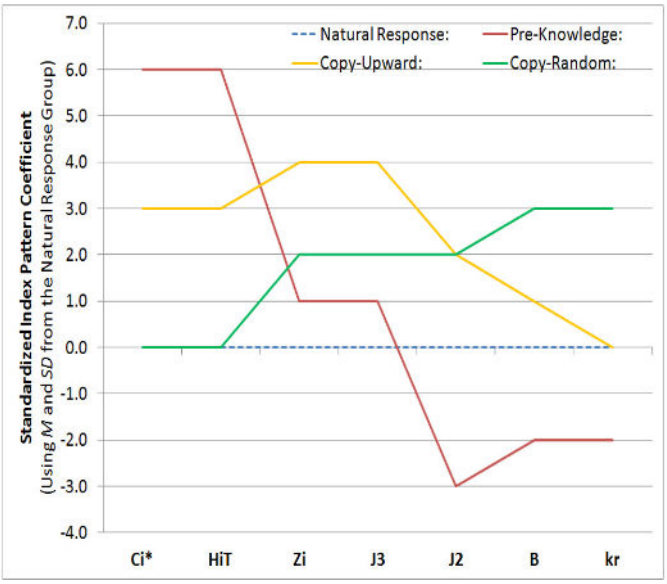




**TABLE 7.** DESCRIPTIVE STATISTICS FOR PROFILE SIMILARITY FIT VALUES

| | Natural-Response Profile | Pre-Knowledge Profile | Copy-Upward Profile | Copy-Random Profile |
|---|---|---|---|---|
| **Group** | ***M(SD)*** | ***M(SD)*** | ***M(SD)*** | ***M(SD)*** |
| Natural-Response | **0.87** (0.50) | 3.68 (0.65) | 2.94 (0.43) | 2.24 (0.51) |
| Pre-Knowledge | 3.79 (0.67) | **0.88** (0.38) | 3.51 (0.53) | 4.83 (0.61) |
| Copy-Upward | 3.17 (0.86) | 3.50 (1.14) | **1.40** (0.62) | 2.72 (0.72) |
| Copy-Random | 1.23 (0.67) | 3.96 (0.61) | 2.61 (0.48) | **1.53** (0.59) |

*Note: Numbers in bold represent the closest-fitting group for each profile.*

**TABLE 8.** CLASSIFICATION ACCURACY OF PROFILE SIMILARITY-BASED GROUP CLASSIFICATIONS

| Type of Cheating | | | Predicted Group | | | | Total |
|---|---|---|---|---|---|---|---|
| | | | Nat-Resp | Pre-Know | Copy-Up | Copy-Rand | |
| Actual Group | *n* | Nat-Resp | 1249 | 3 | 4 | 61 | 1317 |
| | | Pre-Know | 0 | 78 | 0 | 0 | 78 |
| | | Copy-Up | 3 | 6 | 64 | 5 | 78 |
| | | Copy-Rand | 48 | 0 | 2 | 28 | 78 |
| | % | Nat-Resp | 94.8 | 0.2 | 0.3 | 4.6 | 100.0 |
| | | Pre-Know | 0.0 | 100.0 | 0.0 | 0.0 | 100.0 |
| | | Copy-Up | 3.8 | 7.7 | 82.1 | 6.4 | 100.0 |
| | | Copy-Rand | 61.5 | 0.0 | 2.6 | 35.9 | 100.0 |
| | | **Total %** | **83.8** | **5.6** | **4.5** | **6.1** | **100.0** |

the Natural-Response group, so we opted not to maximize fit in this way.

Final classifications are shown in Table 8. In Table 8, each test-taker was assigned to the profile they had the greatest fit to (lowest D value), and these assignments were cross-tabulated with the actual group they belonged to in the simulation. Overall the results were consistent with the earlier discriminant function classifications, except for a slightly higher false positive rate for Natural-Response test-taker's and slightly lower false negative rates for the Copy-Up and Copy-Random groups. Totals along the bottom of the table show that the Copy-Random group is now 1.1% over-classified relative to its expected value of 5%, and the Natural-Response group is now 1.2% under-classified relative to its expected value of 85%. Nevertheless, the results overall are quite similar to those from the discriminant function analysis.

### 5.5 Cross-Validation of Model-Fit Based Classifications

Finally, the profile analysis method was cross-validated in a second exam through replication of the needle-haystack manipulation and computation of fit to the heuristic profile models in Figure 2. The mean D values in the second exam followed the same patterns as shown in Table 7 (and are therefore not provided), with each profile fitting best to its expected group, and all groups except Copy-Random fitting best to their expected profile. Again, the Natural-Response and Copy-Random groups were not as distinct as the others. Classification results in Table 9 again show a similar pattern with the Copy-Random group being slightly over-classified and the Natural-Response group being slightly under-classified, resulting from the lesser distinctiveness of these two groups' profiles.

For the sake of comparison, an additional discriminant function analysis was run in the cross-validation dataset with the reduced subset of predictors identified earlier: $C_i$*, $Z_i$, $\kappa_r$. The same patterns emerged with the three functions being each defined by one of the three indices.

In terms of classification accuracy, the discriminant analysis results showed slightly higher accuracy rates for the Natural Response (97.5 vs. 94.3) and Pre-Knowledge (100.0 vs. 98.3) groups, identical rates for the Copy-Up group (91.4), but substantially lower rates for the Copy-Random group (22.4 vs. 41.4). Further, a subsequent analysis including all seven of the original indices showed a stronger structure coefficient for B vs. $\kappa_r$ meaning that had this sample been used first, a different exemplar of this function would have been selected. This supports the strategy of retaining multiple indices for each pattern.

## 6. Discussion

The purpose of this study was to evaluate the use of a suite of statistical detection indices for differentiating two patterns of cheating on multiple-choice tests: A pattern indicative of the *presence* of cheating and a pattern indicative of cheating that *impacts* test scores. Both are useful for quality-control monitoring of test security practices, such as monitoring test sites for evidence of security breaches, while the latter is of particular importance to situations where scores must be investigated and validated. The results of the discriminant function analysis supported this distinction and the sensitivity of different indices to each pattern, and also suggested that it may be useful to consider general response similarity and more specific error similarity indices separately in order to detect different patterns.

**TABLE 9. CLASSIFICATION ACCURACY OF PROFILE SIMILARITY IN CROSS-VALIDATION EXAM**

| Type of Cheating | | | Predicted Group | | | | Total |
|---|---|---|---|---|---|---|---|
| | | | Nat-Resp | Pre-Know | Copy-Up | Copy-Rand | |
| Actual Group | *n* | Nat-Resp | 938 | 2 | 7 | 48 | 995 |
| | | Pre-Know | 0 | 57 | 1 | 0 | 58 |
| | | Copy-Up | 1 | 3 | 53 | 1 | 58 |
| | | Copy-Rand | 33 | 0 | 1 | 24 | 58 |
| | % | Nat-Resp | 94.3 | 0.2 | 0.7 | 4.8 | 100.0 |
| | | Pre-Know | 0.0 | 98.3 | 1.7 | 0.0 | 100.0 |
| | | Copy-Up | 1.7 | 5.2 | 91.4 | 1.7 | 100.0 |
| | | Copy-Rand | 56.9 | 0.0 | 1.7 | 41.4 | 100.0 |
| | | **Total %** | **83.1** | **5.3** | **5.3** | **6.2** | **100.0** |

In the end, however, high collinearity among measures was deemed a limiting factor in the discriminant function analysis approach unless the purpose is to eliminate indices to include just one per pattern.

The profile similarity strategy revealed a promising alternative that does not require elimination of indices from the suite, where strong fit to either the pre-knowledge or copy-upward models successfully identified cheating with score impact, while more moderate fit to the copy-upward and/or strong fit to the copy-random profiles helped detect the presence of cheating that did not as definitively impact scores. While to some degree this strategy resulted in a loss of distinction between impact and presence that was gained through the partialling in the discriminant function analysis, the practical outcome in terms of classification accuracy was not greatly affected. Further, the response process distinction between fitting the "pre-knowledge" profile more strongly than an answer-copying or "response similarity" profile still provides information on the specific cheating strategy, in terms of the degree to which it appears to involve direct collusion between test-taker's versus individual pre-knowledge of test content or answers.

In this vein, it is also worth noting that for exploratory purposes we submitted the four D values for each test-taker to discriminant function analysis in place of the suite of seven raw indices and found the same three functions to emerge with even somewhat more function clarity than our original results reported in this paper, and with much less collinearity. However, the practical outcome revealed similar classification accuracies as the simpler method of assigning people to groups based directly on their lowest D values. This demonstrates that if there is a theoretical need to partial impact first and estimate presence second via discriminant function analysis, the D values from the profile similarity approach can still be used to this end.

Further development of the profile similarity approach is warranted, especially to better distinguish patterns like our Copy-Random manipulation in order to not only detect this pattern more strongly but also to reduce the false-positive rate where some "natural-responders" are misclassified as having copied off on another (random ability) test-taker. One possibility is to add new or different error similarity indices to the profile, or other indices that are more sensitive to random copying. The profile analysis strategy developed in this study places no limit on the number of indices computed on each test-taker and allows for customization and adaptation to detect specific patterns of interest. Future research would benefit from conceptualizing additional profile models for other aberrant patterns of interest and carefully selecting and adapting the suite of indices to detect and differentiate those patterns. This strategy thus allows for theory development and testing of proposed aberrant response models at a broader level than much of the extant research which focuses on statistical detection rates of individual indices under specific conditions.

An open question for this line of research is how generalizable the specific numeric values of the heuristic model profiles in Figure 2 are across a number of potentially relevant factors such as test length, degree of fit to a particular item response model (e.g., the Rasch model, as in the current study), and any other factors that impact the sensitivity of individual statistical indices within the profile. $H_i^T$, for example, is sensitive to the length of the test (Dimitrov & Smith, 2006), the percentage of aberrant item responses (St-Onge et al., 2011), and the degree to which items have sufficiently monotonic and parallel item response functions (Sijtsma & Meijer, 1992), and other indices not used in the current study often include item probabilities from an item response model (e.g., Frary et al., 1977; Wollack, 1997). To what degree will multidimensionality or other causes of poor fit of a test's items to a response model suppress or render inconsistent the values of these statistical indices, and make it more difficult to detect patterns consistent with the profile models? This is not a unique problem to the approach in the current study, and in fact would be an even greater problem if only a single index were used or even one per pattern, so strengths and weaknesses of multiple indices across such conditions could not be balanced out.

For future development of profile models, it would be prudent to ensure that the specific indices selected for the suite are robust across multiple variations in conditions they will be applied in, and that alternative indices be included that are each robust to different sets of conditions, so they can balance each other out. Likewise, it would be prudent to adjust the heuristic model values if future research (e.g., based on simulations) shows different model values to be more optimal. With some further development and fine-tuning of these types of details, the profile similarity approach holds promise in advancing test fraud detection through forensic data analysis.

# 7. References


Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, 69, 44–49. https://doi.org/10.1080/01621459. 1974.10480126

Bellezza, F. S. & Bellezza, S. F. (1989). Detection of cheating on multiple-choice test by using error-similarity analysis. *Teaching of Psychology*, 16, 151–155. https://doi.org/10.1207/s15328023top1603_15

Belov, D. I. (2016). Comparing the performance of eight item pre-knowledge detection statistics. *Applied Psychological Measurement*, 40, 83–97. https://doi.org/10.1177/0146621615603327 PMid: 29881040, PMCid: PMC5982173.

Bird, C. (1927).The detection of cheating in objective examinations. School and Society, 25, 261–262.

Bird, C. (1929). An improved method of detecting cheating in objective examinations. *The Journal of Educational Research*, 19, 341–348. https://doi.org/10.1080/00220671.1929.10879954.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46. https://doi.org/10.1177/001316446002000104.

Dimitrov, D. M. & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied Measurement*, 7, 170–183. PMid: 16632900.

Donlan, T. F. & Fischer, F. E. (1968). An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement*, 28, 105–113. https://doi.org/10.1177/001316446802800110.

Frary, R. B., Tideman, N. & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 2, 235–256. https://doi.org/10.3102/10769986002004235, https://doi.org/10.2307/1164808.

Harnisch, D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 133–146. https://doi.org/10.1111/j.1745-3984.1981.tb00848.x.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16, 277–298. https://doi.org/10.1207/S15324818AME1604_2.

Næs, T. & Mevik, B. (2001). Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, 15, 413–426. https://doi.org/10.1002/cem.676

Sato, T. (1975). The construction and interpretation of S-P tables. Tokyo: Meiji Tokyo.

Saupe, J. L. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement*, 30, 475–489. https://doi. org/10.1177/001316446002000304.

Sijtsma, K. & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's non-parametric IRT model. *Applied Psychological Measurement*, 16, 149–157. https://doi.org/10.1177/014662169201600204.

Sotaridona, L., van der Linden, W. J. & Meijer, R. R. (2006). Detecting answer copying using the kappa statistic. *Applied Psychological Measurement*, 30, 412–431. https://doi.org/10.1177/0146621606288891.

St-Onge, C., Valois, P., Abdous, B. & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement*, 35, 419–432. https://doi.org/10.1177/0146621610391777.

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. https://doi.org/10.1111/jedm.12046.

van der Linden, W. J. & Sotaridona, L. (2004). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement*, 41, 361–377. https://doi.org/10.1111/j.1745-3984.2004.tb01171.x

Weiner, J., Saiar, A. & Granger, E. (2013, October). An Empirical Method for the Detection of Potential Test Fraud. *Presented at the 2nd annual meeting of the Society for the Detection of Potential Test Fraud* in Madison, Wisconsin.

Weiner, J., Saiar, A. & Hurtz, G. (2014, October). Follow-up study of an empirical method for the detection of potential test fraud. *Presented at the Conference on Test Security,* Iowa City, Iowa. https://cete.ku.edu/sites/cete.ku.edu/files/docs/Conference_on_test_security/2014_conference_on_test_security_full_conference_program.pdf.

Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement*, 21, 307–320. https://doi.org/10.1177/01466216970214002.

Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, 19, 265–288 https://doi.org/10.1207/s15324818ame1904_3.

Wollack, J. A., (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement*, 40, 189–205. https://doi.org/10.1111/j.1745-3984.2003.tb01104.x.