Running Head: Cheating Detection

A Review of Methods for Detecting Cheating

William P. Skorupski

University of Kansas

Karla Egan

CTB/McGraw-Hill

Joseph Fitzpatrick

University of Kansas

DRAFT - DO NOT DISTRIBUTE

Note: This paper is under review for journal article acceptance. Please do not distribute.

Abstract

Statistical methods for detecting cheating are reviewed. This includes a discussion of the concept of cheating, and how it has been operationalized differently by researchers. These methods are classified, reviewed, and discussed in terms of their evidence of sensitivity and specificity. A 2 x 2 classification of cheating methods is established and explicated. All methods are described as either (1) parametric or nonparametric, and (2) focused on either aberrance or answer copying to infer cheating. Parametric methods assume a certain statistical model fits the data, while nonparametric methods do not. Aberrance methods determine the extent to which item response patterns do not fit the statistical model or are dissimilar to the group of test takers with similar scores. Answer copying methods determine the extent to which two or more individuals have an unlikely amount of similarity in their response patterns. The review ends with a summary and set of recommendations for cheating detection practice.

DRAFT - DO NOT DISTRIBUTE

Introduction

The possibility of examinee cheating is always a concern for educational testing programs (Cizek, 1999; Thiessen, 2007). Besides the obvious fairness issues involved with having students cheat on their exams, there are additional operational concerns with cheating, such as the misrepresentation of person and item parameters caused by the contamination of item response data (e.g., those items on which cheating has occurred appear easier, which influences the overall scaling and will ultimately bias ability parameter estimates). Cheating is therefore a validity issue that affects not only individual ability estimates by biasing them upwards, but may in fact result in other fairly earned ability estimates being biased downwards. Thus, for any high-stakes operational testing program it is of vital importance to incorporate powerful methods to detect cheating behavior. The purpose of this review is to summarize the most popular of these methods, classifying them based on their assumptions and approaches, as well as their sensitivity and specificity in identifying cheating behavior. The similarity and differences among these methods is also discussed.

Cheating behavior may occur in multiple ways. One popular way to conceptualize and detect cheating behavior is to detect examinees copying from one another. Often, the paradigm for a cheater is someone with relatively low ability who copies answers or otherwise engages in forbidden behavior, seeking to increase his/her test score. Perhaps the cheater tries to sit near a more-able neighbor, and attempts to copy as many answers as possible (in the case of a very low-ability examinee), or perhaps just the answers to the most difficult questions (in the case of, say, an average-ability examinee). Thus, under this paradigm, the expectation is that a cheater's score pattern will contain an uncharacteristically high number of correct answers to relatively difficult

DRAFT - DO NOT DISTRIBUTE

questions. Of course, there is no guarantee that the neighbor actually has a very high ability level, or that his/her response to any one item is correct.

A cheater who is very effective at copying from a very able neighbor will not appear to be aberrant (unless the neighbor appears to be), but might be detectable by copying methods. However, other cheaters may not look at their neighbor's answers, but rather arrive at the testing site carrying prohibited reference materials, or having memorized some illegally exposed test items. A cheater who is very effective at these strategies will simply appear very able. However, a cheater who brings prohibited materials covering only, say, one difficult topic may still end up with a low score on the test, despite answering a small number of very difficult questions correctly. This examinee may be detectable by aberrance methods, but not by copying methods. Ultimately, cheating may be manifested in a number of different ways, so it is probably reasonable to consider multiple approaches to detection (while keeping the possible inflation of Type I error that may come along in mind).

There are a myriad of methods for detecting cheating behavior, and research interest in the topic remains high In 2012, the first Conference on Statistical Detection of Potential Test Fraud included nearly 20 scholarly presentations of methods and evaluations of statistical approaches for detecting cheating; the conference was popular enough to become an annual event. A Handbook of Test Security (Wollack & Fremer, 2013) was recently published, and includes "best practice" recommendations and advice from testing professionals on the design of test security protocol. Clearly, as the number of high-stakes tests increases, so too does the need for effective and cost-efficient ways to detect potential cases of cheating.

Cheating has generally been recognized as a phenomenon of primary concern with multiple choice test questions, so these approaches have all focused on statistics for

DRAFT - DO NOT DISTRIBUTE

dichotomously scored data (though not all of them require it). This review will individually mention over 20 different statistics (with varying levels of detail) that have been suggested and studied over the years. A number of these represent novel approaches (or what were novel approaches when first developed), but still more of these methods represent modifications to procedures that previously existed. Despite the multitude of methods, however, all of these may be classified according to a 2 x 2 organization. All methods reviewed here either: (1) assume a parametric model fits the item response data or use a nonparametric, "group-as-norm" approach to detection, and (2) approach detection based on person-misfit (the "aberrance" of item response patterns) or attempt to detect answer copying behavior. Table 1 demonstrates this organization.

[Insert Table 1 About Here]

Parametric approaches are categorized by first fitting a psychometric model to the item response data. Then, model-based probabilities are used to test hypotheses about aberrance or answer copying. Since multiple choice questions are assumed, these approaches have focused on item response theory (IRT) models which are appropriate for dichotomously scored data. Researchers have primarily used the Rasch model (e.g., Molenaar and Hoijtink, 1990), the three-parameter logistic (3-PL) IRT model (e.g., Levine & Rubin, 1979; Drasgow, Levine, & Williams,1985) or the Nominal Response Model (e.g., Wollack, 1997) to estimate response probabilities for answer options. See van der Linden and Hambleton (1997) for a review of these psychometric models.

The nonparametric, "group-as-norm" approaches generally have critical values (rules of thumb) used to determine 'significance' of aberrance or copying likelihood. The nonparametric methods all operate by determining empirical group norms from the observed data to establish "likely" and "unlikely" response patterns. Guttman (1950) scaling is often considered an

DRAFT - DO NOT DISTRIBUTE

exemplar for a "non-aberrant" response pattern. That is, one can infer aberrance from the presence of so-called "Guttman errors." A Guttman vector is most easily described by ordering all items in a set from easiest to most difficult (as defined by the proportion correct, or "p-value" for each item, with easier items having higher p-values). For dichotomously scored items, a perfect Guttman vector for a raw score of *r* from among *n* test items will consist of *r* 1s followed by *n-r* 0s. The idea is that an examinee who is just capable of answering the *rth* most difficult item correctly is likely to answer all preceding questions correctly, and all subsequent questions incorrectly. While it is not reasonable to assume everyone with a score of *r* will have this identical response pattern, it is nonetheless the expected response pattern because it makes the most sense given the vector of p-values.

There are basically two interrelated ways that researchers have used empirical data to approach estimation of group norms for aberrance detection, and both are influenced by the definition of a Guttman vector (Guttman, 1950). Researchers in the areas of aberrance and cheating detection have nonparametrically determined group norms from observed data by either (1) considering a Guttman vector as the ideal response vector for a given total score, and evaluating departures from it, or (2) determining conditional distributions of item response patterns for each total score (which will tend to be similar to Guttman vectors for relatively reliable tests). Departures from a "perfect" Guttman vector are (after ordering items from easiest to most difficult) defined as any item score of 1 that follows an item score of 0. Lastly, a "reverse" Guttman vector represents the worst fitting response vector (for a total score of *r*, this is represented by 0s for the *n-r* easiest items and 1s for the *r* most difficult questions).

Person-misfit methods have been applied to cheating detection because this behavior has generally been characterized by aberrance in examinees' response patterns (Meijer & Sijtsma,

DRAFT - DO NOT DISTRIBUTE

1995). The particular pattern of aberrance assumed to exist is that relatively low-ability examinees are correctly answering an unexpected number of relatively difficult test questions (but this could also just be due to guessing). There are a number of issues that go along with this definition. For one, examinees who are exceptionally good at cheating are likely to never be detected, because they end up answering nearly every question correctly. As such, they have very high ability estimates, and missing only a few items does not likely appear aberrant. Most of the person-fit measures compare the observed item vector to an expected one. The expected item vector is either determined by a psychometric model (a parametric approach, such as IRT scaling), or empirically derived based on observed data in the sample (Meijer & Sijtsma, 2001).

It should be noted that this review is not comprehensive with respect to measures of person-fit/aberrance (see Meijer & Sijtsma, 2001, for such a review), but rather includes those person-fit measures which have been used most for the detection of cheating. That is, aberrance of an item response pattern is a general concern in psychometric modeling, one which may have multiple causes, only one of which is cheating. Furthermore, the importance of this topic is such that it has motivated a number of psychometricians to create proprietary methods for cheating detection (e.g., Impara, Kingsbury, Maynes, & Fitzgerald, 2005; Maynes, 2009), such as the "data forensics" packages commercially available through Caveon™ Test Security. While some details of how these methods operate are communicated to the public, many of the algorithms themselves are kept private for the purposes of protecting intellectual property (Maynes, 2009). As such, these methods are not reviewed here, as their theoretical bases cannot be fully evaluated.

Besides general methods for detecting person-misfit, other researchers have approached cheating detection by examining the similarity between pairs of examinees' item responses.

DRAFT - DO NOT DISTRIBUTE

Answer copying behavior is evaluated by how similar the answers for two individuals are. As with aberrance detection, there is still a comparison between observed and expected values, except that the observed and expected values are based on measures of similarity. Most of the differences among these answer copying methods come from how "similarity" and "dissimilarity" are defined, and what aspects of the answering process are considered. Some methods look only at the similarity of incorrect responses and omissions, while others consider all item responses (with the possible pitfall that two very able examinees may have very similar response patterns due to answering nearly every question correctly).

Parametric answer copying methods use model-based probabilities to compare observed item response vectors to expected values and test hypotheses about the likelihood of the observed amount of similarity given a null hypothesis of no cheating. Nonparametric methods proceed in much the same manner, except that the probabilities in these approaches are not model based, but empirically derived from observed proportions in the sample.

Review of Methods

Aberrance Detection

Table 2 contains a list of all methods reviewed here which have approached cheating detection through the definition of person-misfit or aberrance. These methods are categorized as nonparametric or parametric and reviewed in the following two sections.

[Insert Table 2 About Here]

Nonparametric Approaches. Among the earliest of approaches to detecting aberrance of examinee response patterns was Sato's (1975) Caution index. The Caution index, *C*, is an overall measure of person misfit, defined by the formula:

$$C = 1 - \frac{\text{cov}(X, p)}{\text{cov}(X^*, p)},$$

DRAFT - DO NOT DISTRIBUTE

where X is the observed response vector for a given examinee with *r* correct answers to *n* items, X* is a Guttman response vector for *r* of the *n* items answered correctly, and *p* is a vector of p-values (proportion correct) for the *n* items. Sato notes that *C* will be equal to zero for an examinee with a Guttman vector and suggests that values larger than 0.5 are indicative of examinees with aberrant response vectors.

Harnisch and Rubin (1981) noted that for the caution index, values close to zero clearly indicate person-fit (i.e., when an examinee's response vector is close to a "perfect" Guttman vector), but that the index has no upper limit. As a result, it is not immediately clear how to interpret relatively large versus trivial values for *C*. Their modified caution index, *C**, corrects for this by forcing the statistic in between zero and one, as follows:

$$C^* = \frac{\mathrm{cov}(X^*, p) - \mathrm{cov}(X, p)}{\mathrm{cov}(X^*, p) - \mathrm{cov}(X', p)},$$

where X' is a reverse Guttman vector (a string of item responses, *r* of which are correct, where the correct responses are for the *r* most difficult items, leaving the *n*-*r* easiest items incorrect). *C** is equal to zero when an examinee's response vector is a Guttman vector and one when an examinee's response vector is a reverse Guttman vector. Harnisch and Rubin (1981) recommend using values of *C** of 0.3 and greater as suggestive of aberrance. Tatsuoka & Tatsuoka (1982, 1983) recommended another modification to *C*, called the Norm Conformity Index (*NCI*), with a similar approach but different scale. Tatsuoka and Tatsuoka scaled the *NCI* such that values equal to -1 indicate an examinee with a Guttman vector of responses, and values equal to 1 indicate a reverse Guttman vector.

Kane and Brennan (1980) proposed the use of three related statistics to detect aberrance of examinee item responses, "Agreement" (*A*), "Disagreement" (*D*), and "Dependability" (*E*).

DRAFT - DO NOT DISTRIBUTE

For the j^{th} examinee, A_j is equal to the sum of the n scored item responses, each weighted by its p-value:

$$A_j = \sum_{i=1}^{n} X_{ij} p_i$$

Thus the Agreement index is equal to the sum of p-values for items answered correctly by an examinee. Disagreement for an individual, D_j, is defined as the difference between the maximum value A can obtain (for a given total score, r, this is the sum of the p-values for the r easiest items) and the observed value of A_j. Dependability, E_j, is the ratio of A_j,/ D_j. The rescaling in calculating E_j ensures that it will always be a value in between zero and one, with one indicating a perfect Guttman vector, and zero indicating a reverse Guttman vector.

Meijer (1994), in a comparison of multiple person-fit statistics, utilized as one of them the simple number of Guttman errors an examinee made as an indicator of misfit (and therefore, perhaps, cheating behavior). A Guttman error is defined as a correct response to any item more difficult (as measured by p-value) than the easiest item answered incorrectly. Meijer found that using the number of Guttman errors as a measure of person misfit was as powerful as more complicated statistics and relatively uncorrelated with the number-correct score.

Parametric Approaches. Most approaches to person-misfit detection in the literature are based on the idea of fitting a parametric psychometric model to the observed item response data, estimating an ability level for each examinee, and then evaluating the value of the likelihood function for each individual, given the ability estimate. This approach was first introduced by Levine and Rubin (1979) when they proposed the ℓ_0 statistic, based on the 3-PL IRT model (though this could be used for any IRT model). The ℓ_0 statistic is simply the natural log of the likelihood function from the IRT model:

DRAFT - DO NOT DISTRIBUTE

$$\ell_0 = \sum_{i=1}^{n} \{X_i \ln[P_i(\theta)] + (1 - X_i) \ln[1 - P_i(\theta)]\}.$$

Once an estimate of examinee ability, $\hat{\theta}$, has been determined, its value is used in place of θ in the log-likelihood function above to determine to the chances of the observed item response pattern, given the ability estimate. When item responses are close to expected values (i.e., a score of "1" is associated with a relatively high probability, and a score of "0" is associated with a relatively low probability), the ℓ_0 statistic approaches its maximum value of zero. When a number of observed item responses are incongruent with expectation (as when a relatively low ability candidate answers a very difficult question correctly), the likelihood function for that individual becomes increasingly flat and the ℓ_0 statistic becomes increasingly negative.

While this approach to estimating person-fit is model-based and sensible, it has limitations. The ℓ_0 statistic is not standardized, so its value will depend, perhaps in large part, on $\hat{\theta}$. The null distribution of ℓ_0 is not known, so it is unclear how far from zero ℓ_0 should be before classifying a response pattern as "aberrant." To address these concerns, Drasgow, Levine, and Williams (1985) proposed the ℓ_z statistic, a standardized version of the ℓ_0 statistic, which corrects ℓ_0 for its expected value and standard deviation across independent replications:

$$\ell_z = \frac{\ell_0 - E(\ell_0)}{\sigma_{\ell_0}},$$

where

$$E(\ell_0) = \sum_{i=1}^{n} \{P_i(\theta) \ln[P_i(\theta)] + [1 - P_i(\theta)] \ln[1 - P_i(\theta)]\},$$

and

$$\sigma_{\ell_0} = \sqrt{\sum_{i=1}^{n} P_i(\theta)[1 - P_i(\theta)] \left[\ln \frac{P_i(\theta)}{1 - P_i(\theta)}\right]^2}.$$

DRAFT - DO NOT DISTRIBUTE

Drasgow, et al (1985) posited the ℓ_z statistic as approximately standard normal, thus suitable for hypothesis testing with regard to person misfit. When based on the Rasch model, l_z generally outperformed other person-fit indices in terms of rates of detection and false positives (Li & Olejnik, 1997).

Despite the improvements upon ℓ_0 , the ℓ_z statistic is not without limitations. Since ℓ_z was initially purported to asymptotically follow a standard normal distribution (Drasgow et al., 1985), ℓ_z was interpreted like a *z*-score, with decisions regarding the relative aberrance of response patterns based on values from a *z*-score table, given a desired Type I error rate (e.g., ±1.96 for an α level of 0.05). However, a number of researchers have since demonstrated that ℓ_z is standard normal only when using θ in its estimation. When substituting $\hat{\theta}$ for using θ (which is always necessary in practice with real data), the ℓ_z statistic may not be normal (Nering, 1995; Van Krimpen-Stoop & Meijer, 1999) and its variance may be underestimated (Snijders, 2001). Even when using θ in estimation, the guessing parameter of the 3-PL IRT model reduces the power of ℓ_z to detect aberrance at the low end of the ability continuum (Reise & Due, 1991). In addition, ℓ_z was shown to have little power to detect aberrant response patterns caused by item memorization, particularly in a computerized adaptive environment in which the range of item difficulties may be limited (McLeod & Lewis, 1999).

A number of modifications for the ℓ_0 and ℓ_z statistics have been developed to address some of these limitations. The *U3* statistic (Van der Flier, 1980, 1982; Mokken & Lewis, 1982) is the nonparametric version of ℓ_0 and ℓ_z (it is listed in Table 3 with other nonparametric approaches to aberrance detection, but it is mentioned here because its form is so similar to the other likelihood statistics). For *U3*, the likelihood of a response pattern is conditioned not on a latent trait estimate, like $\hat{\theta}$, but on total score. Van der Flier (1982) and Mokken and Lewis

DRAFT - DO NOT DISTRIBUTE

(1982) have demonstrated that the null distribution of *U3* is approximately standard normal, and thus may be used in hypothesis testing of aberrant behavior. Simulation results from St-Onge, Valois, Abdous, and Germain (2011) show that both ℓ_z and *U3* are more effective at detecting spuriously low scores (e.g., due to fatigue or unlucky guessing) than spuriously high scores (due, presumably, to cheating). Molenaar and Hoijtink (1990) proposed the *M* statistic, a simplified version of ℓ_z based on the Rasch model, with properties similar to those purported for ℓ_z with multi-parameter IRT models. That is, Molennaar and Hoijtink's (1990) *M* statistic follows a standard normal distribution, as long as the Rasch model fits the data. More recent research on the ℓ_z statistic with multi-parameter IRT models has suggested employing corrections to $\hat{\theta}$ when used in place of θ (de la Torre & Deng, 2008). For their corrected ℓ_z statistic, de la Torre and Deng (2008) present a version of ℓ_z which corrects the ability estimate for unreliability (using the standard error of $\hat{\theta}$) and generating, through Monte Carlo simulation, a reference distribution suitable for hypothesis testing.

Cumulative sum (CUSUM) statistics are another newer approach, specific to computer-based testing designs (including computerized adaptive testing (CAT)), that allow for detection of aberrance on-the-fly, that is, while a test is being taken by an examinee. These parametric CUSUM statistics were motivated by the fact that more traditional aberrancy measures are not accurate in a CAT environment (van Krimpen-Stoop & Meijer, 2000). CUSUM statistics (van Krimpen-Stoop & Meijer, 2000; Meijer, 2002; Armstrong & Shi, 2009a) accumulate a person-fit measure sequentially during the exam, thus allowing item sequence to affect results. This, in turn, allows psychometricians to monitor trends within item response patterns to determine if certain sub-strings of responses are demonstrating aberrancy. Several CUSUM statistics outperformed ℓ_z in detecting aberrance over longer sequences of aberrant item responses, but the

DRAFT - DO NOT DISTRIBUTE

advantage was lost when the sequences of aberrant responses are short (Tendeiro & Meijer, 2012). Armstrong and Shi (2009b) also introduced a modified version of the van Krimpen-Stoop and Meijer (2000) CUSUM statistic conditioned on number correct (NC) scores rather than ability. (This "model-free" version is listed in the Table 3, but is included here because of its similarity in form to the parametric CUSUM approaches.) Their non-parametric version, $CUSUM_{LR}$, is a likelihood ratio test comparing upward and downward aberrant shifts in responses over sequences of items. In a simulation study, $CUSUM_{LR}$ outperformed other non-parametric approaches, but did not detect aberrant behavior as well as the parametric CUSUM statistic (Armstrong & Shi, 2009b).

A Bayesian approach to aberrance detection was proposed by McLeod, Lewis, and Thissen (2003) to identify examinees who have memorized items on a computerized adaptive test. The log odds that an examinee is using item preknowledge is calculated by incorporating a modified 3-PL IRT model in which the guessing parameter, *c*, is replaced by *c'*, a combined "guessing plus item preknowledge" parameter. The final index is the base 10 log of the ratio between odds of preknowledge after *n* items have been administered and the prior odds. A value of 0, therefore, indicates that the probability of using item preknowledge has not changed after *n* items. The success of this method depends on the prior probability that an item has been memorized, and the index performed best when the prior was calculated empirically for each item using simulated results. Thus, the method is not intended to identify memorization of random items but rather to identify the use of preknowledge on specific items administered during an adaptive test.

Shu (2010) and Shu, Henson, and Luecht (2013) introduced another method, the Deterministic, Gated Item Response Theory Model (DGM), to detect aberrance due to item

DRAFT - DO NOT DISTRIBUTE

exposure. By first classifying items as either compromised or secure (based on previous exposure), examinee performance can be decomposed into responses due to true ability and responses due to cheating ability. True ability is estimated using only the secure items, and cheating ability is estimated using only the exposed items. Under the null hypothesis that an examinee has not cheated, the expected difference between these two abilities is 0. Differences are compared to an empirically selected cut point to flag potential cheaters. Compared with ℓ_z, DGM was better able to detect simulated cheaters when a large proportion of the examinees were cheating. However, because examinees with lower true ability have more "space" to show improvement on exposed items, the DGM is less sensitive to cheaters with relatively high true ability.

Another recent person-fit measure (Clark, 2012) proposed specifically for the detection of cheating treats cheating behavior as multidimensionality. The difference between person-fit statistics is therefore expected to change across nested factor analytic models. Clark uses the one-factor *lco* (Ferrando, 2007) and two-factor *M-lco* (Ferrando, 2009) person fit statistics, and then calculates the "*lco* difference," which is hypothesized to be χ^2 distributed with one degree of freedom. A significant difference indicates significant improvement in fit using the additional factor, which is assumed to represent cheating.

Another Bayesian method (Skorupski & Egan, 2011) was designed to detect cheating at the group level rather than individual cheaters. Using a Bayesian Hierarchical Linear Model (HLM), Skorupski and Egan flag potentially cheating schools by modeling the change over time in individual scores nested within groups (schools). An unusually large group-by-time interaction is treated as evidence of potential cheating. This method has shown promise with both real and simulated data sets, although its success may depend on using baseline group means that are not

DRAFT - DO NOT DISTRIBUTE

influenced by aberrance (Skorupski & Egan, 2012). The method differs from others reviewed here in that cheating is defined as groups of individuals who are unfairly prepared for an upcoming exam. Cannell (1989) described how an overwhelming majority of school districts demonstrated year-to-year improvements in their standardized test score over a period of 15 years, despite the fact that other indicators of achievement did not reflect these gains. The possibility of group-level cheating being a cause of such phenomena is certainly a concern, and there may be a grey area between honest "test preparation" and inappropriate coaching or outright cheating.

A somewhat different approach to aberrance detection that has received recent attention is the analysis of answer changes or erasures. For paper-and-pencil tests, scanners can detect answers that have been erased, and computer-based tests can record initial and subsequent responses to each item. Rather than detecting unlikely response patterns, answer-change analysis methods look for aberrant patterns of answer-changes. For example, van der Linden and Jeon (2012) presented a parametric method designed to detect unexpectedly large numbers of wrong-to-right (WR) answer changes. They first fit a psychometric model (such as the 3-PL IRT model) to the set of initial responses. Then, the same model is fitted to the final responses for each item except that the ability parameters are fixed to their initial estimated values, and the guessing parameter (in the case of the 3-PL) is fixed at 0. By leaving the other item parameters free to vary, van der Linden and Jeon calculate the probability of a WR erasure for a given examinee on a given item. The number of WR erasures follows a generalized binomial distribution, and may be compared with a critical value to flag potential cheaters. Because a large amount of data may be "missing" (i.e., an unchanged answer, or a RW or WW erasure), the method relies on

DRAFT - DO NOT DISTRIBUTE

Bayesian estimation with weakly informative priors. A key assumption of this method is that all examinees have adequate time to review every item on the test (van der Linden & Jeon, 2012).

Answer Copying Detection

Table 3 contains a list of all methods reviewed here which have approached cheating detection through the definition of answer copying. These methods are categorized as nonparametric or parametric and reviewed in the following two sections.

Nonparametric Approaches. Among the first published methods used to identify cheaters was that presented by Angoff in 1974. A combination of logical and empirical evidence is presented to derive satisfactory methods for detecting answer copying. Indices *B* and *H* are deemed the most useful (of the 8 statistics, *A* through *H*, proposed and evaluated in the paper) for cheating detection, primarily due to power considerations (Type I error rates were not explicitly tested in the 1974 paper, merely assumed under the auspices of the *t*-test). The values for these statistics were calculated for every pair of students in the norm group. Then the value was calculated for a suspected cheater and compared, via a one-sample *t*-test, to the mean for that statistic derived in the norm group. To ensure a very conservative Type I error rate, only those examinees with *t*-values greater than or equal to 3 were flagged as cheaters.

Bivariate reference distributions were created for each statistic using the norm group, such that conditional values for each quantity were determined and used as reference points for "typical" levels of agreement expected among non-cheating examinees. That is, each statistic was referenced by a conditional distribution, controlling for some quantity of interest. Index *B* was defined as the number of items that examinees *i* and *j* answer incorrectly in the same way (exact incorrect matches), controlling for the product of the total number of incorrect responses for examinee *i* and examinee *j*. Index *H* was defined as the longest run of identically marked incorrect or omitted responses shared by examinees *i* and *j*, after controlling for the number of

DRAFT - DO NOT DISTRIBUTE

incorrect and omitted responses for either examinee *i* or *j*, whichever was smaller. Angoff (1974) describes that known cheaters have been operationally identified by applying Index *B* and then *H*. That is, Index *H* was calculated only for those examinees not identified as cheating by Index *B*. If both indices fail to identify cheating, it is assumed the examinee has not copied. Those flagged by Index *B* or *H* were then required to follow security procedures for re-testing to determine the validity of the original score (Angoff, 1974). It appears that these indices were used operationally for a number of years until researchers at the Educational Testing Service developed additional statistical criteria (e.g., Holland, 1996).

Building on the work of Angoff (1974), Holland (1996) developed the "K-index" for the detection of answer copying. The K-index is based on the premise that matching incorrect responses for two test takers working independently will have a binomial sampling distribution:

$$K = \sum_{i=WM}^{WB} \binom{WB}{i} P^i (1-P)^{WB-i},$$

where WM is the number of matching incorrect responses for two test takers (A and B), and WA and WB are the numbers of incorrect responses for test takers A and B, respectively, under the definition that WA $\geq$ WB, and thus WM $\leq$ WB. The formula for the "success" probability (the probability of matching incorrect responses), *P*, for the K-index is a continuously piecewise linear function of WA/N, the proportion of incorrect responses for test taker A. It has an intercept, *a*, and a slope, *b*, that control how the probability changes with WA/N. Its formula is:

$$P = \begin{Bmatrix} a + b(WA/N), \text{ if } 0 < (WA/N) \leq .3, \\ [a + b(.3)] + (.4)b[(WA/N) - 0.3], \text{if } .3 < (WA/N) \leq 1 \end{Bmatrix}$$

Holland (1996) and Lewis and Thayer (1998) demonstrate that in practice, values for *a* and *b* must be specified for every test with which the K-index is used. (In the 1996 paper, Holland uses 0.085 for *a* and 0.5 for *b*, though the choice of these values seems to be very sample dependent).

DRAFT - DO NOT DISTRIBUTE

These values are to be chosen empirically to maximize agreement probabilities. Lewis and Thayer (1998) further demonstrate a slight modification to the K-index, called the Probability of Matching Incorrect Responses (PMIR). The PMIR takes advantage of the well-known Bonferroni inequality by multiplying *K* by the number of "sources" being considered for a potential cheater. This modification is intended to control the Type I error rate for each subject.

Building on the work of Holland (1996) and Lewis and Thayer (1998), Sotaridona and Meijer (2002, 2003) developed a series of statistics to improve the detection of answer copying. They introduce the $\overline{K}_2$ index in their 2002 paper, and the S_1 and S_2 indices in their 2003 paper. Like the K-index, $\overline{K}_2$ is based on the number of exactly matching incorrect responses given by a pair of examinees. As with the K-index and PMIR, this quantity is posited to follow a binomial distribution with "success" probability *P*. The difference introduced is how *P* is estimated for the $\overline{K}_2$ index. For the K-index and PMIR, *P* is empirically determined by the average number of matching incorrect answers divided by the number of wrong answers given by the source. For the $\overline{K}_2$ index, *P* is estimated from a quadratic regression model ($P = \beta_0 + \beta_1 Q_r + \beta_2 Q_r^2 + \varepsilon_r$), where Q_r is the proportion of wrong answers for examinees with *r* incorrect answers. Sotaridona and Meijer (2002) showed that this modification yielded acceptable Type I error rates while producing higher detection rates than the K-index, though detection rates were still higher for the ω index (Wollack, 1997, discussed later in this review).

In 2003, Sotaridona and Meijer further advanced this area of research by introducing two additional answer copying detection statistics, the S_1 and S_2 indices. As with the K-index, PMIR, and $\overline{K}_2$ index, S_1 and S_2 are based on exactly matching item response patterns between pairs of examinees. However, while the S_1 is based only on exactly matching incorrect responses (like K,

DRAFT - DO NOT DISTRIBUTE

PMIR, and $\overline{K}_2$), the S_2 also incorporates information about matching correct answers, and is therefore more sensitive. For S_1, the number of exact incorrect matches is not assumed to follow a binomial distribution, but a Poisson distribution. The Poisson parameter, μ, is estimated by a loglinear model based on the number of incorrect answers, *r*. S_2 similarly incorporates the Poisson density function, but to represent the number of exactly matching incorrect answers plus a weighted sum of matching correct answers. The weight of each pair of matching correct answers, δ, is inversely related to the probability of the copier answering the question correctly, with the constraint: $1 \geq \delta \geq 0$ (see Sotaridona & Meijer (2003) for additional estimation details). Thus, suspected copiers with a low probability of correctly answering a question who answer correctly have a higher suspicion of cheating. Sotaridona and Meijer (2003), in their study, found that the K-index, $\overline{K}_2$, S_1 and S_2 all demonstrated acceptable Type I error rates, with S_1 having higher power than $\overline{K}_2$, and S_2 demonstrating higher power than K and $\overline{K}_2$.

In 2004, van der Linden and Sotaridona presented a statistical test to detect answer copying on multiple choice tests that can be used without reference to the examinee population. The test statistic, γ_{js}, the unknown number of matching incorrect item responses examinee *j* copied from examinee *s*, is evaluated without reference to examinee total scores or their distribution. The only assumption made is about the response behavior of the examinees suspected of copying. The approach is classified here as nonparametric, as no IRT model is assumed to fit the data or otherwise explain the response probabilities with which the "source" examinees answered questions. The likelihood of the source and copier examinees sharing γ_{js} matching incorrect item responses is modeled by a family of "shifted binomial" distributions. Power functions for several sets of parameter values can be found in van der Linden and Sotaridona (2004).

DRAFT - DO NOT DISTRIBUTE

Two related statistics were developed by van der Ark, Emons, and Sijtsma (2008). These authors developed a small-sample approach to cheating detection that allows for the use of alternate test forms containing common and unique items and seat location information for detecting answer copying. Their τ_1 statistic measures a normed count of suspicious "pair-scores" (matching answers options, that may or may not both be correct/incorrect, because alternate forms were administered) and τ_2 provides a suggestion of who is copier and who is source, based on estimates from other parts of the test. These indices were developed specifically to deal with small sample situations (N = 230 in their study) where some parametric methods may be untenable.

Belov and colleagues proposed two additional methods for tests with both a fixed and variable part. The first is "Algorithm 1" (Belov & Armstrong, 2010), designed to reduce the Type I error rate of the K-index by first screening examinee pairs using the Kullback-Leibler divergence (KL; Kullback & Leibler, 1951) between the two parts of the test. Examinees with KL values higher than an empirically derived critical value are flagged, and the K-index is computed between flagged examinees and unflagged examinees who took the test on the same date at the same test center and had different variable parts. This algorithm is very conservative, with a near-zero Type I error rate for both simulated and real data (Belov & Armstrong, 2010). The second approach is the Variable Match Index (VM-Index; Belov, 2011). The VM-Index was designed to detect both "blind copying," in which the copier, *c*, and source, *s*, have matching responses but different variable parts of the test, and "shift copying," in which *c* copies answers from *s* but from the next or previous item rather than the current item. Simulation results show the VM-Index has lower Type I and Type II error rates than the K-index and that it is better able to detect copying from a high-ability source (Belov, 2011).

DRAFT - DO NOT DISTRIBUTE

Parametric Approaches. Parametric approaches to the detection of answer copying are relatively recent in the literature. These recent approaches utilize a parametric IRT model in conjunction with a hypothesis testing framework to compare observed and expected amounts of agreement between examinees. One early, nonparametric approach to detecting answer copying was the g_2 statistic presented by Frary, Tideman, & Watts (1977), though this approach was not widely used or studied. Twenty years later, Wollack (1997) presented a closely related, though updated, version of Frary et al's (1977) work, statistic ω. The similarity of these methods (in terms of their hypothesis testing framework) warrants their simultaneous description, though Frary et al's approach is nonparametric (as indicated by its placement in Table 3).

In this notation, examinee *c*, the copier, is suspected of copying answers from examinee *s*, the "source." Both g_2 and ω compute a standardized difference between the number of answer matches between a pair of examinees, h_{cs}, and the number predicted by chance alone, conditioned on *s*'s answers, *c*'s ability level, and the properties of the item:

$$\frac{h_{cs} - \sum_{i=1}^{n} P_c(u_{is})}{\sigma_{h_{cs}}} \sim N(0,1),$$

where $P_c(u_{is})$ is the probability of examinee *c* selecting the answer provided by *s* to item *i*, and

$$\sigma_{h_{cs}} = \sqrt{\sum_{i=1}^{n} \left[P_c(u_{is})\right]\left[1 - P_c(u_{is})\right]}.$$

The only difference between g_2 and ω is with regard to how $P_c(u_{is})$ is estimated. For g_2, $P_c(u_{is})$ is determined by the proportion of examinees choosing each response option to item *i* and the ratio of *c*'s total score to the mean total score. For ω, $P_c(u_{is})$ is estimated from the nominal response model (as previously mentioned, see van der Linden & Hambleton, 1997 for a description), which can be used to explicitly model the response probability for every response option on

DRAFT - DO NOT DISTRIBUTE

multiple choice items. It should be noted that a large distinction between g_2 and ω and other agreement indices (e.g., the K-index) is that most other approaches focus only on matching incorrect response options, while g_2 and ω incorporate matching information for correct and incorrect responses. While g_2 has not received much attention in the literature on cheating, since its inception, Wollack's ω statistic has become the benchmark by which other cheating detection methods are compared (e.g., Sotaridona & Meijer, 2002; Sotaridona & Meijer, 2003; Wollack, 2003; van der Linden & Sotaridona, 2006; van der Linden, 2009).

More recently, van der Linden and colleagues (van der Linden & Sotaridona, 2006; van der Linden & Guo, 2008; van der Linden, 2009) have presented a series of related parametric approaches to conducting hypothesis tests about cheating behavior. As pointed out by van der Linden and Sotaridona (2006), the extant approaches for the detection of answer copying on multiple choice questions are all based on a statistic that in some way counts the number of agreements between pair of examinees, though they may differ with regard to which items may count towards this sum (for example, the K-index only counts exact agreements for incorrect responses, while ω considers all agreements, including omissions). An additional point of differentiation is whether or not the statistic is standardized. The parametric approaches reviewed here are, while many of the nonparametric approaches are not. Lastly, these agreement indices may differ with regard to the null distribution the test statistic is posited to have. This last point is what distinguishes many of the more recent approaches to answer copying detection.

In 2006, van der Linden and Sotaridona developed alternative approaches to Wollack's (1997) ω for hypothesis testing about matching item responses, under the assumption of a known IRT model fitting the data. (Conversely, when a response model is assumed but unknown, Sotaridona, van der Linden, and Meijer (2006) propose a test of answer copying based on

DRAFT - DO NOT DISTRIBUTE

Cohen's κ statistic.) They present an exact conditional hypothesis test for answer copying which has a generalized binomial for its null distribution. The authors point out that the asymptotic normality assumption made for hypothesis testing with the ω statistic will likely hold only for long tests, and that for shorter tests, the normal approximation may be problematic, especially if the generalized binomial distribution is skewed (as it is likely to be when examinees *j* and *s* have very different ability levels). Thus, ω should asymptotically approach γ_{js} (the number of matching item responses examinee *j* copied from examinee *s*) as test length increases. Two versions are offered, unconditional agreement (whereby source and cheater cannot be distinguished) and a conditional version assuming one examinee as source, with probabilities of matching conditional on that examinee's response vector. Simulation results show that the conditional version of the generalized binomial test and ω are about equally effective (Zopluogu & Davenport, 2012), with their power depending in large part upon the number of answers copied and the ability level of the source.

Lastly, a recent approach (van der Linden, 2009) to answer copying detection is based on the additional information available when computer-based testing is employed (as was also the case for the CUSUM statistics presented by van Krimpen-Stoop & Meijer, 2000; Meijer, 2002; Armstrong & Shi, 2009). In his paper, van der Linden (2009) introduces a bivariate lognormal response-time model which may be used to detect collusion (examinees working together to share answers). This approach is especially novel because, unlike any other approach, it does not quantify agreement based on matched item responses, but rather matched response times. Thus, the issue of whether to include correct matching item responses with the matched incorrect responses is completely avoided. Collusion may be inferred by the presence of unusual agreement in response times to the same items. Two parameters from this bivariate model are of primary

DRAFT - DO NOT DISTRIBUTE

interest, τ_j^* and ρ_{jk}, the average speed of labor by examinee *j* on item *i* and the correlation between response times for examinee *j* and *k*, respectively. Since the model includes parameters for the effects of item characteristics on response times (i.e., parametrically how much time is required to answer each item), the correlation parameters automatically correct for the fact that any two random examinees may have correlated response times. That is, because some items require more time to answer than others, it is expected that response times will correlate. The approach presented by van der Linden (2009) determines if a correlation is present after accounting for the item effect on response time.

Summary

This review clearly demonstrates that there is a great availability of methods for detecting cheating, whether that is defined by answer copying behavior or a more general approach to detecting aberrance of item response patterns. Of the nearly 30 methods reviewed here, all are classified as either (1) parametric or nonparametric, and (2) based on person-misfit or answer-copying behaviors. While some comparisons of the performance of these methods in certain situations have been conducted, it is strictly speaking difficult to identify one or more of these methods as clearly superior, as no study to date has comprehensively compared their performance. That being said, a few general guidelines can be reasonably inferred. For one, the choice of a parametric versus nonparametric approach should be based on the scoring model used for scaling item response data. If, for an example, an IRT model is being fit to the data to determine ability estimates (and, importantly, there is adequate model-data fit present), then it would be recommended to approach cheating detection using a parametric approach. For those testing programs without adequate sample sizes – or, perhaps, desire – to employ IRT scaling, nonparametric approaches would make more sense.

DRAFT - DO NOT DISTRIBUTE

The decision to choose an approach based on aberrance versus answer copying may not be as simple as it sounds. Person-misfit is clearly a topic which is broader than simply cheating behavior, but powerful methods that can detect aberrance will likely also identify certain types of cheating. However, as previously stated, if a low ability examinee copies answers from a high ability examinee, and does so very effectively, that examinee's response pattern will not appear aberrant at all; it will simply produce a very high (obviously positively biased) ability estimate. In this situation, only an answer copying procedure could potentially identify the problem behavior. But conversely, there are more ways to cheat than just copying answers. Cheaters who bring prohibited outside materials to the test session, or who may have memorized the answers to a few, very difficult, and previously exposed test questions, will likely answer difficult questions correctly at a much higher than expected rate. This is precisely the kind of cheating behavior that an aberrance index could potentially identify, that an answer copying index could not.

It may be beneficial for us to consider cheating as a kind of multidimensionality, in contrast to the typical assumption of most psychometric models, namely, that a unidimensional construct is solely responsible for all systematic test and item score variance. That is, the behavior of test takers is likely governed by several abilities simultaneously, and cheating behavior and/or aberrance may be a manifestation of those. A necessary venture to further develop the science of cheating detection is to understand the nature of the cheating process itself. How do cheaters cheat, and how can we reliably detect it? Aberrant score patterns can occur for a variety of reasons, and answer copying alone is clearly not sufficient to describe all cheating behaviors.

With these issues in mind, it is recommended that operational testing programs consider multiple indices simultaneously and employ at least one aberrance and answer copying index.

DRAFT - DO NOT DISTRIBUTE

Different kinds of cheaters will likely be detected by these different methods, so both kinds have merit. That being said, it is still a wise policy to protect the non-cheaters from being flagged and harassed simply because they answered questions in an aberrant (if also honest) manner. Examinees, after all, are not familiar with what IRT models expect of them, and are rarely aware that their item response patterns are reverse Guttman vectors! If multiple methods are to be employed, corrections for likely-to-be-inflated Type I error rates must also be considered.

References

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

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

Armstrong, R. D., & Shi, M. (2009b). Model-free CUSUM methods for person fit. *Journal of Educational Measurement*, *46*(4), 408-428.

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Applied Psychological Measurement, 34*(6), 379-392. doi: 10.1177/0146621610370453

Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests: The "Lake Wobegon" report.* Albuquerque, NM: Friends for Education.

Cizek, G. (1999). Cheating on Tests: How to Do It, Detect It, and Prevent It. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Clark III, J. M. (2012). *Nested factor analytic model comparison as a means to detect aberrant response patterns*. Paper presented at the 1st Annual Statistical Detection of Potential Test Fraud Conference, Lawrence, KS.

DRAFT - DO NOT DISTRIBUTE

de la Torre, J. & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45,* 159 – 177.

Drasgow, F., Levine, M.V., Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67 – 86.

Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research, 42,* 481-507.

Ferrando, P. J. (2009). Multidimensional factor-analysis-based procedures for assessing scalability in personality measurement. *Structural Equation Modeling, 16,* 109-133.

Frary, R. B, Tideman, T. N. & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 2,* 235 – 256.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60 – 90). Princeton, NJ: Princeton University Press.

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support.* ETS Research Report, RR-96-7.

Impara, J. C., Kingsbury, G., Maynes, D., & Fitzgerald, C. (2005). *Detecting cheating in computer adaptive tests using data forensics*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, CA.

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

DRAFT - DO NOT DISTRIBUTE

Lewis, C. & Thayer, D. T. (1998). *The power of the K-index (or PMIR) to detect copying.* ETS Research Report, RR-98-49.

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

Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*(3), 215-231. doi: 10.1177/01466216970213002

Maynes, D. (2009). Combing statistical evidence for increased power in detecting cheating. Caveon Test Security. Retrieved online, 5/15/2010, from: http://74.220.207.132/~caveonco/articles/Combining_Statistical_Evidence_for_Increased_Power_in_Detecting_Cheating_2009_Apr_04.pdf

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23*(2), 147-160. doi: 10.1177/01466219922031275

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27*(2), 121-137.

Meijer, R. R. & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8,* 261 – 272.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107 – 135.

Mokken, R. J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417 – 430.

DRAFT - DO NOT DISTRIBUTE

Molenaar, I. W. & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55,* 75 – 106.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19,* 121 – 129.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*(3), 217-226. doi: 10.1177/014662169101500301

Shu, Z. (2010). *Detecting test cheating using a deterministic, gated item response theory model.* (Doctoral Dissertation), The University of North Carolina at Greensboro, ProQuest Dissertations and Theses. Retrieved from http://search.proquest.com/docview/845237696?accountid=14556

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory Model to detect test cheating due to item compromise. *Psychometrika*. doi: 10.1007/S11336-012-9311-3

Skorupski, W. P. & Egan, K. (2011). *Detecting cheating through the use of hierarchical growth models*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Skorupski, W.P. & Egan, K. (2012). *A hierarchical linear modeling approach for detecting cheating and aberrance.* Paper presented at the 1st Annual Statistical Detection of Potential Test Fraud Conference, Lawrence, KS.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66,* 331 – 342.

DRAFT - DO NOT DISTRIBUTE

Sotaridona, L. S. & Meijer, R. R. (2002). Statistical properties of the *K*-Index for detecting answer copying. *Journal of Educational Measurement, 39,* 115 – 132.

Sotaridona, L. S. & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40,* 53 – 69.

Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2006). Detecting answer copying using the kappa statistic. *Applied Psychological Measurement, 30*(5), 412-431. doi: 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*(6), 419-432. doi: 10.1177/0146621610391777

Tatsuoka, K. K. & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7,* 215 – 231.

Tatsuoka, K. K. & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20,* 221 – 230.

Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to Detect Person Misfit A Discussion and Some Alternatives for Existing Procedures. *Applied Psychological Measurement, 36*(5), 420-442. doi: 10.1177/0146621612446305

Thiessen, B. (2007). Case study—policies to address educator cheating. Retrieved from: http://homepage.mac.com/bradthiessen/pubs/format.pdf

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26,* 199 – 218.

DRAFT - DO NOT DISTRIBUTE

van der Ark, L. A., Emons, W. H. M., & Sijtsma, K. (2008). Detecting answer copying using alternate test forms and seat locations in small-scale examinations. *Journal of Educational Measurement, 45,* 99 – 117.

Van der Flier, H. (1980). Vergelijkbaarheid van individuele testprestaties [Comparability of individual test performance]. Lisse: Swets & Zeitlinger.

Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13,* 267 – 298.

van der Linden, W. J., & Hambleton, R. K. (1997). Handbook of modern item response theory. NY: Springer-Verlag.

van der Linden, W. J. & Sotaridona, L. S. (2004). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement, 41,* 361 – 377.

van der Linden, W. J. & Sotaridona, L. S. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31,* 283 – 304.

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

van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics, 34,* 378 – 394.

van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics, 37*(1), 180-199. doi: 10.3102/1076998610396899

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

DRAFT - DO NOT DISTRIBUTE

Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement, 40,* 189 – 205.

DRAFT - DO NOT DISTRIBUTE

Table 1. Organization structure of cheating detection methods reviewed in this paper.

	Aberrance/Person-misfit	**Answer Copying**
Nonparametric	• Group norms establish "likely" response patterns • Observed item responses compared to "expected"	• Group norms establish "likely" response patterns • Similarity of response patterns is compared
Parametric	• Psychometric model used • Observed item responses compared to "expected"	• Psychometric model used • Similarity of response patterns is compared

DRAFT - DO NOT DISTRIBUTE

Table 2. Statistical methods for detecting person-misfit (i.e., aberrance) of examinee item responses. These methods have been employed in cheating detection research, but are not explicitly designed for its detection.

Group of Methods	**Description**	**Variations**	**Reference(s)**
Nonparametric person-fit indices	Item response vector is compared to its similarity with a Guttman vector. Nonparametric "group as norm" approaches to aberrance detection	*C*	Sato (1975)
		A, D, E	Kane & Brenna (1980)
		*C**	Harnisch & Linn (1981)
		NCI	Tatsuoka & Tatsuoka (1982); Tatsuoka & Tatsuoka (1983)
		N Guttman errors	Meijer (1994)
		U3	Van der Flier (1980, 1982); Mokken & Lewis (1982)
		$CUSUM_{LR}$	Armstrong & Shi (2009b)
Parametric person-fit indices	Given an IRT/Rasch model-based ability estimate, test the likelihood of a given response pattern (i.e., person fit to model). General aberrance detection, not specific to cheating	ℓ_0	Levine & Rubin (1979)
		ℓ_z	Drasgow et al (1985)
		M	Molenaar & Hoijtink (1990)
		Bayesian log-odds ratio index	McLeod, Lewis, & Thissen (2003)
		corrected ℓ_z	de la Torre & Deng (2008)
		CUSUM	van Krimpen-Stoop & Meijer (2000); Meijer (2002); Armstrong & Shi (2009a)
		DGM	Shu (2010); Shu, Henson, & Luecht (2013)
		Bayesian HLM	Skorupski & Egan (2011)
		lco difference	Clark (2012)
		WR probability	van der Linden & Jeon (2012)

DRAFT - DO NOT DISTRIBUTE

Table 3. Statistical methods for detecting cheating behavior through answer copying. These methods have been explicitly designed for the detection of unusual agreement between pairs of examinees.

Group of Methods	**Description**	**Variations**	**Reference(s)**
Nonparametric agreement indices	Sampling distribution of chance agreement is compared to actual agreement between pairs of examinees	*Index B, H*	Angoff (1974)
		g_2	Frary, et al (1977)
		K-index	Holland (1996)
		PMIR	Lewis & Thayer (1998)
		$\overline{K}_2$	Sotaridona & Meijer (2002)
		S_1 *and* S_2	Sotaridona & Meijer (2003)
		γ_{js} (binomial)	van der Linden & Sotaridona (2004)
		τ_1 and τ_2	van der Ark, et al (2008)
		Algorithm 1	Belov & Armstrong (2010)
		VM-Index	Belov (2011)
Parametric agreement indices	The differences between model-based, expected chance agreement and observed agreement is analyzed by means of a hypothesis testing framework	ω	Wollack (1997); Wollack (2003)
		γ_{js} (generalized binomial)	van der Linden & Sotaridona (2006)
		τ_j^* and ρ_{jk}	van der Linden & Sotaridona (2009)

DRAFT DO NOT DISTRIBUTE