Application of Bayesian Methods for Detecting Fraudulent Behavior on Tests

Sandip Sinharay, Educational Testing Service

An Updated Version of this document appeared in the journal *Measurement: Interdisciplinary Research and Perspectives* on 3/30/2018. The website for the article is: https://doi.org/10.1080/15366367.2018.1437308

The citation for the article is: Sinharay, S. (2018). Application of Bayesian Methods for Detecting Fraudulent Behavior on Tests. *Measurement: Interdisciplinary Research and Perspectives, 16*(2), 100-113, DOI: 10.1080/15366367.2018.1437308.

Note: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170026. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education or of Educational Testing Service.

Application of Bayesian Methods for Detecting Fraudulent Behavior on Tests

Sandip Sinharay, Educational Testing Service

February 2, 2018

Note: Any opinions expressed in this publication are those of the author and not necessarily of Educational Testing Service.

Application of Bayesian Methods for Detecting Fraudulent Behavior on Tests

Abstract

Producers and consumers of test scores are increasingly concerned about fraudulent behavior before and during the test. There exist several statistical or psychometric methods for detecting fraudulent behavior on tests. This paper provides a review of the Bayesian approaches among them. Four hitherto-unpublished real data examples are provided to demonstrate the application of Bayesian approaches to detect various types of fraudulent behavior on tests. The examples show that Bayesian methods can be useful in detecting several types of test fraud.

Key words: posterior predictive model checking, posterior probability of cheating, test fraud.

Acknowledgements

The author would like to thank Randall Schumacker and the two anonymous reviewers for several helpful comments that led to a significant improvement of the paper. The author would also like to thank Hongwen Guo, Shelby Haberman, Daniel McCaffrey, Wim van der Linden, Howard Wainer, Chun Wang, Xi Wang, and James Wollack for several helpful comments. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170026. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education or of Educational Testing Service.

An increasing concern of producers and consumers of test scores is fraudulent behavior before and during the test. Such behavior is more likely to be observed when the stakes are high, such as in admission, licensing, and certification testing (van der Linden, 2009). A search on the web with words "test fraud" and "cheating" fetches a large number of news materials in the local and national media outlets including that of educator cheating in Atlanta public schools (e.g., Kingston, 2013) and "cram schools" selling items on SATs (e.g., Strauss, 2014). Standard 6.6 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014) includes the recommendation among others that testing programs with high-stakes consequences should have defined procedures for detecting potential testing irregularities. Several guidelines on methods for detecting fraudulent behavior on tests can be found in a report on the results on a nationwide survey of state testing directors conducted by the United States Government Accounting Office (Government Accountability Office, 2013), in the transcript of proceedings of a test integrity symposium (National Center for Education Statistics, 2012), and in a special report on testing and data integrity published by the National Council on Measurement in Education (National Council on Measurement in Education, 2012).

Naturally, there is a growing interest in statistical/psychometric methods for detecting fraudulent behavior on tests, which is evident from the publication of three recent edited volumes (Cizek & Wollack, 2017; Kingston & Clark, 2014; Wollack & Fremer, 2013), numerous journal articles and conference presentations (e.g., Allen & Ghattas, 2016; McLeod, Lewis, & Thissen, 2003; Sinharay, 2017b, 2017a; Sinharay, Duong, & Wood, 2017; van der Linden, 2009; van der Linden & Guo, 2008; van der Linden & Lewis, 2015; van der Linden & Sotaridona, 2006; Wollack, 1997; Wollack, Cohen, & Eckerly, 2015; Wang, Liu, & Hambleton, 2017) on the topic, and a conference (Conference on Test Security) dedicated to the topic. The methods for detecting fraudulent behavior on tests are often referred to as "data forensics" or "data forensics methods" and the analysis using these methods is often referred to as "data forensics analysis". Data forensics analysis usually focuses on one of five areas: analysis of gain scores (or score gains), similarity analysis including detection

of answer-copying, erasure analysis, person-fit analysis, and other analyses including response-time analysis (e.g. Olson & Fremer, 2013).

Among the statistical methods that have been suggested for detecting fraudulent behavior on tests (see, for example Ferrara, 2017, for a list of some of those methods), most are frequentist. However, Bayesian methods have been suggested in each of the five key areas identified by Olson and Fremer (2013). The first goal of this paper is to review some of those methods. The second goal of this paper is to bring attention of those interested in statistical methods for detecting test fraud to the advantages of some of the existing Bayesian methods including posterior probability of answer-copying (Allen & Ghattas, 2016; van der Linden & Lewis, 2015) and posterior probability of cheating (Skorupski & Wainer, 2017).

The next section includes, for each of the five areas in data forensics analysis, a brief description of the area, the most popular approaches under the area, and the Bayesian approaches that have been suggested under the area. The Real Examples section includes hitherto-unpublished applications of Bayesian methods for detecting fraudulent behavior to four types of analyses using two real data sets. The final section includes the conclusions and some recommendations.

Bayesian Methods in the Five Areas of Data Forensics Analysis

Analysis of Gain Scores

Olson and Fremer (2013) stated that the type of data forensics analysis that is the easiest to employ and has a very long history of use in state assessments is making comparisons of scores from one testing occasion to another. A large score gain, that is, a large score difference between an examinee's score on a test and his/her score on a previous administration on the same test, may indicate fraudulent behavior. Holland (1996) and Lewis and Thayer (1998) stated that a large score gain may lead to a test taker being identified for further attention. In 2013, 28 states reported using score gains to detect possible fraudulent test-taking behavior in their K-12 assessments (Government

Accountability Office, 2013). States and testing organizations typically do not use statistical hypothesis tests in analysis of score gains but identify those with a score gain above a predetermined cutoff (equal to, for example, several multiples of the standard deviation of the score) for further attention (e.g., Buss & Novick, 1980).

However, there exist methods to determine whether a score gain is statistically significant. Researchers such as Fischer (2003) and Finkelman, Weiss, and Kim-Kang (2010) suggested frequentist hypothesis tests for determining if a score gain is statistically significant. Skorupski and Egan (2011, 2014) and Skorupski, Fitzpatrick, and Egan (2017) suggested a Bayesian hierarchical linear model to detect group-level cheating based on score gains. In this approach, the change in the individual scores is modeled and the individuals are nested within groups (where "schools" is an example of "groups"). Unusually large group-by-time interaction (that indicates a large score gain for a group on average) is considered as evidence of potential cheating. Let Y_{igt} denote the score of examinee i in group g at time t. Then, the approach of Skorupski and Egan (2011) involves expressing Y_{igt} as

$$Y_{igt} = \beta_0 + \beta_{1g} I(g) + \beta_{2t} I(t) + \beta_{3gt} I(g) I(t) + \epsilon_{igt}, \tag{1}$$

where β_0, β_{1g}, β_{2t}, and β_{3gt} are the model (regression) parameters, ϵ_{igt} is the error term that is assumed to follow a normal distribution, and $I(g)$ and $I(t)$ are indicator variables (where, for example, $I(g)$ is equal to 1 for the examinees belonging to group g and 0 for other examinees). Prior distributions are assumed on the model parameters and the model is fitted using a Markov chain Monte Carlo algorithm (e.g., Gelman, Carlin, Stern, & Rubin, 2003). A large positive value of the estimate of β_{3gt} indicates possible aberrant behavior by examinees in group g. Skorupski and Egan (2011) successfully applied the model to a real data set while Skorupski and Egan (2014) and Skorupski et al. (2017) applied the model to simulated data. Liu, Liu, and Simon (2014) examined score gains using a Bayesian polynomial mixed-effect growth model.

Note that the use of only gain score could falsely accuse many examinees (e.g., Roberts, 1987) and hence gain scores should be used along with other, often nonstatistical, evidence

in an investigation of test fraud.

Similarity Analysis

The goal of *similarity analysis* is to detect potential collusion among a pair or group of examinees by investigating the similarity of their answers. The indices used to perform similarity analysis, which are often used to determine whether an examinee (copier) potentially copied answers from another examinee (source), are often referred to as *answer-copying indices*.

The answer-copying indices that are most popular include the K-index (e.g., Holland, 1996), which, for a pair of examinees, is the chance of observing at least the observed number of matching incorrect responses if the pair were working independently, or the closely related probability of matching incorrect responses (PMIR; e.g., Lewis & Thayer, 1998), the ω index (Wollack, 1997) that is the standardized value of the number of matching responses, and the index based on the generalized binomial model (van der Linden & Sotaridona, 2006) that is the probability of observing at least the observed number of matching responses under the generalized binomial model if the examinee pair were working independently.

van der Linden and Lewis (2015) suggested an approach to compute the posterior probability of answer-copying and the posterior odds of answer-copying between a pair of examinees. The approach does not employ any existing answer-copying index and instead employs combinatorial expressions and a reformulation of the simple recursive algorithm of Lord and Wingersky (1984) for the calculation of the distribution of the number-correct score. The approach is computation-intensive, but they suggested an efficient recursive algorithm for the calculation of the posterior odds of answer-copying. Allen and Ghattas (2016) also suggested an approach to compute the posterior probability of answer-copying when the value of an answer-copying index is available—they used the ω index (Wollack, 1997) to compute the posterior probability in their simulation and real-data examples. Because it is difficult to compute the posterior probability of answer-copying for real data, Allen and Ghattas (2016) provided the following expression of the lower bound of the

posterior probability for someone for whom the observed value of the ω index is W:

$$1 - \frac{\text{Proportion of values of } \omega \text{ near } W \text{ among between-center pairs}}{\text{Proportion of values of } \omega \text{ near } W \text{ among within-center pairs}}.$$

Both van der Linden and Lewis (2015) and Allen and Ghattas (2016) were inspired by the fact that a frequentist p-value is an answer to the question "What is the probability of a significant value of the test statistic given that the examinee did not commit fraud?" that is not the question of interest and the posterior probability of answer-copying is an answer to the question "What is the probability that test fraud occurred, given a significant value of the test statistic?" that is very important in the context of detecting test fraud. The concept of the posterior probability of answer-copying (Allen & Ghattas, 2016; van der Linden & Lewis, 2015) is very similar to that of the posterior probability of cheating (PPoC) that was suggested by Skorupski and Wainer (2017). The PPoC can be computed for any type of cheating and Skorupski and Wainer (2017) provided the following simple approximation for the PPoC for an individual or pair of individuals:

$$\text{PPoC} \approx 1 - \frac{\text{p-value} \times P(\text{non-cheater})}{P(\text{More extreme value in the sample})},$$

where p-value is the frequentist p-value for an examinee or examinee pair for detecting the relevant type of test fraud computed using a frequentist approach, P(non-cheater) is the prior probability of non-cheaters in the sample and P(More extreme value in the sample) is the proportion of individuals in the sample with a more extreme value of the test statistic than the individual in question. For example, if an investigator has computed the value of the ω index (Wollack, 1997) for an examinee pair, he/she can compute the p-value for ω and then use the above approximation to compute a PPoC for the examinee pair.[1]

Erasure Analysis

Erasure analysis gained prominence after the alleged educator cheating in 2009 in Atlanta public schools (e.g., Kingston, 2013). Erasures, or, in general, answer changes, are found roughly for one out of every 50 answers (e.g., Primoli, Liassou, Bishop, &

[1]PPoCs are computed from ω's in the first real data example later in this paper.

Nhouyvanisvong, 2011). While most of these erasures are benign, fraudulent erasures arise when teachers or school administrators help students by improving wrong answers on their answer sheets after the test, often to hide underachievement by their class or school (van der Linden & Lewis, 2015). In erasure analysis, an investigator attempts to find out if the erasures are fraudulent. In 2013, 33 states reported using erasure analysis to detect possible fraudulent test-taking behavior in their K-12 assessments (Government Accountability Office, 2013). However, erasure analysis in several states involves only the use of the average wrong-to-right erasure count for a school or school-district (e.g., Bishop & Egan, 2017, pp. 204-205), a statistic whose power has been shown to be rather low, that is, the statistic would fail to flag many cheaters (e.g., Sinharay, 2017a). Additional indices that have been suggested for performing erasure analysis include the erasure detection index (EDI; Wollack et al., 2015), EDI for groups of examinees (Wollack & Eckerly, 2017), the L-index (Sinharay et al., 2017) and two extensions of the EDI for groups of examinees (Sinharay, 2017a). The EDI is the standardized value, computed under an IRT model, of the wrong-to-right erasure count of an examinee and the L-index is a variation of the likelihood ratio test statistic for the null hypothesis that the erasures found in the answer sheet are not fraudulent.

While the above-mentioned approaches for performing erasure analysis are frequentist, van der Linden and Lewis (2015) suggested an approach to compute the posterior odds of fraudulent erasures—this approach is very similar to their approach to compute the posterior odds of answer-copying. Further, Sinharay and Johnson (2017) suggested a Bayesian approach for performing erasure analysis; the approach involves the computation of the posterior predictive p-value (e.g., Gelman, Meng, & Stern, 1996) for the EDI after implementing a Markov chain Monte Carlo algorithm (e.g., Gelman et al., 2003) that draws values of the examinee ability from the posterior distribution based on the scores on the items on which no erasures were found.

Person-fit Analysis

In person-fit analysis, an investigator tries to determine whether the item scores of an examinee follows an IRT model. In the presence of cheating, the item scores usually do

not follow an IRT model because, for example, an examinee who cheats may answer more difficult items correctly than what an IRT model predicts. There exist several indices for performing person-fit analysis—Meijer and Sijtsma (2001) provided a review of several of these indices.

Drasgow and Guertler (1987) suggested a Bayesian decision-theoretic approach to use person-fit indices; the approach essentially is equivalent to concluding that the item-scores of an examinee are aberrant if the likelihood ratio, that is, the ratio of the likelihood of the item-scores under aberrant behavior and the likelihood of the item-scores under non-aberrant behavior, is too large. Glas and Meijer (2003) suggested performing person-fit analysis using the Bayesian p-values or *posterior predictive p-values* (e.g., Gelman et al., 1996) of several person-fit indices; they found the power of the posterior predictive p-values for the ζ_2 index of Tatsuoka (1984) to be the largest among the person-fit indices that they considered. Sinharay (2015a), Sinharay (2015b) and Sinharay (2016) also performed person-fit analysis using the posterior predictive p-values corresponding to several person-fit indices including l_z (Drasgow, Levine, & Williams, 1985), l_z^* (Snijders, 2001), and ζ_1 and ζ_2 (Tatsuoka, 1984) and found that (a) l_z^* performs at least as well as ζ_2 under a Bayesian approach for dichotomous items, (b) a Bayesian approach leads to more satisfactory false alarm rates compared to a frequentist approach using l_z^* for dichotomous items, and (c) a Bayesian approach leads to a larger power compared to a frequentist approach with l_z for tests that include both dichotomous and polytomous items. The Bayesian approach of using the posterior predictive p-values of person-fit indices is computation-intensive. However, the approach does not depend on asymptotic results, and, as demonstrated by Glas and Meijer (2003), Sinharay (2015a), Sinharay (2015b) and Sinharay (2016), may lead to more satisfactory rate of false positive rate and/or power compared to a frequentist approach in person-fit assessment.

Other Analyses Including Response-time Analysis

There are several other types of analyses that have been suggested for detecting fraudulent behavior on tests. Among them, while there has been sporadic research on

topics such as matching of response patterns (e.g. Haberman & Lee, 2017) and Trojan Horse approach (e.g., Foster, 2013, p. 56), the most popular topics have probably been response-time analysis and detection of item preknowledge.

Response-time Analysis

van der Linden and Guo (2008) suggested a Bayesian approach that first involves the fitting of a combination of a traditional IRT model and a response-time model to a data set that includes both the item-scores and the item-response times of the examinees; then, the examinee-item combinations whose observed response time was too extreme compared to the corresponding posterior predictive density are flagged. An examinee with too many flagged examinee-item combinations is flagged for possible fraudulent behavior. Marianti, Fox, Avetisyan, Veldkamp, and Tijmstra (2014) suggested a Bayesian approach in which the lognormal response-time model of van der Linden (2006) is employed and an aberrant behavior is concluded when the sum of squared standardized residuals is large. Wang, Xu, and Shang (2016) suggested a two-stage approach to detect fraudulent behavior; in the first stage, a mixture hierarchical model is fitted to both the item scores and response times to distinguish solution behavior from aberrant behavior; in the second stage, a Bayesian residual index is constructed to differentiate rapid guessing from fraudulent behavior. Fox and Marianti (2017) suggested a Bayesian approach to person-fit analysis that involves fitting of a joint model to the item-scores and response time; large posterior probabilities of aberrant item-scores, aberrant response times, or both indicate possible fraudulent behavior.

Detection of Item Preknowledge

Examinees benefit from item preknowledge when a "source" shares test questions and/or their answers (where the source could be a teacher, a test preparation company, a website, or individual examinees) and then several beneficiaries memorize the assessment questions and/or answers. The items that are shared are usually referred to as "compromised" items

in the context of detection of item preknowledge. There exist several frequentist approaches for detecting item preknowledge including a variation of the optimal index (Drasgow, Levine, & Zickar, 1996) and the L_s statistic that is a variation of the likelihood ratio test statistic for the null hypothesis that an examinee did not benefit from item preknowledge (Sinharay, 2017b).

McLeod et al. (2003) suggested a Bayesian approach that involves the use of a posterior log-odds ratio as an index for detecting item preknowledge in the computerized adaptive test (CAT) environment. They extended the concept of odds ratios to describe the increased likelihood (based on the item responses) that a response pattern arises from the normal or aberrant models. Segall (2002) and Shu, Henson, and Luecht (2013) suggested new IRT models for detecting item preknowledge. Wang et al. (2017) suggested a (Bayesian) predictive checking method to detect a person's preknowledge on compromised items by using information from the non-compromised items; responses on the non-compromised items are used to estimate a person's proficiency distribution, and then the corresponding predictive distribution for the person's responses on the possibly compromised items is constructed; the use of preknowledge is identified by comparing the observed responses to the predictive distribution. McLeod et al. (2003) did not assume that the set of compromised items is known while Segall (2002), Shu et al. (2013), and Wang et al. (2017) did make such an assumption.

Real Data Examples

In the following discussion, four data examples demonstrate that Bayesian approaches can be used to detect various types of test fraud. Each example also involves at least one frequentist approach that has been found to perform satisfactorily compared to competing frequentist approaches in comparison studies. [2]

[2]For example, Example 2 involves the frequentist L-index that was found to perform satisfactorily compared to other frequentist statistics for erasure analysis by Sinharay et al. (2017).

Example 1: Similarity Analysis

A data set including the item scores of 1,644 examinees on one form of a licensure examination was available. The data set was analyzed in several chapters of Cizek and Wollack (2017) including in Zopluoglu (2017) who computed the values of several answer-copying indices for all examinee pairs within each test center. The examinees took the test in a total of 326 test centers. The test form includes 170 operational items that are dichotomously scored. The licensure organization (that administers the examination) identified as compromised 61 items on the form. The organization also flagged 48 individuals on the form as possible cheaters from a variety of statistical analysis and an investigative process that brought in other information. These 48 individuals can be treated as truly aberrant for all practical purposes because of the rigor of the investigative process. Six of these flagged individuals belonged to test center #2305.

The nominal response model (NRM; Bock, 1972) was fitted to the data using the R package mirt (Chalmers, 2012). The ω index (Wollack, 1997) was computed for all within-center examinee pairs (5,640 of them). The percentage of statistically significant values of ω was much larger than the nominal level in test center #2305.[3] The PPoC was computed using the p-value corresponding to ω for each examinee pair for all test centers under the assumption that the probability of cheating in the population is 0.01. The PPoC was found to be larger than 0.95 for 203 pairs of examinees among the total of 5,640 within-center pairs across all the test centers. Among the 48 examinees flagged by the licensure organization, 10 examinees feature in at least one of these 592 pairs; interestingly, this set of 10 examinees includes all six individuals belonging to test center #2305 who were flagged by the licensure organization.

Figure 1 shows a plot of the PPoC vs the ω index for all within-center examinee pairs in the data set. The PPoC increases as ω increases. The PPoC corresponding to ω of 1.64 and 2.33 (the 95th and 99th percentile of the standard normal distribution; vertical dotted lines are provided in the figure for these values of ω) are about 0.43 and 0.72, respectively.

[3]Zopluoglu (2017) also reported a large percentage of significant values for the test center.

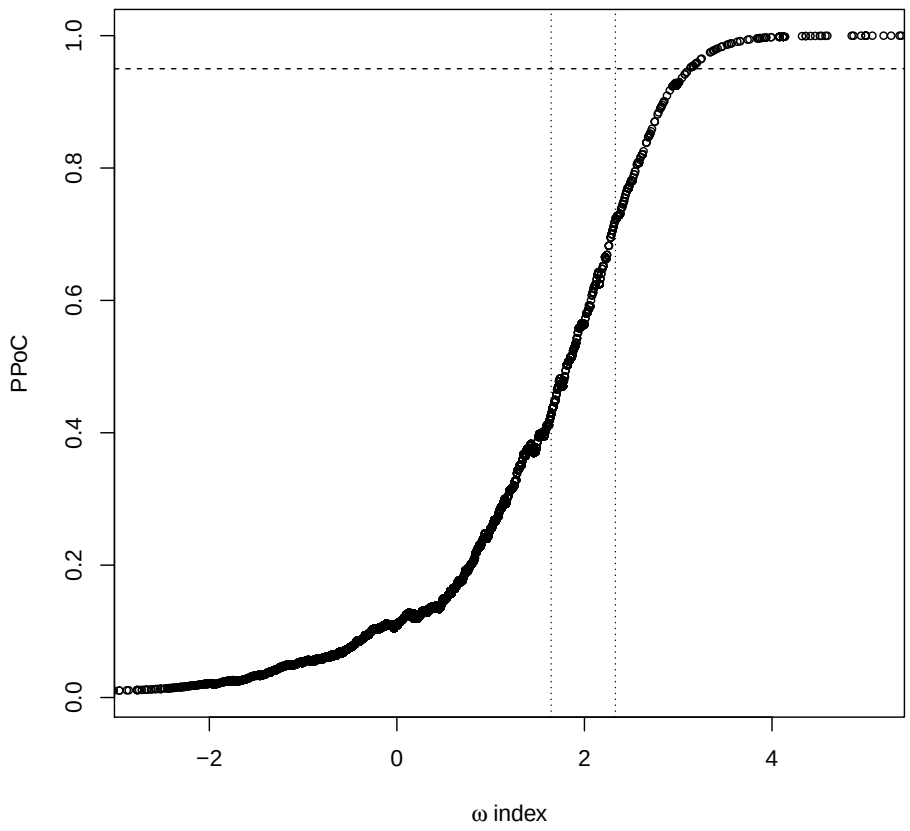

Figure 1: A plot of the PPoC versus the ω index.

The PPoC becomes larger than 0.95 (a horizontal dashed line is provided at the value 0.95 of PPoC) for values of ω larger than about 3.13 (that is the 99.9th percentile of the standard normal distribution). Thus, it seems that the investigator can be confident about fraudulent behavior only for really large (>3.13) values of ω and flagging an examinee based on a p-value of, for example, 0.01, is quite likely to yield a false accusation. Thus, Figure 1 provides a strong justification in favor of the use of the PPoC rather than a p-value associated with ω.

Example 2: Erasure Analysis

The data set consisted of the scores of a random sample of about 3,000 high-school students on 46 dichotomous items on a subject area of a computerized end-of-course state assessment. The test is high-stakes for both the students and teachers. The students in the sample belonged to about 30 schools. The item parameters were estimated under the three-parameter logistic model (3PLM) using marginal maximum likelihood estimation—the 3PLM showed adequate item fit for the data. The first quartile, median, third quartile, and maximum of the number of answer changes among the students in the sample were 0, 1, 2, and 24, respectively; thus, the average number of erasures is in agreement with the prevalence of erasures reported by Primoli et al. (2011).

There were multiple strong sources of external evidence of test fraud for the students from one particular high school, henceforth denoted as school S_1. The data set included 205 students from School S_1 among whom 159 made at least one answer change. The median number of answer changes among the 205 students of School S_1 was 4, which is four times that in the full sample. The (frequentist) L-index (Sinharay et al., 2017), which quantifies the difference between the estimated ability on the erased items and non-erased items and approximately follows a standard normal distribution under no test fraud, was computed for each examinee. The EDI for each examinee was computed using a continuity correction of 0.5 as in Wollack et al. (2015) and the posterior predictive p-value for the EDI with a continuity correction was computed for each examinee as in Sinharay and Johnson (2017); a standard normal prior distribution on the examinee ability was used.[4]

Table 1: The Results of Erasure Analysis.

Examinees	L-index	EDI: Frequentist	EDI: Bayesian
Not in S_1	1.4	0.7	0.9
In S_1	12.6	7.5	10.1

Table 1 shows the percentages of examinees for which the p-values were statistically significant at the 1% level for the three approaches. For the L-index and the EDI under a

[4]Changing the prior did not affect the results in some additional limited analyses.

frequentist approach, a value larger than 2.33 (that is the 99th percentile of the standard normal distribution) of the statistic was considered statistically significant. The first row of numbers in the table provide the percentages among the examinees who were not in School S_1 (that is, in any school other than S_1) and made at least one answer change. The second row of numbers provide the percentages among the 159 examinees in School S_1 who made at least one answer change; thus, for example, the percentage of 12.6 for the L-index in the second row indicates that among the 159 examinees in School S_1 who made at least one answer change, the index was statistically significant at 1% level, that is, larger than 2.33, for 20 examinees (note that $\frac{100\times 20}{159} \approx 12.6$).

The percentage of significant values for either examinee group is largest for the L-index followed by the Bayesian approach with the EDI. One important finding in Table 1 is that the percent significant for each approach is considerably larger among students from school S_1 than among other students. For example, while the EDI with the Bayesian approach was significant for only 0.9% examinees among other schools, the index was significant for about 10% of the students in School S_1. This result is favorable to the approaches including the Bayesian approach in the sense that they seem to work as intended in practice by flagging many students from a school (S_1) that appears suspicious from multiple sources of evidence.

Example 3: Person-fit Analysis

The licensure data set considered in Example 1 was used to perform person-fit analysis. Two parametric person-fit statistics—ζ_2 (Tatsuoka, 1984) and l_z^* (Snijders, 2001)—were used with both the frequentist and Bayesian approaches.

The Rasch model is operationally used in the licensure test; however, the two-parameter logistic model (2PLM) was found to fit the data better than the Rasch model—therefore the 2PLM was used in the analysis. The item parameters under the 2PLM were estimated from the data set and were used to compute the person-fit statistics. The weighted maximum likelihood estimate (WLE; Warm, 1989) was used as the estimate of the examinee ability. For ζ_2 and l_z^*, under a frequentist approach, the critical values at 1% significance level were 2.33 and -2.33, respectively, which are the 99th and 1st percentiles of the standard normal

distribution. To use a Bayesian approach as described in Sinharay (2015b) and Sinharay (2016), the Markov chain Monte Carlo algorithm was used to simulate draws from the posterior distribution of the examinee ability given the scores on the test items, posterior predictive data sets were generated using the draws from the algorithm, and then posterior predictive p-values were computed for each examinee using ζ_2 and l_z^*. A standard normal prior distribution was used on the examinee ability.

Table 2: The Results of Person-fit Analysis for the Licensure Data.

Examinees	l_z^*: Frequentist	l_z^*: Bayesian	ζ_2: Frequentist	ζ_2: Bayesian
Not Flagged	2.2	1.9	1.6	1.6
Flagged	8.3	6.3	10.4	10.4

Table 2 shows the percentages of examinees for which the p-values were statistically significant at the 1% level for the two person-fit statistics under the frequentist and Bayesian approaches. The first row of numbers provides the percentages among the examinees that were not flagged by the licensure organization. The second row of numbers provides the percentages among the 48 examinees who were flagged by the licensure organization. Thus, for example, the percentage of 8.3 for l_z^* in the second row under the frequentist approach indicates that among the 48 examinees flagged by the licensure organization, l_z^* was statistically significant at 1% level, that is, smaller than -2.33, for four examinees (note that 4/48=0.083).

Table 2 shows little difference in the percentage of significant values produced by the frequentist and Bayesian approaches for any statistic. For l_z^*, the Bayesian approach led to fewer significant values for both flagged and non-flagged examinees; this could be related to the inflated Type I error rate of the l_z^* statistic at 1% level that has been found by researchers such as van Krimpen-Stoop and Meijer (1999) and Snijders (2001). The ζ_2 statistic led to more significant values among flagged examinees compared to l_z^* under either a frequentist or Bayesian approach. In addition, for each statistic, the frequentist and Bayesian approaches mostly flagged the same examinees; for example, among the 48 examinees flagged by the licensure organization, the p-values for l_z^* under frequentist and

Bayesian approaches were either both-significant or both-nonsignificant for 47 examinees.

Table 2 also shows that the percent significant for each person-fit statistic is much larger among the examinees flagged by the licensure organization compared to among non-flagged examinees—this result provides some evidence that the person-fit statistics provide useful information by being significant at a larger rate among the examinees who are truly aberrant for all practical purposes.

Example 4: Detection of Item Preknowledge

The licensure data set described above can be used to detect item preknowledge because 61 items were identified as compromised by the licensure organization. A frequentist and a Bayesian approach were used to detect item preknowledge. The 2PLM was used in all the analyses.

The frequentist approach comprised the use of the signed likelihood ratio statistic L_s (Sinharay, 2017b) that quantifies the difference between the estimated ability on the compromised items and non-compromised items—substantial item preknowledge would cause the statistic to be a large positive number. The large-sample distribution of L_s is the standard normal distribution under the null hypothesis of no item preknowledge (Sinharay, 2017b).

The Bayesian approach comprised the use of the predictive p-value (Wang et al., 2017) of the raw score on the compromised items. A Markov chain Monte Carlo algorithm was used to simulate draws from the posterior distribution of the examinee ability given the scores on the non-compromised items, posterior predictive data sets were generated using the draws from the algorithm, and then predictive p-values were computed for each examinee using the raw score on the compromised items as the test statistic. A standard normal prior distribution was used on the examinee ability.

Table 3 shows the percentages of examinees for which the p-values were statistically significant at 1% level for the frequentist and Bayesian approaches. The two row of numbers provide the percentages among the examinees that were not flagged and flagged by the licensure organization. The Bayesian approach leads to slightly more number of significant

Table 3: The Results of Analysis for Detection of Item Preknowledge for the Licensure Data.

Examinees	Frequentist	Bayesian
Not Flagged	3.1	3.8
Flagged	18.8	21.2

p-values for both group of examinees.

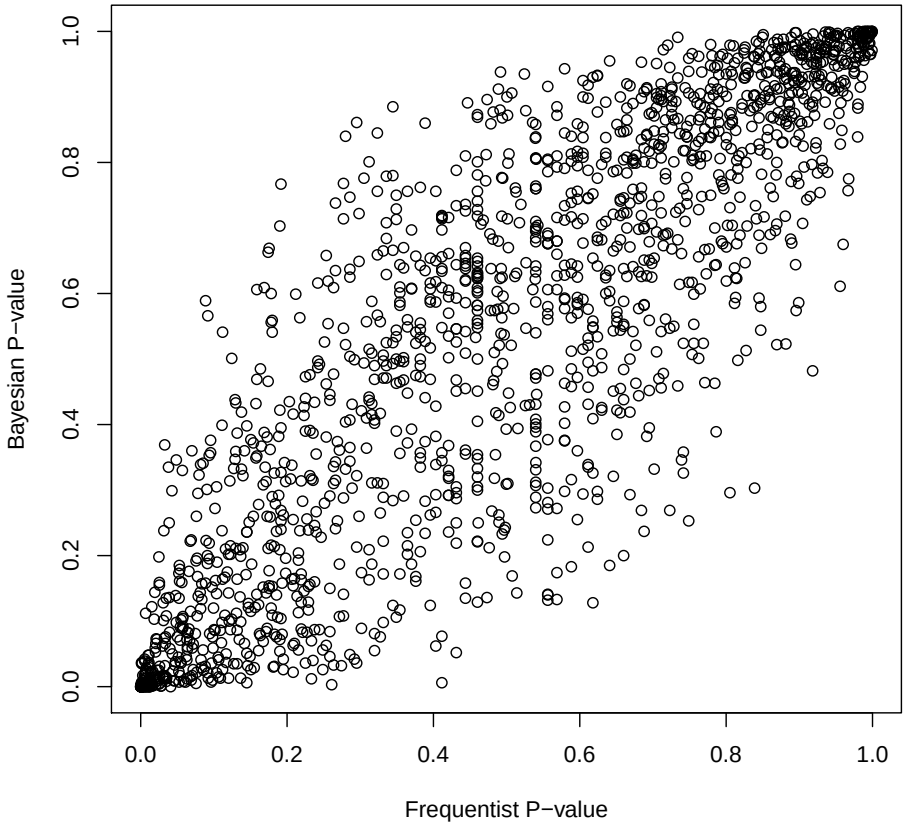

Figure 2: A scatter-plot of the frequentist and Bayesian p-values.

Figure 2 shows a plot of the frequentist p-values (X-axis) versus the (Bayesian) predictive p-values. The frequentist p-value is the probability of a larger value than the observed value of L_s under the standard normal distribution. There seems to be a substantial agreement between the two sets of p-values.

Conclusions and Recommendations

A brief review of the five main areas (identified by Olson & Fremer, 2013) under data forensics analysis is provided. Brief descriptions are included of most of the Bayesian approaches that have been suggested under each of the five areas.

Most of the existing Bayesian approaches for detecting test fraud are used either to compute a Bayesian p-value (e.g. Sinharay, 2015b; Wang et al., 2017) or to compute a posterior probability of cheating/answer-copying (e.g., Allen & Ghattas, 2016; Skorupski & Wainer, 2017; van der Linden & Lewis, 2015). While researchers such as Sinharay (2015b) showed that a Bayesian p-value may be superior than a frequentist p-value in some problems, the posterior probability of cheating/answer-copying has the advantage over frequentist or Bayesian p-values that the former directly answers the question "What is the chance of test fraud given a significant value of the test statistic?" (e.g., Skorupski & Wainer, 2017). Also, the use of a p-value can lead to a large proportion of false positives; Skorupski and Wainer (2017) provided an example where a statistic with a Type I error rate of 0.01 and power of 0.99 is expected to flag 1,386 examinees in a population of 70,000 examinees that includes 1% cheaters, but half of the flags are false positives. The PPoC does not have such a limitation.

There exist several approaches, both frequentist and Bayesian, for detecting test fraud. However, there remain several unanswered questions on how these approaches should be used in practice. For example, how many approaches should be used and which ones for a specific data set? Wollack and Cizek (2017b, p. 397) warned against the use of multiple conceptually similar methods and commented that evidence of cheating from distinctly different indices can be quite compelling. Wollack and Cizek (2017b, p. 397) also recommended controlling for false positives in the context of test fraud where several statistical tests are typically performed on many examinees; one way to limit the number of false positives is to choose a critical value that adjusts for multiple comparisons by controlling the family-wise error rate (using, for example, a Bonferroni correction) or controlling the false discovery rate (using the procedure of Benjamini & Hochberg, 1995). Buss and Novick (1980) asserted that statistical methods of detection should generally not

be the sole basis for a judgment that an examinee cheated (or that an examinees scores are sufficiently questionable to justify non-reporting) in the absence of corroboration from other types of evidence." Other researchers such as Holland (1996), Hanson, Harris, and Brennan (1987), and Tendeiro and Meijer (2014) made similar recommendations. However, test-security experts such as Wollack and Cizek (2017a, p. 200) have recently presented the viewpoint that statistical evidence based on even a single statistic may constitute conclusive proof of cheating provided the statistic has been properly vetted and accepted by the research community and the degree of aberrance is clearly extreme.

Though a substantial amount of research has been performed on Bayesian methods for detection of fraud on tests, there seems to be some gaps in addition to those discussed above. First, there has been no application of Bayes factors (e.g., Kass & Raftery, 1995) in the area although Skorupski and Wainer (2017, p. 351) mentioned the possibility of such applications. It is hoped that Bayesian methods will find more applications in detection of cheating on tests as computers become faster and Bayesian analysis becomes more popular among researchers and practitioners. Second, the number of Bayesian approaches is small in all of the five areas of data forensics analysis and there is a need for further research on Bayesian approaches in each of these areas. Third, there is a severe lack of Bayesian approaches for detecting fraudulent behavior at an aggregate level (that is, at the level of classes or schools that the examinees belong to), that by Skorupski and Egan (2011) being an exception—so more research on this area will be useful. Fourth, Bayesian approaches have the advantage (over frequentist approaches) of being able to use prior information in the form of an informative prior distribution; however, there does not exist any example of the use of an informative prior distribution in the context of Bayesian methods in detection of fraud in tests. Finally, there is a need for more simulation studies comparing Bayesian and frequentist methods for detecting test fraud; such studies could potentially reveal the type of applications where one of these approaches would be preferable over the other.

References

Allen, J., & Ghattas, A. (2016). Estimating the probability of traditional copying, conditional on answer-copying statistics. *Applied Psychological Measurement*, *40*, 258–273.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*, 289–300.

Bishop, S., & Egan, K. (2017). Detecting erasures and unusual gain scores: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 193–213). Washington, DC: Routledge.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Buss, W. G., & Novick, M. R. (1980). The detection of cheating on standardized tests: Statistical and legal analysis. *Journal of Law and Education*, *9*, 1-64.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of detecting cheating on tests*. Washington, DC: Routledge.

Drasgow, F., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology*, *72*, 10–18.

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.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, *9*, 47–64.

Ferrara, S. (2017). A framework for policies and practices to improve test security programs: Prevention, detection, investigation, and resolution (PDIR). *Educational Measurement: Issues and Practice*, *36*(3), 5-24.

Finkelman, M., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement*, *34*, 238–254.

Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement*, *27*, 3–26.

Foster, D. (2013). Security issues in technology-based testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 299–311). New York, NY: Routledge.

Fox, J.-P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, *54*, 243–262.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*.

New York, NY: Chapman and Hall.

Gelman, A., Meng, X., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–807.

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*, 217–233.

Government Accountability Office. (2013). *K-12 education: States' test security policies and procedures varied (GAO-13-495R)* (Tech. Rep.). Washington, DC: Author.

Haberman, S. J., & Lee, Y.-H. (2017). *A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses* (ETS Research Report No. RR-17-23). Princeton, NJ: ETS.

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying (ACT research report series no. 87-15).* Iowa City, IA: American College Testing.

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 No. RR-94-4). Princeton, NJ: ETS.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kingston, N. (2013). Educator testing case studies. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 299–311). New York, NY: Routledge.

Kingston, N., & Clark, A. (2014). *Test fraud: Statistical detection and methodology.* New York, NY: Routledge.

Lewis, C., & Thayer, D. T. (1998). *The power of the K-index (or PMIR) to detect copying* (ETS Research Report No. RR-98-49). Princeton, NJ: ETS.

Liu, X., Liu, F., & Simon, M. (2014, April). *Are the score gains suspicious? a Bayesian growth analysis approach.* Paper presented at the annual meeting of the National Council of Measurement in Education, Philadelphia, PA.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*, 453–461.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*, 426–451.

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

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

National Center for Education Statistics. (2012). *Transcript of proceedings of the testing integrity symposium* (Tech. Rep.). Washington, DC: Institute of Education Science.

National Council on Measurement in Education. (2012). *Testing and data integrity in the administration of statewide student assessment programs* (Tech. Rep.). Madison, WI: Author.

Olson, J. F., & Fremer, J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test securities irregularities.* Washington DC: Council of Chief State

School Officers.

Primoli, V., Liassou, D., Bishop, N. S., & Nhouyvanisvong, A. (2011, April). *Erasure descriptive statistics and covariates.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Roberts, D. M. (1987). Limitations of the score-difference method in detecting cheating in recognition test situations. *Journal of Educational Measurement*, *24*, 77–81.

Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics*, *27*, 163–179.

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, *78*, 481–497.

Sinharay, S. (2015a). Assessment of person fit for mixed-format tests. *Journal of Educational and Behavioral Statistics*, *40*, 343–365.

Sinharay, S. (2015b). Assessing person fit using l_z^* and the posterior predictive model checking method for dichotomous item response theory models. *International Journal of Quantitative Research in Education*, *2*, 265–284.

Sinharay, S. (2016). Assessment of person fit using resampling-based approaches. *Journal of Educational Measurement*, *53*, 63–85.

Sinharay, S. (2017a). Detecting fraudulent erasures at an aggregate level. *Journal of Educational and Behavioral Statistics.* (Advance online publication. doi:10.3102/1076998617739626)

Sinharay, S. (2017b). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*, 46–68.

Sinharay, S., Duong, M. Q., & Wood, S. W. (2017). A new statistic for detection of aberrant answer changes. *Journal of Educational Measurement*, *54*, 200–217.

Sinharay, S., & Johnson, M. S. (2017). Three new methods for analysis of answer changes. *Educational and Psychological Measurement*, *77*, 54–81.

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

Skorupski, W. P., & Egan, K. (2014). A Bayesian hierarchical linear modeling approach for detecting cheating and aberrance. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 121–136). New York, NY: Routledge.

Skorupski, W. P., Fitzpatrick, J., & Egan, K. (2017). A Bayesian hierarchical model for detecting aberraant growth at the group level. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 232–244). Washington, DC: Routledge.

Skorupski, W. P., & Wainer, H. (2017). The case for Bayesian methods when investigating test fraud. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 214–231). Washington, DC: Routledge.

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

Strauss, V. (2014). *The six-step SAT cheating operation in Asia and how to stop it.* (Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2014/11/16/the-six-step-

sat-cheating-operation-in-asia-and-how-to-stop-it/)

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*, 95-110.

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.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.

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

van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika*, *80*, 689–706.

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

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, *23*, 327–345.

Wang, C., Xu, G., & Shang, Z. (2016). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*. (Advance online

publication. doi:10.1007/s11336-016-9525-x)

Wang, X., Liu, Y., & Hambleton, R. K. (2017). Detecting item preknowledge using a predictive checking method. *Applied Psychological Measurement*, *41*, 243–263.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

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

Wollack, J. A., & Cizek, G. J. (2017a). Test security for licensure and certification examination programs. In S. Davis-Becker & C. Buckendahl (Eds.), *Testing in the professions* (pp. 178–209). New York, NY: Routledge/Taylor & Francis.

Wollack, J. A., & Cizek, G. J. (2017b). The future of quantitative methods for detecting cheating. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 390–399). Washington, DC: Routledge.

Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting test tampering using item response theory. *Educational and Psychological Measurement*, *75*, 931–953.

Wollack, J. A., & Eckerly, C. (2017). Detecting test tampering at the group level. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 214–231). Washington, DC: Routledge.

Wollack, J. A., & Fremer, J. J. (2013). *Handbook of test security.* New York, NY: Routledge.

Zopluoglu, C. (2017). Similarity, answer copying, and aberrance: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests*

(pp. 25–46). Washington, DC: Routledge.