Review Article

Sandip Sinharay*

Latent-variable Approaches Utilizing Both Item Scores and Response Times To Detect Test Fraud

<https://doi.org/10.1515/edu-2020-0137>

received August 13, 2020; accepted December 30, 2020.

Abstract: There is a growing interest in approaches based on latent-variable models for detecting fraudulent behavior on educational tests. Wollack and Schoenig (2018) noted the presence of five types of statistical/psychometric approaches to detect the three broad types of test fraud that occur in educational tests. This paper includes a brief review of the five types of statistical/psychometric approaches mentioned by Wollack and Schoenig (2018). This paper then includes a more detailed review of the recent approaches for detecting test fraud using both item scores and response times—all of these approaches are based on latent-variable models. A real data example demonstrates the use of two of the approaches.

Keywords: chi-bar-square distribution; likelihood ratio statistic; Wald statistic.

Fraud and other aberrant test-taking behaviors on educational tests have been known to undermine the reliability of test usage and the validity of test score interpretations and inferences drawn from these scores (e.g., Cizek & Wollack, 2017). 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 that testing programs with high-stakes consequences should have defined procedures for detecting potential test fraud. Guidelines on methods for detecting fraudulent behavior on tests have been provided by other organizations as well (e.g., Government Accountability Office, 2013; National Center for Education Statistics, 2012; National Council on Measurement in Education, 2012).

Consequently, there is a growing interest in statistical/psychometric methods for detecting fraudulent behavior on tests, which is evident from the numerous publications on the topic (e.g., Cizek & Wollack, 2017; Drasgow, Levine, & Zickar, 1996; Ferrara, 2017; Kingston & Clark, 2014; McLeod, Lewis, & Thissen, 2003; Sinharay, 2017, 2018; Sinharay & Johnson, 2020; 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; Wollack & Fremer, 2013; Wang, Liu, & Hambleton, 2017). While most research on detecting test fraud is based only on item scores, the increasing popularity of computerized assessments has allowed the recording of response times, and, subsequently, detection of test fraud using both item scores and response times (e.g., Sinharay & Johnson, 2020). Researchers such as Boughton, Smith, and Ren (2017), Fox and Mariani (2017), Lee and Wollack (2020), Man, Harring, and Sinharay (2019), Sinharay and Johnson (2020), van der Linden and Guo (2008), and Wang, Xu, Shang, and Kuncel (2018) have suggested a variety of approaches that can be used to detect test fraud using both item scores and response times and all of these approaches are based on latent variable models (e.g., Bartholomew, Knott, & Moustaki, 2011). The goal of this paper is to provide a review of these approaches. The next section includes brief reviews of the five types of statistical/psychometric approaches to detect test fraud and the most popular methods under each type of approach. The following section includes a review of the approaches based on both item scores and response times. The Real Data section includes a real data example. The last section includes some conclusions and recommendations.

*Corresponding author: Sandip Sinharay, Educational Testing Service, Princeton, NJ, United States, E-mail: ssinharay@ets.org

Table 1: The Three Types of Test Fraud and Five types of Statistical Approaches to Detect Them.

Approaches	Answer-copying and Collusion	Preknowledge	Test tampering
Score Differencing		✓	✓
Analysis of Erasures & Answer Changes			✓
Copying-detection & Similarity Analysis	✓	✓	✓
Person-fit analysis	✓	✓	✓
Use of response-time models		✓	

1 The Five types of Statistical/psychometric Approaches to Detect Test Fraud

Wollack and Schoenig (2018) noted the occurrence of the following three broad types of test fraud in educational assessments: (a) answer-copying and collusion, (b) preknowledge, and (c) test tampering. They also stated that there exist five types of statistical/psychometric approaches to detect these three types of fraud. Interestingly, almost all of these approaches are based on item response theory (IRT) models (e.g., Hambleton & Swaminathan, 1985), which are latent variable models (e.g., Bartholomew et al., 2011) and are widely used in the reporting of scores to takers of educational and psychological tests.

Table 1 shows the three types of test fraud (the column names) and the five types of statistical/psychometric approaches (row names). A check mark for a cell indicates that the approach in the heading of the corresponding row can be applied to detect the type of fraud corresponding to that column. For example, a check mark for score differencing and preknowledge indicates that score differencing can be used to detect preknowledge.

Brief descriptions of the three types of test fraud mentioned by Wollack and Schoenig (2018) are given below:

- *Answer-copying and collusion:* Examinees often copy answers from another examinee, most often one sitting close to them. Collusion involves two or more examinees working together (when they are not allowed to) to answer the test items.
- *Preknowledge:* Examinees benefit from item preknowledge when a “source” (such as a teacher, a test preparation company, or a website) shares test questions and/or their answers and then several beneficiaries memorize the assessment questions and/or answers. The items that are shared are usually referred to as “compromised” items.

- *Test tampering:* In its most common form, classroom teachers or administrators change student answers after the tests are complete (e.g., Wollack et al., 2015). Individual students may also work on their own or with the help of a companion to change their answers. While less common, test tampering can take place on computerized tests as well (see, e.g., Sinharay et al., 2017).

Note that test fraud may occur at the level of individual examinees (as documented in Sinharay et al., 2017, p. 202) and at the level of groups (such as classrooms or schools) of examinees (e.g., Maynes, 2013).

Brief descriptions of the five types of statistical/psychometric approaches that are used to detect the three abovementioned types of test fraud are given below:

- *Score Differencing:* Score differencing involves a test of the null hypothesis of equal performance of an examinee over two sets of items. The alternative hypothesis is that the performance is better, presumably due to test fraud, on one of the item sets. Score differencing can be used to detect several types of test fraud including item preknowledge (e.g., Sinharay, 2017) and test tampering (e.g., Sinharay et al., 2017). Among the several approaches for score differencing, those that have been found to be most powerful are a variation of the optimal index (Drasgow et al., 1996) and the L_s statistic (Sinharay, 2017) that is a variation of the likelihood ratio test statistic.
- *Analysis of Erasures and Answer Changes:* While erasures are found on answer sheets irrespective of whether test fraud occurred (e.g., Sinharay et al., 2017), fraudulent erasures arise when teachers or school administrators correct students’ wrong answers on their answer sheets after the test, often to hide underachievement by their class or school (van der Linden & Lewis, 2015). In analysis of erasures, an investigator attempts to find out if the erasures found on answer sheets are fraudulent. Popular statistics for analysis of erasures include (a) the erasure detection

index (EDI; Wollack et al., 2015), (b) the L-index (Sinharay et al., 2017), (c) EDI for groups of examinees (Wollack & Eckerly, 2017), and (d) two extensions of the EDI for groups of examinees (Sinharay, 2018). While the EDI and the L-index are intended to detect individual examinees, the EDI for groups of examinees and its extensions are intended to detect groups (such as classrooms or schools) of examinees.

- *Copying-detection and Similarity Analysis:* Copying-detection and similarity analyses are intended to detect potential collusion among a pair or group of examinees by investigating the similarity of their answers and flagging the examinees who produce an observed number of answer matches that significantly exceeds the number of matches predicted under the model. These analyses are performed using *answer-copying indices* and *similarity indices*. While both of these two types of indices quantify the similarity of the answers of pairs of examinees, they differ in the way they are computed. In computing the answer-copying indices for a pair of examinees, one first specifies a source and a copier in the pair and computes the expected number of answer matches conditional on the estimated ability of the copier and the answer string of the source. In computing the similarity indices for a pair of examinees, one does not specify a source or a copier and computes the expected number of answer matches conditional on the estimated abilities of the pair of examinees. The answer-copying indices that are most popular include the K-index (e.g., Holland, 1996), the closely related probability of matching incorrect responses (PMIR; e.g., Lewis & Thayer, 1998), the ω index (Wollack, 1997), and the generalized binomial model approach (van der Linden & Sotaridona, 2006). The popular similarity indices include the M4 index (Maynes, 2014) for detecting pairs of examinees one or more of whom may have been involved in test fraud, the index of Wollack and Maynes (2017) that employs cluster-analysis methods (e.g., Everitt, Landau, Leese, & Stahl, 2011) in conjunction with the M4 index (Maynes, 2014) to detect groups of test-takers engaged in collusion, and the analysis for matching of response patterns based on a specialized IRT model (Haberman & Lee, 2017).
- *Person-fit Analysis:* In person-fit analysis, an investigator tries to determine whether the item scores of an examinee follows an item response theory (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—these are referred to as person-fit statistics. Meijer and Sijtsma (2001) provided a review of several of these indices. Popular person-fit statistics include the caution indices (Tatsuoka, 1984), the H^T statistic (Sijtsma, 1986), and the I_z^* statistic (Snijders, 2001; Sinharay, 2016).

- *Use of Response-time Models:* Response-time models or RTMs (e.g., van der Linden, 2016) are latent variable models that are fitted to data on examinee's response times on educational or psychological tests. While such models have a long history of use in psychology (e.g., Luce, 1986), they are only recently becoming popular tools in educational measurement, primarily propelled by an increasing popularity of computerized tests that allow the recording of response times. There is also an upswing in research on the use of RTMs to detect test fraud. Researchers such as van der Linden and Guo (2008), Marianti, Fox, Avetisyan, Veldkamp, and Tijmstra (2014), Fox and Marianti (2017), Sinharay and Johnson (2020), and Wang et al. (2018) suggested various approaches to detect test fraud using RTMs.

Note that Olson and Fremer (2013) also provided a list of five types of statistical approaches, but the classification provided by Wollack and Schoenig (2018) is slightly different from and a refinement of that of Olson and Fremer (2013). While most of the abovementioned approaches are designed to detect individual examinees who may have committed test fraud, several approaches have been suggested for detecting test fraud at group level (e.g., Sinharay, 2018; Wollack & Eckerly, 2017; Wollack & Maynes, 2017). Finally, some fraud-detection methods such as biometric learner verification (www.ets.org/accelerate/ai-portfolio/biometric-learner-verification/) are not covered by any of the above types of approaches.

2 The Models

This section is intended to provide brief descriptions of the statistical/psychometric models that form the basis of the approaches for detecting test fraud that are discussed in this paper. All of these models are latent variable models and include latent variables that measure the examinees' ability or speed.

2.1 Item Response Theory Models

Item response theory (IRT) models (e.g., Hambleton, 1989; Yen & Fitzpatrick, 2006) are a family of statistical/psychometric models used to analyze item response data.

These models are employed in many educational measurement applications including test construction, adaptive test administration, test scoring, and score reporting. The primary reason for the popularity of IRT models is that the model allows for the estimation of individual item locations (difficulties) and examinee abilities separately, but on the same scale.

The core of an IRT model is the mathematical expression for $P_{ic}(\theta) = P(x_i = c | \theta, \boldsymbol{\delta}_i)$, the probability of observing a particular item score given an examinee's ability and the item parameters, where θ denotes the examinee ability¹, x_i is the score of the examinee on item i , c denotes a specific value of x_i , and $\boldsymbol{\delta}_i$ is the collection of parameters of item i .

For example, for the 3-parameter logistic model (3PLM; Birnbaum, 1968, p. 405), x_i denotes the binary score on item i , and

$$P_{i1}(\theta) = P_i(\theta) = P(x_i = 1 | \theta) = c_i + (1 - c_i) \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]},$$

where a_i , b_i , and c_i respectively are the slope, difficulty, and guessing parameters of item i . The Rasch model (Rasch, 1960) is a special case of the 3PLM with c_i 's being equal to 0 and the a_i 's being the same over all the items. The 2-parameter logistic model (2PLM; Birnbaum, 1968, p. 400) is a special case of the 3PLM with c_i 's being equal to 0. Because θ is a latent variable, IRT models are a type of latent variable models (e.g., Bartholomew et al., 2011). One popular IRT model in detecting test fraud is the nominal response model (NRM; Bock, 1972), which has been heavily used in erasure analysis, copying-detection, and similarity analyses. Under the NRM, examinees do not receive any scores on the items, but choose various response options on them and it is assumed that $P_{ik}(\theta)$, the probability that an examinee of ability θ chooses the response option k on item i , is given by

$$p_{ik}(\theta) = \frac{\exp[\zeta_{ik} + \lambda_{ik}\theta]}{\sum_m \exp[\zeta_{im} + \lambda_{im}\theta]},$$

where ζ_{im} and λ_{im} respectively are the intercept and slope parameters for response option m of item i .

The common IRT models involve the assumption of conditional independence, which implies that conditional on θ , x_i , $i = 1, 2, \dots, I$ are independent, where I is the number of items on a test. Because of the assumption, the likelihood of an examinee on an I -item test under the 2PLM is given by

$$L(\mathbf{x} | \theta) = \prod_{i=1}^I P_i(\theta)^{x_i} (1 - P_i(\theta))^{1-x_i} = \prod_{i=1}^I \frac{\exp[x_i a_i (\theta - b_i)]}{1 + \exp[a_i (\theta - b_i)]},$$

where $\mathbf{x} = (x_1, x_2, \dots, x_I)$. The marginal likelihood of an examinee is the weighted average of $L(\mathbf{x} | \theta)$ with the weight proportional to a population distribution $p(\theta)$ and is typically defined as

$$L(\mathbf{x}) = \int_{\theta} L(\mathbf{x} | \theta) p(\theta) d\theta.$$

The marginal likelihood of a sample of examinees is defined as the product of the marginal likelihood over all examinees. The item parameters are typically estimated by maximizing the marginal likelihood of a sample of examinees. The details of such estimation, which typically involves the EM algorithm (e.g., Dempster, Laird, & Rubin, 1977), can be found in, for example, e.g., Baker and Kim (2004).

In detection of test fraud using IRT models, investigators typically have access to the item parameter estimates computed from a previous estimation/calibration (mostly not affected by test fraud) and perform the required computations using the examinee likelihood $L(\mathbf{x} | \theta)$ so that θ is treated as the only unknown parameter (e.g., Snijders, 2001). Descriptions of approaches for the frequentist estimation of θ given item parameters, which involve an iterative algorithm such as the EM algorithm or Newton-Raphson algorithm (e.g., Thisted, 1988, p. 164), can be found in, for example, Baker and Kim (2004). One can also compute the posterior mean of θ , $\int_{\theta} \theta L(\mathbf{x} | \theta) p(\theta) d\theta$, using numerical integration. Several publicly available software packages including mirt (Chalmers, 2012) and ltm (Rizopoulos, 2006) are available to estimate the IRT item parameters and examinee abilities given the item parameters.

2.2 The Lognormal Model for Response Times

Let t_i denote the response time of a randomly chosen examinee on item i , where $i = 1, 2, \dots, I$. Let us define

¹ θ is assumed to be a scalar quantity in this paper. There exist IRT models that involve vector-valued examinee ability, but those models will not be considered in this paper.

$$y_i = \log(t_i).$$

Under the lognormal model for response times (LNMRT; van der Linden, 2006), y_i , $i = 1, 2, \dots, I$, are independent given τ and

$$y_i|\tau \sim \mathcal{N}\left(\beta_i - \tau, \frac{1}{\alpha_i^2}\right), \quad (1)$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 . That is, the probability density function of y_i is given by

$$f(y_i|\tau, \alpha_i, \beta_i) = \frac{\alpha_i}{\sqrt{2\pi}} e^{-\frac{1}{2}\alpha_i^2(y_i - \beta_i + \tau)^2}. \quad (2)$$

The parameter τ is the examinee's speed parameter; a larger value of the parameter results in smaller expected response times on all the items for the examinee. The parameter β_i is the time-intensity parameter for item i ; a larger value of the parameter results in larger expected response times for all examinees on the item. The parameter α_i is the discrimination parameter for item i ; a larger value of the parameter leads to more information on and hence smaller standard error of the examinee speed parameters. To estimate the item parameters of the LNMRT using a marginal maximum likelihood approach or to perform a Bayesian inference on the examinee ability, one assumes a population distribution $g(\tau)$ on τ . The distribution $g(\tau)$ is assumed to be the normal distribution with mean 0 and variance σ^2 in most applications of the LNMRT (see, for example, van der Linden & Guo, 2008).

The assumption of conditional independence is made, which leads to the following expression of the likelihood of the logarithms of response times of an examinee:

$$L(\mathbf{t}|\tau) = \prod_i \frac{\alpha_i}{\sqrt{2\pi}} e^{-\frac{1}{2}\alpha_i^2(y_i - \beta_i + \tau)^2}.$$

The LNMRT is arguably one of the most popular RTMs. The model was considered, either to analyze only the response times, or to analyze the response times and item scores, by, for example, Boughton et al. (2017), Glas and van der Linden (2010), Qian, Staniewska, Reckase, and Woo (2016), van der Linden (2007), van der Linden (2009), van der Linden (2016), and van der Linden and Guo (2008). Bolsinova and Tijmstra (2018, p. 13) commented that the LNMRT is used in most applications of RTMs.

A Gibbs sampler (e.g., Gelman et al., 2014, p. 276) was suggested by van der Linden (2006) to estimate the item parameters of the LNMRT. That approach has been used in most applications of the model and the R package LNIRT (Fox, Klein Entink, & Klotzke, 2017) can be used to implement the Gibbs sampler. Glas and van der Linden (2010) suggested an approach to compute the marginal maximum likelihood estimates (MMLEs) of the item parameters when the LNMRT is used along with the three-parameter logistic model (3PLM) to jointly analyze both response times and item scores. Finger and Chee (2009) showed how one can use factor analysis to obtain the MMLEs of the item parameters of the LNMRT when it is used as a stand-alone model, as in van der Linden (2006). The R package *lavaan* (Rosseel, 2012), which is used to perform factor analysis and structural equation modeling (SEM), can be used to estimate the item parameters of the LNMRT, as shown by Sinharay and van Rijn (2020).

It was shown by van der Linden (2006) that given α_i 's and β_i 's, the MLE of the person speed parameter τ for the LNMRT can be obtained as

$$\hat{\tau} = \frac{\sum_i \alpha_i^2(\beta_i - y_i)}{\sum_i \alpha_i^2}. \quad (3)$$

An expression for the posterior mean of τ given the α_i 's, β_i 's and y_i 's can be found in van der Linden (2007).

2.3 The Hierarchical Model for Item Scores and Response Times

The hierarchical/joint modeling approach of van der Linden (2007) involves the application of a model for response times in combination with an IRT model for the item scores. In the approach, one typically assumes that

- the response time follows the lognormal model (LNMRT; van der Linden, 2006) defined above.
- the item scores follow an IRT model; the flexibility of the approach allows the use of any IRT model such as the 2PLM described above.

One also assumes a suitable population distribution, typically a bivariate normal distribution with means equal to 0, for the vector of examinee parameters, $(\tau, \theta)^t$. Given that both the ability parameter θ and the speed parameter τ are latent variables, the hierarchical model for item scores and response times is a latent variable model that includes two latent variables.

Klein Entink, Fox, and van der Linden (2009) and van der Linden (2007) suggested Bayesian approaches to estimate the item parameters of the hierarchical model using the Markov chain Monte Carlo (MCMC) algorithm (e.g., Gelman et al., 2014). The estimation approach is implemented in the R package LNIRT (Fox et al., 2017). Glas and van der Linden (2010) suggested an expectation-maximization (EM) algorithm to compute the marginal maximum likelihood estimates (MMLEs) of the item parameters of the hierarchical model of van der Linden (2007); van Rijn and Ali (2017) provided further details (such as expressions of the first and second derivatives of the marginal likelihood function) on the EM algorithm for the hierarchical model involving the LNMRT and the 2PLM. Molenaar, Tuerlinckx, and van der Maas (2015) provided Mplus codes for computing the MMLEs of the item parameters of the hierarchical model.

Once the item parameters have been estimated, the joint distribution of the item scores and the response times of an examinee given the item parameters can be expressed as the product of the distribution of the item scores and that of the response times (e.g., van der Linden, 2007). Therefore, to compute the joint maximum likelihood estimates (MLEs) of τ and θ for a person given the item parameters, one can compute the MLEs separately—the one for τ only using the response times and the LNMRT and the one for θ only using the item scores and the IRT model.

Note that there exist other joint models for item scores and response times some of which are closely related to those used by cognitive psychologists. For example, the diffusion model (e.g., Ratcliff, 1978) and the race model (e.g., Townsend & Ashby, 1978) are popular in cognitive psychology for modeling cognitive processes based on response time and response accuracy; van der Maas, Molenaar, Maris, Kievit, and Borsboom (2011) and Ranger, Kuhn, and Gaviria (2014) respectively developed versions of these two models for use with item scores and response times. However, these models will not be considered in this paper.

3 The Approaches Based on Latent Variable Models Using Both Item Scores and Response Times to Detect Test Fraud

The approaches for detecting test fraud that are based on latent-variable models and both item scores and response

times include: (a) a Bayesian person-fit approach to detect test fraud (Fox & Marianti, 2017), (b) the use of standardized residuals to detect test fraud (van der Linden & Guo, 2008), (c) the use of mixture models to detect test fraud (Wang et al., 2018), (d) the use of data mining methods to detect test fraud (Man et al., 2019), and (e) the use of the χ^2 statistic to detect test fraud (Sinharay & Johnson, 2020). Given that these approaches are based on RTMs, they all belong to the last type of methods listed in Table 1. In addition, the first approach belongs to the fourth type and the last approach belongs to the first type of method listed in Table 1.

3.1 A Bayesian Person-fit Approach

Mariani et al. (2014) suggested a person-fit statistic that is based on response times and the LNMRT (van der Linden, 2006) and is given by

$$l^t = \sum_i \alpha_i^2 (y_i - \beta_i + \tau)^2. \quad (4)$$

Given that $E(y_i) = \beta_i - \tau$ (Equation 1), Equation 4 implies that l^t is a sum of squares of standardized residuals—so large values of the statistic indicate an aberrant response-time pattern often caused by preknowledge of some items. Marianti et al. (2014) also described a Bayesian approach that involves first the use of an MCMC algorithm to fit the abovementioned hierarchical model to the item scores and response times followed by the computation of l^t in each iteration of the algorithm. The proportion of values of l^t that are larger than the 95th percentile of the χ^2 distribution is an estimate of the posterior probability of an aberrant response-time pattern, often due to test fraud. Fox and Marianti (2017) suggested a Bayesian approach to estimate the posterior probability of an aberrant item-score pattern using the l_z statistic (Drasgow, Levine, & Williams, 1985) that is given by

$$l_z = \frac{\ell(\mathbf{x}|\theta) - E(\ell(\mathbf{x}|\theta))}{\sqrt{\text{Var}(\ell(\mathbf{x}|\theta))}},$$

where $\ell(\mathbf{x}|\theta) = \log(L(\mathbf{x}|\theta))$. The statistic l_z is the standardized log likelihood of the item scores of an examinee—so an extreme value of the statistic indicates an aberrant score pattern caused by item preknowledge, guessing etc. Finally, Fox and Marianti (2017) suggested a Bayesian approach using both l^t and l_z to estimate the

posterior probability of both an aberrant item-score pattern and aberrant response-time pattern. The posterior probability is expected to be large (for example, larger than 0.95 or 0.99) for aberrant response patterns. The underlying model in the approach of Fox and Maranti (2017) is the hierarchical model for item scores and response times (van der Linden, 2007). The approach of Fox and Maranti (2017) is designed to detect a variety of aberrant responses including preknowledge. However, the power of person-fit approaches to detect test fraud has been found to be small (e.g., Sinharay & Johnson, 2020). Further, the approach of Fox and Maranti (2017) was found to have inflated Type I error rate and small power in detecting test fraud in the simulations of Sinharay and Johnson (2020).

3.2 The Use of Standardized Residuals

A Bayesian approach for detecting aberrant response times was suggested by van der Linden and Guo (2008). They showed that the posterior distribution of the predicted value of an examinee's log-response time on item i conditional on $\mathbf{y}_{-i} = (y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_I)$ and also on $\mathbf{x}_{-i} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_I)$ under the hierarchical model of van der Linden (2007) is approximately normal. Then, van der Linden and Guo (2008) defined the standardized residual for the examinee on item i , e_i , as

$$e_i = \frac{y_i - E(y_i | \mathbf{y}_{-i}, \mathbf{x}_{-i})}{\sqrt{\text{Var}(y_i | \mathbf{y}_{-i}, \mathbf{x}_{-i})}}. \quad (5)$$

The response time for the examinee on item i is concluded as aberrant at $\alpha\%$ level if the absolute value of e_i is larger than the $100(1-\alpha)$ -th percentile of the standard normal distribution. This approach is designed to detect a variety of aberrant responses. For example, the approach can be used to flag for possible item preknowledge the examinees with several statistically significant and negative e_i 's, as was performed in Boughton et al. (2017, p. 181), Qian et al. (2016), and van der Linden and Guo (2008).

3.3 Use of Mixture Models

Lee and Wollack (2020) and Wang et al. (2018) suggested using mixture hierarchical IRT models, which are fitted using Bayesian estimation approaches, to detect aberrant item scores and response times. These models include a 0/1 aberrance indicator Δ_{ij} (that is unknown and is estimated

from the data) for each examinee-item combination, with $\Delta_{ij} = 1$ indicating an aberrant behavior on item j by examinee i . If examinee i does not show aberrant behavior on item j , then the item score and response time for that examinee-item combination is obtained by the hierarchical model of van der Linden (2007). If examinee i shows aberrant behavior on item j , then the probability of a correct answer for that examinee-item combination does not depend on the examinee ability and is the same (d_j) for all examinees on the item; and the response time for that examinee-item combination follows a lognormal distribution with a mean (μ_c) and variance (σ_c^2), each of which is constant over all items and all examinees. The mixture model is fitted to the data using an MCMC algorithm and the estimated posterior probability of Δ_{ij} being equal to 1 can be used to determine whether the response time and item score on item j by examinee i are aberrant. An examinee with too many aberrant item-examinee combinations may indicate possible test fraud.

3.4 Use of Data-mining Methods

Man et al. (2019) suggested two approaches based on data-mining methods (e.g., Hastie, Tibshirani, & Friedman, 2009) and latent-variable models to detect test fraud. In both of these approaches, one first fits an IRT model and a RTM to the item scores and response times to compute the values of several person-fit statistics using item scores and using response times. Then, in the first approach, one applies an unsupervised learning methods such as the K-means clustering method (e.g., Hastie et al., 2009, p. 460) using the person-fit statistics, item scores, and response times, and any other features that may be related to test fraud, to divide the examinees into multiple clusters and attempts to interpret the groups. If the features are powerful predictors of test fraud, then the clusters will correspond to those committing various types of test fraud and those not committing test fraud. In the second approach, which is applicable only when the investigator has available a set of examinees who are known to have committed test fraud, one applies a supervised learning method such as the random forests (e.g., Hastie et al., 2009, p. 587) using the abovementioned features to predict the examinees who committed test fraud. If the features are powerful predictors of test fraud, then the second approach will lead to a powerful (nonparametric) prediction model. Man et al. (2019) applied both approaches to a data set from Cizek and Wollack (2017) that involved real test fraud and found the second approach (based on supervised learning) to be more powerful in detecting test fraud.

3.5 The Use of the $\bar{\chi}^2$ Statistic

Sinharay and Johnson (2020) suggested an approach that is based on the hierarchical model of van der Linden (2007) and can be applied when the investigator is interested in the types of test fraud that lead to better and/or faster performance on a known set of items. For example, real data sets have been used to demonstrate that in cases of item compromise and preknowledge, the examinees who benefit from item preknowledge perform better (e.g., Sinharay, 2017; Smith & Davis-Becker, 2011) and faster (e.g., Kasli & Zopluoglu, 2018; Smith & Davis-Becker, 2011) on the compromised items than on the non-compromised items.

Let C and $\bar{C}$ respectively denote the set of compromised items and non-compromised items. Let $\mathbf{x}$, $\mathbf{x}_C$, and $\mathbf{x}_{\bar{C}}$ respectively denote the collection of the scores of a randomly chosen examinee on all items, on the items in C and on the items in $\bar{C}$, respectively. Thus, for example, $\mathbf{x}_C = \{x_i, i \in C\}$. Similarly, let $\mathbf{y}$, $\mathbf{y}_C$, and $\mathbf{y}_{\bar{C}}$ respectively denote the collection of the logarithms of response times of the examinee on all items, the items in C and the items in $\bar{C}$. For the examinee, let us denote the true ability based on the whole test, C , and $\bar{C}$ as θ , θ_C , and $\theta_{\bar{C}}$, respectively. Let us denote the MLEs of these parameters as $\hat{\theta}$, $\hat{\theta}_C$, and $\hat{\theta}_{\bar{C}}$, respectively. Let τ_C and $\tau_{\bar{C}}$ respectively denote the examinee's true speed parameters on the compromised and non-compromised items, respectively, and let $\hat{\tau}_C$ and $\hat{\tau}_{\bar{C}}$ denote their MLEs. Let $\hat{\tau}$ denote the MLE of the examinee's true speed parameter based on all the I items on the test.

Sinharay and Johnson (2020) suggested that one can detect whether an examinee performed better and/or faster on the compromised items by testing the null hypothesis $H_0: \tau_C = \tau_{\bar{C}}$ and $\theta_C = \theta_{\bar{C}}$ versus the alternative hypothesis $H_1: \tau_C > \tau_{\bar{C}}$ or $\theta_C > \theta_{\bar{C}}$. The rejection of the null hypothesis may indicate possible item preknowledge. For this hypothesis-testing problem, Sinharay and Johnson (2020) recommended the use of the constrained likelihood ratio test statistic

$$\Lambda_{ST}^* = \Lambda_S I(\hat{\theta}_C \geq \hat{\theta}_{\bar{C}}) + \Lambda_T I(\hat{\tau}_C \geq \hat{\tau}_{\bar{C}}), \quad (6)$$

where

$$\Lambda_S = 2 \left\{ \sum_{i \in C} \log P_i(x_i | \hat{\theta}_C) + \sum_{i \in \bar{C}} \log P_i(x_i | \hat{\theta}_{\bar{C}}) - \sum_{i=1}^I \log P_i(x_i | \hat{\theta}) \right\},$$

$$\Lambda_T = \hat{\tau}_C^2 \sum_{i \in C} \alpha_i^2 + \hat{\tau}_{\bar{C}}^2 \sum_{i \in \bar{C}} \alpha_i^2 - \hat{\tau}^2 \sum_{i=1}^I \alpha_i^2,$$

and $I(\hat{\theta}_C \geq \hat{\theta}_{\bar{C}})$ is the indicator function taking the value of 1 if $\hat{\theta}_C \geq \hat{\theta}_{\bar{C}}$ and 0 otherwise.

Sinharay and Johnson (2020) also proved that the cumulative density function (CDF) of Λ_{ST}^* for large C and $\bar{C}$ and under the null hypothesis of no item preknowledge is given by

$$P(\Lambda_{ST}^* \leq \lambda) = \frac{1}{4}P(\chi_2^2 \leq \lambda) + \frac{1}{2}P(\chi_1^2 \leq \lambda) + \frac{1}{4}I\{\lambda \geq 0\}, \quad (7)$$

where, for example, χ_2^2 denotes a χ^2 random variable with two degrees of freedom. The CDF shown in Equation 7 corresponds to a distribution that is referred to as the chi-bar-square ($\bar{\chi}^2$) distribution (e.g., Dykstra, 1991; Silvapulle & Sen, 2001) that is popular in constrained statistical inference. Consequently, for $\lambda > 0$, the p-value for the test statistic Λ_{ST}^* is calculated as

$$P(\Lambda_{ST}^* > \lambda) = \frac{1}{4}P(\chi_2^2 > \lambda) + \frac{1}{2}P(\chi_1^2 > \lambda). \quad (8)$$

The critical values for Λ_{ST}^* for significance levels of 0.001, 0.01, and 0.05 are 11.76, 7.289, and 4.231, respectively.

If the alternative hypothesis is true, which corresponds to an examinee having item preknowledge, one or both of Λ_S and Λ_T will have a large positive value and, consequently, Λ_{ST}^* will have a large and positive value.

4 Real Data Example

Item scores and response times on one form of a non-adaptive information technology certification examination were available for 1,992 examinees. The examinees are sorted in chronological order, so that the first examinee in the data set took the test before any other examinee. The data set was discussed in Eckerly (2020) and Eckerly, Smith, and Lee (2018). The examination comprises 60 multiple choice and dichotomously scored items. All of these 60 items, along with their answer keys, were found on an internet website, but the answer keys provided on the website were correct for 24 items and incorrect for 36 items. Eckerly et al. (2018) computed the values of several statistics, all using item scores, for detecting test fraud for these data and found evidence of various types of test fraud including potential collusion between groups of examinees who were conjectured to have developed their own answer keys.

The top left panel of Figure 1 shows a scatter plot of the raw score (along the X axis) versus the total time (in minutes; Y axis) of the examinees.² The correlation coefficient between the raw score and total time is -0.20, which indicates a slight tendency of more able examinees answering faster. Data sets showing a similar pattern include one from the computerized uniform CPA examination (e.g., van der Linden, 2009).

The item parameters for the hierarchical model of van der Linden (2007) were estimated from the data set under the assumption that the 2PLM fits the item scores and the LNMRT fits the response times. The values of Λ_{ST}^* were computed from the data set using the estimated item parameters. The set of 24 items with the correct answer keys were considered as the compromised items (the set C) in computing Λ_{ST}^* . The residuals suggested by van der Linden and Guo (2008) were also computed. The appendix includes R code for fitting the hierarchical model, computing Λ_{ST}^* , and computing the residuals suggested by van der Linden and Guo (2008).³

This data set presents somewhat of a unique case of preknowledge, that is, the type of preknowledge for this data set is somewhat different from that in other data sets (e.g., those analyzed by Sinharay & Johnson, 2020). Whereas some of the items in the data sets analyzed by Sinharay and Johnson (2020) were compromised, some were not; in contrast, all the items were compromised for this data set, albeit some with incorrect answer keys. Examinees who trust the answer keys and do not spend any efforts themselves are expected to perform much better on the items with correct answer keys than on those with incorrect answer keys. It is also possible that some examinees will notice that some answer keys are incorrect, spend some time on those items, and yet incorrectly answer those items, causing them to perform worse and slower on the items with incorrect answer keys. While analyzing the same data set, Eckerly (2020) found a cluster of examinees whose responses corresponded closely to the key posted to the website, whereas examinees in the other clusters seemed to recognize that the posted key was incorrect and deviated from the posted key to achieve a higher score.

According to Eckerly (2020), the testing vendors had reason to believe that several response patterns among

Examinees 601-1992 were contaminated by preknowledge of the examination items. Figure 1 shows plots of the raw score (top right panel), total time in minutes (bottom left panel), and the Λ_{ST}^* (bottom right panel) for all the examinees in the data set; the examinee number is shown along the X-axis in each of these panels. Two of these panels include solid lines showing the lowess-smoothed mean⁴ (e.g., Cleveland, 1981) of the raw score and total time, respectively—these lines were created using the R function *loess* (e.g., R Core Team, 2019). The figure shows a tendency of the raw score increasing and total time decreasing with an increase in examinee number, which agrees with the assertion of Eckerly (2020) regarding possible preknowledge among Examinees 601-1992. The bottom right panel shows a dashed horizontal line showing the critical value (11.76) at significance level of 0.001 for Λ_{ST}^* . A conservative level of 0.001 is used here because several experts (e.g., Wollack et al., 2015) recommended the use of conservative hypothesis testing in detection of test fraud. The bottom right panel indicates that Λ_{ST}^* is significant at level 0.001 for only two examinees among the first 600 examinees (0.3%), but for 19 among the last 1,392 examinees (1.4%, which is considerably larger than the nominal level of 0.1%), and points to potential fraud by several of the latter examinees. There exist approaches for detection of preknowledge only based on item scores (e.g., Sinharay, 2017) and only based on response times (e.g., Sinharay, 2020)—both of these types of approaches were found to flag fewer examinees compared to Λ_{ST}^* .⁵

Figure 2 includes a comparison of the response times versus average response times for the whole sample, of two examinees (referred to as Examinees 1 and 2) for whom Λ_{ST}^* was very large and significant at level of 0.001 and one additional randomly chosen examinee (Examinee 3) for whom Λ_{ST}^* was not significant. In each panel, the logarithms of the response times (in minutes) of an examinee on the test items are shown along the Y-axis and the logarithms of the average response time (in minutes) of the items over the whole sample are shown along the X-axis. A plus sign and a minus sign respectively correspond to items that the examinee answered correctly and incorrectly. The black plus or minus signs correspond

² The other three panels of the figure will be discussed below.

³ While the data used in this paper cannot be shared publicly, those interested can try the code on the credentialing data sets discussed in Cizek and Wollack (2017, p. 14)—the data sets are available from James Wollack upon request, include both item scores and response times, and involve known compromised items.

⁴ A lowess-smoothed mean is the estimated value from a particular type of nonparametric regression (referred to as the *locally weighted scatter-plot smoothing*) of the dependent variable on the independent variable.

⁵ While this finding itself does not mean that Λ_{ST}^* is more powerful than these other statistics, simulation studies in Sinharay and Johnson (2020) showed that Λ_{ST}^* is more powerful than approaches that utilize only item scores or only response times when preknowledge leads to better and faster performance on compromised items.

Figure 1: Plot of Sum Score, Average Time, and Λ_{ST}^* Versus Examinee Number.

to the items with correct answer keys (on the website) while the gray plus or minus signs correspond to the items with incorrect answer keys. A diagonal line is provided in each panel for convenience. The logarithms are plotted to prevent the outliers leading to too much vacant space in the plots. Examinee 1 correctly answered all items in C (but incorrectly answered several items in $\bar{C}$), as clear from the black symbols all being plus signs, and answered the former types of items faster in comparison to the latter types of items. Examinee 2 correctly answered all but four items in C, but incorrectly answered all but two items in $\bar{C}$; this examinee did not perform faster on the items in C. Thus, the item scores and response times of Examinee 2 look like those of examinees who trusted the keys on the website and did not try to solve the items themselves

(which is why almost all gray symbols are minuses). Examinee 3 did not perform any better or faster on the items in C; most of the points (both black and gray) for the examinee fall below the diagonal line, which indicates that the examinee is faster than the other examinees on average.

The two panels of Figure 3 show the standardized residuals e_i 's (shown in Equation 5) for all the items for Examinees 1 and 2 of Figure 2. Note that the e_i 's are roughly distributed as standard normal variables if the hierarchical model fits the data (van der Linden & Guo, 2008)—so e_i 's larger than 2.0 in absolute values are statistically significant at the level of 0.05. The item number is shown along the horizontal axis and the e_i 's are shown using a bar along the vertical axis. The items are

Figure 2: Response times of three examinees.

sorted so that the first 24 items (to the left of the vertical dashed line) are those with correct answer keys (C) and the last 36 with incorrect answer keys ($\bar{C}$). Positive e_i 's are indicated by a bar above the 0-line and negative e_i 's are indicated by a bar below the 0-line. A bar is gray if the corresponding answer is incorrect and black if the corresponding answer is correct. Horizontal (dotted) lines are shown at 2 and -2 for convenience of evaluating whether a specific e_i is statistically significant. For Examinee 1, all answers are correct (hence all the bars to the left of the vertical dashed line are black) and all but six residuals are negative (one of them being significantly so) among the items in C. The pattern is different for the items in $\bar{C}$ for the examinee—there are nine incorrect answers and a mix of positive and negative residuals (eight of which are statistically significant). There is no surprise then that the value of Λ_{ST}^* was quite large (17.4) for the examinee. Only one residual is statistically significant for Examinee 2 (so that the response times are not aberrant for the examinee), who performed much better on the items in C (only four incorrect answers) in comparison to those in $\bar{C}$ (only two correct answers). Figure 3 shows how graphical plots of standardized residuals can provide useful information in addition to any other test statistics.

5 Conclusions and Recommendations

This paper begins with a brief review of the five types of statistical/psychometric approaches (identified by

Wollack & Schoenig, 2018) to detect test fraud. More detailed reviews of the recent approaches for detecting test fraud using both item scores and response times are then included—these approaches are all based on latent variable models (e.g., Bartholomew et al., 2011). The approaches increase the chance of detection of test fraud by utilizing response times in addition to the item scores, as was demonstrated by, for example, Sinharay and Johnson (2020). As more information such as keystroke logs (e.g., Leijten & Van Waes, 2013) and eye-tracking data (e.g., Man & Harring, 2019) become available for computerized tests, interest is bound to grow in more approaches based on latent variable models to detect test fraud.

While the existence of the abovementioned approaches is encouraging, existing approaches have several limitations and, consequently, there remain several questions, mainly on how these approaches (and other approaches for detection of test fraud in general) should be used in practice. For example, there are no clear guidelines on how many approaches and which ones should be used for a specific data set. Wollack and Cizek (2017, p. 397) recommended the collection of evidence of cheating from multiple distinctly different indices. They also recommended controlling for false positives in the context of test fraud where several statistical tests are typically performed on many examinees, which could lead to tens of thousands of statistical tests for a data set that is as large as the real data set discussed earlier; one way to limit the number of false positives is to choose a critical value that adjusts for multiple comparisons by controlling

Figure 3: Estimated standardized residuals of log response times of two examinees.

the family-wise error rate (using, for example, a Bonferroni correction) or controlling the false discovery rate (using the procedure of Benjamini & Hochberg, 1995). Experts such as Buss and Novick (1980), Holland (1996), and Hanson, Harris, and Brennan (1987) recommended that statistical methods of detection should generally not form the sole basis for a judgment that an examinee cheated (or that an examinees scores and/or response times are sufficiently questionable to justify non-reporting) in the absence of corroboration from other types of evidence.

There are several topics that could benefit from further research. First, more applications of the abovementioned approaches to real data sets, especially that that involve actual test fraud, will be helpful. Second, there is a lack of research on the use of keystroke logs (e.g., Leijten & Van Waes, 2013) and eye-tracking data (e.g., Man & Haring, 2019), possibly in addition to item scores and response times, to detect test fraud. Third, there is a severe lack of approaches utilizing item scores and response times for detecting test fraud at an aggregate level (that is, at the

level of classes or schools that the examinees belong to)—so more research on this area will be useful. Fourth, while the hierarchical model of van der Linden (2007) was used in most of the approaches discussed in this paper, research on the use of other models for item scores and response times (for example, those by Maris & van der Maas, 2012; van der Maas et al., 2011) to detect test fraud would be useful. Fifth, the item parameters are assumed known in the application of the approaches discussed here and it is possible to explore approaches, possibly Bayesian, to account for the uncertainty of the item parameters in using the approaches. Finally, more research on examining the consequences of misfit of the latent-variable model on the properties of the abovementioned approaches will be helpful.

References

- 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.
- Baker, F. B., & Kim, H. S. (2004). *Item response theory: Parameter estimation techniques (2nd ed.)*. New York, NY: Marcel Dekker.
- Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis*. New York, NY: John Wiley & Sons, Ltd.
- 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.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.
- Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, 71, 13–38.
- Boughton, K., Smith, J., & Ren, H. (2017). Using response time data to detect compromised items and/or people. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 177–190). Washington, DC: Routledge.
- 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.
- Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35, 54.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- 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. (doi=10.1111/j.2044-8317.1985.tb00817.x)
- Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, 9, 47–64. (doi=10.1207/s15324818ame09015)
- Dykstra, R. (1991). Asymptotic normality for chi-bar-square distributions. *Canadian Journal of Statistics*, 19, 297–306. (doi=10.2307/3315395)
- Eckerly, C. (2020). *Answer similarity analysis at the group level*. Manuscript under review.
- Eckerly, C., Smith, R., & Lee, Y. (2018, October). *An introduction to item preknowledge detection with real data applications*. Paper presented at the Conference on Test Security, Park City, UT.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*. John Wiley & Sons, Ltd.
- 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.
- Finger, M. S., & Chee, C. S. (2009, April). *Response-time model estimation via confirmatory factor analysis*. Paper presented at the Annual meeting of the National Council of Measurement in Education, San Diego, CA.
- Fox, J.-P., Klein Entink, R. H., & Klotzke, K. (2017). *LNIRT: Lognormal response time item response theory models*. (R package version 0.2.0)
- Fox, J.-P., & Mariani, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, 54, 243–262. (doi=10.1111/jedm.12143)
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. New York, NY: Chapman and Hall.
- Glas, C. A. W., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, 63, 603–626. (doi=10.1348/000711009x481360)
- 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.
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In *Educational measurement* (pp. 143–200). New York, NY: Macmillan.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Academic Publishers.

- 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.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.
- 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.
- Kasli, M., & Zopluoglu, C. (2018, October). *Do people with item pre-knowledge really respond faster to items they had prior access?* An empirical investigation. Paper presented at the Conference on Test Security, Park City, UT.
- Kingston, N., & Clark, A. (2014). *Test fraud: Statistical detection and methodology*. New York, NY: Routledge.
- Klein Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74, 21–48. (doi=10.1007/s11336-008-9075-y)
- Lee, S. Y., & Wollack, J. (2020). Concurrent use of response time and response accuracy for detecting examinees with item preknowledge. In R. Feinberg & M. Margolis (Eds.), *Integrating timing considerations to improve testing practices*. New York, NY: Routledge.
- Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, 30, 358–392.
- 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.
- Luce, R. D. (1986). *Response times*. New York, NY: Oxford University Press.
- Man, K., & Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items. *Educational and Psychological Measurement*, 79, 617–635.
- Man, K., Harring, J. R., & Sinharay, S. (2019). Use of data mining methods to detect test fraud. *Journal of Educational Measurement*, 56, 251–279.
- 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. (doi=10.3102/1076998614559412)
- Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, 77, 615–633.
- Maynes, D. (2013). Educator cheating and the statistical detection of group-based test security threats. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 173–199). New York, NY: Routledge.
- Maynes, D. (2014). Detection of non-independent test-taking by similarity analysis. In
- N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 53–82). New York, NY: Routledge.
- McLeod, L. D., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, 27, 121–137. (doi=10.1177/0146621602250534)
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107–135.
- Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, 50 (1), 56–74.
- 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.
- Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, 35 (1), 38–47. (doi=10.1111/emip.12102)
- R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria.
- Ranger, J., Kuhn, J.-T., & Gaviria, J.-L. (2014). A race model for responses and response times in tests. *Psychometrika*, 80, 791–810.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17 (5), 1–25.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48 (2), 1–36.
- Sijtsma, K. (1986). A coefficient of deviant response patterns. *Kwantitative Methoden*, 7, 131–145.
- Silvapulle, M. J., & Sen, P. K. (2001). *Constrained statistical inference: Order, inequality, and shape constraints*. New York, NY: John Wiley & Sons, Inc.
- Sinharay, S. (2016). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*, 81, 992–1013.
- Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, 42, 46–68. (doi=10.3102/1076998616673872)
- Sinharay, S. (2018). A new person-fit statistic for the lognormal model for response times. *Journal of Educational Measurement*, 55, 457–476.
- Sinharay, S. (2020). Detection of item preknowledge using response times. *Applied Psychological Measurement*, 44, 376–392.
- 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. (doi=10.1111/jedm.12141)
- Sinharay, S., & Johnson, M. S. (2020). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology*, 73, 397–419.

- Sinharay, S., & van Rijn, P. W. (2020). Assessing fit of the lognormal model for response times. *Journal of Educational and Behavioral Statistics*, 51, 419–440.
- Smith, R. W., & Davis-Becker, S. L. (2011, April). *Detecting suspect examinees: An application of differential person functioning analysis*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.
- Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika*, 66, 331–342.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95–110.
- Thisted, R. A. (1988). *Elements of statistical computing: Numerical computation*. London: Chapman and Hall.
- Townsend, J., & Ashby, F. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), *Cognitive theory*, vol. 3 (pp. 199–239). Hillsdale, NJ: Erlbaum.
- 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. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308. (doi=10.1007/s11336-006-1478-z)
- van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46, 247–272.
- van der Linden, W. J. (2016). Lognormal response-time model. In W. van der Linden (Ed.), *Handbook of item response theory, Volume 1. Models*. Boca Raton, FL: Chapman and Hall/CRC.
- 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 der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, 118, 339–356.
- van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 70, 317–345. (doi=10.1111/bmsp.12101)
- Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, 43, 469–501.
- Wang, X., Liu, Y., & Hambleton, R. K. (2017). Detecting item preknowledge using a predictive checking method. *Applied Psychological Measurement*, 41, 243–263. (doi=10.1177/0146621616687285)
- 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. (2017). 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.
- Wollack, J. A., & Maynes, D. (2017). Detection of test collusion using cluster analysis. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 124–150). Washington, DC: Routledge.
- Wollack, J. A., & Schoenig, R. W. (2018). Cheating. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 260–265). Thousand Oaks, CA: Sage.
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In *Educational measurement* (pp. 111–153). West Port, CT: American Council on Education and Praeger Publishers.

Appendix: R Code for Computing Λ_{ST}^* (Equation 6) and Creating Figure 3

```

library(irtos)
library(lattice)
library(LNIRT)
library(ltm)

# Some R functions that will be needed later
compstat=function(alpha,beta,ltimes)#estimates person speed parameters
{return(rep(sum(alpha*alpha*beta)/sum(alpha*alpha),nmen)-
  ltimes%*(alpha*alpha)/(sum(alpha*alpha)))}
pr2PL=function(t,a,b){return(1/(1+exp(a*(b-t))))}#Computes 2PL probability
logpr=function(y,t,itp) {l=0 #computes examinee likelihood for 2PL model
  for (i in 1:length(y))
    {p=pr2PL(t,itp[i,1],itp[i,2])
    l=l+ifelse(y[i]==1,log(p),log(1-p))}
  return(l)}

Stats=function(y,s2,itparm)#Computes preknowledge statistic from scores
{
  nitem=length(y)
  n2=length(s2)#s2 is the set of compromised items in our context
  n1=nitem-n2
  s1=setdiff(1:nitem,s2)
  itPs1=itparm[s1,] # Item parameters of Item Set 1
  itPs2=itparm[s2,] # Item parameters of Item Set 2
  w1 = wle(y[s1],cbind(itPs1,rep(0,n1)))
  theta1=w1[1] # Ability estimate based on Item Set 1
  L1=logpr(y[s1],theta1,itPs1)# Log-likelihood based on Item Set 1
  w2= wle(y[s2],cbind(itPs2,rep(0,n2)))
  theta2=w2[1] # Ability estimate based on Item Set 2
  L2=logpr(y[s2],theta2,itPs2) # Log-likelihood based on Item Set 2
  w=wle(y,cbind(itparm,rep(0,nitem)))
  theta=w[1] # Ability estimate based on all items
  Lall=logpr(y,theta,itparm) # Log-likelihood based on all items
  LRT=2*(L2+L1-Lall) #The likelihood ratio statistic for 2-sided test
  LRT[LRT<0]=0 # Reset the LRT's with negative values to 0
  Ls=ifelse(theta2>=theta1,sqrt(LRT),-sqrt(LRT)) # Calculate the L-index
  return(Ls)}

# Use the R package LNIRT to estimate the model parameters
nrep=2000
out=LNIRT(RT=ltimes,Y=y,WL=1,XG=nrep)#y: item scores; ltimes: log(response time)
alpha=out$Post.Mean$Sigma2#Time-discrimination parameter
beta=out$Post.Mean$Time.Intensity#Time-intensity parameter
b=out$Post.Mean$Item.Difficulty#Item difficulty parameter
a=out$Post.Mean$Item.Discrimination#Item slope parameter
# Compute the First Component of Equation 6; 'comp'=Set of compromised items

# 'ncomp'=Set of non-compromised items
Ls=NULL
thetas=NULL
itparm=cbind(a,b)
for (i in 1:nrow(scr)) {Ls=c(Ls,Stats(y[i,],comp,itparm))}
# Compute the Second Component of Equation 6
tcomp=compstat(alpha[comp],beta[comp],ltimes[,comp])
tncomp=compstat(alpha[ncomp],beta[ncomp],ltimes[,ncomp])
Lt=(tcomp-tncomp)/sqrt(1/sum((alpha[comp])^2) + 1/sum((alpha[ncomp])^2))
#
JointStat=ifelse(Ls>0,Ls^2,0)+ifelse(Lt>0,Lt^2,0)#This is LambdaST* statistic
#
# Plot of the Residuals for examinee k in the data set
#
tauhat=compstat(alpha,beta,ltimes)
items=1:nitem#nitem is the number of items in the data set
res=alpha*(ltimes[k,]-beta+tauhat[k])/sqrt(1-alpha*alpha/sum(alpha*alpha))
y=scores[k,]
labs<-c("1",rep(" ",8),"10",rep(" ",9),"20",rep(" ",9),"30",rep(" ",9),"40",
  rep(" ",9),"50",rep(" ",10))
barchart(res~items,origin=0,groups=y,main=list("Examinee k",cex=1.2),
  scales=list(y=list(cex=1),x=list(at=items,labels=labs,cex=1,rot=0)),
  col=c(gray(0.9),gray(0.1)),horizontal=FALSE,xlab="Item",
  ylab="Estimated Standardized Residual",axes=FALSE)
addLine(h=2, lty=3)
addLine(h=-2,lty=3)

```