Comparing the Performance of Eight Item Preknowledge Detection Statistics

Applied Psychological Measurement

2016, Vol. 40(2) 83–97

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621615603327

apm.sagepub.com

Dmitry I. Belov¹

Abstract

Item preknowledge describes a situation in which a group of examinees (called *aberrant examinees*) have had access to some items (called *compromised items*) from an administered test prior to the exam. Item preknowledge negatively affects both the corresponding testing program and its users (e.g., universities, companies, government organizations) because scores for aberrant examinees are invalid. In general, item preknowledge is hard to detect due to multiple unknowns: unknown groups of aberrant examinees (at unknown test centers or schools) accessing unknown subsets of items prior to the exam. Recently, multiple statistical methods were developed to detect compromised items. However, the detected subset of items (called the *suspicious subset*) naturally has an uncertainty due to false positives and false negatives. The uncertainty increases when different groups of aberrant examinees had access to different subsets of items; thus, compromised items for one group are uncompromised for another group and vice versa. The impact of uncertainty on the performance of eight statistics (each relying on the suspicious subset) was studied. The measure of performance was based on the receiver operating characteristic curve. Computer simulations demonstrated how uncertainty combined with various independent variables (e.g., type of test, distribution of aberrant examinees) affected the performance of each statistic.

Keywords

test security, item preknowledge, hypothesis testing, Neyman–Pearson lemma, Kullback–Leibler divergence, ROC, person misfit, person fit, lz

Introduction

Item preknowledge describes a situation in which a group of examinees (called *aberrant examinees*) have had access to a subset of items (called *compromised items*) from an administered test prior to the exam. Aberrant examinees perform better on compromised items as compared with uncompromised items. When the number of aberrant examinees is large, the corresponding testing program—paper-and-pencil testing (P&P), computer-based testing (CBT), multiple-stage testing (MST), and computerized adaptive testing (CAT)—and its users (e.g., universities, companies, government organizations) are negatively affected because scores for aberrant examinees are invalid. Item preknowledge is a special case of test collusion that has recently received

¹Law School Admission Council, Newtown, PA, USA

Corresponding Author:

Dmitry I. Belov, Psychometric Research, Law School Admission Council, 662 Penn Street, Newtown, PA 18940, USA.

Email: dbelov@lsac.org

a lot of attention in test security research and practice (Maynes, 2013). Test collusion may be described as large-scale sharing of test materials or answers to test questions. The source of the shared information could be a teacher, a test-preparation company, the Internet, or examinees communicating on the day of the exam (Wollack & Maynes, 2011). Detection of certain types of test collusion can be reduced to the detection of item preknowledge (e.g., teacher correcting answers to hard items for a group of students, students working together on a subset of items).

In general, item preknowledge is hard to detect due to multiple unknowns involved: unknown groups of examinees (at unknown schools or test centers) accessing unknown compromised subsets of items prior to taking the test. Recently, multiple statistical methods were developed to detect compromised items (Belov, 2014; Choe, 2014; Obregon, 2013; O'Leary & Smith, 2013). However, the detected item subset, called here the *suspicious subset*, naturally has uncertainty due to false positives and false negatives. The uncertainty increases when different groups of aberrant examinees had access to different subsets of items; thus, compromised items for one group are uncompromised for another group and vice versa. Therefore, there is a need to analyze the effect of uncertainty on statistics that rely on the suspicious subset.

The analysis concentrates on eight different statistics. Seven of them measure some specific difference between responses to items from the suspicious subset versus responses to other items. The eighth statistic I_z by Drasgow, Levine, and Williams (1985) serves as a baseline because it does not rely on the suspicious subset. Throughout the article, the following notation is used: Lowercase letters $a, b, c, \dots$ denote scalars; lowercase Greek letters $\alpha, \beta, \gamma, \dots$ denote random variables; an estimator of a random variable θ is denoted as $\hat{\theta}$; capital letters $A, B, C, \dots$ denote sets and sequences; $|S|$ denotes the size of S ; and bold capital letters $\mathbf{A}, \mathbf{B}, \mathbf{C}, \dots$ denote functions.

Eight Item Preknowledge Detection Statistics

An examinee is defined by two random variables: unobservable latent trait (ability) θ and observable response $\chi_i \in \{0, 1\}$ to item i . Consider an arbitrary subset of items I administered to the examinee, where I may vary among examinees (e.g., CAT) or be fixed (e.g., P&P). Then several characteristics can be computed. The following random variable is the resultant score:

$$\varphi_I = \sum_{i \in I} \chi_i. \quad (1)$$

Bayes's theorem is applied to compute the discrete posterior distribution of θ with uniform prior:

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

where $\mathbf{F}_I(y)$ is the probability of $\theta=y$, $\mathbf{P}_i(\chi_i|y)$ is the probability of response χ_i to item i conditioned on $\theta=y$, and set Y contains ability levels (this article used $Y = \{-5, -4.9, \dots, 5\}$). Then the expected a posteriori (EAP) estimator of θ is computed as follows:

$$\hat{\theta}_I = \sum_{y \in Y} y \mathbf{F}_I(y). \quad (3)$$

Let S denote the suspicious subset. Consider an examinee taking a test T (where T may vary among examinees), which can be partitioned into two disjoint subtests $C = T \cap S$ (suspicious

Figure 1. Posterior shift.

Note. Dotted unshaded area corresponds to the posterior shift $S(F_C || F_U)$ between F_C and F_U , where the largest right boundary region on which F_C is not lower than F_U is $[0, 3]$.

items) and $U = T \setminus S$ (unsuspicious items). Then the following characteristics can be computed for C , U , and T : scores φ_C , φ_U , $\varphi_T = \varphi_C + \varphi_U$; posteriors of ability F_C , F_U , F_T ; and ability estimates $\hat{\theta}_C$, $\hat{\theta}_U$, $\hat{\theta}_T$. These characteristics are used to describe eight statistics for detecting aberrant examinees: lz (Drasgow et al., 1985); score ratio $\varphi_C / (\varphi_U + 1)$; ability difference $\hat{\theta}_C - \hat{\theta}_U$; Kullback–Leibler divergence $D(F_C || F_U)$ (Kullback & Leibler, 1951); a statistic by Shu, Henson, and Luecht (2013); a modified lz (Armstrong, Stoumbos, Kung, & Shi, 2007); a statistic based on Neyman–Pearson lemma (Levine & Drasgow, 1988); and a new statistic based on posterior shift. Each statistic is computed such that aberrant examinees should be located at the right tail of the corresponding null distribution. A detailed description of seven of the statistics is given in the online appendix; the new statistic based on posterior shift is presented next.

Statistic Based on Posterior Shift

This new statistic is based on the assumption that, for an aberrant examinee, the posterior computed from the responses to compromised items should be shifted more to the higher ability than the posterior computed from the responses to uncompromised items. This statistic measures the difference between the two posteriors in the largest right boundary region where the first posterior is not lower than the second posterior (see Figure 1).

If the corresponding region is empty, then the statistic equals zero. Formally, in Equation 2, which computes the discrete posterior of ability, let us consider set Y as a decreasing sequence of ability levels $Y : y_1 > y_2 > \dots > y_n$, where $0 < h = y_j - y_{j+1}$ is fixed, $j = 1, \dots, n - 1$ (e.g., $Y : 3 > 2.9 > \dots > -3$, $h = 0.1$). Find the largest $1 \leq k \leq n$, such that $F_C(y_j) \geq F_U(y_j)$, $j = 1, 2, \dots, k$; otherwise, assume $k = 0$. Then the following statistic measures how far the posterior F_C is shifted toward the higher ability with respect to the posterior F_U (see Figure 1):

$$S(F_C || F_U) = h \sum_{j=1}^k F_C(y_j) - F_U(y_j). \quad (4)$$

By definition, the divergence $S(F_C||F_U)$ is asymmetric, nonnegative, and equal to zero when $k=0$ or $F_C(y_j)=F_U(y_j)$, $j=1,2,\dots,k$. For the example in Figure 1, $S(F_C||F_U)>0$, but $S(F_U||F_C)=0$ because $k=0$ in this case; at the same time, both $D(F_C||F_U)$ and $D(F_U||F_C)$ are positive. Thus, it is expected that the new statistic $S(F_C||F_U)$ should be more sensitive to item preknowledge than statistic $D(F_C||F_U)$ (see results in the Computer Simulations section below).

Computer Simulations

The performance of the above eight statistics was analyzed via computer simulations. Multiple independent variables potentially influencing the performance were studied. The above statistics and the simulation environment were implemented by the author in standard C++. Therefore, the resultant software is more scalable and runs much faster than existing tools written in script languages like R. The source code for this software can be easily adapted to different item response theory models, types of tests, or types of distributions for nonaberrant and aberrant test takers. The author plans to make the source code available on email request.

Measure of Performance

The measure of performance is based on the receiver operating characteristic (ROC) curve. The ROC curve allows estimating the quality of a binary classifier (Green & Swets, 1966) by showing the relation between empirical detection and Type I error rates under changing significance levels. ROC allows integrating Type I and Type II errors in just one chart. This is more informative for a practitioner than separate reports of Type I and Type II error rates. The quantitative interpretation of ROC is given by area under the ROC curve (*ROC area*): the higher the area, the better the classifier. Figure 2 illustrates two ROC curves computed for two different statistics (Iz and Neyman–Pearson lemma), where the significance level changed from .01 to 1 in increments of .01. Given a significance level, a suspicious subset S , and type of test (CAT or P&P), the critical value of each statistic was estimated from a null distribution computed from the responses of 10,000 nonaberrant examinees drawn from $N(0,1)$.

In the context of test security, one is only interested in the ROC curve for the region of small Type I errors. Therefore, the measure of performance is computed here as the area under the ROC curve for the region of Type I errors below 0.1 divided by 0.1. This measure is called the *truncated ROC area* and it is always between 0 and 1 (see Figure 2).

Stability of the performance across independent variables is important in practice. This article employs the following measure of stability:

$$1 - \sigma, \quad (5)$$

where σ is the standard deviation of the truncated ROC area across all independent variables.

Setup

Nonaberrant examinees were drawn from an $N(0,1)$ distribution. In each scenario (simulating item preknowledge), the number of nonaberrant examinees was 1,000. Each scenario was replicated 10 times. The following four independent variables were considered.

Type of distribution of aberrant examinees. Aberrant examinees were drawn from $N(0,1)$ or $U(-3,0)$ distributions.

Figure 2. Example of two ROC curves computed for significance level changing from .01 to 1 in increments of .01.

Note. The ROC areas are the following: 0.76 for the *Iz* statistic; and 0.97 for the statistic based on the Neyman–Pearson lemma. The truncated ROC areas (*truncated ROC area* is the ROC area for Type I error below 0.1 divided by 0.1) are 0.31 for the *Iz* statistic and 0.86 for the statistic based on the Neyman–Pearson lemma. ROC = receiver operating characteristic.

Amount of aberrancy. The percentages of aberrant examinees were 5%, 10%, and 20% of the size of the nonaberrant population, resulting in 50, 100, and 200 aberrant examinees. Thus, the total number of examinees in each scenario varied as 1,050, 1,100, and 1,200 examinees.

Level of uncertainty in the suspicious subset. The above statistics (except statistic *Iz*) rely on the suspicious subset of items, represented by subset *S*. However, subset *S* naturally has uncertainty due to false positives and false negatives associated with the statistical detection of the suspicious subset. The uncertainty increases when different groups of aberrant examinees had access to different subsets of items; thus, compromised items for one group are uncompromised for another group and vice versa. To simulate uncertainty, the aberrant examinees were partitioned into 10 aberrant groups of equal size, where *i*th aberrant group had preknowledge of a unique random subset of items S_i , $i = 1, \dots, 10$. Each compromised subset S_i had the same size as *S* but was formed in such a way that the following three levels of uncertainty were simulated:

[0% uncertainty] Each compromised subset S_i , $i = 1, \dots, 10$ equals the suspicious subset *S*, which is an ideal situation for all statistics relying on the suspicious subset. This corresponds

to zero uncertainty, where 0% of items in S are uncompromised for each aberrant group (see Figure 3a).

[25% uncertainty] Each compromised subset S_i , $i = 1, \dots, 10$ includes only 75% of the suspicious subset S . This corresponds to a small amount of uncertainty, where 25% of items in S are uncompromised for each aberrant group (see Figure 3b).

[50% uncertainty] Each compromised subset S_i , $i = 1, \dots, 10$ includes only 50% of the suspicious subset S . This corresponds to a large amount of uncertainty, where 50% of items in S are uncompromised for each aberrant group (see Figure 3c).

Type of test. Two types of tests were studied: adaptive (CAT) and nonadaptive (P&P). Simulations were conducted using disclosed items of the Law School Admission Test (LSAT). Each aberrant examinee had a .9 probability of giving a correct response to each compromised item; otherwise, the response probability was modeled by the three-parameter logistic (3PL) model (Lord, 1980).

In all simulations, the parameters of items were known (see Levine & Drasgow, 1988), thus simulating the realistic situation of using pretest parameters. If items are newly developed (and their pretest parameters are estimated from a subpopulation), then item preknowledge cannot exist unless there is a leak in the testing organization.

Type of test: CAT. Multiple simulation studies were conducted using disclosed Logical Reasoning (LR) items of the LSAT. The CAT pool consisted of 500 LR items. The distribution of (a) discrimination, (b) difficulty, and (c) guessing parameters of the items in the CAT pool have the following minimums, maximums, means, and variances, respectively: (a) minimum 0.28, maximum 1.67, mean 0.75, variance 0.06; (b) minimum -2.47 , maximum 2.92, mean 0.49, variance 1.27; and (c) minimum 0.00, maximum 0.52, mean 0.17, variance 0.01.

The item selection criterion for CAT was the maximization of Fisher information at the current estimate of ability $\hat{\theta}$. The test length was fixed at 50 items for each examinee. The estimator of θ was the EAP estimator with a uniform prior (see Equation 3). The ability estimate was initialized at $\hat{\theta} = 0$. There was no item exposure control. The following CAT setup was performed:

Step 1: Simulate CAT without item preknowledge with 1,000 nonaberrant examinees drawn from $N(0,1)$. Then compute item exposure. Form a subset S of potentially compromised items with exposure higher than 0.6. This step resulted in S with 13 items (see Figure 4).

Step 2: Simulate new CAT without item preknowledge with 10,000 nonaberrant examinees drawn from $N(0,1)$. Compute critical values for each statistic, where the significance level was changing from .001 to 1 in increments of .001.

Type of test: P&P. Multiple simulation studies were conducted using a disclosed form of the LSAT with 100 items (partitioned into four sections) of the following item types: Analytical Reasoning (AR), LR, and Reading Comprehension (RC). More information on the LSAT can be found at www.LSAC.org. The following P&P setup was performed:

Step 1: One operational LR section was assumed to be memorized during a previous administration of the LSAT (when this section was pretested) and then later partially distributed to various groups of aberrant examinees. A suspicious subset of potentially compromised items S contains all items from this section (26 items, see Figure 4).

Figure 3. This is an illustration of uncertainty for P&P test with 100 items.
Note. Compromised items for each group of aberrant examinees are highlighted (gray cells). P&P = paper-and-pencil.

Figure 4. Distribution of discrimination and difficulty in suspicious subset S in CAT and P&P.
 Note. CAT = computerized adaptive testing; P&P = paper-and-pencil.

Step 2: Simulate P&P without item preknowledge with 10,000 nonaberrant examinees drawn from $N(0,1)$. Compute critical values for each statistic, where the significance level was changing from 0.001 to 1 in increments of 0.001.

Simulation Procedure

The above setup resulted in 2 (type of distribution for aberrant examinees) $\times$ 3 (amount of aberrancy) $\times$ 3 (level of uncertainty) $\times$ 2 (type of test) = 36 simulation scenarios. For each scenario, the following steps were repeated 10 times to compute average truncated ROC areas:

- Step 1: Assign a unique random compromised subset (formed around S for a given level of uncertainty) to each group of aberrant examinees.
- Step 2: Given the type of distribution for aberrant examinees, the amount of aberrancy, and the type of test, simulate response data with item preknowledge.
- Step 3: Given simulated data, compute the truncated ROC area for each statistic.

Results

Detailed results are presented in Figures 5 to 8. Truncated ROC areas averaged over all independent variables are 0.454 for I_z , 0.556 for score ratio, 0.579 for ability difference, 0.520 for Kullback–Leibler divergence, 0.605 for posterior shift, 0.601 for Shu, 0.560 for modified I_z , and 0.584 for Neyman–Pearson lemma. Stabilities computed over all independent variables are 0.804 for I_z , 0.744 for score ratio, 0.740 for ability difference, 0.707 for Kullback–Leibler divergence, 0.746 for posterior shift, 0.747 for Shu, 0.708 for modified I_z , and 0.690 for Neyman–Pearson lemma.

Figure 5. Truncated ROC areas for CAT where aberrant examinees were drawn from $N(0,1)$.
Note. ROC = receiver operating characteristic; CAT = computerized adaptive testing.

Discussion of Results

This section is structured as a list of questions with answers acquired from the analysis of Figures 5 to 8. Taking into account practical considerations, the following questions are formulated.

Which statistic has the best performance on average?

On average, the statistic based on posterior shift has the largest truncated ROC area. The following two statistics have average truncated ROC areas above 0.6: Shu (0.601) and posterior shift (0.605).

How does the test type affect the performance of the statistics?

Figures 5 to 8 show that all statistics perform better for P&P than for CAT. This is expected because of the higher number of compromised items administered to an aberrant examinee in P&P. The ratio of the number of compromised items to the test length is 0.26 for both CAT and

Figure 6. Truncated ROC areas for CAT where aberrant examinees were drawn from $U(-3,0)$.
Note. ROC = receiver operating characteristic; CAT = computerized adaptive testing.

P&P (for CAT with 50 items, the number of compromised items is 13; for P&P with 100 items, the number of compromised items is 26). However, because of the adaptive nature of CAT, the ratio of the number of compromised items (actually administered to an aberrant examinee) to the test length was smaller in CAT, which increased both Type I and Type II error rates. Indeed, in CAT, the average number of compromised items administered to an aberrant examinee was 8 (the ratio is 0.16). Meanwhile, in P&P, the average number of compromised items administered to an aberrant examinee was 26 (the ratio is 0.26).

How does the distribution of aberrant examinees affect the performance of the statistics?

When the distribution of aberrant examinees is $U(-3,0)$, all statistics demonstrate better performance (see Figures 5-8). This result is expected, and it follows from a larger average difference (than for $N(0,1)$) in responding between compromised and uncompromised items for each aberrant examinee.

Figure 7. Truncated ROC areas for P&P where aberrant examinees were drawn from $N(0,1)$.
Note. ROC = receiver operating characteristic; P&P = paper-and-pencil.

Which independent variable has the most negative effect on the performance of the statistics?

Figures 5 to 8 show that uncertainty has the most negative effect on the performance of all statistics: the larger the uncertainty, the smaller the truncated ROC area. When uncertainty was 0%, all statistics relying on the suspicious subset performed much better than Iz (Figures 5-8).

Which statistic is the most stable?

The Iz statistic has the highest stability due to its low sensitivity to uncertainty (because Iz does not rely on the suspicious subset), which resulted in a smaller drop in performance for Iz than for other statistics (see Figures 5-8).

How are results compatible with results from existing publications?

Figure 8. Truncated ROC areas for P&P where aberrant examinees were drawn from $U(-3,0)$.
Note. ROC = receiver operating characteristic; P&P = paper-and-pencil.

A suspicious subset can be defined by assigning to each item a probability of preknowledge (McLeod, Lewis, & Thissen, 2003). Computer simulations (Hui, 2008) demonstrated that when the number of items with a high probability of preknowledge grows (e.g., due to a large number of aberrant groups, each with a unique compromised subset), the detection rate of the Bayesian method by McLeod et al. (2003) drops, which conforms to the results in Figures 5 to 8.

How can one address the negative effect of uncertainty?

As shown above, uncertainty has the most negative effect on the performance of the statistics (Figures 5-8). There are three approaches to address this problem.

The first approach is to operate without explicitly taking into account the suspicious subset (like the Iz statistic). Multiple parametric and nonparametric statistics (Karabatsos, 2003) can be applied. The CUSUM method (van Krimpen-Stoop & Meijer, 2001) is only applicable when compromised items are positioned sequentially in the test (Tendeiro & Meijer, 2012). Cluster

analysis (Wollack & Maynes, 2011) and factor analysis (Zhang, Searcy, & Horn, 2011) were applied to detect item preknowledge; however, both methods rely on the number of response matches, which is not applicable to MST and CAT, where the actual test varies across examinees.

The second approach is to simultaneously detect groups of aberrant examinees and corresponding compromised subsets of items. Belov (2014) demonstrated that this approach is feasible by exploiting information theory and combinatorial optimization. The resultant method (3D algorithm) is applicable to all testing programs, and in CAT simulations it performed substantially better than IZ. However, it is based on the assumption that a test center cannot have more than one group of aberrant examinees. This limitation is addressed by the formulation of the generalized 3D algorithm (Belov, 2014), but its study is a future work. Other advantages and limitations of the 3D algorithm can be found in Belov (2014).

The third approach is to combine statistics from the previous two approaches. As different statistics are sensitive to different types of item response patterns (related to item preknowledge) and are based on different assumptions, this approach may demonstrate the best performance. One type of combination is based on nested hypothesis testing, which was introduced by Belov and Armstrong (2010) to detect answer copying. They applied two nested stages—Stage 1: test each examinee for being a potential copier; and Stage 2: for each detected copier, test his or her neighbors for being potential sources of answers. The major advantage of nested hypothesis testing is the use of larger significance levels at each stage because then the resultant Type I error rate can be well approximated by the multiplication of significance levels used at each stage.

Summary

Detection of item preknowledge (especially on a large scale) is currently a hot topic of research in test security. Therefore, it is important to know which statistics could be used in practice and how they would perform in different scenarios.

In general, item preknowledge is hard to detect due to multiple unknowns involved: unknown groups of examinees (at unknown schools or test centers) accessing unknown compromised subsets of items prior to taking the test. Recently, multiple statistical methods were developed to detect compromised items. However, the detected subset of items (called the *suspicious subset*) naturally has uncertainty due to false positives and false negatives. The uncertainty increases when different groups of aberrant examinees had access to different subsets of items; thus, compromised items for one group are uncompromised for another group and vice versa. Therefore, there is a need for the analysis of the effect of the uncertainty on statistics that rely on the suspicious subset. Such analysis was done by comparing the performance of eight statistics for the detection of item preknowledge in 36 simulated scenarios controlled by four independent variables: type of distribution for aberrant examinees, amount of aberrancy, level of uncertainty, and type of test. The measure of performance was the truncated ROC area (Figure 2). Computer simulations demonstrated that uncertainty causes the most dramatic loss of performance (Figures 5-8). Three approaches to address this problem were discussed. This article extends the comparison study by Karabatsos (2003) to item preknowledge detection statistics relying on information about compromised items.

Declaration of Conflicting Interests

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

Funding

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

References

- Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the I_2 person-fit statistic. *Practical Assessment Research & Evaluation*, 12(16). Retrieved from <http://pareonline.net/getvn.asp?v=12&n=16>
- Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, 2(3), 37-58.
- Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Applied Psychological Measurement*, 34, 379-392.
- Belov, D. I., & Armstrong, R. D. (2011). Distributions of the Kullback-Leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology*, 64, 291-309.
- Belov, D. I., Pashley, P. J., Lewis, C., & Armstrong, R. D. (2007). Detecting aberrant responses with Kullback-Leibler distance. In K. Shigemasa, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 7-14). Tokyo, Japan: Universal Academy Press.
- Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, 58, 37-52.
- Choe, E. (2014, April). *Utilizing response time in sequential detection of compromised items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.
- 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.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.
- Hui, H.-f. (2008). *Stability and sensitivity of a model-based person-fit index in detecting item preknowledge in computerized adaptive test* (Unpublished dissertation). Hong Kong: The Chinese University of Hong Kong.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16, 277-298.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79-86.
- Lehmann, E. L. (1999). *Elements of large-sample theory*. New York, NY: Springer.
- Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53, 161-176.
- Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Maynes, D. (2013). Educator cheating and the statistical detection of group-based test security threats. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 173-199). New York, NY: Routledge.
- 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.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107-135.
- Obregon, P. (2013, April). *A Bayesian approach to detecting compromised items*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- O'Leary, L. S., & Smith, R. W. (2013, April). *Extending differential person and item functioning to aid in maintenance of exposed exams*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- 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.
- Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitative Methoden*, 7, 131-145.

- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, 66, 331-342.
- Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement*, 36, 420-442.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, 26, 199-217.
- Wollack, J. A., & Maynes, D. (2011, April). *Detection of test collusion using item response data*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Zhang, Y., Searcy, C. A., & Horn, L. (2011, April). *Mapping clusters of aberrant patterns in item responses*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.