Higher-Order asymptotics and its use to test the equality of the examinee ability over two sets of items

Sandip Sinharay, Educational Testing Service
Jens L. Jensen, Aarhus University

An Updated Version of this document appeared in Psychometrika. The website for the article is https://link.springer.com/article/10.1007%2Fs11336-018-9627-8

The citation for the article is: Sinharay, S., & Jensen, J. L. (2018). Higher-Order asymptotics and its use to test the equality of the examinee ability over two sets of items. Advance Online Publication. doi: 10.1007/s11336-018-9627-8.

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

Higher-Order Asymptotics and Its Use to Test the Equality of the Examinee Ability Over Two Sets of Items

Sandip Sinharay, Educational Testing Service

Jens Ledet Jensen, Aarhus University

May 22, 2018

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

Higher-Order Asymptotics and Its Use to Test the Equality of the Examinee Ability Over Two Sets of Items

Abstract

In educational and psychological measurement, researchers and/or practitioners are often interested in examining whether the ability of an examinee is the same over two sets of items. Such problems can arise in measurement of change, detection of cheating on unproctored tests, erasure analysis, detection of item preknowledge etc. Traditional frequentist approaches that are used in such problems include the Wald test, the likelihood ratio test, and the score test (e.g., Fischer, 2003; Finkelman, Weiss, & Kim-Kang, 2010; Glas & Dagohoy, 2007; Guo & Drasgow, 2010; Klauer & Rettig, 1990; Sinharay, 2017). This paper shows that approaches based on higher-order asymptotics (e.g., Barndorff-Nielsen & Cox, 1994; Ghosh, 1994) can also be used to test for the equality of the examinee ability over two sets of items. The modified signed likelihood ratio test (e.g. Barndorff-Nielsen, 1986) and the Lugannani-Rice approximation (Lugannani & Rice, 1980), both of which are based on higher-order asymptotics, are shown to provide some improvement over the traditional frequentist approaches in three simulations. Two real data examples are also provided.

Key words: detection of cheating, item preknowledge, measurement of change.

In applications of item response theory (IRT), measurement researchers and practitioners are often interested in examining whether the ability of an examinee is the same on two sets of items. The following are examples of such problems:

1. Two sets of items are answered at two different time-points and the investigator is interested in examining whether the examinee's ability changed between the two time points. Researchers such as Fischer (2003) and Finkelman et al. (2010) showed that if the two sets of items were calibrated using an IRT model on the same scale, one can perform a statistical hypothesis test to determine whether the examinee ability is equal at the two time points. A significant test statistic would indicate a change in ability.

2. A set of items on an assessment is not known to have been compromised and the remaining items are known to have been compromised. In this case, the investigator may want to test for the equality of the examinee ability over these two sets of items (compromised and non-compromised) to detect if the examinee cheated, that is, benefited from preknowledge of the compromised items. A significantly larger estimated ability on the compromised items may indicate cheating. See Sinharay (2017) for such an example. A similar problem involves two sets of items where the first set includes field-test or pretest items (and have not been used before) and the second includes operational items (and have been used before). Better performance by an examinee on the second set of items may indicate possible cheating. See, for example, Drasgow, Levine, and Zickar (1996), for a case like this.

3. A set of items is answered in a proctored version of the assessment and another set is answered in an unproctored version of the assessment; the investigator may want to test for the equality of the examinee ability over these two assessments to detect if the examinee cheated on the unproctored version. See Guo and Drasgow (2010) for such an example.

4. No erasures were found on the answer sheet on a set of items and erasures were found

on the answer sheet on the remaining items; the investigator may want to test for the equality of the ability over these two sets of items to detect if the examinee benefited from cheating in the form of fraudulent erasures. See Wollack, Cohen, and Eckerly (2015) for such an example.

5. The two sets of items belong to two subsections/subtests of an assessment and the investigator wants to assess person fit by testing the equality of the ability over the two subsections. See Glas and Dagohoy (2007), Klauer and Rettig (1990), and Klauer (1991) for such examples.

In such applications, the investigator is often interested in testing against a one-sided alternative hypothesis because of an interest in detecting cheating on assessments. For example, Wollack and Schoenig (2018) categorized the statistical methods to detect cheating (on tests) into six categories one of which is "score differencing"—this category of methods essentially involves a test of the hypothesis of equal ability of an examinee (or a group of examinees) over two sets of items against a one-sided alternative hypothesis. Cheating would lead to a better performance on the second set of items compared to the first set of items. A one-sided alternative hypothesis maybe more appropriate in the second, third, and fourth of the above examples and also in the first example if the investigator is only interested in a positive change of ability because of, for example, cheating on the assessment (e.g., Lewis & Thayer, 1998, stated that a large positive score change often initiates further investigation on possible cheating by an examinee). This paper focuses on one-sided alternatives.

Traditionally, researchers and practitioners in educational and psychological measurement have applied the Wald test, the likelihood ratio test (LRT), or the score test to test for the equality of the ability over two sets of items. See, for example, Fischer (2003), Finkelman et al. (2010), Glas and Dagohoy (2007), Guo and Drasgow (2010), Klauer and Rettig (1990), and Sinharay (2017). In some of the example problems cited above, researchers have used other methods to test the hypothesis, but those methods have been limited to only one context or one IRT model. For example, Wollack et al. (2015)

used a test statistic called "erasure detection index" in the context of erasure analysis and Klauer (1991) used a uniformly most powerful test under the Rasch model in the context of testing for the equality of abilties over two subsections. These specific methods are not considered henceforth. Instead, the focus will be on the Wald test, the LRT, and the score test, which are often referred to as *methods based on first-order asymptotics* (e.g., Brazzale, Davison, & Reid, 2007, p. 1).

The Wald test, the LRT, and the score test are appropriate for large samples, which, in the context of testing of the equality of the ability over two sets of items, means that these tests are appropriate when they are based on two large sets of items. However, the test of the equality of abilities often has to be performed using at least one small or moderately large set of items. For example, in erasure analysis (e.g., Wollack et al., 2015), the set of erased items is typically small (often consisting of fewer than 10 items). When one or more set of items is small, the methods based on first-order asymptotics often would not be appropriate and would have inflated Type I error rate or low power; for example, Guo and Drasgow (2010) found the Type I error rate of the Wald test to increase as the proctored test became shorter. Thus, there is a scope of further research on hypothesis testing methods that would perform better than those based on first-order asymptotics.

One set of large-sample approaches that have been found to perform better than the methods based on first-order asymptotics in hypothesis testing based on small or moderately-sized samples in several areas of statistics involve higher-order asymptotics (e.g. Barndorff-Nielsen & Cox, 1994; Ghosh, 1994). Specifically, the modified signed likelihood ratio test (MSLRT; e.g. Barndorff-Nielsen, 1986) and the Lugannani-Rice approximation (LRA; Lugannani & Rice, 1980), both of which are based on higher-order asymptotics, have been used in hypothesis-testing problems in several areas of statistics and have been proved, theoretically and empirically, to have better properties compared to the Wald test, the LRT, and the score test (e.g., Barndorff-Nielsen, 1991; Pierce & Peters, 1992). While the MSLRT and the LRA usually perform similarly to the methods based on first-order asymptotics for very large samples, the former methods often perform considerably better than the latter methods for samples that are not very large (e.g., Barndorff-Nielsen, 1991).

Higher-order asymptotics have found a few applications in educational and/or psychological measurement. Bedrick (1997) and von Davier and Molenaar (2003) applied Edgeworth expansions, which are based on higher-order asymptotics, to obtain the distributional form of person-fit indices for the dichotomous and polytomous Rasch models. Biehler, Holling, and Doebler (2015) suggested a saddlepoint approximation of the distribution of the ability parameter for the two-parameter logistic model (2PLM). However, there are no known applications of higher-order asymptotics to test the equality of abilties in the context of educational and/or psychological measurement. Thus, this paper attempts to fill an important void in the literature.

The next section includes a literature review—the existing tests for the equality of the abilities over two sets of items are discussed followed by a review of the MSLRT and LRA. The Methods section includes the derivations of the MSLRT and LRA for the 2PLM and the generalized partial credit model (GPCM; Muraki, 1992). The MSLRT and LRA are compared to the Wald test, the LRT-based test, and the score test using simulated data sets in the Simulation section. Two real data examples are included in the penultimate section. Conclusions and recommendations are provided in the last section.

The item parameters are usually not estimated and assumed known in tests of hypothesis regarding the examinee ability in IRT applications (e.g., Glas & Dagohoy, 2007). The same assumption of known item parameters was made here, one reason being that Glas and Dagohoy (2007) found that accounting for the uncertainty in the (estimated) item parameters had little impact on the testing of the equality of abilities.

Background

Joint Likelihood Over Two Sets of Items

Let us consider two sets of binary items that have been calibrated using the 2PLM. Let the true slope and difficulty parameters of the first set of items be denoted by a_i and b_i's, respectively, where $i = 1, 2, \cdots, n_1$, and let the true slope and difficulty parameters of the second set of items be denoted by $\tilde{a}_j$ and $\tilde{b}_j$'s, respectively, where $j = 1, 2, \cdots, n_2$. Let the true ability of a randomly chosen examinee on the two sets of items be denoted by θ_1

and θ_2, respectively and let the examinee's scores on the two sets of items be denoted by $X_i, i = 1, 2, \cdots, n_1$ and $Y_j, j = 1, 2, \cdots, n_2$, respectively. Let us denote the probability of a correct answer under the 2PLM on the two sets of items as

$$P(X_i = 1) = \frac{\exp[a_i(\theta_1 - b_i)]}{1 + \exp[a_i(\theta_1 - b_i)]} = p_i(\theta_1) \text{ and } P(Y_j = 1) = \frac{\exp[\tilde{a}_j(\theta_2 - \tilde{b}_j)]}{1 + \exp[\tilde{a}_j(\theta_2 - \tilde{b}_j)]} = \tilde{p}_j(\theta_2).$$

The joint log-likelihood $\ell(\theta_1, \theta_2)$ of the two true abilities for an examinee is given by

$$\begin{aligned}
& \ell(\theta_1, \theta_2) \\
= \quad & \sum_i [X_i \log p_i(\theta_1) + (1 - X_i) \log(1 - p_i(\theta_1))] + \sum_j [Y_j \log \tilde{p}_j(\theta_2) + (1 - Y_j) \log(1 - \tilde{p}_j(\theta_2))] \\
= \quad & \sum_i \left[X_i \log \frac{p_i(\theta_1)}{1 - p_i(\theta_1)} + \log(1 - p_i(\theta_1)) \right] + \sum_j \left[Y_j \log \frac{\tilde{p}_j(\theta_2)}{1 - \tilde{p}_j(\theta_2)} + \log(1 - \tilde{p}_j(\theta_2)) \right] \\
= \quad & \sum_i [X_i a_i(\theta_1 - b_i) + \log(1 - p_i(\theta_1))] + \sum_j \left[Y_j \tilde{a}_j(\theta_2 - \tilde{b}_j) + \log(1 - \tilde{p}_j(\theta_2)) \right] \\
= \quad & S_1\theta_1 - \sum_i X_i a_i b_i + \sum_i \log(1 - p_i(\theta_1)) + S_2\theta_2 - \sum_j Y_j \tilde{a}_j \tilde{b}_j + \sum_j \log(1 - \tilde{p}_j(\theta_2)), \qquad (1)
\end{aligned}$$

where $S_1 = \sum_i X_i a_i$ and $S_2 = \sum_j Y_j \tilde{a}_j$.

Existing Methods for Testing the Equality of Ability Over Two Sets of Items

Primarily, three hypothesis testing approaches have been suggested for testing the equality of the true ability on two sets of items, that is, for testing $H_0 : \theta_1 = \theta_2$ in the context of IRT models: the Wald test, the LRT, and the score test. Researchers such as Fischer (2003), Finkelman et al. (2010), and Klauer and Rettig (1990) have suggested the use of the Wald test for testing the hypothesis. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ denote the maximum likelihood estimates (MLE) of θ_1 and θ_2, respectively. Let $\hat{\theta}_0$ denote the MLE of the common ability parameter computed from the examinee's scores on the two sets of items (combined). The Wald test statistic (that leads to the Wald test or the Z-test) is given by

$$Z = \frac{\hat{\theta}_2 - \hat{\theta}_1}{\sqrt{s_2^2(\hat{\theta}_0) + s_1^2(\hat{\theta}_0)}}, \qquad (2)$$

where $s_1^2(\hat{\theta}_0)$ and $s_2^2(\hat{\theta}_0)$ are the estimated variances of $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively, both of which are computed at $\hat{\theta}_0$.[1] For the 2PLM,

$$s_1^2(\hat{\theta}_0) = \left[\sum_i a_i^2 p_i(\hat{\theta}_0)(1 - p_i(\hat{\theta}_0))\right]^{-1} = \left[\sum_i a_i^2 \frac{\exp[a_i(\hat{\theta}_0 - b_i)]}{(1 + \exp[a_i(\hat{\theta}_0 - b_i)])^2}\right]^{-1}. \tag{3}$$

For long tests, the Z statistic approximately follows a standard normal null distribution and its square follows a χ^2 null distribution with one degree of freedom (Guo & Drasgow, 2010; Finkelman et al., 2010; Klauer & Rettig, 1990). Under the alternative $H_1 : \theta_2 > \theta_1$, Z is expected to be a large positive number.

Guo and Drasgow (2010), Finkelman et al. (2010), and Klauer and Rettig (1990) suggested the use of the LRT for testing $H_0 : \theta_1 = \theta_2$. To apply this test, one computes $\mathcal{L}(\theta_1, \theta_2)$, the joint likelihood of the two ability parameters, twice—once each at $\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2$ and $\theta_1 = \hat{\theta}_0, \theta_2 = \hat{\theta}_0$. Then, the LRT statistic is computed as

$$\Lambda = \frac{\mathcal{L}(\hat{\theta}_0, \hat{\theta}_0)}{\mathcal{L}(\hat{\theta}_1, \hat{\theta}_2)}. \tag{4}$$

For long tests, $-2 \log \Lambda$ follows a χ^2 null distribution with one degree of freedom (Guo & Drasgow, 2010; Finkelman et al., 2010).

If the alternative hypothesis is two-sided, the statistic $-2 \log \Lambda$ can be used as it is. If, however, the alternative hypothesis is one-sided and is given by $H_1 : \theta_2 > \theta_1$, Sinharay (2017) showed that the signed likelihood ratio (SLR) statistic defined as

$$L_s = \begin{cases} \sqrt{-2 \log \Lambda} \text{ if } \hat{\theta}_2 \geq \hat{\theta}_1, \\ -\sqrt{-2 \log \Lambda} \text{ if } \hat{\theta}_2 < \hat{\theta}_1, \end{cases} \tag{5}$$

and suggested by, for example, Cox and Hinkley (1974, p. 315), is more appropriate. The asymptotic distribution of L_s is standard normal under the null hypothesis (e.g., Sinharay, 2017).

In addition to the Wald test and the LRT-based test, the score test (e.g., Rao, 1973, p. 417) has also been used to test for equal abilities, for example, by Glas and Dagohoy

[1]Finkelman et al. (2010) suggested the computation of $s_1^2(\theta_1)$ and $s_2^2(\theta_2)$ using $\theta_1 = \theta_2 = \hat{\theta}_0$. Instead, one can use $\theta_1 = \hat{\theta}_1$ and $\theta_2 = \hat{\theta}_2$ to perform the Wald test—this variation did not produce results that are much different in a limited simulation. Therefore, results using $\theta_1 = \theta_2 = \hat{\theta}_0$ are reported in this paper.

(2007), Klauer and Rettig (1990), and Sinharay (2017). Denoting $\log(\mathcal{L}(\theta_1, \theta_2)) = \ell(\theta_1, \theta_2)$, the score statistic is given by

$$R = \left(\frac{\partial \ell(\theta_1, \theta_2)}{\partial \theta_1} |_{\theta_1 = \theta_2 = \hat{\theta}_0} \right)^2 s_1^2(\hat{\theta}_0) + \left(\frac{\partial \ell(\theta_1, \theta_2)}{\partial \theta_2} |_{\theta_1 = \theta_2 = \hat{\theta}_0} \right)^2 s_2^2(\hat{\theta}_0), \tag{6}$$

where, for example, $\frac{\partial \ell(\theta_1, \theta_2)}{\partial \theta_1} |_{\theta_1 = \theta_2 = \hat{\theta}_0}$, is the first partial derivative of $\ell(\theta_1, \theta_2)$ with respect to θ_1 at $\theta_1 = \theta_2 = \hat{\theta}_0$. From Equation 1,

$$\frac{\partial \ell(\theta_1, \theta_2)}{\partial \theta_1} |_{\theta_1 = \theta_2 = \hat{\theta}_0} = S_1 - \sum_i a_i \frac{\exp[a_i(\hat{\theta}_0 - b_i)]}{(1 + \exp[a_i(\hat{\theta}_0 - b_i)])}.$$

The R statistic is asymptotically equivalent to $-2 \log \Lambda$ and has an asymptotic χ^2 null distribution with one degree of freedom (e.g., Cox & Hinkley, 1974; Rao, 1973). For one-sided alternative hypothesis, Sinharay (2017) suggested the signed score (SS) statistic

$$R_s = \begin{cases} \sqrt{R} \text{ if } \hat{\theta}_2 \geq \hat{\theta}_1, \\ -\sqrt{R} \text{ if } \hat{\theta}_2 < \hat{\theta}_1, \end{cases} \tag{7}$$

and showed the asymptotic null distribution of R_s to be the standard normal distribution.

Modified Signed Likelihood Ratio Test

Barndorff-Nielsen (1986) suggested the modified signed likelihood ratio test (MSLRT) that can be used to test for the equality of two parameter vectors. Barndorff-Nielsen (1991), Barndorff-Nielsen and Cox (1994), Brazzale et al. (2007), and Reid (2003) provided accessible descriptions of the MSLRT. Let us consider the simple (and the most directly applicable to our case) application of the MSLRT when the probability model used to describe the data involves two parameters: ψ and λ. Let the (scalar) parameter of interest be denoted as ψ and let the hypothesis of interest be $H_0 : \psi = \psi_0$. Let λ denote the (scalar) nuisance parameter so that one has to estimate λ but one is not interested in testing any hypotheses regarding λ. This framework subsumes the case of testing of equality of two abilities for $\psi = \theta_2 - \theta_1; \lambda = \theta_1$, or, equivalently, $\theta_1 = \lambda; \theta_2 = \psi + \lambda$, and $\psi_0 = 0$.

Let $\ell(\psi, \lambda)$ denote the logarithm of the joint likelihood of ψ and λ for a set of data. Let $\hat{\psi}$ and $\hat{\lambda}$ jointly maximize $\ell(\psi, \lambda)$ with respect to ψ and λ, respectively, that is, $\hat{\psi}$ and $\hat{\lambda}$ are

the joint MLEs. Let $j(\psi, \lambda)$ denote the observed information matrix, or, equivalently, the negative of the matrix of the second derivatives of $\ell(\psi, \lambda)$. That is,

$$j(\psi, \lambda) = \begin{pmatrix} j_{\psi\psi}(\psi, \lambda) & j_{\psi\lambda}(\psi, \lambda) \\ j_{\lambda\psi}(\psi, \lambda) & j_{\lambda\lambda}(\psi, \lambda) \end{pmatrix} = - \begin{pmatrix} \frac{\partial^2 l(\psi,\lambda)}{\partial \psi^2} & \frac{\partial^2 l(\psi,\lambda)}{\partial \psi \partial \lambda} \\ \frac{\partial^2 l(\psi,\lambda)}{\partial \lambda \partial \psi} & \frac{\partial^2 l(\psi,\lambda)}{\partial \lambda^2} \end{pmatrix}, \tag{8}$$

where the individual elements of $j(\psi, \lambda)$ are denoted as $j_{\psi\psi}(\psi, \lambda)$, $j_{\psi\lambda}(\psi, \lambda)$, $j_{\lambda\psi}(\psi, \lambda)$, and $j_{\lambda\lambda}(\psi, \lambda)$.

Let $\ell_p(\psi)$ denote the *profile log likelihood* that is defined as the maximum value of the joint likelihood of ψ and λ when maximized with respect to λ for a fixed ψ, that is,

$$\ell_p(\psi) = \max_{\lambda} \ell(\psi, \lambda) = \ell(\psi, \hat{\lambda}_\psi) \cdot \tag{9}$$

Thus, $\hat{\lambda}_\psi$ is the constrained maximum likelihood estimate of λ, where ψ has been constrained to be equal to a fixed value. Note that the maximum profile likelihood estimate is the same as $\hat{\psi}$, that is,

$$\max_{\psi} \ell_p(\psi) = \ell_p(\hat{\psi}) \tag{10}$$

(e.g., Barndorff-Nielsen & Cox, 1994, p. 90).

To test $H_o : \psi = \psi_0$ against a one-sided alternative $H_1 : \psi > \psi_0$, one can use the SLR statistic or the likelihood-root statistic

$$r(\psi_0) = \text{sign}(\hat{\psi} - \psi_0) \sqrt{2[\ell_p(\hat{\psi}) - \ell_p(\psi_0)]} \tag{11}$$

(e.g., Brazzale et al., 2007, p. 139).

Some algebra shows that the expression for $r(\psi_0)$ for the 2PLM becomes identical to the expression for the L_s statistic provided in Equation 5. If H_0 is true, then $r(\psi_0)$ has a standard normal asymptotic distribution (e.g., Brazzale et al., 2007, p. 139).

Barndorff-Nielsen (1986) and Barndorff-Nielsen (1991) suggested the statistic

$$r^*(\psi_0) = r(\psi_0) + \frac{1}{r(\psi_0)} \log \frac{q(\psi_0)}{r(\psi_0)}, \tag{12}$$

referred to as the modified signed likelihood ratio (MSLR) statistic, where expressions of $q(\psi_0)$ are provided in Barndorff-Nielsen (1986) and Barndorff-Nielsen (1991); they also

proved that the statistic has a standard normal null distribution asymptotically. Further, several researchers such as Barndorff-Nielsen and Cox (1994, p. 203), Brazzale et al. (2007, p. 11), Jensen (1997), and Reid (2003, p. 1722) showed that for probability distributions that belong to the exponential family of distributions, $q(\psi_0)$ can be computed as

$$q(\psi_0) = (\hat{\psi} - \psi_0)\sqrt{\frac{|j(\hat{\psi}, \hat{\lambda})|}{j_{\lambda\lambda}(\psi_0, \hat{\lambda}_{\psi_0})}}. \tag{13}$$

The statistic $q(\psi_0)$ is a version of the Wald statistic (e.g., Brazzale et al., 2007, p. 12). Therefore, $r^*(\psi_0)$ can be considered to be an adjusted version of the SLR statistic $r(\psi_0)$, where the extent of adjustment depends on the relative magnitude of a version of the Wald statistic and the SLR statistic.

Lugannani-Rice Approximation

An alternative approach to the use of the MSLRT is the use of the Lugannani-Rice approximation (LRA; Lugannani & Rice, 1980) that expresses the probability of the SLR statistic being smaller than $r(\psi_0)$ under the null hypothesis as

$$\Phi^*(r(\psi_0)) = \Phi(r(\psi_0)) + \left[\frac{1}{r(\psi_0)} - \frac{1}{q(\psi_0)}\right]\phi(r(\psi_0)), \tag{14}$$

where ϕ denotes the standard normal density function and Φ denotes the standard normal cumulative distribution function. Thus, to test $H_o : \psi = \psi_0$ against $H_1 : \psi > \psi_0$, one using the LRA obtains the p-value as $(1 - \Phi^*(r(\psi_0)))$ whereas one using the MSLRT obtains the p-value as $(1 - \Phi(r^*(\psi_0)))$. Asymptotically, these two p-values are equivalent for probability distributions that belong to the exponential family of distributions (e.g., Brazzale et al., 2007; Jensen, 1992). The LRA and MSLRT approaches are often referred to as the p^* and the r^* approaches, respectively, in the literature on higher-order asymptotics.

The Advantages of the MSLRT and the Lugannani-Rice Approximation

The MSLRT and the LRA are covered under the umbrella of methods that are referred to as higher-order asymptotics (e.g., Barndorff-Nielsen & Cox, 1994; Ghosh, 1994) and have been found to lead to more accurate asymptotic approximations than the Wald test,

the LRT, and the score test that are based on first-order asymptotic theory (e.g. Pierce & Peters, 1992; Brazzale et al., 2007, p. 1).

While the asymptotic distribution of both $r^*(\psi_0)$ and $r(\psi_0)$ is standard normal under the null hypothesis, researchers such as Barndorff-Nielsen (1991) showed that for continuous response variables, the asymptotic result for $r^*(\psi_0)$ holds with a higher degree of approximation compared to $r(\psi_0)$. Specifically, when the response variable is continuous, the relative error from the use of $r^*(\psi_0)$ is typically $O(n^{-3/2})$, which means, for example, that if $\hat{p}$ denotes an estimated p-value computed using $r^*(\psi_0)$ and p denotes the corresponding true p-value, then

$$\left|\frac{\hat{p}-p}{p}\right| \leq \frac{M_1}{n^{3/2}} \text{ as } n \to \infty$$

(Barndorff-Nielsen, 1991), where n is the sample size and M_1 is a finite number. In contrast, the relative error from the use of $r(\psi_0)$ is typically $O(n^{-1/2})$, which means that if $\hat{p'}$ denotes an estimated p-value computed using $r(\psi_0)$, then

$$\left|\frac{\hat{p'}-p}{p}\right| \leq \frac{M_2}{n^{1/2}} \text{ as } n \to \infty$$

(Barndorff-Nielsen, 1991), where M_2 is a finite number. Because $\frac{M_1}{n^{3/2}}$ would typically be considerably smaller than $\frac{M_2}{n^{1/2}}$ and converges to 0 much faster than $\frac{M_2}{n^{1/2}}$ as $n \to \infty$, the p-value from the use of $r^*(\psi_0)$ is expected to be more accurate (that is, be close to the true p-value) compared to that from $r(\psi_0)$ for continuous response variables. Because the use of the LRA (Lugannani & Rice, 1980) is asymptotically equivalent to the use of $r^*(\psi_0)$ for the exponential family of distributions, the LRA has the same advantages as $r^*(\psi_0)$ over $r(\psi_0)$.

To test the null hypothesis $H_o : \psi = \psi_0$, it is also possible to employ the Wald statistic or the SS statistic (e.g., Brazzale et al., 2007, p. 139), whose expressions for the 2PLM are provided in Equations 2 and 7, respectively. The relative error of these two statistics is the same as that of $r(\psi_0)$ and is $O(n^{-1/2})$. Therefore, if the response variable is continous, the p-values originating from $r^*(\psi_0)$ and LRA are expected to be more accurate compared to those originating from the Wald statistic and the SS statistic as well.

The response variables in applications of IRT models are the item scores, which are discrete. No general result on the relative error for $r^*(\psi_0)$ or the LRA is available for such variables. However, if the probability distribution of a discrete response variable is a special case of the exponential family of distributions, it is possible to compute the MSLR statistic and the LRA using equations 12 and 14. Further, researchers such as Brazzale et al. (2007, Chapter 4) have found methods based on higher-order asymptotics to provide satisfactory results for discrete response variables. Therefore, even though the MSLR statistic and the LRA may not have relative error of $O(n^{-3/2})$, they may still lead to better results compared to the Wald test, SLR test and score test for IRT models. The following section includes descriptions of the computation of the MSLR statistic and the LRA for IRT models.

Methods: MSLRT and LRA for IRT Models

As mentioned above, the test for the equality of two ability parameters, that is, the test $H_0 : \theta_1 = \theta_2$ for IRT models can be placed in the framework discussed above by using the transformations $\psi = \theta_2 - \theta_1$ and $\lambda = \theta_1$ and letting ψ_0 to be equal to 0. These transformations imply that

$$\theta_2 = \psi + \lambda \text{ and } \theta_1 = \lambda\cdot \tag{15}$$

MSLRT and LRA to Test for the Equality of Two Abilities for the 2PLM

Equations 1 and 15 imply that the log-likelihood of ψ and λ, $\ell(\psi, \lambda)$, is equal to

$$\begin{aligned} & S_1\lambda - \sum_i X_i a_i b_i + \sum_i \log(1 - p_i(\lambda)) + S_2(\psi + \lambda) - \sum_j Y_j \tilde{a}_j \tilde{b}_j + \sum_j \log(1 - \tilde{p}_j(\psi + \lambda)) \\ = \; & S_2\psi + (S_1 + S_2)\lambda + \sum_i \log(1 - p_i(\lambda)) + \sum_j \log(1 - \tilde{p}_j(\psi + \lambda)) \\ & - \sum_i X_i a_i b_i - \sum_j Y_j \tilde{a}_j \tilde{b}_j \end{aligned} \tag{16}$$

for the 2PLM. The log-likelihood is a special case of the log-likelihood of the exponential family of distributions with canonical parameters ψ and λ and sufficient statistics S_2 and $(S_1 + S_2)$ because the first two terms above depend only on S_2, $(S_1 + S_2)$, and the

parameters, Terms 3-4 depend only on the parameters, and Terms 5-6 depend only on the data.[2] Therefore, the earlier discussion on the application of the MSLRT and the LRA to the exponential family of distributions implies that it is possible to apply the MSLRT and the LRA to test the hypothesis $H_0 : \theta_1 = \theta_2$ and to obtain an expression of $q(\psi_0)$ using Equation 13 for the 2PLM.

The first derivatives of $\ell(\psi, \lambda)$ are given by

$$\frac{\partial \ell(\psi,\lambda)}{\partial \psi} = S_2 - \sum_j \tilde{a}_j \frac{\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)])},$$
$$\frac{\partial \ell(\psi,\lambda)}{\partial \lambda} = S_1 + S_2 - \sum_i a_i \frac{\exp[a_i(\lambda-b_i)]}{(1+\exp[a_i(\lambda-b_i)])} - \sum_j \tilde{a}_j \frac{\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)])}.$$

Then, the matrix $j(\psi, \lambda)$ of the negative of the second derivatives of $\ell(\psi, \lambda)$ can be computed as

$$j(\psi,\lambda) = -\begin{pmatrix} \sum_j \tilde{a}_j^2 \frac{\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)])^2} & \sum_j \tilde{a}_j^2 \frac{\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)])^2} \\ \sum_j \tilde{a}_j^2 \frac{\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)])^2} & \sum_i a_i^2 \frac{\exp[a_i(\lambda-b_i)]}{(1+\exp[a_i(\lambda-b_i)])^2} + \sum_j \tilde{a}_j^2 \frac{\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)])^2} \end{pmatrix} \cdot \quad (17)$$

Noting that the off-diagonal elements of $j(\psi, \lambda)$ are the same as its first diagonal, the determinant of $j(\psi, \lambda)$ is obtained as

$$\begin{aligned} |j(\psi,\lambda)| &= \sum_i a_i^2 \frac{\exp[a_i(\lambda-b_i)]}{(1+\exp[a_i(\lambda-b_i)])^2} \sum_j \tilde{a}_j^2 \frac{\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\psi+\lambda-\tilde{b}_j)])^2} \\ &= \sum_i a_i^2 \frac{\exp[a_i(\theta_1-b_i)]}{(1+\exp[a_i(\theta_1-b_i)])^2} \sum_j \tilde{a}_j^2 \frac{\exp[\tilde{a}_j(\theta_2-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\theta_2-\tilde{b}_j)])^2}. \end{aligned} \quad (18)$$

To implement the MSLRT and LRA for the 2PLM, one has to compute the MLEs of θ_1 and θ_2, denoted $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively, and set the MLEs of ψ and λ as $\hat{\psi} = \hat{\theta}_2 - \hat{\theta}_1$ and $\hat{\lambda} = \hat{\theta}_1$, respectively. One also has to compute, under the restriction $\psi = \psi_0$, which is equivalent to the restriction $\theta_1 = \theta_2 = \theta_0$, the MLE of the common ability parameter θ_0—let us denote this MLE as $\hat{\theta}_0$. Note that $\hat{\theta}_0$ is computed from all the item scores (X_i's and Y_j's) for an examinee. The estimate $\hat{\theta}_0$ can also be denoted as $\hat{\lambda}_{\psi_0}$ according to the notation introduced earlier.

[2]Note that the item parameters are assumed known throughout this paper.

Then, one obtains $|j(\hat{\psi},\hat{\lambda})|$ by replacing θ_1 and θ_2 by $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively, in Equation 18, and obtains $j_{\lambda\lambda}(\psi_0,\hat{\lambda}_{\psi_0})$ from Equations 8 and 17 as

$$\begin{aligned} & j_{\lambda\lambda}(\psi_0,\hat{\lambda}_{\psi_0}) = \text{the second diagonal of } j(\psi,\lambda) \text{ computed at } \psi=\psi_0, \lambda=\hat{\lambda}_{\psi_0} \\ = & \left[\sum_i a_i^2 \frac{\exp[a_i(\hat{\lambda}_{\psi_0}-b_i)]}{(1+\exp[a_i(\hat{\lambda}_{\psi_0}-b_i)])^2}\right] + \left[\sum_j \tilde{a}_j^2 \frac{\exp[\tilde{a}_j(\hat{\lambda}_{\psi_0}-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\hat{\lambda}_{\psi_0}-\tilde{b}_j)])^2}\right] \\ = & \left[\sum_i a_i^2 \frac{\exp[a_i(\hat{\theta}_0-b_i)]}{(1+\exp[a_i(\hat{\theta}_0-b_i)])^2}\right] + \left[\sum_j \tilde{a}_j^2 \frac{\exp[\tilde{a}_j(\hat{\theta}_0-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\hat{\theta}_0-\tilde{b}_j)])^2}\right]. \end{aligned}$$

Then, from Equation 13, $q(\psi_0)$ is given by

$$q(\psi_0) = (\hat{\theta}_2-\hat{\theta}_1)\sqrt{\frac{\sum_i a_i^2 \frac{\exp[a_i(\hat{\theta}_1-b_i)]}{(1+\exp[a_i(\hat{\theta}_1-b_i)])^2} \sum_j \tilde{a}_j^2 \frac{\exp[\tilde{a}_j(\hat{\theta}_2-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\hat{\theta}_2-\tilde{b}_j)])^2}}{\sum_i a_i^2 \frac{\exp[a_i(\hat{\theta}_0-b_i)]}{(1+\exp[a_i(\hat{\theta}_0-b_i)])^2} + \sum_j \tilde{a}_j^2 \frac{\exp[\tilde{a}_j(\hat{\theta}_0-\tilde{b}_j)]}{(1+\exp[\tilde{a}_j(\hat{\theta}_0-\tilde{b}_j)])^2}}}, \tag{19}$$

which looks very similar to the Wald statistic given in Equation 2.[3]

Further, Equations 9-11 imply that $r(\psi_0)$, the SLR statistic, is obtained as

$$\begin{aligned} r(\psi_0) &= \text{sign}(\hat{\psi}-\psi_0)\left[2[\ell(\hat{\psi},\hat{\lambda})-\ell(\psi_0,\hat{\lambda}_{\psi_0})]\right]^{1/2} \\ &= \text{sign}(\hat{\psi}-\psi_0)\sqrt{2}\left[S_2(\hat{\psi}-\psi_0)+(S_1+S_2)(\hat{\lambda}-\hat{\lambda}_{\psi_0}))+\sum_i \log(1-p_i(\hat{\lambda}))\right. \\ &\quad \left. +\sum_j \log(1-\tilde{p}_j(\hat{\psi}+\hat{\lambda}))-\sum_i \log(1-p_i(\hat{\lambda}_{\psi_0}))-\sum_j \log(1-\tilde{p}_j(\hat{\lambda}_{\psi_0}))\right]^{1/2} \\ &= \text{sign}(\hat{\theta}_2-\hat{\theta}_1)\sqrt{2}\left[S_1(\hat{\theta}_1-\hat{\theta}_0)+S_2(\hat{\theta}_2-\hat{\theta}_0)-\sum_i \log(1+\exp[a_i(\hat{\theta}_1-b_i)]\right. \\ &- \sum_j \log(1+\exp[\tilde{a}_j(\hat{\theta}_2-\tilde{b}_j)])+\sum_i \log(1+\exp[a_i(\hat{\theta}_0-b_i)]) \\ &+ \left.\sum_j \log(1+\exp[\tilde{a}_j(\hat{\theta}_0-\tilde{b}_j)])\right]^{1/2}. \end{aligned} \tag{20}$$

Once $q(\psi_0)$ is computed using Equation 19 and $r(\psi_0)$ is computed using Equation 20, one can compute the MSLR statistic $r^*(\psi_0)$, which is a function of $q(\psi_0)$ and $r(\psi_0)$, using Equation 12, and can compute the LRA using Equation 14.

[3]If both $\hat{\theta}_1$ and $\hat{\theta}_2$ under the square root sign in the expression of $q(\psi_0)$ are replaced by $\hat{\theta}_0$, $q(\psi_0)$ would become identical to the Wald statistic.

If $q(\psi_0)$ and $r(\psi_0)$ are both close to zero, $r^*(\psi_0)$ involves the logarithm of the ratio of two very small numbers and hence becomes unstable (e.g., Jensen, 1995, p. 136). An approximation of $r^*(\psi_0)$ in such a case was obtained by performing calculations similar to that in Example 5.3.4 of Jensen (1995, p. 136) and was employed whenever $|r(\psi_0)| < 0.05$. The approximation involves tedious algebra and is not described here—it can be obtained upon request from the authors.

An Example

Let us consider the application of the MSLRT and LRA to detect item preknowledge for two examinees belonging to a data set that would be later described in the Real Data Section.

Table 1. The Values of Several Quantities for Two Examinees.

Examinee	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_0$	$r(\psi_0)$	$q(\psi_0)$	$r^*(\psi_0)$	$\Phi(r^*(\psi_0))$	$\Phi^*(r(\psi_0))$
1	-1.80	-0.20	-1.43	3.09	2.77	3.05	0.9989	0.9989
2	-1.44	-0.83	-1.28	1.22	1.16	1.18	0.8810	0.8810

Table 1 lists the values of $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_0$, $r(\psi_0)$, $q(\psi_0)$, $r^*(\psi_0)$, $\Phi(r^*(\psi_0))$ and $\Phi^*(r(\psi_0))$ for the two examinees. The null hypothesis corresponds to no cheating (or no item preknowledge) while the alternative hypothesis corresponds to potential cheating. The p-values obtained from the MSLR statistic and the LRA were identical up to 4 decimal places for both the examinees. For the first examinee, $\hat{\theta}_2$ is considerably larger than $\hat{\theta}_1$; naturally, H_0 is rejected at level 0.01 by all of SLR, MSLRT, and LRA. For the second examinee, $\hat{\theta}_2$ is not much larger than $\hat{\theta}_1$ and H_0 is not rejected at level 0.05 by any of the statistics.

Like in this example, the MSLRT and LRA led to very similar results in the analysis of simulated data and real data in this paper. Therefore, only the results of the LRA among these two statistics are discussed henceforth.

MSLRT and LRA to Test for the Equality of Two Abilities for the GPCM

Let us consider the GPCM (Muraki, 1992) for which the log-likelihood of the examinee ability on a test with n polytomous items is given by

$$\ell(\theta) = \sum_{i=1}^{n} \sum_{k=0}^{m_i} d_k(X_i) \log P_{ik}(\theta), \tag{21}$$

where X_i, the examinee's score on item i, is an integer between 0 and m_i,

$$d_k(X_i) \;\; = \;\; \begin{cases} 1 \text{ if } X_i = k, \\ 0 \text{ otherwise,} \end{cases}$$

$$P_{ik}(\theta) = P(d_k(X_i) = 1) = E(d_k(X_i))$$

is the probability that the examinee's score on item i is equal to k and is given by

$$P_{ik}(\theta) = \frac{\exp[\sum_{h=0}^{k} a_i(\theta - b_{ih})]}{\sum_{c=0}^{m_i} \exp[\sum_{h=0}^{c} a_i(\theta - b_{ih})]} = \frac{\exp[\sum_{h=0}^{k} a_i(\theta - b_{ih})]}{\Gamma_i(\theta)}, \tag{22}$$

where a_i's and b_{ih}'s are the slope and location parameters, respectively, and

$$\Gamma_i(\theta) = \sum_{c=0}^{m_i} \exp\left[\sum_{h=0}^{c} a_i(\theta - b_{ih})\right].$$

Let us denote an examinee's scores on two sets of polytomous items as $X_i, i = 1, 2, \ldots, n_1$, and $Y_j, j = 1, 2, \ldots, n_2$, and the underlying true abilities as θ_1 and θ_2, respectively. Let us further assume that the possible scores on item i range between $k = 0, 1, \ldots, m_i$, and those on item j range between $k = 0, 1, \ldots, m_j$. Let us also assume that the probabilities given by Equation 22 are denoted by $P_{ik}(\theta_1)$ for the first set of items and by $\tilde{P}_{jk}(\theta_2)$ for the second set of items. It is proved in the appendix that for the GPCM, one can compute the MSLR statistic $r^*(\psi_0)$ using Equation 12 and the LRA using Equation 14, where $q(\psi_0)$ is given by

$$q(\psi_0) = (\hat{\theta}_2 - \hat{\theta}_1) \sqrt{\frac{\sum_{i=1}^{n_1} a_i^2 \operatorname{Var}(X_i|\hat{\theta}_1) \sum_{j=1}^{n_2} \tilde{a}_j^2 \operatorname{Var}(Y_j|\hat{\theta}_2)}{\sum_{i=1}^{n_1} a_i^2 \operatorname{Var}(X_i|\hat{\theta}_0) + \sum_{j=1}^{n_2} \tilde{a}_j^2 \operatorname{Var}(Y_j|\hat{\theta}_0)}}, \tag{23}$$

where, for example,

$$\operatorname{Var}(X_i|\hat{\theta}_1) = \sum_{k=0}^{m_i} k^2 P_{ik}(\hat{\theta}_1) - \left[\sum_{k=0}^{m_i} k P_{ik}(\hat{\theta}_1)\right]^2,$$

and $r(\psi_0)$ is given by

$$r(\psi_0) = \text{sign}(\hat{\theta}_2 - \hat{\theta}_1)\sqrt{2}\left[\sum_{i=1}^{n_1}\sum_{k=0}^{m_i} d_k(X_i)\log P_{ik}(\hat{\theta}_1) + \sum_{j=1}^{n_2}\sum_{k=0}^{m_j} d_k(Y_j)\log \tilde{P}_{jk}(\hat{\theta}_2) - \sum_{i=1}^{n_1}\sum_{k=0}^{m_i} d_k(X_i)\log P_{ik}(\hat{\theta}_0) - \sum_{j=1}^{n_2}\sum_{k=0}^{m_j} d_k(Y_j)\log \tilde{P}_{jk}(\hat{\theta}_0)\right]^{1/2}. \quad (24)$$

If X_i's and Y_j's are dichotomous (that is, $m_i = m_j = 1$), then, for example, $\text{Var}(X_i|\hat{\theta}_1) = p_{i1}(\hat{\theta}_1)[1 - p_{i1}(\hat{\theta}_1)] = \frac{\exp[a_i(\hat{\theta}_1 - b_i)]}{(1+\exp[a_i(\hat{\theta}_1 - b_i)])^2}$, and Equation 23 becomes equal to Equation 19, the corresponding expression for the 2PLM, and Equation 11 becomes equal to Equation 20. Thus, the expression of the MSLR statistic for the 2PLM is a special case of that for the GPCM.

Simulation Studies

Simulation 1: Measurement of Change Using Dichotomous Items

A simulation somewhat similar to one in Finkelman et al. (2010) was performed to compare the performances of the statistics in the context of measurement of change. It was assumed that the null hypothesis is that the ability θ_1 at time point 1 is equal to that (θ_2) at time point 2, that is, $H_0 : \theta_1 = \theta_2$, and the alternative hypothesis is $H_1 : \theta_2 > \theta_1$.

Two non-adaptive assessments, each with 10, 20, 30, or 50 dichotomous items, were used as the assessments administered at time points 1 and 2. As in Finkelman et al. (2010), the true value of either of θ_1 or θ_2 was considered to be equal to one among -2, -1.5, ...,1.5, 2. Item-scores of 100,000 examinees were simulated for each possible combination of true θ_1 and true θ_2 where $\theta_1 \leq \theta_2 \leq \min(\theta_1 + 1.5, 2.0)$. Thus, for example, when true θ_1 is 0.5, true θ_2 can take only one of the four values 0.5, 1, 1.5, and 2.0; this strategy limits the maximum change in ability to 1.5. The nine combinations with true θ_1 = true θ_2, that is, the combinations (-2,-2), (-1.5,-1.5), ... ,(2,2), represent the "no change" condition and were used to study the Type I error rates of the statistics. The remaining 21 conditions with true $\theta_2 >$ true θ_1 represent the "positive change" condition and were used to study the power of the statistics.

The 2PLM was used in the analysis. The sets of true item parameters for the two assessments were non-overlapping and were randomly drawn from a set of estimated item parameters from a language test that employs the 2PLM operationally. The MLE of ability, restricted to the range between -4.0 and 4.0, was used in the computations.[4] For each simulated examinee, item scores were simulated on the two assessments, the MLE of the ability was separately computed on the first assessment, second assessment, combined assessment, and then the Wald, LRT, SLR, and SS statistics and the LRA were computed using Equations 2, 4, 5, 7, and 14, respectively.

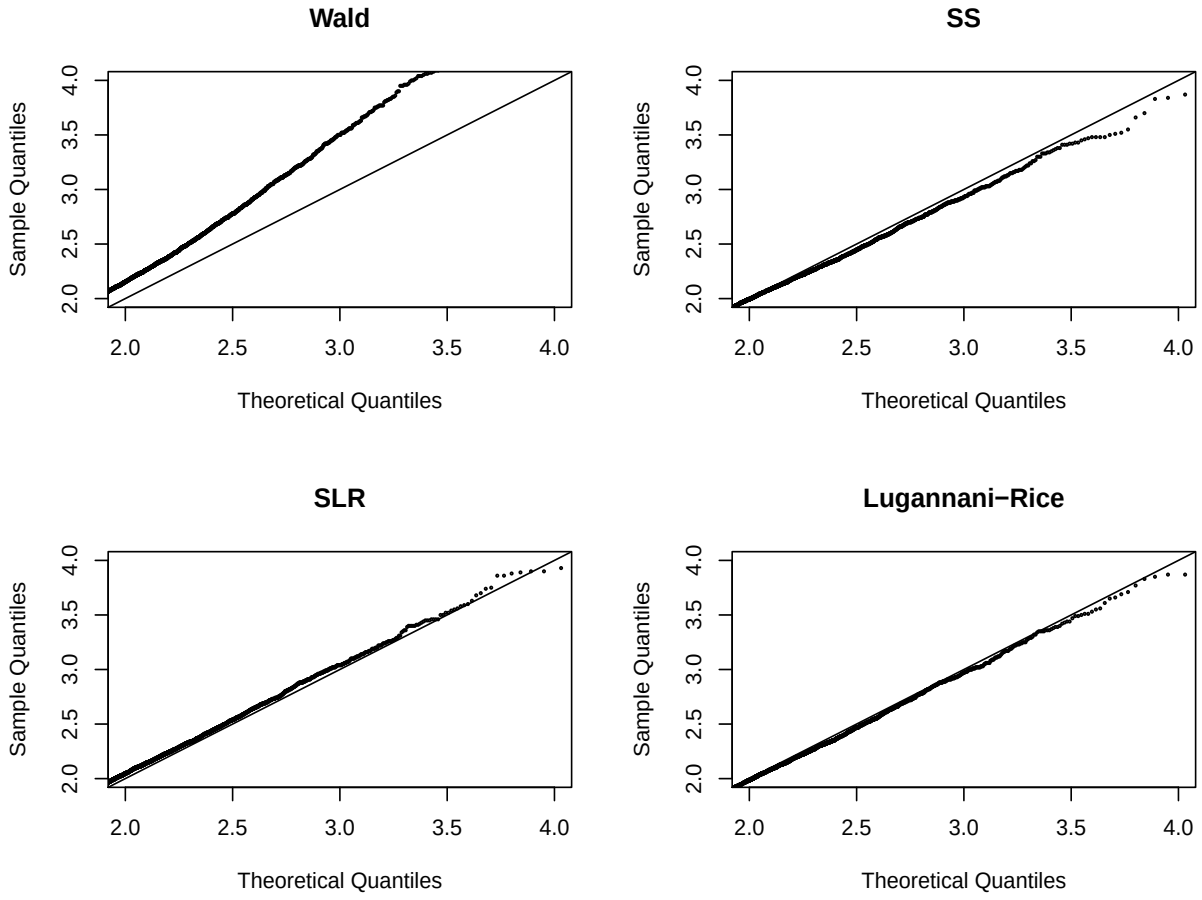

Figure 1. Normal Quantile Plots for the Four Statistics.

[4]The results using the weighted maximum likelihood estimator (WLE; Warm, 1989) of ability were very similar.

Figure 1 shows the normal quantile plots, created using the function qqnorm in the R software (R Core Team, 2017)[5] for the Wald statistic, the SS statistic, the SLR statistic, and the LRA for all the examinees under the "no change" condition for test length of 30. The theoretical quantiles of the standard normal distribution are shown along the X-axis and the sample quantiles are shown along the Y-axis. In each panel, a diagonal line is also provided. The closeness of each curve to the diagonal line is a measure of the closness of the null distribution of the corresponding statistic to the standard normal approximation. Only the region where the quantiles are between 2 and 4 is shown in each panel.[6]

Figure 1 shows that for the Wald statistic, the sample quantiles are larger than the theoretical quantiles—this phenomenon suggests that the Wald statistic would have an inflated Type I error rate under a standard normal null distribution assumption. The curves for the SS and SLR statistics are comparatively closer to the diagonal line except for values larger than about 2.5 where sample quantiles are slightly smaller than the theoretical quantiles for the SS statistic and larger for the SLR statistic. The curve for the LRA is closest, among the four statistics, to the diagonal line for large values; this result implies that under the standard normal null distribution assumption, the Type I error rate for the LRA is

- closest to the nominal level among these four statistics,
- slightly more satisfactory than that of the SLR and SS statistics, and
- considerably more satisfactory than that of the Wald statistic, especially at small levels of significance.

In addition, the sample quantiles for the LRA are slightly smaller than or equal to the

[5]To create the plot for LRA, which lies between 0 to 1, the standard normal quantile of the LRA was used as the input; that is because the use of the LRA provided by Equation 14 to test H_0 is equivalent to the use of the standard normal quantile of the LRA as a statistic along with a standard normal null distribution assumption.

[6]For quantiles between -4 and 2, the curves for the SS and SLR statistics and the LRA were very close to the diagonal line.

Table 2. Summaries of the Distributions of the Four Statistics Under the Null Hypothesis for Test Length of 30.

Statistic	Moments				Percentiles				
	Mean	SD	Skewness	Kurtosis	25	50	75	95	99
$\mathcal{N}(0,1)$	0.00	1.00	0.00	0.00	-0.67	0.00	0.67	1.64	2.33
Wald	0.00	1.07	-0.02	0.47	-0.69	0.00	0.69	1.74	2.54
SS	.00	1.01	0.00	-0.14	-0.69	0.00	0.69	1.66	2.29
SLR	0.00	1.02	0.00	-0.02	-0.69	0.00	0.69	1.69	2.36
LRA	0.00	0.99	0.00	-0.02	-0.68	-0.01	0.67	1.63	2.30

theoretical quantiles through out, which implies that the Type I error rate for the statistic would not typically exceed the nominal level even for very small levels.

Table 2 provides the first four moments (mean, SD, skewness, and kurtosis[7]) and five percentiles (25th, median, 75th, 95th, and 99th) of the standard normal distribution and the distributions of the four statistics for all the examinees under the "no change" condition and test length 30. Table 2 shows that the summary statistics for the Wald statistic are the farthest from those of the $\mathcal{N}(0,1)$ distribution and those of the LRA are overall closest to those of the $\mathcal{N}(0,1)$ distribution.

Figure 2 shows the average Type I error rate (three panels on the left) and power (three panels on the right) for different test lengths of the Wald statistic, the SS statistic, the SLR statistic, and the LRA, averaged over all the true values of θ_1 and θ_2. The top two, middle two, and bottom two panels show the results for significance levels of 0.001, 0.01, and 0.05, respectively. The Type I error rate of a statistic was computed as the proportion of statistically significant values of the statistic among the examinees with true θ_1 equal to true θ_2. The power of a statistic was computed as the proportion of statistically significant values of the statistic among the examinees with true θ_2 larger than true θ_1. Note that the smallest significance level (0.001) that is considered in the figure is important because some of the applications of the test of equality of abilities involve detection of cheating on

[7]Note that 3 has been subtracted from the formula of kurtosis so that the kurtosis of the standard normal distribution is 0 according to the formula used in this paper.

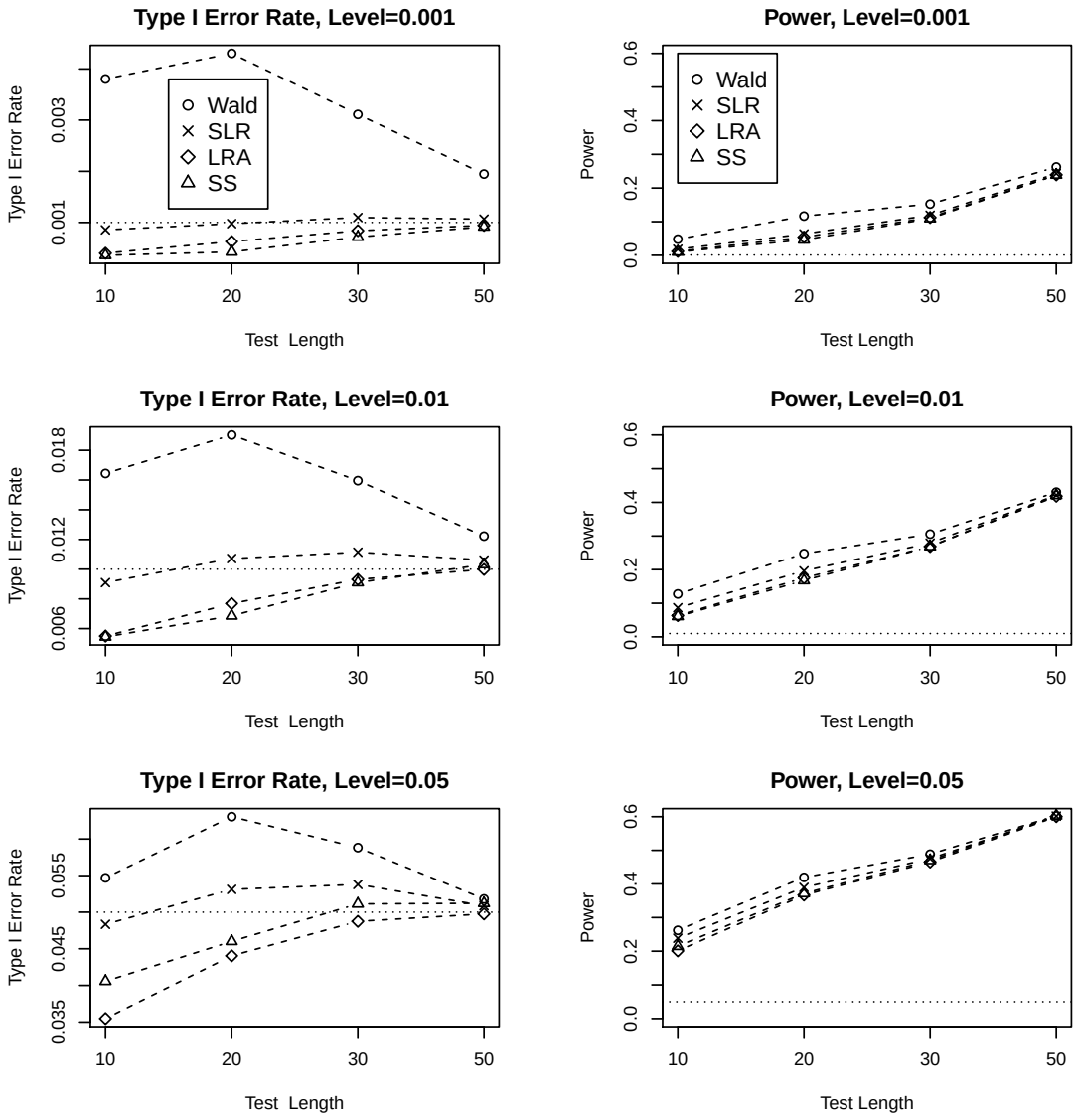

Figure 2. The Type I error rate and power for the four statistics for Simulation 1.

assessments where very small significance levels are often recommended (e.g., Wollack et al., 2015) to avoid potential adverse consequences of a false detection. In any panel of Figure 2, the test length is shown along the X-axis and the average Type I error rates or power for each test length is shown along the Y-axis. The values for the Wald statistic, SS statistic, SLR statistic, and LRA are shown using circles, triangles, cross signs, and diamond signs, respectively. A horizontal dotted line in each of the three panels on the left indicate the

nominal level. Although the range of the vertical axis varies over the three panels on the left, they are the same in the three panels on the right.

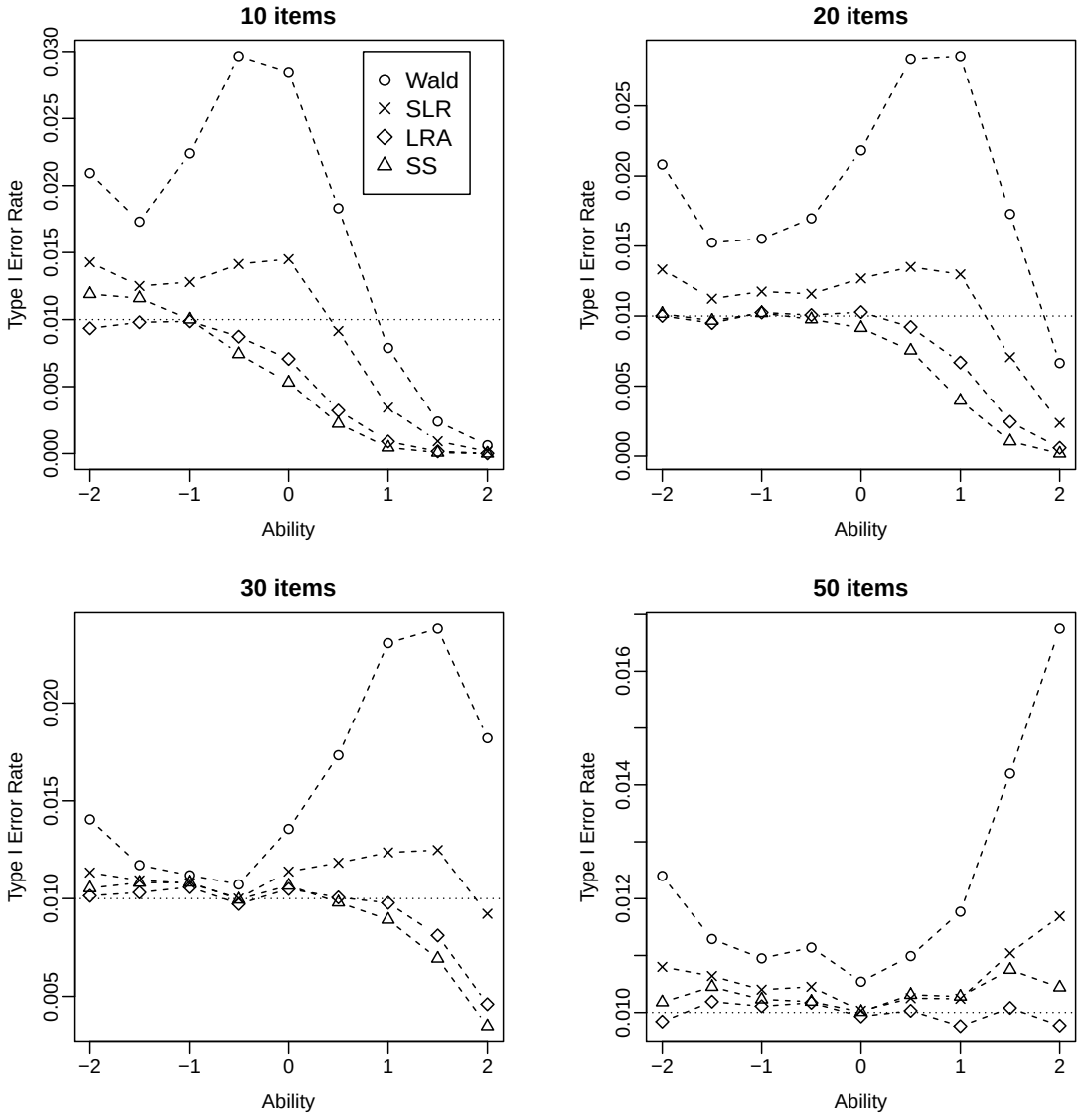

Figure 3. How Type I Error Rates Vary Over the True Abilities for the Four Statistics.

The four panels of Figure 3 show the Type I error rates of the Wald statistic, the SS statistic, the SLR statistic, and the LRA for different values of $\theta_1 = \theta_2$ for the significance level of 0.01. A dotted horizontal line in each panel shows the values of the level.

Figures 2 and 3 show that among the statistics, the LRA has the most satisfactory

Type I error rates overall. The Type I error rates for the LRA are slightly smaller than or equal to the nominal level for all combinations of test length and significance level. The Type I error rate of the Wald statistic is severely inflated at the level 0.001 and inflated at the other levels. This is expected from the top left panel of Figure 1. The Wald statistic was found to have inflated Type I error rate by Guo and Drasgow (2010) as well. The Type I error rates of the SS and SLR statistics are slightly inflated in some cases such as the bottom right panel of Figure 3. As test length increases, the Type I error rate of each statistic converges to the nominal level in all the three panels on the left of Figure 2[8], but the rate for the Wald statistic is larger than the nominal level even for the test length of 50. For the test length of 50, the Type I error rate of the LRA is identical to the significance level in each of the three panels on the left of Figure 2 and in the bottom right panel of Figure 3 while that of the SS and SLR statistics is slightly larger than the nominal level in the two bottom panels on the left of Figure 2 and in the bottom right panel of Figure 3.

Figure 2 shows that the power of each statistic increases steadily as test length increases. The power of the Wald statistic is the largest for all combinations of test length and significance level. The values of power of the other three statistics including the LRA are very close to those of each other. While Figure 2 shows the power after averaging over the three possible values of the ability difference, a separate analyses (whose results are not shown here and can be obtained from the authors upon request) revealed that the power of each statistic increases as the difference in ability increases. For example, for test length 50 and significance level of 0.05, the power of each statistic is close to 0.3, 0.7, and 0.9, respectively, when the ability difference is 0.5, 1.0, and 1.5.

Overall, it seems that the LRA achieves the nominal Type I error rate without losing too much power in comparison to the Wald, SLR, or SS statistics. Also, as test length increases, the Type I error rates of the other statistics converge to slightly above the nominal level in some cases, but that of the LRA converges to the exact significance level.

[8]that is expected given that the null distribution of all these statistics converges to the standard normal distribution as test length increases

Simulation 2: Measurement of Change Using Polytomous Items

A simulation like the earlier one was performed to compare the performances of the statistics for testing $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_2 > \theta_1$ in the context of measurement of change using polytomous items. Two non-adaptive assessments, each with 5, 10, 20, or 40 5-category polytomous items, were used as the assessments administered at time points 1 and 2. As in the earlier simulation, the true value of either of θ_1 or θ_2 was considered to be equal to one among -2, -1.5, ...,1.5, 2. Item-scores of 100,000 examinees were simulated for each possible combination of true θ_1 and true θ_2 where $\theta_1 \leq \theta_2 \leq \min(\theta_1 + 1.5, 2.0)$.

The GPCM (Muraki, 1992) was used in the analysis. The sets of true item parameters for the two assessments were randomly drawn from a set of estimated item parameters from a data set from the NEO Personality Inventory that is considered in the second real data example in this paper. The MLE of ability, restricted to the range -4.0 and 4.0, was used in the computations.

Figure 4 shows the average Type I error rate (the three panels on the left) and power (the three panels on the right) for different test lengths of the four statistics, averaged over all the true values of θ_1 and θ_2, at three significance levels: 0.001, 0.01, and 0.05.

The LRA has the most satisfactory Type I error rates (that are very close to the nominal level) followed by the SLR statistic in Figure 4. The Wald statistic, again, has inflated Type I error rates. The SS statistic occasionally has slightly inflated Type I error rates (for example, for test length of 5 and 10 in the bottom left panel). As in Simulation 1, the power of the Wald statistic is the largest for all cases and the values of power of the other three statistics are very close to those of each other.

Simulation 3: Detection of Item Preknowledge

A simulation somewhat similar to that in Sinharay (2017) was performed to compare the performance of the statistics in the context of detecting item preknowledge when a known set of items has been compromised. In these simulations, a set of items was assumed to have been compromised. It was assumed that the null hypothesis is that the ability over the non-compromised items (θ_1) is equal to that over the compromised items (θ_2), that is,

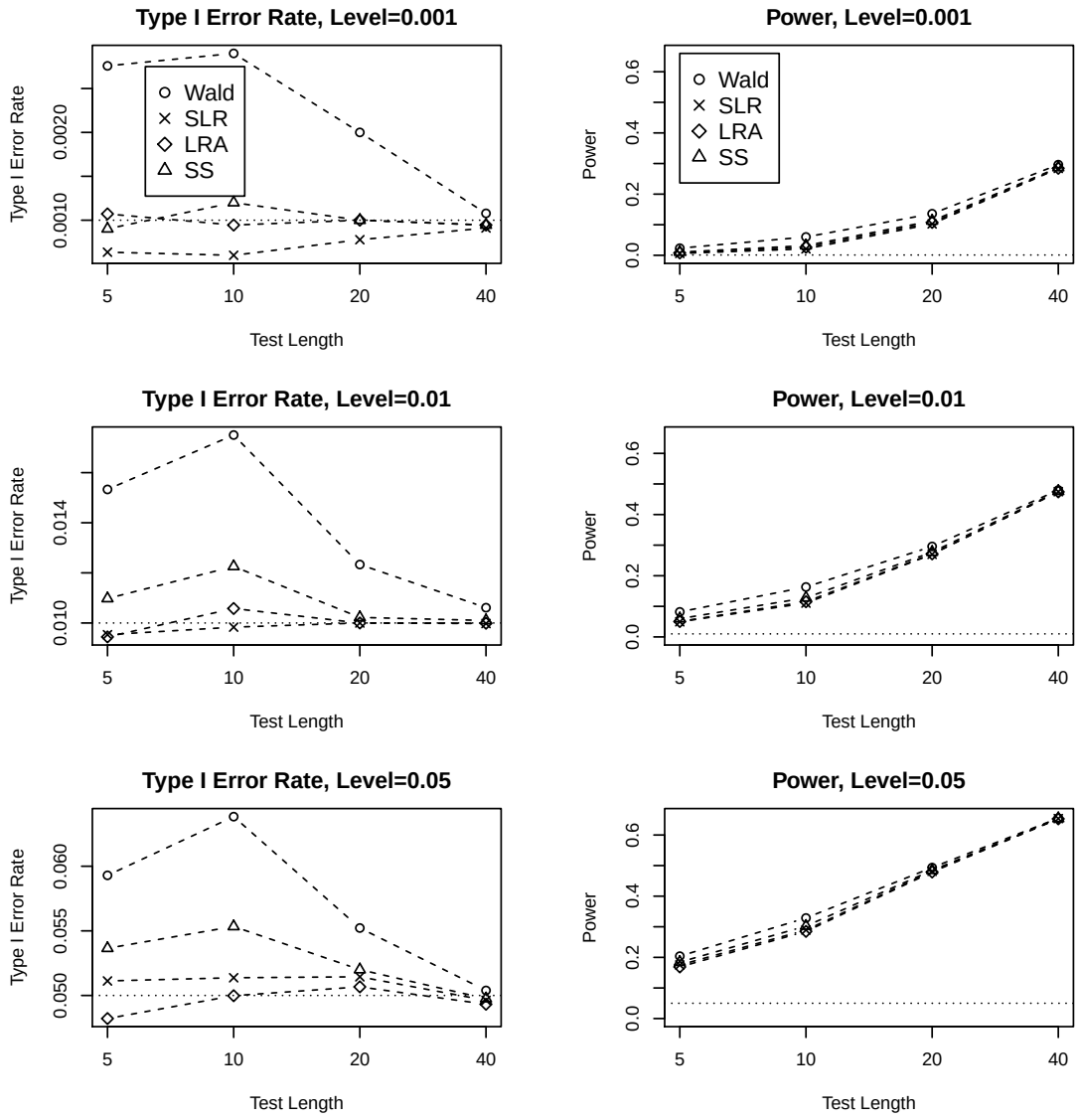

Figure 4. The Type I error rate and power for the four statistics for Simulation 2.

$H_0 : \theta_1 = \theta_2$; the alternative hypothesis was $H_1 : \theta_2 > \theta_1$.

A non-adaptive assessment with 100 dichotomous items was used as the whole assessment. The true values of the ability of those who did not benefit from item preknowledge (non-cheaters) were simulated from a standard normal distribution. The true values of the ability of those who benefited from item preknowledge (cheaters) were simulated from a standard normal distribution or a $\mathcal{N}(-0.5, 1)$ distribution; the first

distribution represents the case when the cheaters are the same as the non-cheaters in ability and the second represents the case when the cheaters have lower ability on average than the non-cheaters. The set of compromised items was assumed to be a subset of size 10, 20, or 30 of the whole assessment; thus, the set of non-compromised items included the remaining 90, 80, or 70 items of the assessment. The proportion of cheaters was assumed to be 0.05, 0.10, or 0.20 of the number of non-cheaters. It was assumed that for each simulation condition, the number of non-cheaters was 10,000; this means that the number of cheaters was 500, 1000, or 2000 in the various simulation conditions. Thus, 18 simulation conditions (involving all combinations of two ability distributions of the cheaters, three sizes of the set of compromised items, and three proportions of cheaters) were used.

The 2PLM was used in the analysis. The item-scores of all examinees on the non-compromised items and the item-scores of the non-cheaters on the compromised items were simulated from the 2PLM. It was assumed, as in Sinharay (2017), that if an examinee has preknowledge of an item, his/her probability of a correct answer on the item was 0.90; therefore, the item-scores of the cheaters on the compromised items were simulated as draws from a Bernoulli distribution with a success probability of 0.90.

The true item parameters were randomly drawn from the set of estimated item parameters from a real data set that is discussed later in this paper. The MLE of ability, restricted to the range -4.0 and 4.0, was used in the computations. For each examinee, the true ability and the item scores were simulated (where the simulating distributions depend on whether the examinee cheated or not), the ability estimate was separately computed on the compromised items, non-compromised items, and all items, and the Wald test statistic, the SLR statistic, the SS statistic, and the LRA were computed.

Table 3 shows the Type I error rates and power of the four statistics, averaged over all the simulation conditions, at three significance levels: 0.001, 0.01, and 0.05. The Type I error rate of a statistic was computed as the proportion of statistically significant values of the statistic among the non-cheaters. The power of a statistic was computed as the proportion of statistically significant values of the statistic among the cheaters.

Table 3 shows that the Type I error rates for the LRA are very close to the nominal

Table 3. The Type I error rate and power for the three statistics for Simulation 3.

Significance	Type I Error Rate				Power			
Level	Wald	SS	SLR	LRA	Wald	SS	SLR	LRA
0.001	.0233	.0005	.0012	.0007	.34	.17	.22	.19
0.01	.0340	.0066	.0141	.0096	.45	.31	.40	.36
0.05	.0727	.0382	.0620	.0500	.58	.50	.56	.54

level at levels 0.01 and 0.05 and slightly conservative at the level of 0.001. The Type I error rates of the Wald statistic are severely inflated at all the levels. Guo and Drasgow (2010) found the Wald statistic to often have inflated Type I error rates in the context of detection of cheating on unproctored tests. The Type I error rates of the SLR statistic are slightly inflated at all the levels. The Type I error rates of the SS statistic are smaller than the nominal level in all cases. The power of the Wald statistic was the largest for all significance levels and that of the SS statistic is the smallest for all levels. The power of the LRA is the second smallest for all the three levels although its power is within 0.04 of that of the SLR. Overall, it seems that as in the earlier simulations, the LRA achieves the nominal Type I error rate without losing too much power in comparison to the existing statistics.

Figure 5 shows the average Type I error rates and power of the Wald, SS, and SLR statistics and the LRA for different number of compromised items (10, 20, or 30). The three panels on the left show the Type I error rates and the three panels on the right show power. Figure 5 shows that the Type I error rate of the Wald statistic decreases to the nominal level as the number of compromised items increases, but is considerably larger than the nominal level even for 30 compromised items. The figure also shows that the Type I error rate of the SLR statistic decreases to the nominal level as the number of compromised items increases, but is slightly larger than the nominal level even for 30 compromised items for levels of 0.01 and 0.05. In contrast, the Type I error rate for the LRA is very close to the nominal level in all cases. This result provides an empirical proof of a much faster convergence of the LRA to the true p-value compared to the Wald statistic and a slightly faster convergence of the LRA to the true p-value compared to the SLR and SS statistics.

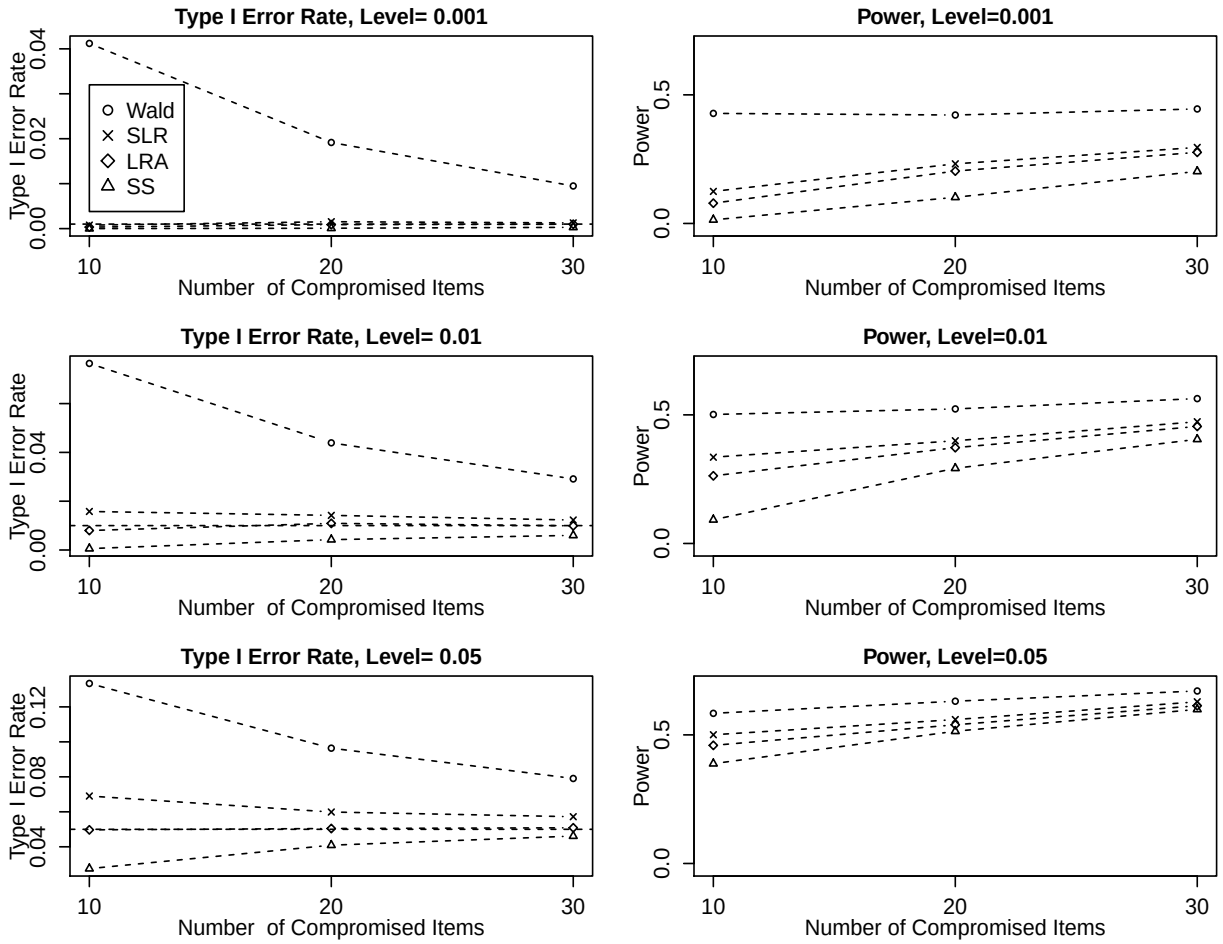

Figure 5. Type I Error Rate and Power for Different Number of Compromised Items.

At a given level, the power of all statistics increases with an increase in the number of compromised items.

Real Data Examples

Example 1: Detection of Item Preknowledge

Let us consider item-response data from one form of a non-adaptive licensure assessment. The data set was analyzed in several chapters of Cizek and Wollack (2017). The form includes 170 operational items that are dichotomously scored. Item scores were available for 1,644 examinees for the form. The licensure organization who provided the data identified 61 items on the form as compromised. The organization also flagged 48

individuals on the form as possible cheaters from a variety of statistical analysis and a rigorous investigative process that brought in other information; given the rigor of the investigative process, these examinees will be treated as true cheaters.

Table 4. The Proportion Significant for the Various Statistics for the Real Data Set.

Group of	Statistic	.001	.01	.05
Not flagged	Wald	.014	.038	.084
Not flagged	SS	.006	.033	.079
Not flagged	SLR	.005	.031	.078
Not flagged	LRA	.004	.028	.074
Flagged	Wald	.125	.271	.312
Flagged	SS	.125	.188	.312
Flagged	SLR	.125	.188	.312
Flagged	LRA	.125	.188	.312

The values of the Wald statistic, SS statistic, and the SLR statistic were computed for each individual in the data set. The LRA was also used to compute a p-value for each examinee. The set of 109 non-compromised items was considered as the first set of items and the set of 61 compromised items were considered as the second set of items. The null hypothesis of equal abilities over the two set of items ($H_0 : \theta_1 = \theta_2$) corresponds to no cheating while the alternative hypothesis corresponds to potential cheating ($H_1 : \theta_2 > \theta_1$). The Rasch model is operationally used in the assessment—the 2PLM was found to fit the data better and was used for the analysis. The item parameters were estimated using the marginal maximum likelihood estimation procedure from the data set and used in the computation of the statistics. The MLEs of the abilities, restricted to the range -4.0 and 4.0, were used to compute the statistics.

The proportion of examinees for which the statistics were significant at significance levels of 0.001, 0.01, and 0.05 are provided in Table 4. The first four rows of Table 4 include the proportions significant among the examinees who were not flagged by the licensure organization. The last four rows of the table include the proportions significant only among the 48 examinees who were flagged by the licensure organization; thus, for example, the

proportion 0.125 for the Wald statistic at level=0.001 in the fifth row of numbers implies that among the 48 examinees flagged by the licensure organization, the Wald statistic was significant at level=0.001 for six examinees (note that 6/48=0.125).

Table 4 shows that the proportion of significant values for the Wald statistic is larger than or equal to that for the other statistics in all cases, which is in agreement with the simulation studies. Table 4 also shows that the proportions of significant values are close for the SLR and SS statistics and LRA in all cases; among these three statistics, the LRA leads to a smaller percentage of significant values than the SLR and SS statistics among the non-flagged examinees and leads to an equal percentage of significant values as the SLR and SS statistics among the flagged examinees.

Table 4 also shows that the proportion significant for each statistic is much larger among the examinees flagged by the licensure organization (bottom four rows of the table) than among those not flagged (top four rows of the table)—this result provides some evidence that the statistics are somewhat successful—they are significant at a larger rate among the examinees who are true cheaters.

Example 2: Comparison of Performance Over Two Subtests

Let us consider a data set from NEO Personality Inventory that was analyzed by Glas and Dagohoy (2007). The NEO Personality Inventory is a personality test designed to provide a general description of normal personality that is relevant to clinical, counseling, and educational situations. The inventory is based on the five-factor model of personality (Costa & McCrae, 1992) and consists of five broad domains. For each domain, six facet scores have been developed to provide specific levels of information. Each facet is measured by eight items each of which is rated on a five-point scale. Data from 1,168 individuals on the neuroticism domain was analyzed in Glas and Dagohoy (2007) who split the 48 items on the domain into three sub-tests so that Items 1-8 and 9-16 in each sub-test relate to different facets; they also found the data within each sub-test to be unidimensional. The unidimensional GPCM was separately fitted to each sub-test and the estimated item parameters were used to test the null hypothesis that the examinee ability is the same over

Items 1-8 and 9-16 (that is, same over the two facets under the sub-test) against a one-sided alternative hypothesis. The MLE of examinee ability, restricted between -4 and 4, was used in the computations.

Table 5. The Proportion Significant for the statistics for the Second Real Data Example.

Sub-test	Wald	SS	SLR	LRA
1	.138	.081	.119	.118
2	.132	.091	.125	.124
3	.129	.084	.110	.109

Table 5 shows the proportions of statistically significant p-values at 5% level for the three sub-tests for the Wald statistic, SS statistic[9], SLR statistic, and LRA. The proportions are largest for the Wald statistic followed by the SLR statistic. The proportions for the SLR statistic and LRA are very close.

Conclusions

Hypothesis-testing approaches based on higher-order asymptotics (Barndorff-Nielsen & Cox, 1994; Ghosh, 1994) were applied to the problem of testing of whether the ability of an examinee is the same over two sets of items. Such problems arise in various contexts in educational and psychological measurement including measurement of change (e.g., Fischer, 2003) and detection of test cheating (e.g., Guo & Drasgow, 2010; Wollack & Schoenig, 2018). The modified signed likelihood ratio test (MSLRT; Barndorff-Nielsen, 1986) and the Lugannani-Rice approximation (LRA; Lugannani & Rice, 1980) were found to perform better than the signed likelihood ratio (SLR) statistic, the signed score (SS) statistic (e.g., Cox & Hinkley, 1974; Sinharay, 2017), and the Wald statistic (e.g., Cox & Hinkley, 1974; Finkelman et al., 2010). In the simulations, the MSLRT and LRA led to Type I error rates that are quite close to the nominal level, even when these tests are based on a few items, and not larger the nominal level in general; this result is encouraging because false positives

[9]the SS statistic in this case is a signed square root of the statistic that Glas and Dagohoy (2007) used to test against a two-sided alternative

in the context of detection of test cheating may have dire consequences (e.g., Skorupski & Wainer, 2017) and should be minimized (e.g., Ferrara, 2017) and the MSLRT and LRA would lead to the fewest false positives among the statistics considered here; especially, the satisfactory Type I error rate of the LRA for small significance levels is promising because of the typical use of conservative significance levels (such as 0.001) in detection of test cheating (e.g., Wollack et al., 2015).

The suggested statistics can be considered as person-fit statistics in the same way that the Langrangian multiplier test statistic of Glas and Dagohoy (2007) or the three statistics of Klauer and Rettig (1990) are person-fit statistics—the suggested statistics are computed for each individual examinee and they can be used to detect one specific type of person misfit—one characterized by a difference in performance over two sets of items.

The choice of the significance level to be used with the suggested statistics is an important issue. Wollack and Eckerly (2017, p. 227) used the significance level of 0.001 in their real data example to limit the number of false positives and commented that states or test sponsors would apply a conservative criterion in practice. Another option to limit the number of false positives is to choose a critical value that adjusts for multiple comparisons by controlling the family-wise error rate using a Bonferroni correction, or, controlling the false discovery rate using the procedure of Benjamini and Hochberg (1995).

The suggested approaches were derived for the 2PLM and the GPCM. They can also be applied to the dichotomous or polytomous Rasch models that are special cases of the 2PLM and GPCM, respectively. Unfortunately, the likelihood distribution of the ability for the three-parameter logistic model (3PLM) does not belong to the exponential family of distributions (e.g., Biehler et al., 2015)—so the methods in this paper do not apply to the 3PLM. Researchers such as Skovgaard (1990) suggested approaches to apply the MSLRT and LRA to distributions that do not belong to the exponential family of distributions, but the application of those approaches does not seem to be straightforward to the 3PLM. While application of the MSLRT and the LRA to the 3PLM remains a topic for further research, the suggested approaches (that can be used in applications of Rasch models, 2PLM, and GPCM) should be useful given that (i) the Rasch models, 2PLM, and GPCM are widely

used, (ii) Haberman (2006) found the 3PLM to not provide much gain for real data over the 2PLM,[10] (iii) researchers such as Maris and Bechger (2009) and Martín, González, and Tuerlinckx (2015) have recently unearthed some problems with the identifiability of the 3PLM.

This paper has several additional limitations and hence leaves scope for further related research. First, more simulations to more cases of testing of the equality of abilities can be performed to further explore the performance of the MSLRT and the LRA. Similarly, more real data applications might provide deeper understanding of the approaches. Second, only a one-sided alternative hypothesis was considered—it is possible to explore two-sided alternatives; the MSLRT and LRA would then involve two one-sided tests while the traditional approaches would be the LRT and the square of the Wald or score statistic. Some limited examination shows that the suggested methods perform slightly better compared to existing tests when two-sided alternatives are of interest. Third, hypothesis tests involving a unidimensional ability parameter was considered—tests involving multidimensional ability parameters would be an obvious next step. Fourth, because the approaches considered in this paper are based on item-scores of one examinee at a time, their power is expected to be low, as seen from the simulation studies; in contrast, methods such as matching analysis (Haberman & Lee, 2017) or aggregate-level erasure analysis (Wollack & Eckerly, 2017) would have larger power because those methods are based on the whole sample or a group of examinees; still, the approaches suggested in this paper promise to be helpful because hypothesis tests based on item-scores of one examinee at a time are routinely performed by researchers and testing organizations.

References

Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, *73*, 307–322.

[10]The method used by Haberman (2006) was applied to our first real data example—the 3PLM did not provide a substantial gain over the 2PLM for that data set either.

Barndorff-Nielsen, O. E. (1991). Modified signed log likelihood ratio. *Biometrika*, *78*, 557–563.

Barndorff-Nielsen, O. E., & Cox, D. R. (1994). *Inference and asymptotics*. London, UK: Springer Nature.

Bedrick, E. J. (1997). Approximating the conditional distribution of person fit indexes for checking the Rasch model. *Psychometrika*, *62*, 191–199.

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.

Biehler, M., Holling, H., & Doebler, P. (2015). Saddlepoint approximations of the distribution of the person parameter in the two parameter logistic model. *Psychometrika*, *80*, 665–688.

Brazzale, A. R., Davison, A. C., & Reid, N. (2007). *Applied asymptotics*. Oxford, UK: Cambridge University Press.

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

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO personality inventory. *Psychological Assessment*, *4*, 5-13.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London, UK: Chapman and Hall.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*, 295–311.

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

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

Finkelman, M., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis

testing for the adaptive measurement of change. *Applied Psychological Measurement*, *34*, 238–254.

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

Ghosh, J. K. (1994). *Higher order asymptotics*. Hayward, CA: Institute of Mathematical Statistics.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*, 159–180.

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored internet tests: The z-test and the likelihood ratio test. *International Journal of Selection and Assessment*, *18*, 351–364.

Haberman, S. J. (2006). *An elementary test of the normal 2PL model against the normal 3PL alternative* (ETS Research Report No. RR-06-14). Princeton, NJ: ETS.

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.

Jensen, J. L. (1992). The modified signed likelihood statistic and saddlepoint approximations. *Biometrika*, *79*, 693–703.

Jensen, J. L. (1995). *Saddlepoint approximations*. Oxford, UK: Clarendon Press.

Jensen, J. L. (1997). A simple derivation of r^* for curved exponential families. *Scandinavian Journal of Statistics*, *24*, 33–46.

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika*, *56*, 213-228.

Klauer, K. C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology*, *43*, 193–206.

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

Lugannani, R., & Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*, *12*, 475–490.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement: Interdisciplinary Research & Perspective*, *7*(2), 75–88.

Martín, E. S., González, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *80*, 450–467.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Pierce, D. A., & Peters, D. (1992). Practical use of higher-order asymptotics for multiparameter exponential families. *Journal of Royal Statistical Society, Series B*, *54*, 701-738.

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley.

Reid, N. (2003). Asymptotics and the theory of inference. *The Annals of Statistics*, *31*, 1695–1731.

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

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

Skovgaard, I. M. (1990). On the density of minimum contrast estimators. *The Annals of Statistics*, *18*, 779–789.

von Davier, M., & Molenaar, I. W. (2003). A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika*, *68*, 213–228.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory.

Psychometrika, *54*, 427–450.

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 quantitative methods for detecting cheating on tests* (pp. 214–231). 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.

Appendix: The MSLRT and the LRA for the GPCM

Donoghue (1994) provided the result that for log likelihood given by Equation 21,

$$\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = \sum_{i=1}^{n} a_i^2 \left(\left[\sum_{k=0}^{m_i} k P_{ik}(\theta) \right]^2 - \sum_{k=0}^{m_i} k^2 P_{ik}(\theta) \right). \tag{25}$$

Noting that

$$\mathrm{Var}(X_i|\theta) = \sum_{k=0}^{m_i} k^2 P_{ik}(\theta) - \left[\sum_{k=0}^{m_i} k P_{ik}(\theta) \right]^2,$$

Equation 25 can be rewritten as

$$\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = -\sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i|\theta). \tag{26}$$

For two sets of items, using notations introduced below Equation 22, the joint log likelihood of θ_1 and θ_2 is given by

$$\begin{aligned}
& \ell(\theta_1, \theta_2) \\
= & \sum_{i=1}^{n_1} \sum_{k=0}^{m_i} d_k(X_i) \log P_{ik}(\theta_1) + \sum_{j=1}^{n_2} \sum_{k=0}^{m_j} d_k(Y_j) \log \tilde{P}_{jk}(\theta_2) \qquad (27) \\
= & \sum_{i=1}^{n_1} \sum_{k=0}^{m_i} d_k(X_i) \left\{ \sum_{h=0}^{k} a_i(\theta_1 - b_{ih}) - \log(\Gamma_i(\theta_1)) \right\} \\
& + \sum_{j=1}^{n_2} \sum_{k=0}^{m_j} d_k(Y_j) \left\{ \sum_{h=0}^{k} \tilde{a}_j(\theta_2 - \tilde{b}_{jh}) - \log(\tilde{\Gamma}_j(\theta_2)) \right\} \\
= & \theta_1 \sum_{i=1}^{n_1} a_i \sum_{k=0}^{m_i} (k+1) d_k(X_i) - \sum_{i=1}^{n_1} a_i \sum_{k=0}^{m_i} d_k(X_i) \sum_{h=0}^{k} b_{ih} - \sum_{i=1}^{n_1} \log(\Gamma_i(\theta_1)) \\
& + \theta_2 \sum_{j=1}^{n_2} \tilde{a}_j \sum_{k=0}^{m_j} (k+1) d_k(Y_j) - \sum_{j=1}^{n_2} \tilde{a}_j \sum_{k=0}^{m_j} d_k(Y_j) \sum_{h=0}^{k} \tilde{b}_{jh} - \sum_{j=1}^{n_2} \log(\tilde{\Gamma}_j(\theta_2))
\end{aligned}$$

The last equality holds because $\sum_{k=0}^{m_i} d_k(X_i) = \sum_{k=0}^{m_j} d_k(Y_j) = 1$ under the assumption that no data are missing, which means, for example, that $\sum_{k=0}^{m_i} d_k(X_i) \log(\Gamma_i(\theta_1)) = \log(\Gamma_i(\theta_1))$. Let us apply the transformations $\psi = \theta_2 - \theta_1$ and $\lambda = \theta_1$, which means that $\theta_1 = \lambda$ and $\theta_2 = \psi + \lambda$. Let us also denote

$$S_1 = \sum_{i=1}^{n_1} a_i \sum_{k=0}^{m_i} (k+1) d_k(X_i), \text{ and } S_2 = \sum_{j=1}^{n_2} \tilde{a}_j \sum_{k=0}^{m_j} (k+1) d_k(Y_j).$$

Note that both S_1 and S_2 are functions of the data (X_i's and Y_j's) and not of the parameters (θ_1 and θ_2). The above log-likelihood, $\ell(\theta_1, \theta_2)$, or, $\ell(\psi, \lambda)$, then is given by

$$\begin{aligned}
\ell(\psi,\lambda) &= \sum_{i=1}^{n_1}\sum_{k=0}^{m_i} d_k(X_i)\log P_{ik}(\lambda) + \sum_{j=1}^{n_2}\sum_{k=0}^{m_j} d_k(Y_j)\log \tilde{P}_{jk}(\psi+\lambda) \qquad (28)\\
&= \lambda S_1 - \sum_{i=1}^{n_1} a_i \sum_{k=0}^{m_i} d_k(X_i)\sum_{h=0}^{k} b_{ih} - \sum_{i=1}^{n_1}\log(\Gamma_i(\lambda))\\
&\quad +(\psi+\lambda)S_2 - \sum_{j=1}^{n_2}\tilde{a}_j\sum_{k=0}^{m_j} d_k(Y_j)\sum_{h=0}^{k}\tilde{b}_{jh} - \sum_{j=1}^{n_2}\log(\tilde{\Gamma}_j(\psi+\lambda))\\
&= S_2\psi + (S_1+S_2)\lambda - \sum_{i=1}^{n_1}\log(\Gamma_i(\lambda)) - \sum_{j=1}^{n_2}\log(\tilde{\Gamma}_j(\psi+\lambda))\\
&\quad -\sum_{i=1}^{n_1} a_i\sum_{k=0}^{m_i} d_k(X_i)\sum_{h=0}^{k} b_{ih} - \sum_{j=1}^{n_2}\tilde{a}_j\sum_{k=0}^{m_j} d_k(Y_j)\sum_{h=0}^{k}\tilde{b}_{jh}. \qquad (29)
\end{aligned}$$

The above log-likelihood belongs to the exponential family of distributions with canonical parameters ψ and λ and joint sufficient statistics S_1 and $(S_1 + S_2)$.

Then, given the discussion on the applicability of the MSLRT and LRA to the exponential family of distributions, the MSLRT and LRA can be applied to test $H_0 : \psi = 0$, or, $H_0 : \theta_1 = \theta_2$, in applications of the GPCM. The SLR statistic is given by

$$\begin{aligned}
r(\psi_0) &= \text{sign}(\hat{\psi}-\psi_0)\left[2[\ell(\hat{\psi},\hat{\lambda}) - \ell(\psi_0,\hat{\lambda}_{\psi_0})]\right]^{1/2}\\
&= \text{sign}(\hat{\theta}_2-\hat{\theta}_1)\sqrt{2}\left[\sum_{i=1}^{n_1}\sum_{k=0}^{m_i} d_k(X_i)\log P_{ik}(\hat{\theta}_1) + \sum_{j=1}^{n_2}\sum_{k=0}^{m_j} d_k(Y_j)\log\tilde{P}_{jk}(\hat{\theta}_2)\right.\\
&\quad \left. - \sum_{i=1}^{n_1}\sum_{k=0}^{m_i} d_k(X_i)\log P_{ik}(\hat{\theta}_0) - \sum_{j=1}^{n_2}\sum_{k=0}^{m_j} d_k(Y_j)\log\tilde{P}_{jk}(\hat{\theta}_0)\right]^{1/2}. \qquad (30)
\end{aligned}$$

Then, using the result provided in Equation 26 (or, by differentiating the joint log likelihood provided in Equation 29 twice),

$$\begin{aligned}
j_{\psi\psi}(\psi,\lambda) &= \sum_{j=1}^{n_2}\tilde{a}_j^2\,\text{Var}(Y_j|\psi+\lambda),\\
j_{\lambda\lambda}(\psi,\lambda) &= \sum_{i=1}^{n_1} a_i^2\,\text{Var}(X_i|\lambda) + \sum_{j=1}^{n_2}\tilde{a}_j^2\,\text{Var}(Y_j|\psi+\lambda), \qquad (31)\\
j_{\psi\lambda}(\psi,\lambda) &= j_{\lambda\psi}(\psi,\lambda) = j_{\psi\psi}(\psi,\lambda).
\end{aligned}$$

Then,

$$\begin{aligned} |j(\psi,\lambda)| &= j_{\psi\psi}(\psi,\lambda)j_{\lambda\lambda}(\psi,\lambda) - [j_{\psi\psi}(\psi,\lambda)]^2 = j_{\psi\psi}(\psi,\lambda)[j_{\lambda\lambda}(\psi,\lambda) - j_{\psi\psi}(\psi,\lambda] \\ &= \sum_{j=1}^{n_2} \tilde{a}_j^2 \operatorname{Var}(Y_j|\psi+\lambda) \sum_{i=1}^{n_1} a_i^2 \operatorname{Var}(X_i|\lambda), \end{aligned}$$

which implies that

$$|j(\hat{\psi},\hat{\lambda})| = \sum_{i=1}^{n_1} a_i^2 \operatorname{Var}(X_i|\hat{\theta}_1) \sum_{j=1}^{n_2} \tilde{a}_j^2 \operatorname{Var}(Y_j|\hat{\theta}_2)\cdot \tag{32}$$

One can obtain $j_{\lambda\lambda}(\psi_0, \hat{\lambda}_{\psi_0})$ from Equation 31 as

$$j_{\lambda\lambda}(\psi_0,\hat{\lambda}_{\psi_0}) = \sum_{i=1}^{n_1} a_i^2 \operatorname{Var}(X_i|\hat{\theta}_0) + \sum_{j=1}^{n_2} \tilde{a}_j^2 \operatorname{Var}(Y_j|\hat{\theta}_0) \tag{33}$$

Then, $q(\psi_0)$ is given by

$$\begin{aligned} q(\psi_0) &= (\hat{\theta}_2 - \hat{\theta}_1)\sqrt{\frac{|j(\hat{\psi},\hat{\lambda})|}{j_{\lambda\lambda}(\psi_0,\hat{\lambda}_{\psi_0})}} \\ &= (\hat{\theta}_2 - \hat{\theta}_1)\sqrt{\frac{\sum_{i=1}^{n_1} a_i^2 \operatorname{Var}(X_i|\hat{\theta}_1) \sum_{j=1}^{n_2} \tilde{a}_j^2 \operatorname{Var}(Y_j|\hat{\theta}_2)}{\sum_{i=1}^{n_1} a_i^2 \operatorname{Var}(X_i|\hat{\theta}_0) + \sum_{j=1}^{n_2} \tilde{a}_j^2 \operatorname{Var}(Y_j|\hat{\theta}_0)}}. \end{aligned} \tag{34}$$

Once $q(\psi_0)$ is computed using Equation 34 and $r(\psi_0)$ is computed using Equation 30, one can compute the MSLR statistic $r^*(\psi_0)$ using Equation 12, and can compute the LRA using Equation 14.