Multivariate Behavioral Research, 49:354–371, 2014

ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2014.910744

Routledge
Taylor & Francis Group

Identifying the Source of Misfit in Item Response Theory Models

Yang Liu
The University of North Carolina at Chapel Hill

Alberto Maydeu-Olivares
Faculty of Psychology, University of Barcelona

When an item response theory model fails to fit adequately, the items for which the model provides a good fit and those for which it does not must be determined. To this end, we compare the performance of several fit statistics for item pairs with known asymptotic distributions under maximum likelihood estimation of the item parameters: (a) a mean and variance adjustment to bivariate Pearson's X^2, (b) a bivariate subtable analog to Reiser's (1996) overall goodness-of-fit test, (c) a z statistic for the bivariate residual cross product, and (d) Maydeu-Olivares and Joe's (2006) M_2 statistic applied to bivariate subtables. The unadjusted Pearson's X^2 with heuristically determined degrees of freedom is also included in the comparison. For binary and ordinal data, our simulation results suggest that the z statistic has the best Type I error and power behavior among all the statistics under investigation when the observed information matrix is used in its computation. However, if one has to use the cross-product information, the mean and variance adjusted X^2 is recommended. We illustrate the use of pairwise fit statistics in 2 real-data examples and discuss possible extensions of the current research in various directions.

Item response theory (IRT) modeling involves fitting a latent variable model to discrete responses obtained from questionnaire/test items designed to measure personality, attitudes, patient-reported health outcomes, and educational achievement, among other things. Before any inferences can be drawn from the fitted model, the model's fit must be assessed, given that any conclusions derived from poorly fitting models can be potentially misleading. To this end, a number of procedures can be reliably used to assess the overall goodness of fit (GOF) of IRT models (for reviews, see Mavridis, Moustaki, & Knott, 2007; Maydeu-Olivares & Joe, 2008). When a model does not fit well, alternative IRT models might be fitted. However, more often than not no such model provides a good fit; this is to be expected, given that IRT modeling involves many degrees of freedom. Facing this situation, researchers often resort to selecting the best fitting model and then seek to improve its fit by using item-level fine-tuning. Hence, in a context of item calibration and selection, researchers have to differentiate well-fitting items from poorly fitting ones; and on the basis of these outcomes they may decide to retain only the former set or to apply an alternative IRT model to the latter set. Note that such piecewise assessment should be performed even when the overall fit is acceptable, as there may be parts of the model whose fit can be improved. Once the model has been modified, or once items have been removed, the overall fit of the resulting model needs to be reassessed.

A substantial body of literature has been published on the identification of misfits in IRT modeling (e.g., Andersen, 1973; Bock, 1972; Cagnone & Mignani, 2007; Chen & Thissen, 1997; Drasgow, Levine, Tsien, Williams, & Mead, 1995; Glas, 1988; Glas & Verhest, 1989, 1995; Glas & Suarez-Falcón, 2003; Kelderman, 1984; McKinley & Mills, 1985; Orlando & Thissen, 2000, 2003; Stone & Zhang, 2003; Tay & Drasgow, 2011; Toribio & Albert, 2011; van den Wollenberg, 1982; Yen, 1984). However, some of the statistics proposed in these studies have unknown sampling distributions and their use relies on heuristics; another group appears to be valid only for certain models, whereas others appear to serve solely to detect certain types of misfit. Maydeu-Olivares

Correspondence concerning this article should be addressed to Yang Liu, Department of Psychology, The University of North Carolina at Chapel Hill, 352 Davie Hall, Chapel Hill, NC 27599. E-mail: liuy0811@live.unc.edu

and Liu (2012) recently identified the main challenges faced when attempting to identify poorly fitting items: first, the sampling distribution of the test statistic needs to be well approximated by a known reference distribution when the model is correctly specified; second, some tests of interest cannot be applied owing to the lack of degrees of freedom; and third, some statistics may lack power to detect alternative models of interest.

Reference distributions are usually obtained using asymptotic methods. If the reference distribution does not closely match the sampling distribution of the statistic for correctly specified models, then the researcher may over- or underreject well-fitting items. Overrejection is particularly undesirable as good items are generally expensive to develop; likewise, underrejection should be avoided because it prevents the researcher from determining the power of the statistic under alternatives of interest.

Quadratic form statistics, such as Pearson's X^2 applied to item pairs, are often used for piecewise goodness-of-fit assessment. For chi-square distributed statistics, degrees of freedom often equal the number of parameters in the saturated model minus the number of parameters in the restricted model. For a binary variable, the saturated model involves two probabilities that must add up to one; as such, there is no degree of freedom available for testing at the item level. For polytomous items, testing at the item level using chi-square distributed statistics is only possible for IRT models with fewer item parameters than the number of response alternatives minus one. This means that testing the source of misfit may require pairs of variables if a chi-square distributed statistic is used and even triplets or quads if the items are binary (Maydeu-Olivares & Liu, 2012).

One way to overcome the problem of the lack of degrees of freedom is to use a large sample z statistic (i.e., an asymptotically normal statistic divided by its standard error). Another way is to use a statistic that draws information from the sum score, for instance, Orlando and Thissen's (2000, 2003) heuristic statistic $S\text{-}X^2$. Reiser (1996) and Maydeu-Olivares and Joe (2005) suggested using bivariate z statistics to assess the source of misfit in two-way marginal subtables for binary item response data. Recently, Maydeu-Olivares and Liu (2012) proposed an extension of Reiser's z statistic suitable for polytomous ordinal items. However, the computation of z statistics involves obtaining an estimate of the asymptotic covariance matrix of all item parameters, which is challenging when the maximum likelihood estimator is used (in this case it amounts to the inverse of the Fisher information matrix). The three most widely used approaches to estimate the Fisher information are usually referred to as the expected, the observed, and the cross-product (XPD) information matrices. When expected information is used, the distribution of bivariate z statistics can be well approximated even in small samples (Maydeu-Olivares & Liu, 2012); however, when XPD information is used, the approximation is presumed to be much poorer. Note that Liu and Maydeu-Olivares (2013) observed this trend for a score test statistic. Unfortunately, for computational reasons the expected information matrix can only be calculated when the number of binary items is under 20. If there are five or more response alternatives, then this matrix can only be computed with 6 or so items. Hence, either the observed or the XPD information matrix has to be used in most real applications.

For a work-around to both the lack of degrees of freedom and the computation of the information matrix, Liu and Maydeu-Olivares (2013) proposed a statistic, $R_{2,ij}$, that involves a pair of item and conditions on sum score levels/groups, as inspired by Orlando and Thissen's (2000, 2003) $S\text{-}X^2$ statistic and Glas's (1988) R_2 statistic. Drawing on the results of Joe and Maydeu-Olivares (2010), Liu and Maydeu-Olivares (2013) were able to derive the asymptotic distribution of $R_{2,ij}$. Alternatively, Maydeu-Olivares and Joe's (2006) M_2 test statistic applied to item pairs can be used for piecewise goodness-of-fit assessment. Under the null hypothesis of a correctly specified model, this statistic follows asymptotically a chi-square distribution. However, Liu and Maydeu-Olivares (2013) and Maydeu-Olivares and Liu (2012) found that statistics for pairs and triplets of items that do not require the computation of an information matrix (e.g., M_2,$R_{2,ij}$) tend to have low power for detecting multidimensionality of the latent trait, even though they may have excellent Type I errors even in small samples.

In summary, to assess the source of misfit in IRT models a number of statistics with known asymptotic distribution are now available. However, if their computation does not involve the information matrix, they appear to lack power to detect certain alternatives of interest in applications. On the other hand, if their computation does involve estimating the information matrix, most research has been undertaken using the expected information matrix, which can only be computed for small models. Extant research suggests that when XPD information is used instead of the expected information, larger samples are needed for the asymptotic distribution to ensure a good approximation of the sampling distribution of the statistic. In most applications, however, expected information matrix cannot be computed because the model is large. Hence, it is of interest to investigate statistics whose sampling distribution is well approximated by the asymptotic theory when either the observed or XPD information matrix is used.

We propose two new quadratic form statistics. The first statistic is a mean and variance correction to Pearson's X^2 so that it can be approximated asymptotically by a chi-square distribution. The second statistic differs from X^2 in that the weight matrix in the quadratic form is chosen to be an estimate of the Moore-Penrose pseudoinverse of the asymptotic covariance matrix of the bivariate residuals. This is similar in spirit to Reiser's (1996) overall GOF statistic; with this choice of weight matrix, the resulting statistic is asymptotically chi-square. Apart from the two new statistics, we also include the bivariate residual z statistic for binary data

(Reiser, 1996) and a natural extension of it suitable for ordinal data (Maydeu-Olivares & Liu, 2012). We only consider multinomial maximum likelihood (ML) estimation of the IRT parameters, which is often referred to as marginal maximum likelihood estimation (e.g., Bock & Aitkin, 1981) in the IRT literature and also as full information maximum likelihood in the item factor analysis literature (e.g., Jöreskog & Moustaki, 2001). Whenever the asymptotic covariance matrix of the item parameter estimates is involved in the computation of the goodness-of-fit statistic, it is estimated by the inverse of the information matrix. Both the observed and XPD information matrix are investigated in this study.

For succinctness, we focus our attention on detecting sources of misfit using pairs of items. However, the methodology presented in this article can be generalized to goodness-of-fit assessment in marginal subtables of arbitrary orders (single items, triplets of items, etc.). By using pairs of items, we may detect whether there is misfit in the associations between items, and they are considerably easier to interpret than item triplets. Interested readers are referred to Maydeu-Olivares and Liu (2012) for testing using single items and to Liu and Maydeu-Olivares (2012) for testing using triplets of items.

The rest of this article is organized as follows: In outlining the motivation and focus of our presentation, we first describe two applications. Next, we describe the statistics under investigation and their asymptotic reference distribution. We then report a simulation study designed to determine whether the sampling distribution of the statistics is well approximated by their reference distribution when the fitted model is correctly specified. For statistics with an adequate empirical Type I error rate, the eventual selection depends on their power to reject alternatives of interest, which is examined by conducting additional simulations. Both binary and polytomous rating data are considered. In polytomous conditions, we also compute as benchmarks the bivariate version of Maydeu-Olivares and Joe's (2006) M_2 and Pearson's X^2 using heuristically the same degrees of freedom as M_2. Because a bivariate X^2 computed from the full table ML parameter estimates is not asymptotically chi-square (Maydeu-Olivares & Joe, 2006), the statistics investigated here should have better Type I error rates than those of unadjusted X^2 and hopefully as good as those of M_2. We also hope they will be more powerful than M_2. In the binary case, M_2 cannot be computed for pairs of items due to lacking degrees of freedom. On the other hand, X^2 is still computable; Chen and Thissen (1997) suggested for it a heuristic chi-square reference distribution with the same degrees of freedom as the independence model (i.e., $df = 1$). It has been found that Chen and Thissen's proposal tends to underreject (see, e.g., Liu & Maydeu-Olivares, 2013); therefore, we expect the statistics proposed here to perform better.

The article concludes with a discussion of the findings and some recommendations for applied researchers. These recommendations are illustrated by using the statistics described in the article in the two applications that we now introduce.

TWO APPLICATIONS

PROMIS Depression Short Form

Pilkonis et al. (2011) described the emotional distress item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS) in detail. Respondents were asked to report the frequency with which they had experienced certain feelings in the past 7 days using a $K = 5$-point rating scale ranging from *never* to *always;* the responses were coded from 0 to 4 for the analyses. The short form of the depression scale (consisting of $n = 8$ items) is analyzed here so that we can provide detailed results for all items. We used the $N = 768$ complete responses in the data kindly provided by the authors, who fitted a logistic graded response model (Samejima, 1969) with a single normally distributed latent trait.

There are $K^n = 5^8 = 390{,}625$ possible response patterns in this example. Under the fitted model, the probability of observing a response pattern is given by[1]

$$\Pr(Y_1 = k_1, \cdots, Y_n = k_n) = \int_{-\infty}^{\infty} \prod_{i=1}^{n} \Pr(Y_i = k_i \,|\eta)\, f(h)\, d\eta, \tag{1}$$

where Y_i denotes the random variable associated to the responses to item *i;* η denotes the latent trait; and $\phi(\eta)$ denotes the latent trait's density, which is assumed to be standard normal. Also, under this model,

$$\Pr(Y_i = k \,|\eta) = \begin{cases} 1 - \Psi\left(\eta\,; \alpha_{i,1}, \beta_i\right) & \text{if} \quad k = 0, \\ \Psi\left(\eta\,; \alpha_{i,k}, \beta_i\right) - \Psi\left(\eta\,; \alpha_{i,k+1}, \beta_i\right) & \text{if} \quad 0 < k < K-1, \\ \Psi\left(\eta\,; \alpha_{i,K-1}, \beta_i\right) & \text{if} \quad k = K-1, \end{cases} \tag{2}$$

where

$$\Psi\left(\eta\,; \alpha_{i,k}, \beta_i\right) = \frac{1}{1 + \exp\left[-\left(\alpha_{i,k} + \beta_i \eta\right)\right]} \tag{3}$$

denotes a standard logistic distribution function evaluated at $\alpha_{i,k} + \beta_i \eta$. Applying this model to the depression data, we have five parameters per item (four intercepts $\alpha_{i,k}$ and one slope β_i) and $q = 5 \times 8 = 40$ parameters in total. We estimated this model by ML using Mplus 7.0 (Muthén & Muthén, 2012); 48 rectangular quadrature points from –5 to 5 were used to approximate the intractable integral in Equation (1).

[1]The notation adopted in this article is different from the standard IRT convention. As noted by a reviewer, it is more consistent with that used in the statistical literature.

The overall GOF of the logistic graded model to these data cannot be assessed using Pearson's X^2 or the likelihood ratio G^2 statistic, as the data are very sparse. In fact, X^2 yields a p-value of zero and G^2 a p-value of one. A more accurate p-value for the overall fit can be obtained using Maydeu-Olivares and Joe's (2005, 2006) M_2 statistic. We obtain $M_2 = 805.44$ on 440 degrees of freedom; $p < .01$; and we conclude that, taking into account sampling variability, the fitted model is not likely to be the data-generating model. We computed an RMSEA statistic using M_2 (denoted RMSEA_2) to assess whether the fitted model provides a close approximation to the true and unknown data-generating model. A 90% confidence interval for the RMSEA_2 yields [0.03, 0.04]. Maydeu-Olivares and Joe (2014) suggest that IRT models with an RMSEA_2 less than or equal to 0.05 provide a close approximation to the data-generating model and that those with an RMSEA_2 less than or equal to $0.05 / (K - 1)$ provide an excellent approximation. Because $K = 5$, their criterion for an excellent approximation is $\text{RMSEA}_2 \leq 0.0125$. We conclude that the fitted model provides a close approximation to the data-generating model but falls short of their criterion for an excellent approximation. Thus, there is room for modifying the model and for improving its fit to these data. To achieve this, we first need to locate the model misfits using a piecewise fit assessment.

EPQ-R Extraversion Scale Short Form

Binary item response data differ from polytomous data with regard to assessing the source of misfit. When the GOF for each pair of items is of interest, there are no degrees of freedom available for some of the statistics considered in this article. Hence, this special case needs to be considered in some detail.

When data are binary, Samejima's (1969) graded model (2) reduces to a two-parameter logistic (2PL) model. Maydeu-Olivares and Liu (2012) fitted a 2PL model using the ML estimator to data provided by the 824 respondents in the female United Kingdom normative sample to the short form of the extraversion scale of Eysenck's Personality Questionnaire-Revised (EPQ-R; Eysenck, Eysenck, & Barrett, 1985). These data are reanalyzed here. The scale consists of 12 binary items: a typical item is "Are you a talkative person?" The response categories are "Yes" and "No," coded as 1 and 0, respectively. In this case, there are $2^{12} = 4{,}096$ possible response patterns. These data are also sparse, so Pearson's X^2 and the likelihood ratio G^2 statistic should not be used. The estimated M_2 statistic is 474.23 on 54 degrees of freedom, which indicates that a one-dimensional 2PL model fits rather poorly. A 90% confidence interval for RMSEA_2 yields [0.09, 0.11] and we conclude that the fitted model is not close to the true data-generating model. What statistics should be used to locate misfits in this case? In the following section, we discuss some proposals.

GOODNESS-OF-FIT STATISTICS FOR BIVARIATE MARGINAL SUBTABLES

Denote the set of q IRT item parameters by $\boldsymbol{\theta}$. For the graded response model considered in this article, $\boldsymbol{\theta}$ consists of slopes and intercepts that are related to item discrimination and difficulty, respectively. Also, let $\boldsymbol{\pi}$ be the $C = K^n$ dimensional vector of response pattern probabilities. We write $\boldsymbol{\pi}(\boldsymbol{\theta})$ to denote the multinomial probabilities expressed as a function of the model parameters. For any pattern, $\boldsymbol{\pi}(\boldsymbol{\theta})$ is given by Equations (1) and (2) for the graded response model. Then, the null hypothesis of overall GOF is $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ versus $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$. After performing an overall GOF test, we wish to examine GOF in a piecewise fashion. More specifically, we seek to assess how well the model reproduces each pair of items, which is very similar in spirit to examining z statistics for residual covariances in structural equation modeling (SEM).

Quadratic Form Statistics: M_{ij}, R_{ij}, and X^2_{ij}

To determine whether a particular pair of items shows model misfit, one natural statistic to use is the quadratic form in bivariate residuals

$$Q_{ij} = N\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right)' \hat{\mathbf{W}}_{ij}\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right). \tag{4}$$

When both items have K response categories, $\mathbf{p}_{ij}$ and $\hat{\boldsymbol{\pi}}_{ij}$ are $C_{ij} = K^2$ dimensional vectors of observed and expected bivariate proportions, respectively. These bivariate probabilities only involve a q_{ij}-dimensional subset of all parameters, denoted as $\boldsymbol{\theta}_{ij,}$ and for simplicity we write $\hat{\boldsymbol{\pi}}_{ij} = \boldsymbol{\pi}_{ij}\left(\hat{\boldsymbol{\theta}}_{ij}\right)$ in which $\hat{\boldsymbol{\theta}}_{ij}$is the ML estimate. For the graded model considered here, $\boldsymbol{\theta}_{ij}$ amounts to a set of two slopes and $2 \times (K - 1)$ intercepts. Finally, $\hat{\mathbf{W}}_{ij}$ is some $C_{ij} \times C_{ij}$ real symmetric matrix that may depend on parameter estimates but converges in probability to some constant matrix:$\hat{\mathbf{W}}_{ij} \xrightarrow{p} \mathbf{W}_{ij}$.

When ML is used to estimate the IRT model parameters, the residuals for a pair of items $\sqrt{N}\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right)$ are asymptotically normally distributed with mean zero and covariance matrix (Maydeu-Olivares & Liu, 2012)

$$\boldsymbol{\Sigma}_{ij} = \mathbf{D}_{ij} - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}'_{ij} - \boldsymbol{\Delta}_{ij}\mathcal{I}^{-1}_{(ij)}\boldsymbol{\Delta}'_{ij}. \tag{5}$$

In Equation (5), $\mathbf{D}_{ij} = \text{diag}(\pi_{ij})$ is a diagonal matrix of the bivariate probabilities, $\boldsymbol{\Delta}_{ij} = \partial\boldsymbol{\pi}_{ij}\left(\boldsymbol{\theta}_{ij}\right)/\partial\boldsymbol{\theta}'_{ij}$ denotes the $C_{ij} \times q_{ij}$ matrix of derivatives of the bivariate probabilities with respect to the parameters involved in the bivariate subtable, $\mathcal{I}$ denotes the Fisher information matrix of all item parameters, and $\mathcal{I}^{-1}_{(ij)}$ denotes the $q_{ij} \times q_{ij}$ submatrix of $\mathcal{I}^{-1}$ obtained by selecting the q_{ij} rows and columns corresponding to $\boldsymbol{\theta}_{ij}$ (Maydeu-Olivares & Liu, 2012).

When the model is correctly specified, Q_{ij} is asymptotically distributed as a mixture of d independent c^2_1 random variables where d is the rank of $\mathbf{W}_{ij}\boldsymbol{\Sigma}_{ij}$ by the general theory of quadratic form statistics in normal random variables (e.g.,

Box, 1954). In particular, if

$$\boldsymbol{\Sigma}_{ij}\mathbf{W}_{ij}\boldsymbol{\Sigma}_{ij}\mathbf{W}_{ij}\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}_{ij}\mathbf{W}_{ij}\boldsymbol{\Sigma}_{ij} \tag{6}$$

is satisfied, then Equation (4) is asymptotically distributed as a chi-square with d degrees of freedom (e.g., Schott, 1997, Theorem 9.10).

Equation (6) holds when $\boldsymbol{\Sigma}_{ij}$ is a generalized inverse of $\mathbf{W}_{ij}$; that is, $\mathbf{W}_{ij}$ satisfies $\boldsymbol{\Sigma}_{ij}\mathbf{W}_{ij}\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}_{ij}$. This is the approach taken by Maydeu-Olivares and Joe (2006), who proposed using the statistic

$$\begin{aligned} M_{ij} &= N\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right)' \hat{\mathbf{C}}_{ij}\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right), \\ \hat{\mathbf{C}}_{ij} &= \hat{\mathbf{D}}_{ij}^{-1} - \hat{\mathbf{D}}_{ij}^{-1}\hat{\boldsymbol{\Delta}}_{ij}\left(\hat{\boldsymbol{\Delta}}_{ij}'\hat{\mathbf{D}}_{ij}^{-1}\hat{\boldsymbol{\Delta}}_{ij}\right)^{-1}\hat{\boldsymbol{\Delta}}_{ij}'\hat{\mathbf{D}}_{ij}^{-1} \end{aligned} \tag{7}$$

to assess the source of misfit in bivariate tables. In Equation (7), all vectors and matrices with a hat are evaluated at the parameter estimates $\hat{\boldsymbol{\theta}}_{ij}$. Provided $\boldsymbol{\Delta}_{ij}$ is of full rank (i.e., $\boldsymbol{\theta}_{ij}$ is estimable only from $\mathbf{p}_{ij}$), M_{ij} is asymptotically distributed as a chi-square distribution with degrees of freedom equal to the number of parameters in the saturated model of the bivariate subtable $C_{ij} - 1$ minus the number of item parameters involved in the same subtable q_{ij}; thus, $df_{ij} = C_{ij} - q_{ij} - 1$. $M_{ij}$is simply the M_2 statistic applied to the bivariate subtable for items i and j; however, Equation (7) differs from the formula of M_2 applied as an overall GOF test. See Maydeu-Olivares and Joe (2006) and Maydeu-Olivares and Liu (2012) for further details.

Another way to satisfy Equation (6) is to use, as the weight matrix, the Moore-Penrose pseudoinverse of Equation (5) evaluated at the parameter estimates, $\hat{\boldsymbol{\Sigma}}_{ij}^{+}$, leading to

$$R_{ij} = N\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right)' \hat{\boldsymbol{\Sigma}}_{ij}^{+}\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right). \tag{8}$$

This statistic is a bivariate subtable counterpart of the overall GOF statistic proposed by Reiser (1996) for binary items. The reference degrees of freedom of R_{ij} is given by the rank of $\boldsymbol{\Sigma}_{ij}^{+}\boldsymbol{\Sigma}_{ij}$, which further equals the rank of $\boldsymbol{\Sigma}_{ij}$. However, as Reiser has noted in the case of overall goodness-of-fit tests, the rank of $\boldsymbol{\Sigma}_{ij}$ may depend on the true parameter values. As a result, in applications, its degrees of freedom must be estimated by determining the rank of $\hat{\boldsymbol{\Sigma}}_{ij}$, for example, using an eigendecomposition. Hence, the value of R_{ij} and its p-value will depend on how many eigenvalues are numerically judged to be zero. This is by no means straightforward in IRT applications because numerical integration is involved, and thus the computation of small eigenvalues is vulnerable to numerical errors (for an illustration of this point, see Maydeu-Olivares & Joe, 2008). Nevertheless, the simulation results of Mavridis et al. (2007) using the overall GOF counterpart of Equation (8) suggest that it is safe to use this statistic in practice. It should be noted that the expected information matrix was used in their study; therefore, the results might not be generalizable to the case where the XPD or the observed information must be used (e.g., for long tests).

Pearson's X^2 applied to a bivariate subtable

$$X_{ij}^2 = N\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right)' \hat{\mathbf{D}}_{ij}^{-1}\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right) \tag{9}$$

does not possess an asymptotic chi-square distribution because with the choice of weight matrix, $\hat{\mathbf{W}}_{ij} = \hat{\mathbf{D}}_{ij}^{-1}$, Equation (6) is not satisfied. Furthermore, Equation (7) implies that $X_{ij}^2 > M_{ij}$, and as a consequence, if we use the same reference distribution for X_{ij}^2 as for M_{ij} we would be rejecting well-fitting items (Maydeu-Olivares & Joe, 2006; Maydeu-Olivares & Liu, 2012).

There are at least two ways to obtain asymptotically correct p-values for Pearson's X_{ij}^2 given in Equation (9). One way is by computing p-values for the mixture of chi-square distributions via the inversion formula given in Imhof (1961); another way is to adjust X_{ij}^2 by its mean and variance so that its asymptotic distribution can be approximated by a chi-square distribution. Liu and Maydeu-Olivares (2012) empirically compared the two approaches and concluded that there was little difference between them. As a result, we only consider the mean and variance adjustments because the computation using Imhof's inversion method is more involved but does not yield a more accurate p-value.

To compute p-values for X_{ij}^2 using a mean and variance adjustment, we assume that the distribution of X_{ij}^2 can be approximated by a $b\chi_a^2$ distribution. The first two asymptotic moments of X_{ij}^2 are

$$\mu_1 = \mathrm{tr}\left(\mathbf{D}_{ij}^{-1}\Sigma_{ij}\right), \quad \mu_2 = 2\mathrm{tr}\left(\mathbf{D}_{ij}^{-1}\Sigma_{ij}\right)^2. \tag{10}$$

Solving for the two unknown constants a and b and evaluating m_1 and m_2 at the parameter estimates (denoted with a hat), we obtain the mean and variance corrected $\bar{X}_{ij}^2$ statistic

$$\bar{X}_{ij}^2 = \frac{X_{ij}^2}{b} = \frac{2\hat{\mu}_1}{\hat{\mu}_2}X_{ij}^2, \tag{11}$$

which has an approximate reference chi-square distribution with degrees of freedom

$$a = \frac{2\hat{\mu}_1^2}{\hat{\mu}_2}. \tag{12}$$

This approach originated in Satterthwaite (1946) and has been applied in the SEM literature to overall GOF testing by Satorra and Bentler (1994). For binary item response data, Cai, Maydeu-Olivares, Coffman, and Thissen (2006) used this method to approximate the asymptotic distribution of several overall GOF test statistics.

Following Asparouhov and Muthén (2010), it is possible to define an alternative mean and variance corrected X_{ij}^2 which, unlike Equation (11), has $df_{ij} = C_{ij} - q_{ij} - 1$ degrees of Freedom (provided of course that $df_{ij} > 0$). Their method entails writing the statistic $\bar{\bar{X}}_{ij}{}^2 = a^* + b^* X_{ij}{}^2$ where a^* and b^* are chosen so that the mean and variance of $\bar{\bar{X}}_{ij}^2$ are df_{ij} and $2\ df_{ij}$, respectively. Solving for the two unknown constants

a^* and b^*, we obtain

$$\bar{\bar{X}}^2_{ij} = X^2_{ij}\sqrt{\frac{2df_{ij}}{\hat{\mu}_2}} + df_{ij} - \sqrt{\frac{2df_{ij}\hat{\mu}_1^2}{\hat{\mu}_2}}. \quad (13)$$

Asparouhov and Muthén's (2010) simulation results in the context of overall GOF tests for SEM models suggested that the difference between Equation (11) and Equation (13) is negligible.

Large Sample *z* Statistics

For binary data, M_{ij} and $\bar{\bar{X}}^2_{ij}$ cannot be computed for item pairs due to the lack of degrees of freedom (i.e., $df_{ij} \leq 0$). $\bar{X}^2_{ij}$, however, can still be used because the degrees of freedom in Equation (12) are estimated as a real number. R_{ij} can also be used with binary data as its (integer-valued) degrees of freedom are estimated as well, unless the estimate is exactly zero.

An attractive alternative for binary data is the standardized residual

$$z_{ij} = \frac{p^{(11)}_{ij} - \hat{\pi}^{(11)}_{ij}}{\mathrm{SE}\left(p^{(11)}_{ij} - \hat{\pi}^{(11)}_{ij}\right)} = \frac{p^{(11)}_{ij} - \hat{\pi}^{(11)}_{ij}}{\sqrt{\hat{\sigma}^{(11)}_{ij}/N}}, \quad (14)$$

as suggested by Reiser (1996) and Maydeu-Olivares and Joe (2005). Here, $\pi^{(11)}_{ij} = \Pr\left(Y_i = 1, Y_j = 1\right)$ and $p^{(11)}_{ij}$ is its corresponding observed proportion. It turns out that $p^{(11)}_{ij}$is simply one of the four probabilities in π_{ij}, and $\hat{\sigma}^{(11)}_{ij}$ is its corresponding diagonal element in Equation (5). The asymptotic distribution of this z statistic is standard normal. Notice that z^2_{ij} is asymptotically chi-square with one degree of freedom; in fact, it can also be expressed as a quadratic form statistic in bivariate marginal residuals.

Because for binary data $EY_iY_j = \pi^{(11)}_{ij}$, $\pi^{(11)}_{ij}$is also the population cross-product moment of the two items, and $p^{(11)}_{ij}$ is the sample counterpart. The cross-product moment of two items can also be computed analogously in the polytomous case (Maydeu-Olivares & Liu, 2012):

$$\kappa_{ij} := EY_iY_j = \sum_{y_i=0}^{K-1}\sum_{y_j=0}^{K-1} y_i y_j \pi^{(y_i y_j)}_{ij} \quad (15)$$

in which $\pi^{(y_i y_j)}_{ij} = \Pr\left(Y_i = y_i, Y_j = y_j\right)$. The corresponding sample estimate is $k_{ij} = \mathbf{y}_i'\mathbf{y}_j/N$, where $\mathbf{y}_i$ denotes the N observations on item i coded using categories (0, 1, … K − 1). Consequently, it is possible to define a z statistic for polytomous response variables as

$$z_{ord} = \frac{k_{ij} - \hat{\kappa}_{ij}}{\mathrm{SE}\left(k_{ij} - \hat{\kappa}_{ij}\right)} = \frac{k_{ij} - \hat{\kappa}_{ij}}{\sqrt{\hat{\sigma}^2_{ord}/N}}. \quad (16)$$

However, the computation of the cross product in Equation (15) is not meaningful if the response categories are not ordered; so in Equation (16) we use the subscript *ord* to remind users that this statistic is for ordinal data only. Also, in Equation (16),

$$\hat{\sigma}^2_{ord} = \mathbf{v}'\hat{\Sigma}_{ij}\mathbf{v}, \quad (17)$$

where, from Equation (15), $\mathbf{v}'$ is the $1 \times K^2$ vector

$$\mathbf{v}' = \Big(0 \times 0, 0 \times 1, \ldots 0 \times (K-1), \ldots, (K-1) \times 0, (K-1) \times 1, \ldots, (K-1) \times (K-1)\Big). \quad (18)$$

The statistic z_{ord} also asymptotically follows a standard normal distribution for correctly specified models. Similarly, z^2_{ord} is a quadratic form statistic of bivariate marginal residuals with a weight matrix leading asymptotically to a chi-square statistic with one degree of freedom.

Estimation of the Asymptotic Covariance Matrix of the Item Parameter Estimates

The computation of $\bar{X}^2_{ij}$, $\bar{\bar{X}}^2_{ij}$, $z_{ij,}$ and its generalization z_{ord} requires an estimate of the $q \times q$ information matrix $\boldsymbol{\mathcal{I}}$. When the item parameters are estimated by ML, two commonly used estimates of $\boldsymbol{\mathcal{I}}$ are the expected information matrix

$$\hat{\boldsymbol{\mathcal{I}}}_E = \hat{\boldsymbol{\Delta}}'\hat{\mathbf{D}}^{-1}\hat{\boldsymbol{\Delta}} \quad (19)$$

and the XPD information matrix

$$\hat{\boldsymbol{\mathcal{I}}}_{XPD} = \hat{\boldsymbol{\Delta}}'_O \mathrm{diag}\left(\mathbf{p}_O/\hat{\boldsymbol{\pi}}^2_O\right)\hat{\boldsymbol{\Delta}}_O. \quad (20)$$

In Equation (19), $\boldsymbol{\Delta} = \partial\boldsymbol{\pi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$denotes the $C \times q$ matrix of derivatives of all possible response pattern probabilities with respect to the item parameters, and $\mathbf{D} = \mathrm{diag}(\boldsymbol{\pi})$ is a diagonal matrix of all pattern probabilities. Both are evaluated at the ML estimates. For polytomous data, the expected information matrix can only be computed in models involving a few items. For instance, for the PROMIS depression data this matrix is very difficult to compute because the dimension of $\boldsymbol{\Delta}$ is 390,625 × 40. In contrast, $\mathbf{p}_O$ and $\boldsymbol{\pi}_O$ in Equation (20) denote, respectively, the observed and expected proportions of the C_O observed patterns, and $\boldsymbol{\Delta}_O$ is the $C_O \times q$ matrix of derivatives of the observed patterns with respect to the model parameters. Because $C_O \leq N$, the number of observations, the XPD estimate of the information matrix can always be computed because the dimension of vectors and matrices involved in Equation (20) does not increase as a function of test length.

A third alternative is the observed information matrix, which can be written as

$$\mathcal{I}_O = N\sum_{c=1}^{C_O}\frac{p_c}{(\pi_c(\boldsymbol{\theta}))^2}\left[\frac{\partial\pi_c(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\pi_c(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} - \pi_c(\boldsymbol{\theta})\frac{\partial^2\pi_c(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]$$

$$= \mathcal{I}_{XPD} - N\sum_{c=1}^{C_O}\frac{p_c}{\pi_c(\boldsymbol{\theta})}\frac{\partial^2\pi_c(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \quad (21)$$

in which π_c and p_c are elements in vectors $\boldsymbol{p}_0$ and $\boldsymbol{\pi}_0$, and it differs from the previous two expressions in that it involves second-order derivatives of the pattern probabilities with respect to the model parameters. Equation (21) is computed directly later in our simulations and empirical examples; however, as pointed out by one referee, numerical approximation of Equation (21) via differentiating the EM map, namely, the supplemented EM algorithm, is another possibility (Cai, 2008; Meng & Rubin, 1991). In the sequel, we use only the observed and the XPD information matrices for the computation of the proposed statistics.

Previous Research

Cagnone and Mignani (2007) investigated the empirical Type I errors and power of a statistic closely related to R_{ij}:

$$G_{ij} = N\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right)' \hat{\mathbf{D}}_{ij}^{-1/2}\left(\hat{\mathbf{D}}_{ij}^{-1/2}\hat{\Sigma}_{ij}\hat{\mathbf{D}}_{ij}^{-1/2}\right)^{+}\hat{\mathbf{D}}_{ij}^{-1/2}\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right). \tag{22}$$

Their fitted model was a two-dimensional logistic graded model, which is also referred to as a proportional odds model (McCullagh, 1980) in some of the statistical literature. The alternative model in their power simulations was a unidimensional logistic graded model. With a sample size of 500, G_{ij} showed reasonable Type I errors (between 4 and 9% at the 5% significance level) and good power (between 30 and 90% for $K = 3$ and over 90% for $K = 4$). However, because they used the expected information in their computation, the largest model they considered involved $n = 6$ variables with $K = 4$ categories ($C = 4^6 = 4{,}096$).

Maydeu-Olivares and Liu (2012) investigated the empirical Type I errors of M_{ij} and z_{ord} when fitting Samejima's (1969) graded model and those of z_{ij} when fitting a 2PL. Empirical rejection rates of M_{ij} were right on target even at the smallest sample size considered ($N = 100$). Using the expected information matrix, empirical rejection rates of z_{ij} and z_{ord} were between 4 and 6% at this sample size, which is very close to the 5% nominal level. With a sample of size 1,000, z_{ij} and z_{ord} showed maximum power for rejecting a multidimensional model, whereas the power of M_{ij} was only marginally higher than the nominal rate (between 5 and 12%). The power of M_{ij} was higher for rejecting a 50/50 unidimensional mixture model with a three standard deviation mean difference (power was between 45 and 60%).

Liu and Maydeu-Olivares (2013) studied the empirical Type I errors and power of several statistics as local dependence diagnostics for the 2PL model. The pairwise statistics involved were X_{ij}^2 using as reference a chi-square distribution with degrees of freedom equal to an independence model (Chen & Thissen, 1997), $z_{ij;}$ two score test statistics (Glas & Suárez-Falcón, 2003; Liu & Thissen, 2012); and the statistic

$$\begin{aligned} R_{2,ij} &= N\left(\mathbf{p}_{ij,S} - \hat{\boldsymbol{\pi}}_{ij,S}\right)' \hat{\mathbf{C}}_{ij,S}\left(\mathbf{p}_{ij,S} - \hat{\boldsymbol{\pi}}_{ij,S}\right), \ \hat{\mathbf{C}}_{ij,S} \\ &= \hat{\boldsymbol{\Xi}}_{ij,S}^{-1} - \hat{\boldsymbol{\Xi}}_{ij,S}^{-1}\hat{\boldsymbol{\Delta}}_{ij,S}\left(\hat{\boldsymbol{\Delta}}_{ij,S}'\hat{\boldsymbol{\Xi}}_{ij,S}^{-1}\hat{\boldsymbol{\Delta}}_{ij,S}\right)^{-1}\hat{\boldsymbol{\Delta}}_{ij,S}'\hat{\boldsymbol{\Xi}}_{ij,S}^{-1}, \end{aligned} \tag{23}$$

which is based on the residual bivariate proportions given each sum-score level, $\mathbf{p}_{ij,S} - \boldsymbol{\pi}_{ij,S}$. In Equation (23), $\boldsymbol{\Xi}_{ij,S}$ denotes the asymptotic covariance matrix of the observed proportions, and $\boldsymbol{\Delta}_{ij,S}$ the corresponding Jacobian matrix with respect to the model parameters. $R_{2,ij}$ belongs to the family of test statistics M_κ(Joe & Maydeu-Olivares, 2010), and it asymptotically follows a chi-square distribution with n – 3 degrees of freedom for the 2PL model. Liu and Maydeu-Olivares (2013) found that in terms of Type I errors, Chen and Thissen's X_{ij}^2 was too conservative, whereas the other statistics behaved well with a sufficiently large sample size ($N = 1{,}000$). Power was investigated for a bifactor and a two-dimensional alternative. There were only marginal differences in power between z_{ij} and the score tests (power as high as 95% for some pairs at the 5% level), but $R_{2,ij}$ showed only slightly higher power than nominal α levels (at most 9%). Liu and Maydeu-Olivares (2013) also considered tripletwise statistics; however, they concluded that it is generally not easy to draw useful inferences from them. They also investigated the choice of estimate of the information matrix and found that the behavior of the score test statistics improved markedly when the expected information matrix is used compared with the XPD approximation.

The present study extends previous research by introducing the mean and variance corrected X_{ij}^2 statistics in Equations (11) and (13) and the R_{ij} statistic in Equation (8). Via simulations we investigate their Type I error rate and power together with the z_{ord}/z_{ij} statistics. The observed and XPD information matrices are used for all statistics to gauge the effect of choice of information matrix estimate on the behavior of the statistics. For benchmark purposes, the results for M_{ij}(for ordinal data only), whose computation does not involve the information matrix, and for X_{ij}^2 are also reported.

SIMULATION PART I: EMPIRICAL TYPE I ERROR RATES

Ordinal Data

We used a graded response model with $n = 10$ items and $K = 5$ categories to simulate the data. The true intercept and slope values were

$$\alpha = \begin{pmatrix} 1.60 & 1.79 & 2.13 & 1.60 & 1.79 & 2.13 & 1.60 & 1.79 & 2.13 & 1.60 \\ 0.53 & 0.60 & 0.71 & 0.53 & 0.60 & 0.71 & 0.53 & 0.60 & 0.71 & 0.53 \\ -0.53 & -0.60 & -0.71 & -0.53 & -0.60 & -0.71 & -0.53 & -0.60 & -0.71 & -0.53 \\ -1.60 & -1.79 & -2.13 & -1.60 & -1.79 & -2.13 & -1.60 & -1.79 & -2.13 & -1.60 \end{pmatrix}',$$

$$\beta = \begin{pmatrix} 1.28 & 1.67 & 2.27 & 1.28 & 1.67 & 2.27 & 1.28 & 1.67 & 2.27 & 1.28 \end{pmatrix}'. \tag{24}$$

Normally shaped response distributions are resulted with this choice of intercept values, whereas skewed ones are typically found in practice. As a reviewer pointed out, further research should investigate the performance of the test statistics when item responses are skewed as well. In this case, item parameters may be poorly estimated; however, the degree to

which it may affect the behavior of the proposed tests remains unknown. Preliminary evidence on this can be found in Liu and Maydeu-Olivares (2012) in which they considered asymmetrically placed intercept values and obtained results mostly comparable with those observed in the present study.

Two sample sizes were considered, $N = 300$ and 1,000; 1,000 replications per condition were used. Estimation of the item parameters was performed using Mplus 7.0 (Muthén & Muthén, 2012); 48 rectangular quadrature points from –5 to 5 were used to approximate numerically the marginal likelihood function. All GOF statistics were computed using R (R Development Core Team, 2012). The reference distributions for the test statistics are (a) χ^2_{14} for M_{ij}, X^2_{ij}, and $\bar{\bar{X}}^2_{ij}$; (b) (0, 1)for z_{ord}; (c) a chi-square distribution with degrees of freedom equal to the number of eigenvalues of $\hat{\Sigma}_{ij}$ greater than 10^{-5} for R_{ij}; and (d) Equation (12) for $\bar{X}^2_{ij}$. Given that the model is correctly specified, empirical rejection rates should be close to the nominal α level for all statistics except for the unadjusted X^2_{ij}, whose distribution is not χ^2_{14} but a mixture of independent χ^2_1.

The mean and variance of the statistics across the 1,000 replications are shown in Table 1 for both sample sizes. For chi-square distributed statistics, the mean of the statistics should be equal to the degrees of freedom, and the variance of the statistics should be 2 times the degrees of freedom. For z statistics, the mean and variance of the statistics should be zero and one, respectively. For conciseness, only the results for pairs (1, 2), (1, 3), and (2, 3) are presented, which cover all three slope combinations. Rejection rates at $\alpha = 0.01$, 0.05, and 0.10 are shown graphically in Figure 1, again for both sample sizes. Rejection rates for $\bar{X}^2_{ij}$ are not shown in Figure 1 as they are almost identical to those of $\bar{\bar{X}}^2_{ij}$.

In Figure 1 we see that the empirical Type I error rates of M_{ij} are right on target even when the sample size is 300, whereas the unadjusted X^2_{ij} rejects slightly more often (8–9% at 5% level); the pattern is similar to what was reported in Maydeu-Olivares and Liu (2012). When the observed information matrix is used, the empirical Type I error rates of $\bar{\bar{X}}^2_{ij}$, R_{ij}, and z_{ord} are also accurate. In contrast, when the XPD information matrix is used z_{ord} becomes unusable: due to negative variance estimates, z_{ord} could only be computed across 300, 189, and 164 replications for these three pairs as $N =$ 300 and 986, 959, and 948 replications as $N = 1{,}000$. Even for $N = 1{,}000$, the statistic rejects too many ($> 15\%$ at 5% level) well-fitting items. In Figure 1, we also see that when the XPD information matrix is used, the empirical rejection rates of $\bar{\bar{X}}^2_{ij}$ are somewhat inflated (about 8% at 5% level) when $N = 300$, whereas those of R_{ij} are not too adversely affected.

However, in Table 1 we see that the empirical variance of R_{ij} can be very large (maximum 432.57; in general larger than 2 times the mean). This means that in correctly specified models one may observe extremely large values of this statistic in applications, which leads researchers to believe that the model grossly misfits one or more pairs. This is an extremely undesirable feature. Also in Table 1 we see that the variance of the statistic generally increases as sample size increases and that it is present for both information matrix estimates, although the variances are generally larger when the XPD information is used.

Note that even in a model of this size (which is small by IRT standards), the expected information matrix cannot be computed. If the computation of the statistic involves the information matrix, the observed information is preferred over the XPD information as suggested by our simulation results. We also conclude that in terms of retaining well-fitting ordinal items, M_{ij} is the best statistic, very closely followed by z_{ord} and the mean-and-variance adjusted X^2_{ij} if

TABLE 1
Estimated Mean and Variance of Bivariate GOF Statistics: Graded Model, Correctly Specified

			$N = 300$			$N = 1{,}000$		
Pair (i, j)	Stat	Info.	Mean	Variance	*df*	Mean	Variance	*df*
(1, 2)	M_{ij}	—	14.05	26.07	14	13.78	27.51	14
	X^2_{ij}	—	15.67	29.51	14	15.37	28.27	14
	$\bar{X}^2_{ij}$	OBS	16.06	30.98	16.00	15.76	29.70	16.01
		XPD	14.16	22.87	12.81	15.34	27.84	15.22
	$\bar{\bar{X}}^2_{ij}$	OBS	14.06	27.13	14	13.77	25.98	14
		XPD	15.44	26.84	14	14.12	26.04	14
	R_{ij}	OBS	22.30	91.48	21.85	22.10	53.78	22.14
		XPD	18.86	157.29	17.53	20.09	181.78	19.06
	z_{ord}	OBS	0.01	1.03	—	0.01	1.06	—
		XPD	0.61	9.94	—	–0.03	2.09	—
(1, 3)	M_{ij}	—	14.02	27.29	14	14.14	29.14	14
	X^2_{ij}	—	15.56	30.15	14	15.65	31.05	14
	$\bar{X}^2_{ij}$	OBS	15.96	31.69	15.94	16.06	32.69	15.95
		XPD	14.07	23.46	12.76	15.63	30.60	15.16
	$\bar{\bar{X}}^2_{ij}$	OBS	14.02	27.85	14	14.10	28.69	14
		XPD	15.40	27.55	14	14.46	28.77	14
	R_{ij}	OBS	22.07	78.16	21.81	23.20	116.51	22.30
		XPD	18.64	77.79	17.60	20.27	138.28	19.16
	z_{ord}	OBS	–0.05	0.96	—	–0.07	0.93	—
		XPD	–0.11	7.77	—	–0.13	3.26	—
(2, 3)	M_{ij}	—	13.95	27.67	14	14.05	28.54	14
	X^2_{ij}	—	15.50	30.30	14	15.56	29.23	14
	$\bar{X}^2_{ij}$	OBS	15.90	31.84	15.92	15.96	30.77	15.93
		XPD	13.99	23.45	12.71	15.53	28.78	15.13
	$\bar{\bar{X}}^2_{ij}$	OBS	13.98	28.01	14	14.03	27.04	14
		XPD	15.38	27.61	14	14.39	27.09	14
	R_{ij}	OBS	22.32	89.33	21.85	23.22	81.99	22.31
		XPD	19.46	210.22	17.62	22.08	432.57	19.22
	z_{ord}	OBS	0.02	0.94	—	–0.03	1.01	—
		XPD	0.03	11.61	—	0.03	4.08	—

Note. df reported for $\bar{X}^2_{ij}$ and R_{ij} are averages across the 1,000 replications. GOF = goodness of fit; OBS = observed information matrix; XPD = cross-product information. All results are based on 1,000 replications except for z_{ord} when using cross-product information. In this case, due to negative variance estimates, z_{ord} could only be computed across 300, 189, and 164 replications for these three pairs as $N = 300$ and 986, 959, and 948 replications as $N = 1{,}000$. — = not applicable.

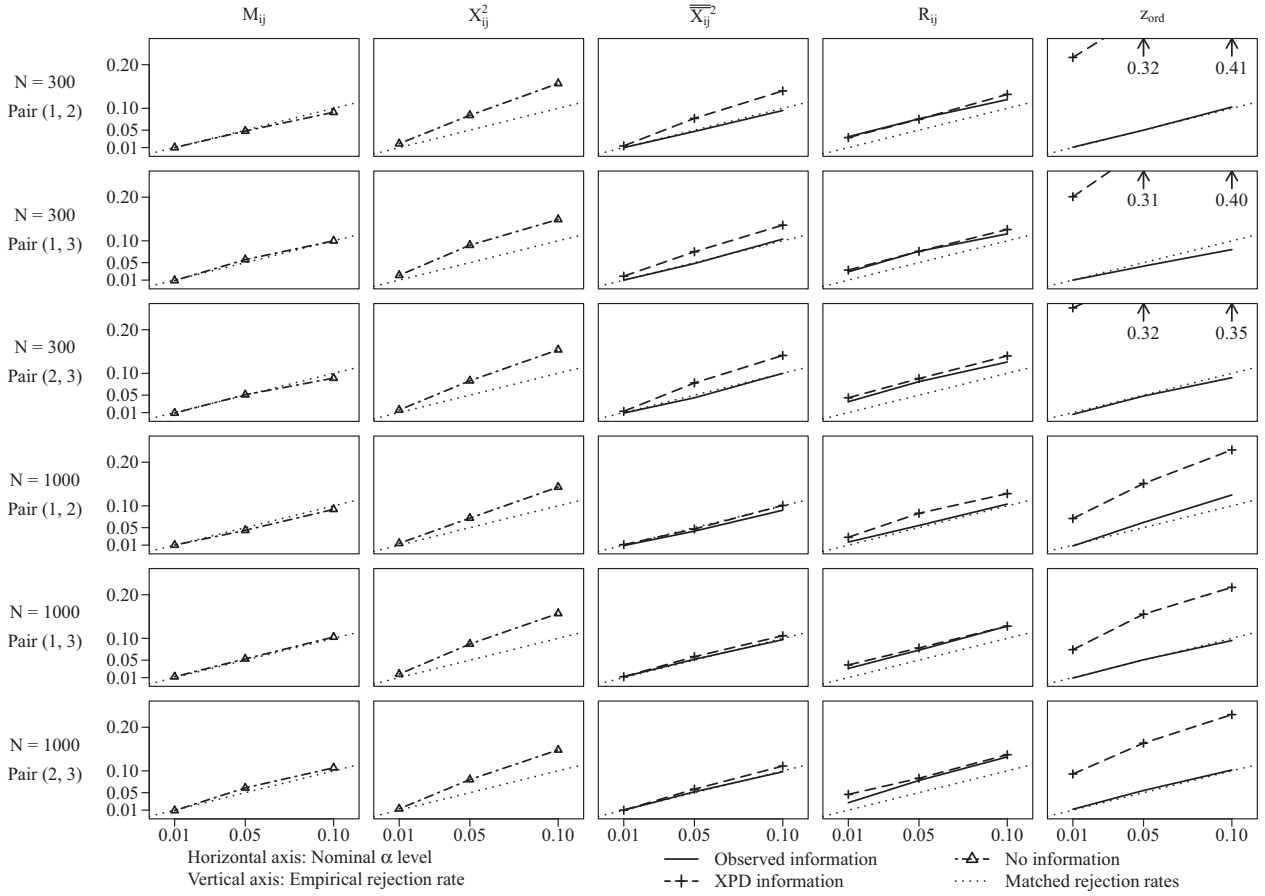

FIGURE 1 Type I error results for bivariate GOF statistics: $N = 300$ and 1,000; true model—graded, fitted model—graded. The statistics are M_{ij}, the unadjusted X^2_{ij}(with the same *df* as M_{ij}), the mean and variance adjusted $\bar{\bar{X}}^2_{ij}$, R_{ij}, and the residual cross product $z_{ord.}$ GOF = goodness of fit; XPD = cross-product information.

the observed information is used. If only XPD information is available, z_{ord} should be avoided. In both cases, R_{ij} should be used with caution due to its problematic performance for some pairs in the simulation. On the other hand, the unadjusted X^2_{ij} should be avoided because of its liberality.

Binary Data

In this section, binary item response data were generated from a 2PL model. The simulation setup was the same as in the graded model conditions; specifically, once again we used 10 items and two sample size conditions: $N = 300$ and 1,000. The true parameter values in this case were $\alpha = 0$ and

$$\beta = \begin{pmatrix} 1.28 & 1.67 & 2.27 & 1.28 & 1.67 & 2.27 & 1.28 & 1.67 & 2.27 & 1.28 \end{pmatrix}'. \tag{25}$$

In this simulation setting, the M_{ij} and $\bar{\bar{X}}^2_{ij}$ statistics cannot be computed due to the lack of degrees of freedom. The reference distributions for the remaining statistics are (a) (0, 1)for z_{ij}; (b) a chi-square distribution with degrees of freedom equal to the number of eigenvalues of $\hat{\Sigma}_{ij}$ greater than 10^{-5} for R_{ij}; and (c) Equation (12) for $\bar{X}^2_{ij}$. For comparative purposes, we also provide results for the unadjusted X^2_{ij} using Chen and Thissen's (1997) proposal of using a chi-square distribution with degrees of freedom equal to those of the independence model. For this setup, this amounts to using a χ^2_1 reference distribution.

The mean and variance of the statistics under consideration are reported in Table 2 for both sample sizes. Empirical rejection rates at selected α levels, again for both sample sizes, are shown in Figure 2.

Figure 2 shows that Pearson's X^2_{ij} with Chen and Thissen's (1997) reference distribution is over-conservative (less than 1% at 5% level) for all three pairs, as has frequently been observed in previous simulation studies. In contrast, when observed information is used, z_{ij}, R_{ij}, and the mean and variance adjusted $\bar{X}^2_{ij}$ have rather accurate empirical Type I errors even at the smallest sample size considered. However, when the XPD information matrix is used, $\bar{X}^2_{ij}$ tends to overreject slightly (13–15% at 10% level) at small sample sizes and underrejects with low slopes (5% at 10% level for Items 1 and 2); z_{ij} is over-liberal (over 10% at 5% level) for all three pairs at small sample sizes. Furthermore, in this case, due to negative variance estimates, z_{ij} could only be computed

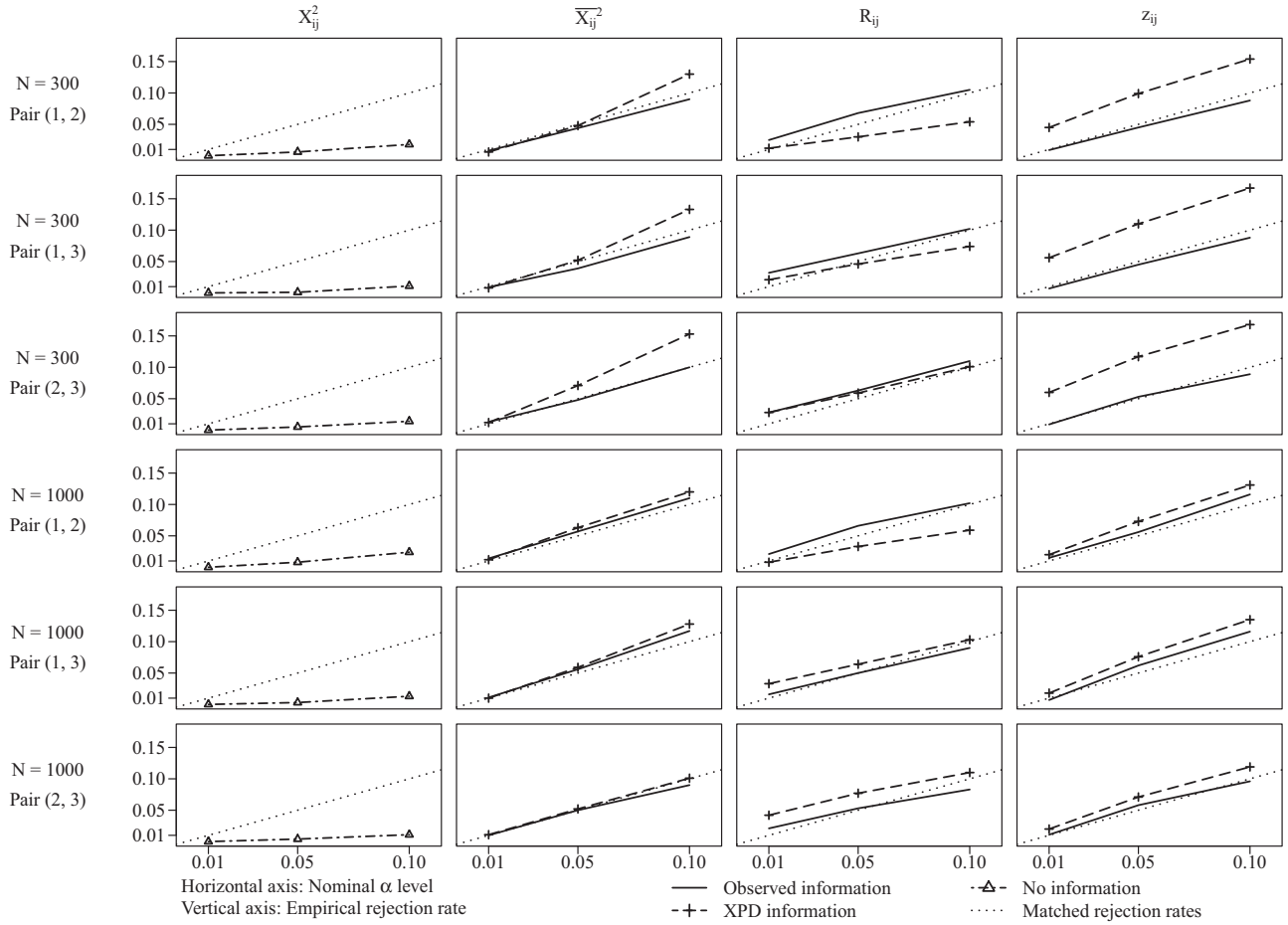

FIGURE 2 Type I error results for bivariate GOF statistics: $N = 300$ and 1,000; true model—2PL, fitted model—2PL. The statistics are the unadjusted X^2_{ij}(with the same *df* as the independence model), the mean and variance adjusted$\bar{X}^2_{ij}$, R_{ij}, and the residual cross product z_{ij}. GOF = goodness of fit; XPD = cross-product information; 2PL = two-parameter logistic.

across 993, 990, and 978 replications for these three pairs as $N = 300$.

Inspecting Table 2, we see again that R_{ij} can take very large values in applications, as its empirical variance is very large, particularly when XPD information is used. As before, we conclude that the observed information matrix should be used whenever possible. We also conclude that in terms of retaining well-fitting binary items, z_{ij} or the mean-and-variance $\bar{X}^2_{ij}$ statistic should be used. If only XPD information is available, $\bar{X}^2_{ij}$ should be used. In both cases, R_{ij} should be used with caution.

SIMULATION PART II: POWER

In this section we examine the power of the statistics under investigation for detecting model misspecification. Only statistics with adequate empirical Type I error rates were investigated: These are M_{ij}, $\bar{X}^2_{ij}$, $\bar{\bar{X}}^2_{ij}$, z_{ord}, and R_{ij} for the ordinal case and $\bar{X}^2$, R_{ij}, and z_{ij} for the binary case. As in the previous section, the fitted model was graded for ordinal data and the 2PL for binary data. We investigated power for detecting (a) a multidimensional model, (b) a unidimensional model with the latent variable distributed as a mixture of normal distributions, and (c) a unidimensional model with a guessing parameter. For conciseness, only power results for $N = 1{,}000$ are reported. A thousand replications were used for each condition.

To generate multidimensional data, we used an independent cluster two-dimensional graded response model. The same intercepts as in the previous simulation were used; that is, we used the ones reported in Equation (23) for the graded case and $\alpha = 0$ in the binary case. The latent traits correlation was set to 0.7 and the slopes used for both the graded and binary case were

$$\beta = \begin{pmatrix} 1.28 & 1.67 & 2.27 & 1.28 & 1.67 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.28 & 1.67 & 2.27 & 1.28 & 1.67 \end{pmatrix}'. \quad (26)$$

TABLE 2
Estimated Mean and Variance of Bivariate GOF Statistics: 2PL Model, Correctly Specified

			$N = 300$			$N = 1{,}000$		
Pair (i, j)	Stat	Info.	Mean	Variance	*df*	Mean	Variance	*df*
(1, 2)	X^2_{ij}	—	0.49	0.45	1	0.55	0.56	1
	$\bar{X}^2_{ij}$	OBS	0.96	1.76	1.02	1.09	2.18	1.02
		XPD	0.74	1.01	0.64	1.01	1.74	0.90
	R_{ij}	OBS	2.35	13.07	2.44	2.21	6.24	2.41
		XPD	1.29	7.75	1.34	1.86	94.05	1.70
	z_{ij}	OBS	0.03	0.95	—	0.02	1.07	—
		XPD	–0.23	7.28	—	–0.05	1.19	—
(1, 3)	X^2_{ij}	—	0.42	0.31	1	0.46	0.41	1
	$\bar{X}^2_{ij}$	OBS	0.99	1.80	1.05	1.09	2.23	1.04
		XPD	0.74	1.02	0.62	1.01	1.85	0.91
	R_{ij}	OBS	2.65	40.44	2.47	2.16	4.99	2.42
		XPD	1.80	44.69	1.43	2.45	47.44	1.86
	z_{ij}	OBS	0.03	0.94	—	0.04	1.04	—
		XPD	–0.24	5.46	—	–0.04	1.22	—
(2, 3)	X^2_{ij}	—	0.42	0.39	1	0.41	0.34	1
	$\bar{X}^2_{ij}$	OBS	1.08	2.56	1.06	1.04	2.20	1.05
		XPD	0.79	1.41	0.60	0.97	1.76	0.92
	R_{ij}	OBS	2.44	11.91	2.48	2.13	5.83	2.42
		XPD	1.90	24.25	1.46	2.62	82.12	1.88
	z_{ij}	OBS	–0.02	1.00	—	–0.05	1.00	—
		XPD	–0.34	7.76	—	–0.12	1.15	—

Note. df reported for $\bar{X}^2_{ij}$ and R_{ij} are averages across the 1,000 replications. GOF = goodness of fit; 2PL = two-parameter logistic; OBS = observed information; XPD = cross-product information. All results are based on 1,000 replications except for z_{ij} when using cross-product information. In this case, due to negative variance estimates, z_{ij} could only be computed across 993, 990, and 978 replications for these three pairs as $N = 300$. — = not applicable.

To generate mixture data, we used the same setup as in simulations for correctly specified models except that the latent variable was drawn from a 50/50 mixture of (0, 1) and $\mathcal{N}(2, 1)$ distributions. Finally, to generate data with a guessing parameter, we used the same setup as in the simulations for correctly specified models, except that we replaced the 2PL formula of Equation (3) with a three-parameter logistic (3PL) model:

$$\Psi\left(\eta\,;\alpha_{i,k},\beta_i,\gamma_i\right)=\gamma_i+\frac{1-\gamma_i}{1+\exp\left[-\left(\alpha_{i,k}+\beta_i\eta\right)\right]}, \tag{27}$$

where γ_i is the guessing parameter. We used $\gamma_i = 0.2$ for all items (in both binary and ordinal conditions). We note that fitting a guessing parameter in conjunction with the graded response model is seldom pursued in practice; it is used here only for data generation purposes.

Ordinal Data

Empirical power rates at $\alpha = 0.05$ using both estimates of the information matrix are shown graphically in Figure 3 for the same item pairs as in the previous section: (1, 2), (1, 3), and (2, 3), where the average slopes are low, medium, and high, respectively. For the multidimensional condition, these pairs correspond to items belonging to the same factor. Thus, for this condition we also provide results for pairs (1, 7), (1, 8), and (2, 8): each comprises one item from the first factor and one from the second factor; all possible combinations of slope are covered by the choice of these six pairs. Also, for z_{ord}, based on the results of the previous section, only results using observed information are provided.

It is clear in Figure 3 that M_{ij} exhibits very low power (almost identical to the nominal level) to detect a multidimensional model even at this sample size, as pointed out previously by Maydeu-Olivares and Liu (2012). However, its power for detecting a mixture model and the presence of a lower asymptote in the response function is similar to the power of the mean and variance adjusted $\bar{\bar{X}}^2_{ij}$. $\bar{\bar{X}}^2_{ij}$ has high power (about 90%) for detecting the presence of multidimensionality in high slope items belonging to the same dimension. However, its power decreases as the item slopes decrease. Furthermore, it has low power (less than 20%) for detecting the presence of multidimensionality in between-factor item pairs.

We also see in this figure that all statistics have problems in detecting the presence of a mixture model. Power at the 5% level is less than 30%. Power increases with increasing slopes, and the most powerful statistic is the residual cross product z_{ord}, followed by R_{ij}. In fact, the residual cross product is also the most powerful statistic (reject 90–100%) for detecting multidimensionality. R_{ij} is the most powerful statistic (almost always reject) for detecting the presence of a lower asymptote in the response function (guessing), although in this condition the power decreases substantially when the XPD information matrix is used instead of the observed information matrix.

Binary Data

Empirical power rates at $\alpha = 0.05$ using both estimates of the information matrix are shown graphically in Figure 4 for the same item pairs as in the ordinal condition. In this case, results using XPD information for the residual cross product statistic z_{ij} are provided due to its acceptable Type I errors.

In this figure we see that R_{ij} (47%–93% for items in the same factor and 22%–45% for items in different factors) is slightly less powerful than z_{ij} and $\bar{X}^2_{ij}$ (67%–98% for items in the same factor and 34%–63% for items in different factors) for detecting multidimensionality in the binary case. Even the most powerful statistic, z_{ij}, exhibits only moderate power for detecting the presence of multidimensionality in pairs composed of items belonging to different dimensions when the items have low slopes.

As in the case of the previous results with ordinal data, the power of these statistics for detecting the presence of mixtures is low (less than 30% at the 5% level); and also when detecting the presence of guessing, the power of $\bar{X}^2_{ij}$

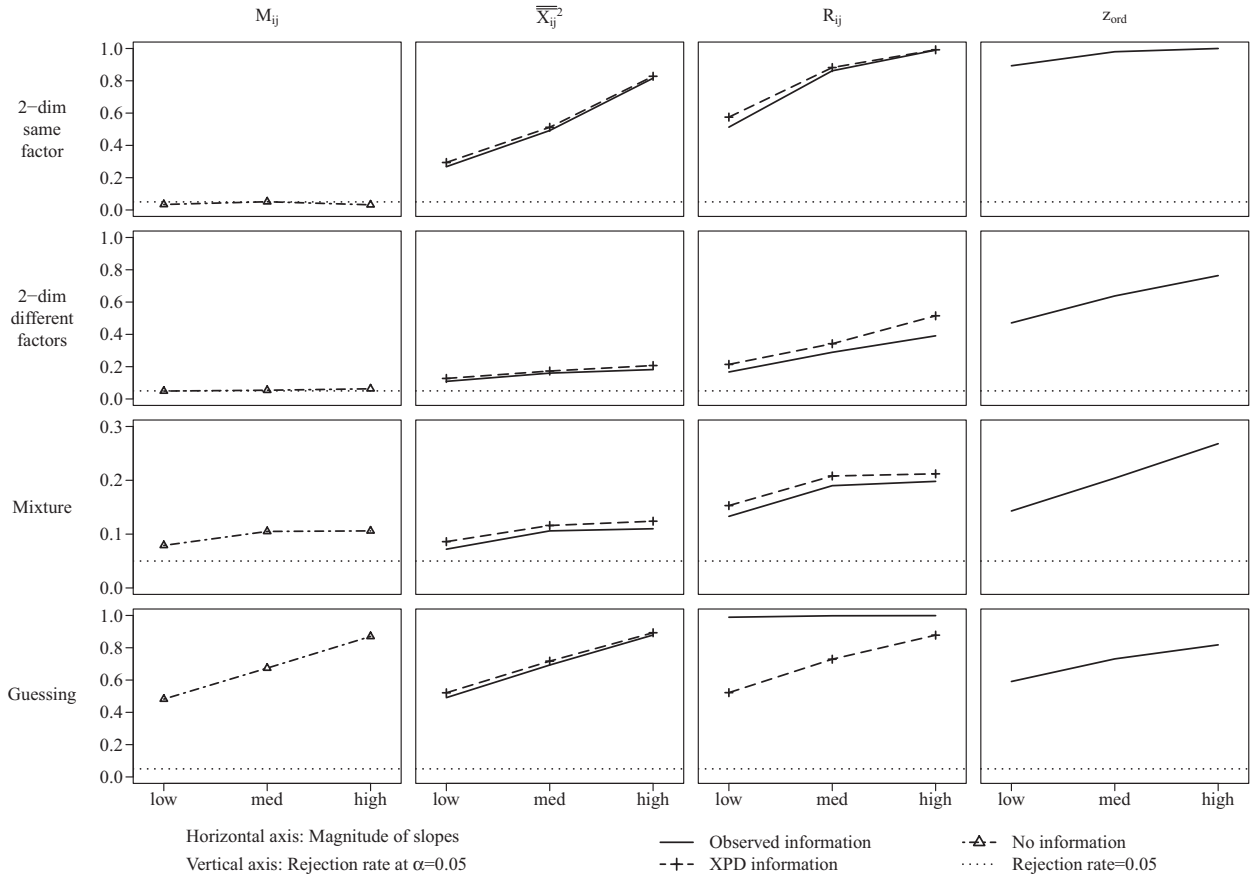

FIGURE 3 Power results at $\alpha = 0.05$ for bivariate GOF statistics: $N = 1{,}000$, true model—two-dimensional graded, mixture of graded models, and graded model with guessing, fitted model—graded. The statistics are M_{ij}, the mean and variance adjusted $\bar{\bar{X}}^2_{ij}$, R_{ij}, and the residual cross product z_{ord}. GOF = goodness of fit; XPD = cross-product information.

is only slightly higher than the nominal level. R_{ij} is the only statistic that has power for detecting the presence of guessing but only when the observed information is used. When XPD information is used, however, the power of R_{ij} is only slightly higher than the nominal level and lower than the power of z_{ij}. In fact, comparing the results shown in Figures 3 and 4 we see that the choice of information matrix has a much larger impact in binary data than in ordinal data. For binary data, generally the power is higher when the XPD information matrix is used to compute the R_{ij} and z_{ij} statistics, although one should also take the differential empirical Type I errors into consideration.

DISCUSSION

The behavior of the statistics varies for different types of misfit and for different choices of information matrix estimate (observed vs. XPD); as such, it is not always easy to provide an overall recommendation as to which statistic should be used. M_{ij} has obvious computational advantages in that it does not require a covariance matrix of the item parameter estimates and exhibits the best Type I error rates. However, its power is lower than that of the other statistics under the three conditions simulated here. Hence, its use cannot be recommended. The remaining statistics require computing the information matrix (used to obtain the covariance matrix of the item parameters). In applications, the choice of information matrix may be limited by the software used to perform the analysis. If both approaches are available, we strongly recommend using the observed information matrix, as this choice yields more accurate Type I errors.

When only the XPD information is available, we suggest computing the mean and variance adjusted statistics $\bar{X}^2_{ij}$ and $\bar{\bar{X}}^2_{ij}$. When the observed information matrix is available, the use of the cross-product residual (z_{ord} or z_{ij}) is recommended because its Type I errors are precisely controlled and it is generally the most powerful statistic. However, it is not meaningful to compute this statistic when items are polytomous and nominal. In this case, we suggest computing the mean and variance adjusted X^2 statistics because the theory discussed in this article still applies; however, additional simulation studies are needed to evaluate the empirical Type I error and power behavior of the statistics.

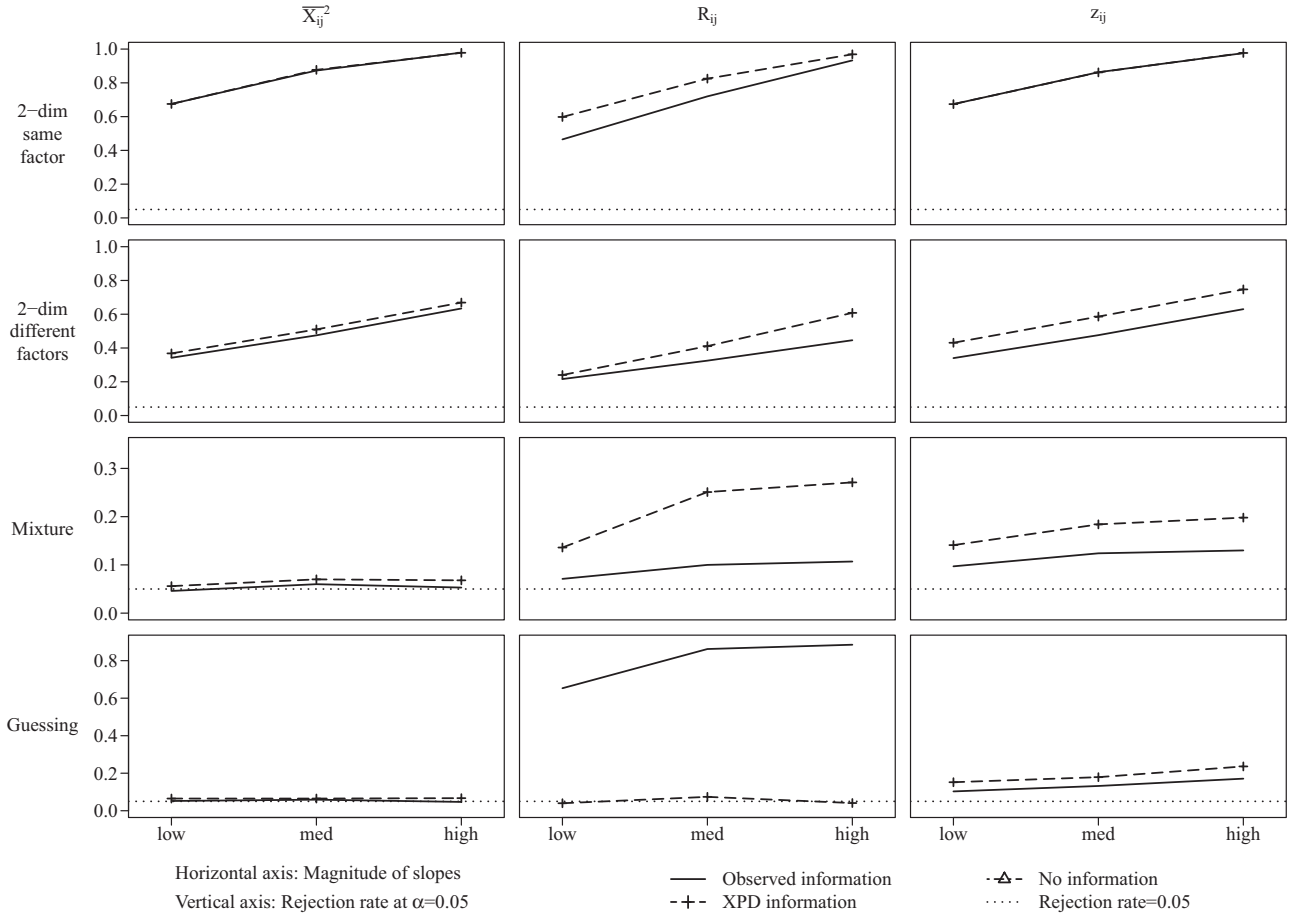

FIGURE 4 Power results at $\alpha = 0.05$ for bivariate GOF statistics: $N = 1{,}000$, true model—two-dimensional 2PL, mixture of 2PLs, and 3PL, fitted model—2PL. The statistics are the mean and variance adjusted $\bar{X}^2_{ij}$, R_{ij}, and the residual cross product z_i. GOF = goodness of fit; 2PL = two-parameter logistic; 3PL = three-parameter logistic; XPD = cross-product information.

In line with the results reported by Asparouhov and Muthén (2010) for structural equation models, we found negligible differences between the two mean-and-variance-adjusted statistics $\bar{X}^2_{ij}$ and $\bar{\bar{X}}^2_{ij}$. From an applied perspective, the latter is preferable because its degrees of freedom are integer valued and determined by the usual formula, that is, the number of parameters in the saturated model minus the number of parameters in the restricted model. However, $\bar{\bar{X}}^2_{ij}$ is not applicable to test the fit of the 2PL model to pairs of items, as the degrees of freedom computed in this fashion are negative. $\bar{X}^2_{ij}$, on the other hand, can still be used, as its degrees of freedom are estimated as a real number. However, values of $\bar{X}^2_{ij}$ for different item pairs cannot be directly compared as they are on a different scale (their estimated *df*). Only the *p*-values can be directly compared across item pairs. This is undesirable in actual applications as it forces researchers to inspect tables of *p*-values with a large number of decimals in order to determine the item pairs with the greatest magnitude of misfit. To improve the reporting of the results, we suggest reporting $\bar{\bar{X}}^2_{ij}$ using one degree of freedom in applications. As an alternative, the standardized $\bar{X}^2_{ij}$,

$$\frac{\bar{X}^2_{ij} - a}{\sqrt{2a}}, \tag{28}$$

could be used, where a denotes the degrees of freedom estimated using Equation (12). One advantage of $\bar{\bar{X}}^2_{ij}$ and Equation (28) is that comparisons of the magnitude of statistics across item pairs that have different estimated degrees of freedom (i.e., the value of a), as well as those composed of different numbers of categories, are facilitated. In fact, selecting one or the other of the two statistics to display binary data results is irrelevant, as it can be easily verified that the two are mathematically related by the formula

$$\frac{\bar{X}^2_{ij} - a}{\sqrt{2a}} = \frac{-1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\bar{\bar{X}}^2_{ij}. \tag{29}$$

If the presence of guessing or a similar phenomenon is of concern, then R_{ij} should be used, as this statistics exhibits

the highest power in our simulations for detecting this kind of misfit. Throughout the conditions investigated here, R_{ij} exhibits good behavior in terms of Type I errors and power rates. Our main concern regarding this statistic is its very large sampling variance across all conditions. This means that in actual applications, very large R_{ij} values are likely to be encountered; this may lead applied researchers to conclude that a certain pair of items fits extremely poorly, although in fact this may simply be the result of the large sampling variability of the statistic. It is for this reason that we recommend the use of the mean and variance adjusted Pearson's X^2 statistic rather than R_{ij}, except when the presence of guessing is a concern. This recommendation, however, involves trading off a small loss in power for a statistic with no excessive variance.

Here we compute the R_{ij} statistic after obtaining a pseudoinverse of $\hat{\Sigma}_{ij}$, the estimated asymptotic covariance matrix of the bivariate residuals using an eigendecomposition. Because the rank of Σ_{ij} may depend on the true parameter values (Reiser, 1996), we obtain $\hat{\Sigma}_{ij}^{+}$ by setting to zero all eigenvalues of $\hat{\Sigma}_{ij}$ smaller than 10^{-5} and taking the number of nonzero eigenvalues as the degrees of freedom of the statistic. In so doing, the degrees of freedom in our simulations may change from replication to replication. Because numerical integration is used in the IRT models considered, the eigenvalues of $\hat{\Sigma}_{ij}$ are not estimated precisely. We conjecture that this is the source of the large sampling variability of R_{ij} in our simulations. If so, this variability may be reduced if a larger cutoff, say 10^{-4}, is used. Regardless of the number of eigenvalues used to compute $\hat{\Sigma}_{ij}^{+}$, the distribution of R_{ij} remains asymptotically chi-square under the null hypothesis. Therefore, an alternative approach for computing R_{ij} is to use a predetermined number of degrees of freedom df_{ij} and set the smallest $C_{ij} - df_{ij}$ eigenvalues to zero in the computation of $\hat{\Sigma}_{ij}^{+}$. We conjecture that the empirical Type I errors of the resulting R_{ij} statistic will remain close to the nominal rates for any selected df_{ij} smaller than the true but unknown df_{ij} and that the empirical variance of the statistic will be close to its expected variance under the reference chi-square distribution for a small enough df_{ij}. However, the selected df_{ij} should not be much smaller than the true and unknown df_{ij} because the smaller the selected df_{ij} the smaller the value of the estimated R_{ij} statistic; consequently, we also conjecture that when the predetermined df_{ij} is too small, the statistic may suffer from a loss of power. Because we have observed in our simulations that most often df_{ij} estimated using a cutoff 10^{-5} for the eigenvalues is larger than the usual formula for degrees of freedom in chi-square statistics, $df_{ij} = C_{ij} - q_{ij} - 1$, it is likely that using this number of predetermined eigenvalues in the computation of R_{ij} may yield the best combination of power and sampling variance of the statistic. However, the question of how to improve the performance of R_{ij} is left for further research.

APPLICATIONS

We now return to the two data examples introduced earlier. Because the overall M_2 statistic and the RMSEA_2 suggested that the model fit could be improved, we assess the misfit of the model to each pair of items. The observed information matrix, obtained from Mplus 7.0, is used in the computation of the asymptotic covariance matrix of bivariate residuals. Based on the simulation results, we mainly refer to the residual cross-product z statistics (z_{ord} or z_{ij}) for identifying the source of misfit; however, we also provide $\bar{\bar{X}}_{ij}^2$ and R_{ij} statistics in order to evaluate the extent to which different statistics agree.

PROMIS Depression Short Form

z_{ord} statistics for this example are presented in Table 3. To control for multiple testing, we used the Bonferroni procedure. Because there are 28 pairs of items, we use a significance level of $\alpha = 0.05/28 = 0.0018$. Using a standard normal reference distribution, the critical value is 3.12. We mark in boldface all values of z_{ord} larger in absolute value than this critical value. It is widely known that the Bonferroni procedure is conservative; that is, it flags fewer misfitting item pairs than it should. If a more precise joint significance level is desired, we recommend using the Benjamini-Hochberg procedure (Thissen, Steinberg, & Kuang, 2002). The results are displayed in Table 3. A column has been added to the table that includes the average of the absolute value of the z_{ord} values. The row average suggests that the poorest item is Item 4, followed by Item 7. Table 3 also reveals that after applying a Bonferroni correction, there are only four statistically significant z statistics. They involve item pairs (5,1), (7,4), (8,3), and (7,6).

For comparison, we also provide the values of the $\bar{\bar{X}}_{ij}^2$ and R_{ij} statistics for these data, which are shown in Table 4. The

TABLE 3
z_{ord} Statistics (z Statistics for Residual Cross Product) for the PROMIS Depression Data

Item	1	2	3	4	5	6	7	8	Average
1		1.44	1.49	–0.24	**3.29**	–0.27	–0.13	1.77	1.23
2	1.44		0.03	–1.23	0.82	–1.40	0.99	1.02	0.99
3	1.49	0.03		0.13	0.71	0.90	–0.21	**4.88**	1.19
4	–0.24	–1.23	0.13		–2.81	3.11	**7.01**	0.24	2.11
5	**3.29**	0.82	0.71	–2.81		–0.55	–0.63	1.66	1.50
6	–0.27	–1.40	0.90	3.11	–0.55		**4.08**	0.39	1.53
7	–0.13	0.99	–0.21	**7.01**	–0.63	**4.08**		0.12	1.88
8	1.77	1.02	**4.88**	0.24	1.66	0.39	0.12		1.44

Note. Statistics in bold are statistically significant at the 5% significance level using the Bonferroni correction. A column has been added to the table that includes the average of the absolute value of the z_{ord} values. PROMIS = Patient-Reported Outcomes Measurement Information System.

TABLE 4
Mean and Variance Adjusted $\bar{\bar{X}}^2_{ij}$ (Above the Diagonal) and R_{ij} (Below the Diagonal) Statistics for the PROMIS Depression Data

Item	1	2	3	4	5	6	7	8	Average
1		23.02	21.64	**36.42**	22.42	25.17	**37.45**	17.54	26.24
2	33.68		16.13	21.97	**38.14**	21.02	29.66	23.82	24.82
3	44.34	27.75		**34.49**	32.97	18.56	16.66	26.74	23.88
4	**54.57**	**50.88**	**60.69**		31.66	32.81	33.25	**37.74**	32.62
5	**64.10**	**70.47**	43.57	**88.49**		20.42	**56.57**	22.30	32.07
6	41.72	41.21	37.02	**54.23**	36.10		22.26	15.90	22.30
7	**55.78**	**70.35**	43.57	**132.99**	**73.92**	42.56		23.97	31.40
8	**53.26**	37.01	**79.89**	**53.43**	36.20	39.98	**52.98**		24.00

Note. Statistics in bold are statistically significant at the 5% significance level using the Bonferroni correction. A column has been added to the table that includes the average of the $\bar{\bar{X}}^2_{ij}$ values across all eight items. PROMIS = Patient-Reported Outcomes Measurement Information System.

estimated degrees of freedom for R_{ij} were 20 for 4 pairs, 21 for 16 pairs, and 22 for 8 pairs. In contrast, the degrees of freedom used for $\bar{\bar{X}}^2_{ij}$ are $5^2 - 2 \times 5 - 1 = 14$. A column has been added to the table that includes the average of the $\bar{\bar{X}}^2_{ij}$ values across all eight items. Averages of R_{ij} statistics are not computed as they are based on different degrees of freedom.

After applying a Bonferroni correction, the $\bar{\bar{X}}^2_{ij}$ values suggest that the model misfits pairs (4,1), (4,3), (5,2), (7,1), (7,5), and (8,4). Thus, although our simulation results reveal that z_{ord} is generally more powerful than $\bar{\bar{X}}^2_{ij}$ under the conditions investigated, in this example $\bar{\bar{X}}^2_{ij}$ identifies more misfitting pairs than z_{ord}. Furthermore, the pairs flagged by these procedures only partially overlap. In our experience, this is generally the case. We conjecture that it is because $\bar{\bar{X}}^2_{ij}$ and z_{ord} summarize the information contained in the bivariate residuals differently, as is manifested by their different degrees of freedom (consider squared z_{ord} having asymptotically $df = 1$). Specifically, z_{ord} provides a more concentrated summary than $\bar{\bar{X}}^2_{ij}$. If the concentration of the information is along the direction of the misfit, z_{ord} will be more powerful than $\bar{\bar{X}}^2_{ij}$; otherwise, it will be as powerful, or even less powerful, than $\bar{\bar{X}}^2_{ij}$. In this article, we have seen that the concentration of information performed by z_{ord} helps to detect the presence of multidimensionality of the latent trait. However, it does not help to improve the detection of mixtures of latent traits or the detection of guessing.

As for R_{ij}, after applying a Bonferroni correction, in Table 4 we see that this statistic suggests that the model misfits 15 of 23 possible item pairs. Furthermore, all R_{ij} statistics involving Item 8 are statistically significant. Although this may suggest that this statistic is the most powerful in this example, given our simulation results we cannot be sure how far these results are due to the variability of the R_{ij} statistic.

Because no item fits the model much worse than the others, and because the scale has already been in short form, we advise against removing any item to improve the fit. Deleting an item when there are so few may sharply reduce the precision of the measurement. Our advice in this application is to attempt to find a better fitting model. Failing to do so, the fitted model may be used as it provides a close fit to the data using the criteria of Maydeu-Olivares and Joe (2014). Here is a word of caution: a piecewise assessment shall be performed regardless of the value of the $RMSEA_2$ (or similar overall measure of fit). It is not hard to find tests where the $RMSEA_2$ is low but one or more items fit poorly.

EPQ-R Extraversion Scale Short Form

z_{ij} statistics for this example are presented in Table 5. The results are displayed in this table after grouping the items according to their misfit. This is achieved by performing a

TABLE 5
z_{ij} Statistics (z Statistics for Residual Cross Product) for the EPQ-R Extraversion Data

Item	1	5	9	2	4	7	3	10	6	8	12	11	Average
1		–0.01	–1.79	0.35	–2.59	**8.28**	**–4.67**	**–5.11**	–0.65	**–4.82**	–0.07	–1.27	2.69
5	–0.01		–0.93	–3.51	–0.46	3.62	2.29	–3.39	–1.06	**–3.67**	–3.34	–0.75	2.10
9	–1.79	–0.93		–2.43	**6.38**	–1.32	1.80	–3.34	1.29	**–12.33**	–3.33	**4.38**	3.57
2	0.35	–3.51	–2.43		–1.05	–2.25	–0.52	**–4.32**	–0.54	**–4.46**	**7.63**	–0.50	2.50
4	–2.59	–0.46	**6.38**	–1.05		–0.66	0.60	–3.34	2.19	**–8.38**	–3.07	1.16	2.72
7	**8.28**	3.62	–1.32	–2.25	–0.66		–2.28	**–5.61**	–0.27	**–5.78**	–1.68	–2.88	3.15
3	**–4.67**	2.29	1.80	–0.52	0.60	–2.28		0.06	–2.96	–0.30	–2.49	3.02	1.91
10	**–5.11**	–3.39	–3.34	**–4.32**	–3.34	**–5.61**	0.06		–0.93	**31.60**	–3.64	–2.94	5.84
6	–0.65	–1.06	1.29	–0.54	2.19	–0.27	–2.96	–0.93		–0.07	–2.11	–0.80	1.17
8	**–4.82**	**–3.67**	**–12.33**	**–4.46**	**–8.38**	**–5.78**	–0.30	**31.60**	–0.07		–1.66	–3.50	6.96
12	–0.07	–3.34	–3.33	**7.63**	–3.07	–1.68	–2.49	–3.64	–2.11	–1.66		0.08	2.65
11	–1.27	–0.75	**4.38**	–0.50	1.16	–2.88	3.02	–2.94	–0.80	–3.50	0.08		1.93

Note. Statistics in bold are statistically significant at the 5% significance level using the Bonferroni correction. EPQ-R = Eysenck's Personality Questionnaire-Revised.

TABLE 6
Mean and Variance Adjusted $\bar{X}^2_{ij}$ (Above the Diagonal) and R_{ij} (Below the Diagonal) Statistics for the EPQ-R Extraversion Data

Item	1	5	9	2	4	7	3	10	6	8	12	11	Average
1		0.58	3.40	0.30	7.48	**86.74**	**22.40**	**15.82**	0.12	9.67	0.17	1.20	13.44
5	2.79		0.12	10.02	0.43	**22.53**	10.14	4.89	0.31	4.08	7.24	0.10	5.49
9	3.95	2.65		9.24	**41.04**	1.61	3.60	3.12	2.49	**20.82**	**11.35**	**21.11**	10.72
2	0.78	13.71	9.45		1.81	5.12	0.19	9.55	0.08	4.37	**89.02**	0.15	11.81
4	8.17	2.77	**40.87**	1.88		0.24	0.34	3.31	6.25	9.39	9.08	1.42	7.34
7	**87.17**	**23.66**	2.18	5.78	0.79		4.83	**19.88**	−0.03	**14.48**	1.57	7.72	14.98
3	**22.95**	12.06	3.70	0.29	0.38	5.35		0.86	7.96	0.61	5.00	10.26	6.02
10	**21.23**	11.68	7.08	13.25	**134.03**	**25.76**	0.52		0.17	**254.98**	6.17	4.54	29.39
6	1.28	3.16	3.28	1.13	6.89	1.08	9.24	2.72		0.55	3.01	0.36	1.94
8	13.73	10.06	**36.45**	7.12	**18.64**	**19.75**	0.14	**225.20**	0.94		0.23	5.45	29.51
12	1.45	11.52	13.31	**90.16**	11.10	3.20	6.50	11.26	5.21	1.99		0.24	12.10
11	1.74	2.62	**21.32**	0.35	1.56	8.34	10.44	6.77	1.39	7.35	1.52		4.78

Note. $\bar{\bar{X}}^2_{ij}$ has been computed with 1 *df*; statistics in bold are statistically significant at the 5% significance level using the Bonferroni correction. A column has been added to the table that includes the average of the $\bar{\bar{X}}^2_{ij}$ values across all 12 items. EPQ-R = Eysenck's Personality Questionnaire-Revised.

cluster analysis of the squared z_{ij} values, the Ward procedure being used in this case. To control for multiple testing, we use the Bonferroni procedure and we mark in boldface all values z_{ij} that are statistically significant at the 5% level after applying this correction. We see that there is considerable misfit in this application and that some z statistics are extremely large. The two worst fitting items are clearly Items 8 and 10. In particular, the model appears to predict rather poorly the association between these two items. Because the z statistic is positive, we infer that the model underestimates the association between these two items.

In Table 6, we also report the results of the mean and variance adjusted statistics and R_{ij}. To facilitate the presentation of results, $\bar{\bar{X}}^2_{ij}$ statistics computed with one degree of freedom are displayed in Table 6; however, p-values of $\bar{X}^2_{ij}$ are used for significance tests. We have boldfaced the item pairs flagged by statistically significant statistics at $\alpha = 0.05$ after a Bonferroni correction. Table 6 also displays the R_{ij} statistics: The estimated degrees of freedom were 1 for 4 pairs, 2 for 32 pairs, and 3 for 26 pairs. As for consistency among statistics, z_{ij} identifies 20 misfitting pairs, more than those detected by $\bar{X}^2_{ij}$ (12 pairs) or R_{ij} (13 pairs); 11 item pairs are identified by all three statistics. Nevertheless, all three statistics suggest the same conclusion: Items 8 and 10 are fitted poorly by the 2PL model.

In view of the results of this piecewise analysis, one should attempt to find a better fitting model. A bifactor model was fitted next, with four second-tier factors specified for four pairs of items showing large bivariate statistics, that is, (1, 7), (2, 12), (4, 9), and (8, 10); within each pair, the secondary slopes were constrained to be equal for model identification. The resulting model fit was greatly improved: $M_2 = 146.46$, $df = 50$, $p < .01$; a 90% confidence interval for RMSEA_2 is [0.04, 0.06]. On the other hand, we note that simply removing the worst fitting items, that is, Items 8 and 10, did not improve the fit much: $M_2 = 287.00$, $df = 35$, $p < .01$; a 90% confidence interval for RMSEA_2 is [0.08, 0.10].

CONCLUSIONS

In this article, we have examined the empirical behavior of a number of test statistics for identifying misfitting items in IRT modeling. We have focused solely on the graded model (and its special case, the 2PL) because this is the most widely used model in IRT applications and because there is some evidence (Maydeu-Olivares, 2005) that it may be the best fitting parametric model for rating data. Our simulation results suggest that for this model, the z statistic based on the residual cross-product (z_{ord} or z_{ij}) is the method of choice if the observed information matrix is available to estimate the covariance matrix of the item parameter estimates. When only the XPD information matrix is available, we recommend the use of mean and variance adjusted X^2 statistics. Finally, in applications where detecting the presence of guessing or a similar phenomenon is of interest, we recommend the use of the R_{ij} statistic, as it is the most powerful statistic in this case.

Future research should also compare the behavior of the statistics investigated in this article with that of score tests (Glas & Suárez-Falcón, 2003; Liu & Thissen, 2012, in press). In score tests, the source of misfit is derived from nested model comparisons. Specifically, researchers can specify a less restrictive model for a subset of items that has the current model as a special case and yet retain the current model for the remaining items. In this way, the resulting model for all items has the current one nested within, and standard procedures for nested model comparison can then be used to provide evidence of whether or not the alternative model is more appropriate. This approach is particularly helpful if researchers have a reasonable conjecture of the misfit

mechanism. In this article, however, we have focused our attention on the use of marginal subtable statistics, which are more appropriate when researchers have no a priori knowledge regarding model misspecification.

Our piecewise fit testing approach can be easily modified and adapted to other $\sqrt{n}$-consistent point estimators such as various types of weighted least squares (e.g., B. Muthén, 1978), composite likelihood (e.g., Jöreskog & Moustaki, 2001; Katsikatsou, Moustaki, Yang-Wallentin, & Jöreskog, 2012), and Bayesian estimators (e.g., Albert, 1992; Mislevy, 1986). For estimators that are not asymptotically optimal (i.e., with minimum variance), the asymptotic covariance matrix of the bivariate residuals [Equation (5)] should be modified according to Equation (2.5) of Maydeu-Olivares and Joe (2006). The Bayesian expected *a posteriori* estimator (typically obtained via MCMC sampling), on the other hand, is usually asymptotically equivalent to the ML solution (i.e., the Bernstein-von Mises theorem) and thus it is asymptotically optimal; therefore, the procedures described here can be used for goodness-of-fit testing in a fashion identical to what has been discussed in this article.

In short, our results suggest that the source of misfit in graded models can be safely evaluated using the same sample size as that needed to obtain an accurate estimate of the item parameters (Forero & Maydeu-Olivares, 2009). Clearly, additional research is required to evaluate the behavior of the statistics discussed here when used with other IRT models. Indeed, the statistics considered in this article are quite general and can be applied to other models for categorical data, such as latent class models.

FUNDING

This research was supported by an ICREA-Academia Award and Grant SGR 2009 74 from the Catalan Government and Grants PSI2009-07726 and PR2010-0252 from the Spanish Ministry of Education awarded to the second author.

SUPPLEMENTAL MATERIAL

Source code for this article can be accessed on the publisher's website.

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*(3), 251–269.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.

Asparouhov, T., & Muthén, B. O. (2010). *Simple second order chi-square correction scaled chi-square statistics*. Unpublished manuscript retrieved from www.statmodel.com

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*, 290–302.

Cagnone, S., & Mignani, S. (2007). Assessing the goodness of fit of a latent variable model for ordinal data. *Metron: International Journal of Statistics*, *65*, 337–361.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309–329.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited information goodness of fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173–194.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polychotomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*, 143–165.

Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. T. (1985). A revised version of the Psychoticism scale. *Personality and Individual Differences*, *6*, 21–29.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded models for rating data: Limited vs. full information methods. *Psychological Methods*, *14*, 275–299.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*, 525–546.

Glas, C. A. W., & Suárez-Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87–106.

Glas, C. A. W., & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika*, *54*, 635–659.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Their foundations, recent developments and applications* (pp. 69–96). New York, NY: Springer.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, *48*, 419–426.

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*, 393–419.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*, 347–387.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis*, *56*(12), 4243–4258.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, *51*, 357–373.

Liu, Y., & Maydeu-Olivares, A. (2012, July). *The use of quadratic form statistics of residuals to identify IRT model misfit in marginal subtables*. Paper presented at the annual meeting of the Psychometric Society, Lincoln, NE.

Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*, 254–274.

Liu, Y., & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement*, *36*, 670–688.

Liu, Y., & Thissen, D. (in press). Comparing score tests and other local dependence diagnostics for the graded response model. *British Journal of Mathematical and Statistical Psychology.*

Mavridis, D., Moustaki, I., & Knott, M. (2007). Goodness-of-fit measures for latent variable models for binary data. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 135–161). Amsterdam, The Netherlands: Elsevier.

Maydeu-Olivares, A. (2005). Further empirical results on parametric vs. non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research*, *40*, 275–293.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732.

Maydeu-Olivares, A., & Joe, H. (2008). An overview of limited information goodness-of-fit testing in multidimensional contingency tables. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 253–262). Tokyo, Japan: Universal Academy Press.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*, 305–328.

Maydeu-Olivares, A., & Liu, Y. (2012). *Item diagnostics in multivariate discrete data*. Under review.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B*, *42*, 109–142.

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*, 49–57.

Meng, X. L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*(416), 899–909.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177–195.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551–560.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item-fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment*, *18*, 263–283.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika*, *61*, 509–528.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph No. 17.*

Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrical Bulletin*, *2*, 110–114.

Schott, J. R. (1997). *Matrix analysis for statistics*. New York, NY: Wiley.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*, 331–352.

Tay, L., & Drasgow, F. (2011). Adjusting the adjusted X^2/df ratio statistic for dichotomous item response theory analyses: Does the model fit? *Educational and Psychological Measurement*, *72*, 510–528.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*, 77–83.

Toribio, S. G., & Albert, J. H. (2011). Discrepancy measures for item fit analysis in item response theory. *Journal of Statistical Computation and Simulation*, *81*, 1345–1360.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123–139.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.