Measurement, 11: 71–101, 2013

ISSN: 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366367.2013.831680

Routledge
Taylor & Francis Group

FOCUS ARTICLE

Goodness-of-Fit Assessment of Item Response Theory Models

Alberto Maydeu-Olivares
Faculty of Psychology, University of Barcelona

The article provides an overview of goodness-of-fit assessment methods for item response theory (IRT) models. It is now possible to obtain accurate p-values of the overall fit of the model if bivariate information statistics are used. Several alternative approaches are described. As the validity of inferences drawn on the fitted model depends on the magnitude of the misfit, if the model is rejected it is necessary to assess the goodness of approximation. With this aim in mind, a class of root mean squared error of approximation (RMSEA) is described, which makes it possible to test whether the model misfit is below a specific cutoff value. Also, regardless of the outcome of the overall goodness-of-fit assessment, a piece-wise assessment of fit should be performed to detect parts of the model whose fit can be improved. A number of statistics for this purpose are described, including a z statistic for residual means, a mean-and-variance correction to Pearson's X^2 statistic applied to each bivariate subtable separately, and the use of z statistics for residual cross-products.

Keywords: maximum likelihood, latent class, categorical data, discrete data, model selection, ordinal factor analysis, polychoric correlation

Item response theory (IRT) modeling involves fitting a latent variable model to discrete responses obtained from questionnaire/test items intended to measure educational achievement, personality, attitudes, and so on. As in any other modeling endeavor, after an IRT model has been fitted, it is necessary to quantify the discrepancy between the model and the data (i.e., the absolute goodness-of-fit of the model). A goodness-of-fit (GOF) *index* summarizes the discrepancy between the values observed in the data and the values expected under a statistical model. A goodness-of-fit *statistic* is a GOF index with a known sampling distribution. As such, a GOF statistic may be used to test the hypothesis of whether the fitted model could be the data-generating model. This

Correspondence should be addressed to Alberto Maydeu-Olivares, Faculty of Psychology, University of Barcelona, P. Vall d'Hebron, 171, 08035 Barcelona. E-mail: amaydeu@ub.edu

is important, because if we cannot reject this hypothesis, then we can be reasonably confident about the validity of the inferences drawn from our fitted model.

In practice, it is likely that the fitted model be rejected using an overall GOF statistic. Simply, it is not easy to find the data-generating model. In this case, we expect the item parameter estimates to be biased, but we do not know the magnitude or the direction of the bias. Any inference drawn on a poorly fitting model is potentially invalid. The extent to which inferences drawn on poorly fitting models are invalid will depend on a number of factors such as the nature of the inference, the nature of the true data-generating and fitted models, and so on, but clearly it will depend on the degree of misfit between the true and fitted models, that is, on the goodness of approximation of the fitted model.

The goodness of approximation of an IRT model should be regarded as the effect size of its misfit. As such, it is convenient that the goodness of approximation statistic can be interpreted qualitatively, for if a model is rejected, the researcher can judge whether the discrepancy is of substantive interest. Furthermore, detailed studies are needed to investigate the validity of inferences drawn for different degrees of model misspecification. This is important, because IRT applications often involve so many degrees of freedom that it is unlikely that any fitted model be the data-generating model. But, if inferences of interest are shown to be valid for some degree and direction of model misspecification, then testing the exact fit for the model can be replaced by a test of this nonzero degree of model misspecification. These tests of approximate fit are of most interest in IRT applications involving large degrees of freedom.

Unfortunately, these important considerations have been largely absent in the IRT literature. Until recently the assessment of the overall goodness-of-fit of IRT models has not been on the IRT research agenda because of the lack of overall goodness-of-fit statistics with accurate *p*-values in models with more than a few degrees of freedom. With the recent introduction of new overall GOF statistics that make use only of low-order associations among the items, this undesirable situation has begun to change.

It is not that there is a lack of literature on goodness of fit in IRT modeling. Quite the opposite, in fact: the body of literature on the topic is very large. However, much of it focuses on piece-wise assessment of the model (i.e., how well the IRT model fits a particular item or a particular pair of items). Many of the statistics proposed for piece-wise fit assessment of IRT models have unknown sampling distributions and their use relies on heuristics, others appear to be valid only for certain models, and still others appear to be valid only for detecting certain types of misfit. In many ways, IRT piece-wise statistics are analogous to *z* statistics for residual means and covariances in structural equation modeling (SEM). Another useful analogy for IRT piece-wise statistics is the use of Bonferroni-corrected–post hoc *t*-tests in ANOVA. But post hoc tests in ANOVA are only meaningful after a statistically significant *F* statistic. And it is necessary to control for the overall Type I error of *z* statistics in SEM, which is done by using an overall SEM goodness-of-fit statistic. In the same fashion, IRT researchers should use an overall GOF statistic before performing a piece-wise assessment of fit. But a piecewise goodness-of-fit assessment should also be performed in addition to (as opposed to instead of) an overall GOF assessment, regardless of the latter's outcome. This is because a model may fit well overall (i.e., on average), but some parts of the data may be poorly reproduced, suggesting that an alternative model should be used. Also, piece-wise GOF assessment may reveal the source of misfit in poorly fitting models.

The aim of this article is to provide a comprehensive framework for goodness-of-fit assessment in IRT modeling. The account presented here reflects my personal view on the topic. In addition,

I focus on procedures that can be applied, in principle, to any IRT model. In fact, the procedures described here do not make use of any of the specific properties of IRT models and can therefore be applied more generally to any model for multivariate discrete data such as latent class models. Finally, the procedures described here are the result of ongoing research: how to best assess the fit of IRT models remains an open question, and more research in this area is needed.

The article is organized as follows. First, I review the classical statistics for assessing the overall fit of categorical data models and their limitations for IRT model-fit testing. Next, I describe the new limited information goodness of fit statistics that have been proposed in the literature in order to overcome the shortcomings of classical statistics. The third section introduces methods for assessing approximate fit. The fourth section describes methods for piece-wise assessment of fit. The fifth section includes 2 numerical examples. Most of the presentation focuses on methods for maximum likelihood (ML) estimation. but estimation of IRT models from polychoric correlations is also widely used. The sixth section discusses methods for estimators based on polychorics. I conclude with a discussion and some recommendations for applied users.

CLASSICAL GOODNESS-OF-FIT STATISTICS

Consider the responses given by N individuals to n test items, each with K categories coded as 0, 1, . . . , $K - 1$. The resulting data can be gathered in an n-dimensional contingency table with $C = K^n$ cells. Each cell corresponds to one of the C possible response patterns. Within this setting, assessing the goodness of fit of a model involves assessing the discrepancy between the observed proportions and the probabilities expected under the model across all cells of the contingency table. More formally, let π_c be the probability of one such cell and let p_c be the observed proportion, $c = 1, \ldots, C$. Also, let $\boldsymbol{\pi}(\boldsymbol{\theta})$ be the C-dimensional vector of model probabilities expressed as a function of the, say, q model parameters to be estimated from the data. Then, the null hypothesis to be tested is $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ against $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$. For instance, if Samejima's (1969) graded-response model with a single latent trait is fitted to the responses to n rating items each with K response categories, then $\boldsymbol{\theta}$ denotes the $n\,(K-1)$ intercepts and n slopes of the model.

The two best known goodness-of-fit statistics for discrete data are Pearson's statistic $X^2 = N\sum_c (p_c - \hat{\pi}_c)^2/\hat{\pi}_c$, and the likelihood ratio statistic $G^2 = 2N\sum_c p_c \ln(p_c/\hat{\pi}_c)$ where $\hat{\pi}_c = \pi_c\left(\hat{\boldsymbol{\theta}}\right)$. Asymptotic p-values for both statistics can be obtained using a chi-square distribution with $C - q - 1$ degrees of freedom when maximum likelihood estimation is used. However, these asymptotic p-values are only correct when all expected frequencies are large (>5 is the usual rule of thumb). A practical way to evaluate whether the asymptotic p-values for X^2 and G^2 are valid is to compare them. If the p-values are similar, then both are likely to be correct. If they are slightly dissimilar, then X^2 yields the most accurate p-value (Koehler & Larntz, 1980). If they are very different, it is most likely that both p-values are incorrect.

Unfortunately, as the number of possible response patterns increases, the expected frequencies must be small because the sum of all C probabilities must be equal to 1 (Bartholomew & Tzamourani, 1999). As a result, in IRT modeling of the p-values for these statistics cannot normally be used. In fact, when the number of categories is large (say $K > 4$), the asymptotic p-values almost invariably become inaccurate as soon as $n > 5$. To overcome the problem of the inaccuracy

of the asymptotic *p*-values for these statistics, two general methods have been proposed: resampling methods (e.g. bootstrap), and pooling cells. Unfortunately, mixed results have been reported (Tollenaar & Mooijaart, 2003; von Davier, 1997) on the accuracy of *p*-values for the X^2 and G^2 statistics obtained by resampling methods and further research on this topic is needed.

Pooling cells results in statistics whose asymptotic distribution may be well approximated by asymptotic methods, because pooled cells must have larger expected frequencies. However, pooling must be performed before the analysis is made to obtain a statistic with the appropriate asymptotic reference distribution. A straightforward way to pool cells a priori for goodness-of-fit testing is to use low-order margins, that is, univariate, bivariate, and so forth, proportions and probabilities. Goodness-of-fit statistics based on low-order margins are referred to in the literature as limited information statistics because they do not use all the information available in the data for testing the overall goodness-of-fit of the model. Because they are based on pooled cells, the *p*-values of limited information statistics are accurate in very large models even with very small samples (Maydeu-Olivares & Joe, 2005, 2006). Furthermore, because they "concentrate" the information available for testing, they are most often more powerful than full information statistics such as Pearson's X^2 for detecting alternatives of interest (Joe & Maydeu-Olivares, 2010; Reiser, 2008).

OVERALL GOODNESS-OF-FIT TESTING USING LIMITED INFORMATION STATISTICS

To understand what limited information methods are, consider the following 2×3 contingency table:

	$Y_2 = 0$	$Y_2 = 1$	$Y_2 = 2$
$Y_1 = 0$	π_{00}	π_{01}	π_{02}
$Y_1 = 1$	π_{11}	π_{11}	π_{12}

This table can be characterized using the cell probabilities $\boldsymbol{\pi}' = (\pi_{00}, \cdots, \pi_{12})$. Alternatively, it can be characterized using the univariate $\dot{\boldsymbol{\pi}}_1' = \left(\pi_1^{(1)}, \pi_2^{(1)}, \pi_2^{(2)}\right)$ and bivariate $\dot{\boldsymbol{\pi}}_2' = \left(\pi_{1\ 2}^{(1)(1)}, \pi_{1\ 2}^{(1)(2)}\right)$ probabilities, where

	$Y_2 = 0$	$Y_2 = 1$	$Y_2 = 2$	
$Y_1 = 0$				
$Y_1 = 1$		$\pi_{1\ 2}^{(1)(1)}$	$\pi_{1\ 2}^{(1)(2)}$	$\pi_1^{(1)}$
		$\pi_2^{(1)}$	$\pi_2^{(2)}$	

and $\pi_2^{(2)} = \Pr(Y_2 = 2)$ and $\pi_{1\,2}^{(1)(2)} = \Pr(Y_1 = 1, Y_2 = 2)$. The 2 characterizations are equivalent. We refer to the representation using $\boldsymbol{\pi}_2{}' = \left(\dot{\boldsymbol{\pi}}_1', \dot{\boldsymbol{\pi}}_2'\right)$ as moment representation. The elements of $\dot{\boldsymbol{\pi}}_1$ and $\dot{\boldsymbol{\pi}}_2$ are clearly univariate and bivariate moments if the variables are binary, for $\Pr(Y = 1) = E(Y)$ and $\Pr\left(Y_i = 1, Y_j = 1\right) = E\left(Y_i Y_j\right)$. They are also moments when the items are polytomous; in this case they are moments of indicator variables used to denote each category except the zero category (Maydeu-Olivares & Joe, 2006). When all variables consist of the same number of categories, K, there are $n(K-1)$ univariate moments $\dot{\boldsymbol{\pi}}_1$ and $\frac{n(n-1)}{2}(K-1)^2$ bivariate moments $\dot{\boldsymbol{\pi}}_2$. The equivalence between the C probabilities $\boldsymbol{\pi}$ and the $C - 1 = \sum_{i=1}^{n} \binom{n}{i}(K-1)^n$ moments $\boldsymbol{\pi}_n{}' = \left(\dot{\boldsymbol{\pi}}_1', \dot{\boldsymbol{\pi}}_2', \cdots, \dot{\boldsymbol{\pi}}_n'\right)$, exemplified in the above 2×3 contingency table, extends to contingency tables of any dimension.

Limited information test statistics simply disregard some of the higher order moments. Thus, in the above 2×3 example, a statistic that only uses the univariate moments is a limited information statistic. In contrast, full information statistics use all moments (up to order n or $\boldsymbol{\pi}_n$) for testing. Pearson's X^2 statistic is a full information statistic, and it can therefore be written as a function of the cell probabilities or as a function of the moments up to order n. In matrix form, X^2 can be written as a function of the cell proportions and probabilities as

$$X^2 = N\left(\mathbf{p} - \hat{\boldsymbol{\pi}}\right)' \hat{\mathbf{D}}^{-1} \left(\mathbf{p} - \hat{\boldsymbol{\pi}}\right), \tag{1}$$

where $\mathbf{p} - \hat{\boldsymbol{\pi}}$ are the cell residuals, and $\hat{\mathbf{D}} = diag\left(\boldsymbol{\pi}\left(\hat{\boldsymbol{\theta}}\right)\right)$ is a diagonal matrix of estimated cell probabilities. On the other hand, regardless of the number of variables and categories, Pearson's X^2 statistic as a function of the sample and expected moments is

$$X^2 = N\left(\mathbf{p}_n - \hat{\boldsymbol{\pi}}_n\right)' \hat{\Xi}_n^{-1} \left(\mathbf{p}_n - \hat{\boldsymbol{\pi}}_n\right), \tag{2}$$

where $\mathbf{p}_n - \hat{\boldsymbol{\pi}}_n$ are the residual moments, and $N\hat{\Xi}_n$ is the asymptotic covariance matrix of the sample moments up to order n, $\mathbf{p}_n$, evaluated at the parameter estimates.

In limited information test statistics only moments up to order $r < n$ are used for testing. Most often, $r = 2$ (or only univariate and bivariate moments are used for testing), but sometimes $r = 3$ needs to be used. For instance, a statistic analogous to Pearson's X^2 statistic but that only involves univariate and bivariate moments is

$$L_2 = N\left(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2\right)' \hat{\Xi}_2^{-1} \left(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2\right). \tag{3}$$

The set of univariate and bivariate moments $\mathbf{p}_2$ is one possible set of statistics that summarizes the information contained in the margins of the contingency table. But other choices of summary statistics could be used instead. Also, given a choice of summary statistics (e.g. $\mathbf{p}_2$) one can construct different test statistics. Finally, the asymptotic distribution of the test statistic will depend as well on how the item parameters have been estimated. I now discuss each of these three topics in turn.

Choice of Summary Statistic

Consider testing a model using the information contained in the bivariate margins. In this case, one can use as summary statistics the set of univariate and bivariate moments $\mathbf{p}_2$ or the set of all bivariate proportions, say, $\hat{\mathbf{p}}_2$, with population counterpart $\hat{\boldsymbol{\pi}}_2$. For instance, we could consider a quadratic form analogous to (3) using the residuals $\hat{\mathbf{p}}_2 - \hat{\boldsymbol{\pi}}$, instead of the residuals $\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2$. Because there are $\frac{n(n-1)}{2}$ bivariate tables and each table is of dimension K^2, $\hat{\boldsymbol{\pi}}_2$ is of dimension $\frac{n(n-1)}{2}K^2$. However, because the probabilities in each bivariate table must add up to 1, there are only $n(K-1) + \frac{n(n-1)}{2}(K-1)^2 + 1$ mathematically independent bivariate probabilities in $\widehat{\boldsymbol{\pi}}_2$. In contrast, $\boldsymbol{\pi}_2$ is of length $\sum_{i=1}^{2} \binom{n}{i}(K-1)^n = n(K-1) + \frac{n(n-1)}{2}(K-1)^2$. It is generally preferable to use the set of moments (e.g. $\boldsymbol{\pi}_2$) instead of the full set of marginal probabilities (e.g. $\widehat{\boldsymbol{\pi}}_2$) because the former leads to smaller matrices and vectors, and, most importantly, there are no redundancies among its elements. Thus, the asymptotic covariance matrix of $\mathbf{p}_2$ is of full rank; whereas, there are $n(n-2)(K-1) + \frac{n(n-1)}{2} + 1$ zero eigenvalues in the asymptotic covariance matrix of the full set of bivariate proportions $\widehat{\mathbf{p}}$. One can see the elements of $\boldsymbol{\pi}_2$ as a way to obtain the mathematically independent elements in $\widehat{\boldsymbol{\pi}}_2$.

Another choice to be made involves the order of margins used for testing. For IRT applications, Maydeu-Olivares and Joe (2005, 2006) suggested testing using the smallest possible order (i.e. the smallest possible r). This is because the lower the order of moments used, the more accurate the p-values and (generally) the higher the power. In this article I focus on the logistic graded-response model (Samejima, 1969) and its special cases, the 2-parameter and 1-parameter logistic (2PL and 1PL) models, as these are the most widely used models in applications. These models can be identified (i.e. estimated) using only bivariate information, and so they can be tested using only this information ($r = 2$).

Choice of Estimation Method

In general, the asymptotic distribution of a test statistic varies according to the choice of the estimation method for the item parameters. For instance, Pearson's X^2 follows a chi-square distribution when item parameters have been estimated using an asymptotically optimal (i.e., minimum variance) full information estimator, such as the ML estimator, but not for other estimators. For instance, it does not follow a chi-square distribution for item parameters estimated using bivariate (aka pairwise) composite likelihood methods (BCL: Katsikatsou, Moustaki, Yang-Wallentin, & Jöreskog, 2012; Maydeu-Olivares & Joe, 2006). The variable X^2 does not follow a chi-square distribution either when the item parameters have been estimated using tetrachoric/polychoric correlations (Jöreskog, 1994; Muthén, 1978, 1984, 1993). In contrast, the statistic L_2 given in (3) does not follow a chi-square distribution for the ML estimator but I conjecture that it does for the BCL estimator. This is just a hypothesis at this stage, which remains to be investigated.

To simplify the exposition, in this article I focus on the ML estimator. This estimator is also referred to in the IRT literature as marginal ML estimator (MML: Bock & Aitkin, 1981). To simplify the exposition further, I assume that all items consist of the same number of response alternatives, K. At the end of the article I provide results for another widely used class of estimators, those based on polychorics.

For the ML estimator, the asymptotic distribution of the univariate and bivariate residual moments $\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2$ is asymptotically normal with mean zero and covariance matrix

$$\boldsymbol{\Sigma}_2 = \Xi_2 - \boldsymbol{\Delta}_2 \mathcal{I}^{-1} \boldsymbol{\Delta}_2{}'. \tag{4}$$

In (4), $\boldsymbol{\Delta}_2 = \frac{\partial \boldsymbol{\pi}_2(\theta)}{\partial \theta'}$ denotes the matrix of derivatives of the univariate and bivariate moments with respect to the parameter vector $\boldsymbol{\theta}$, and $N\Xi_2$ denotes the asymptotic covariance matrix of the univariate and bivariate sample moments $\mathbf{p}_2$. These matrices are evaluated at the parameter estimates, $\hat{\boldsymbol{\theta}}$.

Also, in (4) $\mathcal{I}^{-1}$ divided by sample size is the asymptotic covariance matrix of the item parameter estimates $\hat{\boldsymbol{\theta}}$, and $\mathcal{I}$ denotes the information matrix. The three best known approaches to estimate the information matrix are the expected information matrix, the observed information matrix, and the cross-product information matrix. The expected information matrix is

$$\mathcal{I}_E = \boldsymbol{\Delta}' \operatorname{diag}(\boldsymbol{\pi})\, \boldsymbol{\Delta}, \tag{5}$$

where $\boldsymbol{\Delta} = \frac{\partial \boldsymbol{\pi}(\theta)}{\partial \theta'}$ is a $C \times q$ matrix. Because $C = K^n$, when the items are binary the expected information matrix can only be computed when the number of items is 19 or so. If the response alternatives are 5, this matrix can only be computed with 8 items or so. In contrast, the observed and cross-product information matrices only involve the observed patterns, which are necessarily fewer in number than the number of observations. As a result, either the observed or the cross-product information matrices have to be used in most actual applications. The cross-products information matrix is

$$\mathcal{I}_{XP} = \boldsymbol{\Delta}_O' \operatorname{diag}\left(\mathbf{p}_O / \boldsymbol{\pi}_O^2\right) \boldsymbol{\Delta}_O, \tag{6}$$

where $\mathbf{p}_O$ and $\boldsymbol{\pi}_O$ denote the proportions and probabilities of the C_O observed patterns, and $\boldsymbol{\Delta}_O$ is a $C_O \times q$ matrix of derivatives of the observed patterns with respect to the full set of q item parameters.

The observed information matrix cannot be written easily in matrix form. In scalar form it can be written as

$$\mathcal{I}_O = N \sum_{c=1}^{C_O} \frac{p_c}{(\pi_c(\boldsymbol{\theta}))^2} \left[\frac{\partial \pi_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \pi_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} - \pi_c(\boldsymbol{\theta}) \frac{\partial^2 \pi_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] = \mathcal{I}_{XP} - N \sum_{c=1}^{C_O} \frac{p_c}{\pi_c(\boldsymbol{\theta})} \frac{\partial^2 \pi_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}. \tag{7}$$

Choice of Test Statistic

Assume the model's fit is to be assessed using the vector of univariate and bivariate residual moments $\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2$ and recall that this is simply the vector of univariate and bivariate residual proportions that do not include the lowest category (category zero). To test the overall goodness-of-fit of an IRT model using only bivariate information, we can construct a quadratic form statistic

$$Q = N\left(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2\right)' \hat{\mathbf{W}} \left(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2\right), \tag{8}$$

where $\hat{\mathbf{W}}$ is some real symmetric weight matrix that may depend on the model parameters but converges in probability to some constant matrix: $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$. In general, the asymptotic distribution of Q is a mixture of independent chi-square variates. However, when $\hat{\mathbf{W}}$ is chosen so that

$$\boldsymbol{\Sigma}_2\mathbf{W}\boldsymbol{\Sigma}_2\mathbf{W}\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_2\mathbf{W}\boldsymbol{\Sigma}_2, \tag{9}$$

then (8) is asymptotically distributed as a chi-square with degrees of freedom equal to the rank of $\mathbf{W}\boldsymbol{\Sigma}_2$. There are 2 ways to choose $\hat{\mathbf{W}}$ so that (9) is satisfied.

One choice involves using a weight matrix such that $\boldsymbol{\Sigma}_2$ is a generalized inverse of $\mathbf{W}$; that is, $\mathbf{W}$ satisfies $\boldsymbol{\Sigma}_2\mathbf{W}\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_2$. This is the approach taken by Maydeu-Olivares and Joe (2005, 2006), who proposed using the statistic

$$M_2 = N\left(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2\right)'\hat{\mathbf{C}}_2\left(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2\right),\ C_2 = \Xi_2^{-1} - \Xi_2^{-1}\boldsymbol{\Delta}_2(\boldsymbol{\Delta}_2'\Xi_2^{-1}\boldsymbol{\Delta}_2)^{-1}\boldsymbol{\Delta}_2'\Xi_2^{-1} \tag{10}$$

to assess the overall goodness-of-fit of IRT models. M_2 is asymptotically chi-square equal to the number of univariate and bivariate moments minus the number of estimated parameters, i.e.,

$$df_2 = n(K-1) + \frac{n(n-1)}{2}(K-1)^2 - q. \tag{11}$$

This can be readily verified by noting that $\mathbf{C}_2$ in (10) can be alternatively written as $\mathrm{C}_2 = \boldsymbol{\Delta}_2^{(c)}\left(\boldsymbol{\Delta}_2^{(c)'}\Xi_c\boldsymbol{\Delta}_2^{(c)}\right)^{-1}\boldsymbol{\Delta}_2^{(c)\prime}$, where $\boldsymbol{\Delta}_2^{(c)\prime}$ is an orthogonal complement to $\boldsymbol{\Delta}_2^{(c)\prime}$, that is $\boldsymbol{\Delta}_2^{(c)\prime}\boldsymbol{\Delta}_2 = \mathbf{0}$.

Another way to satisfy (9) is to use a weight matrix such that $\mathbf{W}$ is a generalized inverse of $\boldsymbol{\Sigma}_2$; that is, $\mathbf{W}$ satisfies $\mathbf{W}\boldsymbol{\Sigma}_2\mathbf{W} = \mathbf{W}$. This approach leads to the choice $\hat{\mathbf{W}} = \hat{\boldsymbol{\Sigma}}_2^{+}$ and the statistic

$$R_2 = N\left(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2\right)'\hat{\Sigma}_2^{+}\left(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2\right). \tag{12}$$

For binary data, this statistic was proposed by Reiser (1996, 2008).

M_2 has a computational advantage over R_2 in that it does not require the computation of the asymptotic covariance matrix of the item parameter estimates. Thus, for ML estimation, the information matrix need not be computed; only its diagonal elements are needed to obtain the standard errors of the parameter estimates. Also, a single implementation suits all estimators, since M_2 follows the above chi-square distribution for any consistent estimator.

In contrast, the degrees of freedom of R_2 equal rank $\left(\boldsymbol{\Sigma}_2^{+}\boldsymbol{\Sigma}_2\right)$ = rank $(\boldsymbol{\Sigma}_2)$. However, the rank of $\boldsymbol{\Sigma}_2$ is unknown and may depend on the parameter values (Reiser, 1996). As a result, currently, in applications the degrees of freedom involved when using R_2 must be estimated by determining the rank of $\hat{\boldsymbol{\Sigma}}_2$, for example, using an eigen decomposition. Hence, the p-value of R_2 will depend on how many eigenvalues are numerically judged to be zero. This is tricky in IRT applications, as numerical integration is involved, and, as a result, it may be difficult to judge whether an eigenvalue is zero; for an illustration of this point see Maydeu-Olivares and Joe (2008).

An alternative way to obtain an overall goodness-of-fit statistic is to use as weight matrix in the quadratic form Q given in equation (8) a matrix that is easily computed. In this case, the resulting statistic will asymptotically follow a mixture of chi-square distributions. A p-value in this case can be obtained using the inversion formula given in Imhof (1961). Another way to

obtain a p-value in this case is by adjusting the test statistic by its asymptotic mean-and-variance so that the asymptotic distribution of the adjusted test statistic can be approximated by a chi-square distribution. This approach dates back to Satterthwaite (1946) and was made popular in the structural equation modeling literature by Satorra and Bentler (1994). Imhof's inversion method is slightly more involved computationally but does not yield a more accurate p-value (Liu & Maydeu-Olivares, 2012a). As a result, here I focus on mean-and-variance corrections.

Two obvious choices of weight matrix in (8) that do not lead to a chi-square distributed statistic are $\hat{\mathbf{W}} = \Xi_2^{-1}$ and $\hat{\mathbf{W}} = (\text{diag}\,(\Xi_2))^{-1}$. The first leads to the statistic L_2 in (3). L_2 was proposed by Maydeu-Olivares and Joe (2005, 2006) for testing simple null hypotheses (i.e., known parameter values), in which case it follows an asymptotic chi-square distribution. The latter leads to the statistic

$$Y_2 = N\left(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2\right)'(\text{diag}\,(\Xi_2))^{-1}\left(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2\right), \tag{13}$$

introduced by Bartholomew and Leung (2002) for testing IRT models for binary data. The mean-and-variance adjustment for Bartholomew and Leung's (2002) statistic for ML item parameter estimates was given by Cai, Maydeu-Olivares, Coffman, and Thissen (2006).

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

$$\mu_1 = \text{tr}\,(\mathbf{W}\boldsymbol{\Sigma}_2), \quad \mu_2 = 2\text{tr}(\mathbf{W}\boldsymbol{\Sigma}_2)^2. \tag{14}$$

Solving for the two unknown constants a and b and evaluating μ_1 and μ_2 at the parameter estimates, we obtain the mean-and-variance corrected Q statistic

$$\overline{Q} = \frac{Q}{b} = \frac{\hat{\mu}_2}{2\hat{\mu}_1}Q, \tag{15}$$

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

$$a = \frac{2\hat{\mu}_1^2}{\hat{\mu}_2}. \tag{16}$$

This is the approach used by Cai et al. (2006) to approximate the asymptotic distribution of Y_2 in binary IRT models. However, following Asparouhov and Muthén (2010), it is possible to define an alternative mean-and-variance corrected $\overline{\overline{Q}}$ that, unlike (15), has the same degrees of freedom as M_2. Their method entails writing the statistic as $\overline{\overline{Q}} = a^* + b^*Q$, where a^* and b^* are chosen so that the mean and variance of $\overline{\overline{Q}}$ are df_2 and 2 df_2, respectively. Solving for the 2 unknown constants a^* and b^* we obtain

$$\overline{\overline{Q}} = Q\sqrt{\frac{2df_2}{\hat{\mu}_2}} + df_2 - \sqrt{\frac{2df_2\hat{\mu}_1^2}{\hat{\mu}_2}}. \tag{17}$$

Our simulation results, consistent with the results of Asparouhov and Muthén (2010), suggest that there is a negligible difference in the p-values obtained using (15) and (17). Therefore, we make use of $\overline{\overline{L}}_2$ and $\overline{\overline{Y}}_2$ as they are more intuitive: degrees of freedom are the familiar formula equal to the number of statistics in $\mathbf{p}_2$ minus the number of item parameter estimates. Note that mean-and-variance corrected statistics, as well as the use of Imhof's inversion method, require the computation of an estimate of $\boldsymbol{\Sigma}_2$, the asymptotic covariance matrix of the bivariate residual moments. Therefore, for a computational viewpoint, M_2 is also preferable to the use of quadratic forms, which are asymptotically distributed as a mixture of chi-square variables.

Testing Models for Large and Sparse Ordinal Data

When the number of categories per item is large, M_2 may not be computable for large numbers of items due to the size of the matrices that need to be stored in memory. Using theory from Joe and Maydeu-Olivares (2010), a statistic analogous to M_2 may be computed using as summary statistics residual means and cross-products instead of the residuals $\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2$. The population means and cross-products are

$$\kappa_i = E[Y_i] = 0 \times \Pr(Y_i = 0) + \ldots + (K_i - 1) \times \Pr(Y_i = K_i - 1), \tag{18}$$

$$\begin{aligned}\kappa_{ij} = E\left[Y_i Y_j\right] = 0 \times 0 \times \Pr\left(Y_i = 0, Y_j = 0\right) + \ldots + (K_i - 1) \times (K_j - 1) \\ \times \Pr\left(Y_i = K_i - 1, Y_j = K_j - 1\right),\end{aligned} \tag{19}$$

with sample counterparts $k_i = \bar{y}_i$ (the sample mean) and $k_{ij} = \mathbf{y}_i'\mathbf{y}_j/N$ (the sample cross-product), respectively. In particular, for our previous 2×3 example, the elements of $\boldsymbol{\kappa}$ are

$$\begin{aligned}\kappa_1 &= E[Y_1] = 1 \Pr(Y_1 = 1) = \pi_1^{(1)} \\ \kappa_2 &= E[Y_2] = 1 \Pr(Y_2 = 1) + 2 \Pr(Y_2 = 1) = \pi_2^{(1)} + 2\pi_2^{(2)} \\ \kappa_{12} &= E[Y_1 Y_2] = 1 \times 1 \Pr(Y_1 = 1, Y_2 = 1) + 1 \times 2 \Pr(Y_1 = 1, Y_2 = 2) = \pi_{1\ 2}^{(1)(1)} + 2\pi_{1\ 2}^{(1)(2)}.\end{aligned} \tag{20}$$

Using these statistics we can construct the goodness-of-fit statistic

$$M_{ord} = N\left(\mathbf{k} - \hat{\kappa}\right)' \hat{\mathbf{C}}_{ord} \left(\mathbf{k} - \hat{\kappa}\right), \; \mathrm{C}_{ord} = \Xi_{ord}^{-1} - \Xi_{ord}^{-1}\boldsymbol{\Delta}_{ord}(\boldsymbol{\Delta}_r'\Xi_{ord}^{-1}\boldsymbol{\Delta}_{ord})^{-1}\boldsymbol{\Delta}_{ord}'\Xi_{ord}^{-1}, \tag{21}$$

where now $N\Xi_{ord}$ is the asymptotic covariance matrix of the sample means and cross-products $\mathbf{k}$, $\boldsymbol{\Delta}_{ord}$ is the matrix of derivatives of the population means and cross-products $\boldsymbol{\kappa}$ with respect to the model parameters, and $\mathbf{C}_{ord}$ is evaluated at the parameter estimates.

Notice that M_{ord} has the same form as M_2. However, M_{ord} uses fewer statistics than M_2, and the statistics used in the former are a linear combination of the statistics used in the latter. Thus, for our 2×3 example, M_2 is a quadratic form involving $\boldsymbol{\pi}_2' = \left(\pi_1^{(1)}, \pi_2^{(1)}, \pi_2^{(2)}, \pi_{1\ 2}^{(1)(1)}, \pi_{1\ 2}^{(1)(2)}\right)$, whereas M_{ord} is a quadratic form involving $\boldsymbol{\kappa}$ given in (20). Clearly, $\boldsymbol{\kappa}$ is obtained as a linear

combination of $\boldsymbol{\pi}_2$, where the weights used correspond to the coding of the categories. However, it only makes sense to use means and cross-products, and, therefore M_{ord}, when the item categories are ordered (hence its name). Furthermore, when the data are binary, M_{ord} equals M_2.

M_{ord} is asymptotically distributed as a chi-square with $df_{ord} = \frac{n(n+1)}{2} - q$ degrees of freedom. This means that M_{ord} cannot be used when the number of categories is large and the number of items is small due to the lack of degrees of freedom for testing. For instance, for a graded-response model with a single latent trait, the number of items must be larger than the number of categories times 2 (i.e., $n \geq K \times 2$) for the degrees of freedom M_{ord} to be positive. Thus, for ordinal data, if the model involves a large number of variables and categories one must resort to M_{ord}, as M_2 cannot be computed. On the other hand, when the number of categories is large and the number of items is small, M_{ord} cannot be computed due to a lack of degrees of freedom. In some medium-size models for ordinal data, there is a choice between M_2 and M_{ord}. Because $\boldsymbol{\kappa}$ concentrates the information available in $\boldsymbol{\pi}_2$, M_{ord} may be more powerful than M_2 along most alternatives of interest (Joe & Maydeu-Olivares, 2010) and Cai and Hansen (2013) report simulations showing that this is the case. Note that Cai and Hansen refer to M_{ord} as M_2^*. On the other hand, if the concentration of the information is not along the alternative of interest, M_2 will be more powerful than M_{ord} along that direction. Similarly, $\boldsymbol{\pi}_2$ concentrates the information available in $\boldsymbol{\pi}$, the cell probabilities, and M_2 may be more powerful than X^2 along most alternatives of interest, but less powerful than X^2 if it does not concentrate the information along the alternative of interest. For instance, M_2 will be less powerful than X^2 if the misfit only appears in 3-way and higher associations.

TESTING FOR APPROXIMATE FIT

In many IRT applications, degrees of freedom are so large that it is unrealistic to expect that any model will fit the data. In other words, it is unrealistic to expect that the fitted IRT model is the data-generating mechanism. Therefore, it is more reasonable to test whether the model fits approximately rather than testing whether it fits exactly. By this we simply mean testing whether some statistic is smaller than some cutoff. Drawing on work in the structural equations modeling literature by Browne and Cudeck (1993), Maydeu-Olivares and Joe (in press) have recently proposed a family of population discrepancy parameters between the fitted model and the true and unknown data-generating model

$$F_r = \left(\boldsymbol{\pi}_r^T - \boldsymbol{\pi}_r^0\right)' \mathbf{C}_r^0 \left(\boldsymbol{\pi}_r^T - \boldsymbol{\pi}_r^0\right), \tag{22}$$

with $\boldsymbol{\pi}_r^0$ being the moments up to order r under the fitted (i.e. null) model, $\boldsymbol{\pi}_r^T$ under the true model and $\mathbf{C}_r^0$ being

$$C_r = \Xi_r^{-1} - \Xi_r^{-1}\boldsymbol{\Delta}_r(\boldsymbol{\Delta}_r'\Xi_r^{-1}\boldsymbol{\Delta}_r)^{-1}\boldsymbol{\Delta}_r'\Xi_r^{-1} = \boldsymbol{\Delta}_r^{(c)}(\boldsymbol{\Delta}_r^{(c)'}\Xi_r\boldsymbol{\Delta}_r^{(c)})^{-1}\boldsymbol{\Delta}_r^{(c)'}, \tag{23}$$

based on the fitted (null) model. In this family of population discrepancies, F_1 is the population discrepancy between the univariate moments under the true and null models, F_2 is the population discrepancy between the univariate and bivariate moments under the true and null models, and so

forth up to F_n, a population discrepancy involving all moments, which can be rewritten using the cell probabilities as

$$F_n = (\boldsymbol{\pi}_T - \boldsymbol{\pi}_0)' \mathbf{D}_0^{-1} (\boldsymbol{\pi}_T - \boldsymbol{\pi}_0) \tag{24}$$

To take into account model parsimony, a Root Mean Square Error of approximation (RMSEA; Steiger & Lind, 1980) can be constructed for each of the members of the family of population discrepancies (24) leading to a family of root mean square error of approximation RMSEA$_r$s, given by

$$\varepsilon_r = \sqrt{\frac{F_r}{df_r}} \tag{25}$$

where $df_r = s_r - q$ denotes the degrees of freedom available for testing when only up to r^{th}-way moments are used.

Maydeu-Olivares and Joe (in press) also show that under a sequence of local alternatives, an asymptotically unbiased estimate of the RMSEA$_r$ is

$$\hat{\varepsilon}_r = \sqrt{\text{Max}\left(\frac{M_r - df_r}{N \times df_r}, 0\right)}, \tag{26}$$

where the M_r statistics are of the form (10) involving moments up to order $r = 1, 2, \ldots, n$, and $df_r = \left(\sum_{i=1}^{r} \binom{n}{i} (K-1)^i\right) - q$. Since for ML estimation $M_n = X^2$ (Maydeu-Olivares & Joe, 2005), the full information RMSEA (i.e. RMSEA$_n$) can be estimated as $\hat{\varepsilon}_n = \sqrt{\text{Max}\left(\frac{X^2 - df}{N \times df}, 0\right)}$, with $df = C - q - 1$.

A 90% confidence interval for ε_r is given by

$$\left(\sqrt{\frac{\hat{L}_r}{N \times df_r}}; \sqrt{\frac{\hat{U}_r}{N \times df_r}}\right), \tag{27}$$

with $\hat{L}_r$ and $\hat{U}_r$ being the solution to

$$\Pr\left(\chi^2_{df_r}\left(\hat{L}_r\right) \leq M_r\right) = 0.95, \text{ and } \Pr\left(\chi^2_{df_r}\left(\hat{U}_r\right) \leq M_r\right) = 0.05, \tag{28}$$

respectively.

Finally, researchers may be interested in performing a test of close fit of the type

$$H_0 : \varepsilon_r \leq c_r \text{ vs. } H_1 : \varepsilon_r > c_r, \tag{29}$$

where c_r is an arbitrary cutoff value that depends on r, the highest level of association used. P-values for (29) are obtained using

$$p = 1 - \Pr\left(\chi^2_{df_r}\left(N \times df_r \times c_r^2\right) \leq M_r\right). \tag{30}$$

Which member of this family should be used? Maydeu-Olivares and Joe (in press) argue that the RMSEA that should be used is the one whose sampling distribution can be best approximated in small samples. This leads to using the smallest r at which the model is identified, generally $r = 2$. That is, they recommend using RMSEA_2 in applications. This RMSEA can be estimated from the M_2 statistic using equation (26).

Using simulations, they actually show that the distribution of the sample RMSEA_2 can be well approximated in small samples even for large models, whereas the distribution of the full information sample RMSEA_n can only be well approximated in small models. Furthermore, using the unidimensional graded-response model and its special cases, the 2PL and 1PL models as null (fitted) models, they illustrate the relationship between the population RMSEA_2 and RMSEA_n showing that for all the true models they investigated (which included conditions of multidimensionality and lower asymptote parameters) $\text{RMSEA}_2 > \text{RMSEA}_n$. This is simply a reflection of M_2 being generally more powerful than X^2. Importantly, it implies that if the same cutoff point is used in a RMSEA test of close fit (30), it is harder to retain a model when only bivariate information is used than when full information is used.

Maydeu-Olivares and Joe (in press) then explore what cutoff should be used when testing for close fit using the RMSEA_2. Interestingly, they show that regardless of the number of variables being modeled, a cutoff of $\varepsilon_2 \leq 0.05$ separates quite well misspecified IRT models with correctly specified latent trait dimensionality from IRT models with misspecified latent trait dimensionality. They argue that for IRT modeling, a correctly specified latent trait dimensionality is the most important consideration, and consequently, they suggest that a cutoff of $\text{RMSEA}_2 \leq 0.05$ indicates adequate fit. They also show that the population RMSEA_2 is strongly affected by the number of categories: the larger the number of categories, the smaller the value of the population RMSEA_2. They also show that dividing the RMSEA_2 by the number of categories minus 1, one obtains an RMSEA_2 relatively unaffected by the number of categories. Consequently, they suggest using $\varepsilon_2 \leq 0.05/(K-1)$ as a cutoff for excellent fit.

When the size of the model is so large that M_2 cannot be computed, Maydeu-Olivares and Joe (in press) suggest using

$$F_{ord} = \left(\boldsymbol{\kappa}^T - \boldsymbol{\kappa}^0\right)' \mathbf{C}^0_{ord}\left(\boldsymbol{\kappa}^T - \boldsymbol{\kappa}^0\right) \tag{31}$$

to assess the discrepancy between the true and fitted models. In (31) $\boldsymbol{\kappa}^0$ and $\boldsymbol{\kappa}^T$ denote the population means and cross-products (18) and (19) under the fitted (i.e. null) and true models, respectively, and $\mathbf{C}^0_{ord}$ as given in equation (21) based on the fitted (null) model. An RMSEA can be constructed using the parameter

$$\varepsilon_{ord} = \sqrt{\frac{F_{ord}}{df_{ord}}}, \tag{32}$$

with an asymptotically unbiased estimate

$$\hat{\varepsilon}_{ord} = \sqrt{\text{Max}\left(\frac{M_{ord} - df_{ord}}{N \times df_{ord}}, 0\right)}. \tag{33}$$

However, if M_{ord} is more powerful than M_2, then the RMSEA_{ord} must be larger than the RMSEA_2, as the RMSEAs are a function of the estimated noncentrality parameters. Thus, a larger cutoff must be used for RMSEA_{ord} than for RMSEA_2. Most importantly, the population RMSEA_{ord} depends on the number of variables: the larger the number of variables, the smaller the population RMSEA_{ord}, all other factors being constant.

To assess the approximate fit of large models for ordinal data while overcoming the dependence of RMSEA_{ord} on the number of items and, hence, the difficulty of establishing a cutoff criterion, I recommend the use of the Standardized Root Mean Square Residual (SRMSR) borrowed from the factor analysis literature. For a pair of items i and j, the standardized residual is defined as the sample (product-moment or Pearson) correlation minus the expected correlation. In turn, the expected correlation simply equals the expected covariance divided by the expected standard deviations:

$$r_{ij} - \hat{\rho}_{ij} = r_{ij} - \frac{\hat{\kappa}_{ij} - \hat{\kappa}_i\hat{\kappa}_i}{\sqrt{\hat{\kappa}_{ii} - \hat{\kappa}_i^2}\sqrt{\hat{\kappa}_{jj} - \hat{\kappa}_j^2}}, \tag{34}$$

where the means (κ_i and κ_j) and the cross-product κ_{ij} were given in (18) and (19), and κ_{ii} is

$$\kappa_{ii} = E\left[Y_i^2\right] = 0^2 \times \Pr\left(Y_i = 0\right) + \ldots + \left(K_i - 1\right)^2 \times \Pr\left(Y_i = K_i - 1\right). \tag{35}$$

The SRMSR is simply the square root of the average of these squared residual correlations

$$SRMSR = \sqrt{\sum_{i<j} \frac{\left(r_{ij} - \hat{\rho}_{ij}\right)^2}{n(n-1)/2}}. \tag{36}$$

Being an average of standardized residuals, the SRMSR is not affected by the number of items, all other factors being held constant. In addition, the interpretation of the SRMSR is straightforward and intuitive. In contrast, the RMSEAs cannot be readily interpreted. An advantage of the RMSEAs over the SRMSR is that it is straightforward to compute confidence intervals for them, and to perform hypothesis testing, since they are simply transformations of the M_2 and M_{ord} statistics, respectively, which are chi-square distributed when the fitted model is correctly specified (Maydeu-Olivares & Joe, in press). In contrast, the computation of confidence intervals for the SRMSR, and hypothesis testing, is cumbersome as the asymptotic distribution of the SRMSR is a mixture of independent chi-squares when the model is correctly specified. Thus, the SRMSR is best used as a goodness-of-fit index. Any substantively motivated cutoff may be used with the SRMSR, as its interpretation is straightforward. I feel that residual correlations smaller than 0.05 indicate a substantively negligible amount of misfit, and therefore I suggest using SRMSR $\leq$ 0.05 as a cutoff for well-fitting IRT models for ordinal data. Of course, the

SRMSR is just a summary measure, and therefore it is good practice to report the largest standardized residuals (36) in the model in addition to the SRMSR or to provide the matrix of residuals if the model does not involve too many items. A disadvantage of the SRMSR compared with the RMSEA is that it does not take model complexity into account. However, model complexity is only of interest when different models are fitted to a data set, for instance models with a different number of latent traits.

PIECE-WISE ASSESSMENT OF FIT

After examining the overall fit of a model, it is necessary to perform a piece-wise goodness-of-fit assessment. If the overall fit is poor, a piece-wise fit assessment may suggest how to modify the model. Even if the model fits well overall, a piece-wise goodness-of-fit assessment may reveal parts of the model that misfit.

A useful starting point for our discussion of piece-wise fit assessment is the bivariate Pearson's X^2 statistic. After the IRT model parameters have been estimated using the full data, a X^2 statistic may be computed for each bivariate subtable. In this case, it is convenient to write:

$$X_{ij}^2 = N\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right)' \hat{\mathbf{D}}_{ij}^{-1} \left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right). \tag{37}$$

This is just the standard X^2 statistic (1) applied to the bivariate subtable involving variables i and j. Thus, for a model fitted to K category items, $\mathbf{p}_{ij}$ is the K^2 vector of observed bivariate proportions; $\hat{\boldsymbol{\pi}}_{ij} = \boldsymbol{\pi}_{ij}\left(\hat{\boldsymbol{\theta}}_{ij}\right)$ is the vector of expected probabilities which depend only on the q_{ij} parameters involved in the bivariate table, $\hat{\boldsymbol{\theta}}_{ij}$; and $\mathbf{D}_{ij} = diag\left(\hat{\boldsymbol{\pi}}_{ij}\right)$. For instance, in the case of a graded-response model with a single latent trait, $\hat{\boldsymbol{\theta}}_{ij}$ are the 2 slopes and $2 \times (K - 1)$ intercepts. It is tempting to refer X_{ij}^2 to a chi-square distribution degrees of freedom equal to the number of parameters in the unrestricted model $\boldsymbol{\pi}_{ij}$, $K^2 - 1$, minus the number of parameters in the restricted model $\boldsymbol{\pi}_{ij}\left(\theta_{ij}\right)$, q_{ij}, so that $df_{ij} = K^2 - q_{ij} - 1$. However, Maydeu-Olivares and Joe (2006) showed that the asymptotic distribution of the subtable X_{ij}^2 is stochastically larger than this reference distribution. This means that referring X_{ij}^2 to a chi-square distribution with df_{ij} degrees of freedom may lead to rejecting well-fitting items. They also showed that the M_2 statistic (10) applied to a bivariate subtable is asymptotically distributed as a chi-square with df_{ij} degrees of freedom. Finally, they also showed that the bivariate subtable M_2 can be written in terms of the bivariate cell residuals as

$$M_{ij} = X_{ij}^2 - N\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right)' \hat{\mathbf{D}}_{ij}^{-1} \hat{\boldsymbol{\Delta}}_{ij} \left(\hat{\boldsymbol{\Delta}}_{ij}{}' \hat{\mathbf{D}}_{ij}^{-1} \hat{\boldsymbol{\Delta}}_{ij}\right)^{-1} \hat{\boldsymbol{\Delta}}_{ij}{}' \hat{\mathbf{D}}_{ij}^{-1} \left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right) \tag{38}$$

where $\boldsymbol{\Delta}_{ij}$ denotes the matrix of derivatives of the bivariate probabilities $\boldsymbol{\pi}_{ij}$ with respect to the parameters involved in the bivariate table, $\boldsymbol{\theta}_{ij}$.

Unfortunately, Maydeu-Olivares and Liu (2012) have recently shown that M_{ij} does not have much power to detect multidimensionality, and consequently, alternatives to M_{ij} are needed. From equation (38), M_{ij} can be seen as a correction to X_{ij}^2. An alternative way to correct X_{ij}^2, so that it

can be referred to a chi-square distribution with $df_{ij} = K^2 \times q_{ij} - 1$ degrees of freedom, is to correct it by its asymptotic mean and variance. The mean-and-variance corrected statistic is

$$\overline{\overline{X}}^2_{ij} = X^2_{ij}\sqrt{\frac{df_{ij}}{tr_2}} + df_{ij} - \sqrt{\frac{df_{ij}tr_1^2}{tr_2}}. \tag{39}$$

In (39), $tr_1 = \text{tr}\left(\hat{\mathbf{D}}_{ij}^{-1}\hat{\mathbf{\Sigma}}_{ij}\right)$, $tr_2 = \text{tr}\left(\hat{\mathbf{D}}_{ij}^{-1}\hat{\mathbf{\Sigma}}_{ij}\hat{\mathbf{D}}_{ij}^{-1}\hat{\mathbf{\Sigma}}_{ij}\right)$, where for the MLE

$$\mathbf{\Sigma}_{ij} = \mathbf{D}_{ij} - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{ij}' - \mathbf{\Delta}_{ij}\left(\mathcal{I}^{-1}\right)_{ij}\mathbf{\Delta}_{ij}' \tag{40}$$

multiplied by sample size is the asymptotic covariance matrix of the cell residuals for the pair of variables i and j when the model parameters have been estimated by maximum likelihood using the full table. In (40), $\left(\mathcal{I}^{-1}\right)_{ij}$ denotes the rows and columns of the information matrix corresponding to the item parameters involved in the subtable for variables i and j. As an alternative to (39), we can compute a mean-and-variance corrected X^2_{ij} with degrees of freedom estimated as a real number, $\overline{X}^2_{ij}$, analogous to the overall test statistic computed using (15) and (16). More specifically, $\overline{X}^2_{ij}$ is computed as

$$\overline{X}^2_{ij} = \frac{\text{tr}\left(\hat{\mathbf{D}}_{ij}\hat{\mathbf{\Sigma}}_{ij}\hat{\mathbf{D}}_{ij}\hat{\mathbf{\Sigma}}_{ij}\right)}{\text{tr}\left(\hat{\mathbf{D}}_{ij}\hat{\mathbf{\Sigma}}_{ij}\right)}X^2_{ij}, \tag{41}$$

and it is referred to a chi-square distribution with

$$\frac{\left(\text{tr}\left(\hat{\mathbf{D}}_{ij}\hat{\mathbf{\Sigma}}_{ij}\right)\right)^2}{\text{tr}\left(\hat{\mathbf{D}}_{ij}\hat{\mathbf{\Sigma}}_{ij}\hat{\mathbf{D}}_{ij}\hat{\mathbf{\Sigma}}_{ij}\right)} \tag{42}$$

degrees of freedom.

Similarly, we can also compute a bivariate subtable counterpart of the overall statistic proposed by Reiser (1996, 2008) given in equation (12)

$$R_{ij} = N\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right)'\hat{\mathbf{\Sigma}}^{+}_{ij}\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right). \tag{43}$$

The degrees of freedom of R_{ij} are given by the rank of $\mathbf{\Sigma}_{ij}$, which may be estimated from the data as the number of eigenvalues of $\hat{\mathbf{\Sigma}}_{ij}$, which are nonzero. In estimating the rank of $\hat{\mathbf{\Sigma}}_{ij}$ and of $\hat{\mathbf{\Sigma}}_2$ in (12) we use 10^{-5} as a cutoff.

A drawback of M_{ij} is that it cannot be used with binary data due to the lack of degrees of freedom (Maydeu-Olivares & Liu, 2012). The parameter $\overline{X}^2_{ij}$ may be used with binary data since the degrees of freedom are estimated as a real number using (42), and R_{ij} may also be used with binary data as its (integer-valued) degrees of freedom are estimated as well, unless the estimate were exactly zero.

An attractive alternative for binary data consists in using the z statistics for the residual cross-product of both items,

$$z_{ij} = \frac{p_{ij} - \hat{\pi}_{ij}}{\mathrm{SE}\left(p_{ij} - \hat{\pi}_{ij}\right)} = \frac{p_{ij} - \hat{\pi}_{ij}}{\sqrt{\hat{\sigma}_{ij}/N}}, \tag{44}$$

as suggested by Reiser (1996). Here, $\pi_{ij} = \Pr\left(Y_i = 1, Y_j = 1\right)$ and p_{ij} is its corresponding proportion. Thus, π_{ij} is simply 1 of the 4 probabilities in π_{ij}, and $\hat{\sigma}_{ij}$ is its corresponding diagonal element in (40). The asymptotic distribution of z_{ij} is standard normal, and of z_{ij}^2 is chi-square with 1 degree of freedom.

This statistic can be extended to polytomous ordinal data as

$$z_{ij} = \frac{k_{ij} - \hat{\kappa}_{ij}}{\mathrm{SE}\left(k_{ij} - \hat{\kappa}_{ij}\right)} = \frac{k_{ij} - \hat{\kappa}_{ij}}{\sqrt{\hat{\sigma}_{ij}^2/N}}, \tag{45}$$

where now $\hat{\sigma}_{ij}^2 = \mathbf{v}_{ij}{}'\hat{\Sigma}_{ij}\mathbf{v}_{ij}$, with

$$\mathbf{v}_{ij}{}' = (0 \times 0, 0 \times 1, \ldots 0 \times (K-1), \ldots, (K-1) \times 0, (K-1) \times 1, \ldots, (K-1) \times (K-1)), \tag{46}$$

and $k_{ij} - \hat{\kappa}_{ij} = \mathbf{v}_{ij}{}'\left(\mathbf{p}_{ij} - \hat{\boldsymbol{\pi}}_{ij}\right)$ is the residual cross-product. In the binary case (45) reduces to (44).

Using z statistics, one can also assess the fit of the model to single items. Using k_i to denote the sample mean and $\hat{\kappa}_i$ to denote the expected mean, a z statistic for a polytomous variable can be computed as the residual mean divided by its standard error

$$z_i = \frac{k_i - \hat{\kappa}_i}{SE(k_i - \hat{\kappa}_i)} = \frac{k_i - \hat{\kappa}_i}{\sqrt{\hat{\sigma}_i^2/N}} \tag{47}$$

where $\hat{\sigma}_i^2 = \mathbf{v}_i{}'\hat{\Sigma}_i\mathbf{v}_i$, with $\mathbf{v}_i{}' = (0, 1, \ldots K-1)$, and the residual mean is a linear function of the residual univariate proportions, $k_i - \hat{\kappa}_i = \mathbf{v}_i{}'\left(\mathbf{p}_i - \hat{\boldsymbol{\pi}}_i\right)$. The latter have asymptotic covariance matrix (for the MLE)

$$\boldsymbol{\Sigma}_i = \mathbf{D}_i - \boldsymbol{\pi}_i\boldsymbol{\pi}_i{}' - \boldsymbol{\Delta}_i\left(\mathcal{I}^{-1}\right)_i\boldsymbol{\Delta}_i{}' \tag{48}$$

multiplied by sample size. Here, $\mathbf{p}_j$ is the K–dimensional vector of observed univariate proportions and $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i\left(\hat{\boldsymbol{\theta}}_i\right)$ is the vector of expected probabilities, which depend only on the q_j parameters involved in the univariate subtable. Also, $\mathbf{D}_i = diag\left(\hat{\boldsymbol{\pi}}_i\right)$ and $\boldsymbol{\Delta}_i$ denotes the matrix of derivatives of the univariate probabilities $\boldsymbol{\pi}_i$ with respect to the parameters involved in the bivariate table, θ_i. In the binary case, the z statistic for the residual mean reduces to the univariate z_i statistic proposed by Reiser (1996)

$$z_i = \frac{p_i - \hat{\pi}_i}{SE(p_i - \hat{\pi}_i)} = \frac{p_i - \hat{\pi}_i}{\sqrt{\hat{\sigma}_i^2/N}}, \tag{49}$$

where $\hat{\pi}_i$ is the expected probability of endorsing the item, p_i is the observed proportion, and σ_i^2 is its corresponding diagonal element in (48).

Further research is needed comparing the empirical Type I errors and power against alternatives of interest of the array of statistics described above as well as of other statistics for assessing the source of misfit with a known asymptotic distribution such as score tests (Glas, 1999; Glas & Suárez-Falcón, 2003; Liu & Thissen, 2012, 2013). More details on score tests are given in the discussion section.

Extant results on the small sample behavior of these methods suggest that the choice of approximation of the information matrix has a strong impact on the behavior of the statistics in small samples. Thus, Liu and Maydeu-Olivares (2012b) report that the use of the cross-products approximation to compute score tests for binary data leads to rejecting well-fitting items in small samples ($<$1000 observations). In contrast, they report that the use of the expected information matrix leads to empirical Type I errors that are right on target even with 300 observations (they did not consider smaller samples). However, the expected information matrix can only be used in applications involving a manageable number of possible response patterns (19 or so items if they are binary, but only 8 or so items if they consist of 5 response categories). In our recent preliminary comparison of the behavior of $\overline{\overline{X}}_{ij}^2$ $\left(\text{and } \overline{X}_{ij}^2\right)$, R_{ij}, and z_{ij} we found that the observed information matrix gives excellent results in small samples; whereas, the cross-products approximation may be used only in large samples ($>$1000 observations). Thus, in small samples, z_{ij} should not be computed using the cross-products approximation to the information matrix as $\hat{\sigma}_{ij}^2$ in (45) is often negative. Also, when using the cross-products approximation, the sampling variability of R_{ij} is very large. In the applications below, the observed information matrix is used.

NUMERICAL EXAMPLES

PROMIS Anxiety Short Form

To illustrate the procedures described above, I use the $n = 7$ item PROMIS anxiety short form (Pilkonis et al., 2011). Respondents are asked to report the frequency with which they experienced certain feelings in the past 7 days using a $K = 5$ point rating scale ranging from "never" to "always." I use the $N = 767$ complete responses to these data kindly provided by the authors. Since the observed frequencies of the highest category were rather small for all items {6, 14, 13, 4, 5, 9, 13}, I merged the 2 highest categories prior to the analysis, so that $K = 4$. A unidimensional logistic graded-response model (Samejima, 1969) with a normally distributed latent trait was estimated by maximum likelihood using Mplus 6 (Muthén & Muthén, 2011); 48 Gauss-Hermite quadrature points were used and standard errors were computed using the observed information matrix (7). There are $q = 4 \times 7 = 28$ estimated parameters in this example.

The model does not fit the data exactly, as the value of the statistic M_2 given by equation (10) is 346.34 on 182 degrees of freedom, $p < 0.01$. M_{ord} cannot be used in this example, as the number of degrees of freedom is zero. A larger number of items, or a smaller number of categories with the same number of items, would be needed to estimate M_{ord}. For completeness, Table 1 displays all the remaining overall test statistics described in this article. In this table we see that the estimated degrees of freedom of R_2, 196, differs substantially from the number of

TABLE 1
Overall Goodness-of-Fit Tests for the Fit of a Graded Response Model to the Anxiety Data

stat	*value*	*df*	*RMSEA*
M_2	346.34	182	0.034
R_2	386.45	196	0.036
$\overline{Y}_2$	152.17	55.23	0.048
$\overline{L}_2$	346.28	181.63	0.034
$\overline{\overline{Y}}_2$	357.98	182	0.036
$\overline{\overline{L}}_2$	346.81	182	0.034

Notes: $N = 767$; $n = 7$; $K = 4$.

degrees of freedom of M_2, which in turns equals the number of statistics used for testing minus the number of estimated parameters. Also, the estimated degrees of freedom of the mean-and-variance corrected statistics $\overline{Y}_2$ and $\overline{L}_2$ differ substantially: 55.23 and 181.83, respectively. As the statistics are on different scales (their degrees of freedom), I compute RMSEAs for each of the statistics to gauge the (dis)similarity of the results obtained with the different statistics. As we can see in Table 1, the RMSEAs for all statistics are remarkably close: they range from 0.034 to 0.036 with the exception of the RMSEA of $\overline{Y}_2$, which yields a larger value, 0.048. It is also worth pointing out the remarkable closeness of the value of the mean and variance statistic $\overline{\overline{L}}_2$ to the value of the M_2 statistic.

I also computed a 90% confidence interval for the population RMSEA based on M_2, obtaining (0.029; 0.040). Thus, the fit of the model is adequate ($\text{RMSEA}_2 \leq 0.05$) but falls short of Maydeu-Olivares and Joe's (in press) criterion for an excellent approximation, $\text{RMSEA}_2 \leq (0.05/(K-1) = 0.017$.

If we are interested in detecting where the model misfits, with a view to modifying the model we could use the M_{ij}, R_{ij}, or mean adjusted X^2_{ij} statistics. Of the 2 mean and variance statistics I prefer the one with df_{ij} degrees of freedom, $\overline{\overline{X}}^2_{ij}$. These are displayed in matrix form in Table 2. To control for multiple testing, I use a Bonferroni adjustment. Since there are $(7 \times 6) / 2 = 21$

TABLE 2
Mean-and-Variance Adjusted Bivariate $\overline{\overline{X}}^2_{ij}$ statistics After Fitting a Graded Response Model to the Anxiety Data

Item	*1*	*2*	*5*	*6*	*3*	*4*	*7*	*Average*
1		20.77	19.94	20.67	16.93	11.89	**35.37**	17.94
2	20.77		14.69	**25.39**	15.71	15.44	**23.26**	16.47
5	19.94	14.69		7.70	**22.38**	13.20	**26.26**	14.88
6	20.67	**25.39**	7.70		**22.87**	15.42	15.03	15.30
3	16.93	15.71	**22.38**	**22.87**		4.40	19.90	14.60
4	11.89	15.44	13.20	15.42	4.40		8.54	9.84
7	**35.37**	**23.26**	**26.26**	15.03	19.90	8.54		18.34

Notes: df = 7; Statistics larger than 22.16 are significant at the 5% level with a Bonferroni adjustment and are in bold. The row averages of the statistics have been appended as an additional column.

TABLE 3
M_{ij} and R_{ij} Bivariate Statistics After Fitting a Graded Response Model to the Anxiety Data

Item	*1*	*2*	*5*	*6*	*3*	*4*	*7*	*Average*
1		12.18	16.60	16.52	16.64	6.14	**36.02**	14.87
2	**39.42**		13.73	**27.61**	15.97	12.61	20.62	14.67
5	27.26	25.55		8.37	**24.38**	14.69	**22.84**	14.37
6	24.75	31.32	10.44		**24.60**	12.05	11.50	14.38
3	**47.62**	28.44	29.02	31.11		5.03	10.47	13.87
4	**79.66**	**34.06**	19.50	**46.31**	5.03		4.51	7.86
7	**42.60**	30.08	30.66	28.37	10.47	18.58		15.14

Notes: M_{ij} statistics are displayed above the diagonal; R_{ij} statistics, below the diagonal. df = 7 for M_{ij}, df range from 12 to 14 for R_{ij}. Consequently, R_{ij} values cannot be compared across pairs, and row averages are given for M_{ij} only. Statistics significant at the 5% level with a Bonferroni adjustment are in bold. The averages of the M_{ij} statistics across the 7 items have been appended as an additional column.

statistics the cut-off p-value used is $0.05/21 = 0.002$. The critical value for a chi-square distribution with $df_{ij} = 4^2 - 2 \times 4 - 1 = 7$ degrees of freedom yielding this p-value is 22.16. I have highlighted the statistics that are larger than this critical value with boldface. I have also appended to the table a column containing the average of the $\overline{\overline{X}}^2_{ij}$ statistics for each item. These row averages can be used to identify the best- and worst-fitting items.

In presenting tables of bivariate fit statistics such as $\overline{\overline{X}}^2_{ij}$, I find it useful to apply a cluster analysis to the matrix of $\overline{\overline{X}}^2_{ij}$ statistics (I use Ward's method) and to display the bivariate statistics with items reordered according to the results of the cluster analysis. I do this in Table 2. In this table we see that the model misfit involves the associations between items {1, 2, 5} and item 7, the associations between items {5, 6} and item 3, and the association between items 2 and 6. For comparison, in Table 3 I provide the results obtained using M_{ij} and R_{ij}. Comparing the results presented in tables 2 and 3, we see that generally the values of M_{ij} are smaller than the values of $\overline{\overline{X}}^2_{ij}$. As a result, the row averages obtained using M_{ij} are smaller than using $\overline{\overline{X}}^2_{ij}$. This may indicate that M_{ij} has lower power than $\overline{\overline{X}}^2_{ij}$. However, similar conclusions are reached in this example when using M_{ij} and $\overline{\overline{X}}^2_{ij}$.

The use of R_{ij} involves some complications. Degrees of freedom for different pairs of items differ and as a result, the values of R_{ij} across item pairs cannot be compared directly: only their p-values can be compared. In this application, the estimated degrees of freedom of R_{ij} were 12 for 3 item pairs, 13 for 9 item pairs, and 14 for the remaining item pairs. As a result, row averages of R_{ij} statistics cannot be computed. Most importantly, I observed that although R_{ij} has reasonable Type I error rates and high power, the sampling variability of the statistic is large even when the observed information matrix is used. As a result, in applications, we are likely to encounter R_{ij} values that are very large. We see this in this example. The estimated R_{ij} statistic for item pair {1, 4} is 79.66 on 13 df. On seeing such a large residual statistic, a researcher will believe that there is a serious misfit in the model when, in fact, this high value of R_{ij} may be the result of the large sampling variability of the statistic. Actually, for this item pair, the mean and variance statistic $\overline{\overline{X}}^2_{ij}$ is not statistically significant.

Currently our preferred method to assess the source of misfit is using z statistics. On the one hand, they enable us to test the fit of the model at the item level. On the other hand, preliminary results reveal that these statistics are more powerful for detecting alternatives of interest that the other statistics discussed in this article. More specifically, z_{ij} statistics concentrate the information of the bivariate table of residuals into a 1 degree of freedom statistic (z_{ij}^2 follows a 1 degree of freedom chi-square distribution). If the concentration of the information is along the direction of the misfit, then z_{ij} will be more powerful for detecting that direction of misfit; otherwise, it will be less powerful than alternative statistics. In any case, there need not be a strong relation between the results obtained using z_{ij} and the alternative statistics. The z_{ij} statistics obtained in this example are presented in Table 4. Negative values of the statistics indicate that the model overestimates the association between the items; whereas, positive values indicate that the model underestimates the association. In this table, I have also included the univariate statistics z_i. Using a Bonferroni adjustment, the critical value for this standard normal statistic is |3.12| and in the table I have highlighted the statistics that are larger than this critical value in bold. As we can see in Table 4, the z_{ij} statistics clearly suggest that the misfit is located in the association between items {1,3} with {2,5}. In all cases the residual is negative, suggesting that the model overestimates the associations between these 2 pairs of items.

The residual cross-products also yield a relatively natural way to assess the magnitude of misfit when the data are ordinal through its relation to the residual correlations (34). The residual correlations for this example are reported in Table 4. The correlation between the z_{ij} statistics and the residual correlations in this example is 0.877. As judged by the size of the residual correlations, the magnitude of the misfit of the graded-response model to the PROMIS anxiety data is rather small. The largest residual correlation is 0.036, and it involves items 4 and 2. Consequently, the standardized square root mean squared residual is also very low: SRMSR = 0.016. The model fits these data rather well, although it is unlikely to be the data-generating model.

TABLE 4
z_i Statistics for Residual Means, Z_{ij} Statistics for Residual Cross-Products, and Residual Correlations After Fitting a Graded Response Model to the Anxiety Data

Item	*1*	*2*	*5*	*6*	*3*	*4*	*7*	*Average*
1	**3.91**	**−3.48**	**−3.57**	−1.68	−1.54	−0.94	−1.98	2.44
2	−0.03	−0.77	0.50	−2.10	**−3.21**	1.53	−0.87	1.78
5	−0.02	0.02	−0.96	−2.41	**−3.27**	−1.37	−1.57	1.95
6	<0.01	−0.02	−0.01	−0.62	−0.79	−2.03	0.24	1.41
3	0.01	−0.03	−0.02	<0.01	−0.55	−2.20	−0.54	1.73
4	<0.01	0.04	−0.01	−0.02	−0.02	−2.20	−1.26	1.62
7	−0.01	−0.01	−0.02	<0.01	−0.01	−0.02	−1.26	1.09

Notes: z_{ij} statistics are displayed above the diagonal; z_i statistics, along the diagonal; and residual correlations below the diagonal. z statistics larger than |3.12| are significant at the 5% level with a Bonferroni adjustment and are in bold. The averages of the absolute values of the z statistics across the 7 items have been appended as an additional column.

If we wish to modify the model, how should we use the information obtained in this piece-wise analysis? It is not clear in this example. The misfit does not appear to be located in a particular cluster of items, which would suggest the need of a multidimensional model. Also, the misfit does not appear to be associated to any particular item, which would suggest dropping the item from further analysis. Rather, in this example there appear to be mild deviations from the fitted model. The easiest course of action in this case is to retain the fitted model as a close enough approximation to the true data-generating model. Other alternatives include (a) finding an alternative better-fitting IRT model (with a different response function), possibly including a mixture distribution and (b) identifying outlying individuals whose responses are not well fit by the model.

CHILEAN MATHEMATICAL PROFICIENCY DATA

Due to the lack of degrees of freedom, the application of tests to bivariate subtables when the items are binary presents certain peculiarities. For this reason, it is of interest to consider a binary data example here as well. The data used in this example are the responses of 3,000 individuals to a 15-item test aimed at measuring mathematical proficiency in Chilean adults. A 1-parameter logistic model (1PL, aka random effects Rasch model) was applied to these data. Estimation was again performed using ML with 50 Gauss-Hermite quadrature points; standard errors were again computed using the observed information matrix.

Table 5 displays the values of the different goodness-of-fit statistics applied, along with their degrees of freedom and *p*-values. As we can see in this table, using a significance level of 5%, we cannot reject the hypothesis that the 1PL is the generating model for these data with any of the statistics. This is surprising given the large sample size used and the restrictiveness of the fitted model. In fact, all test statistics but $\overline{Y}_2$ and $\overline{\overline{Y}}_2$ yield very similar *p*-values, between 0.09 and 0.12. $\overline{Y}_2$ and $\overline{\overline{Y}}_2$ yield much larger *p*-values, around 0.50. More research is needed to explain this discrepancy.

When fitting a 1PL model, 3 parameters are involved in each bivariate table (2 intercepts and 1 common slope) and there are 3 mathematically independent probabilities. So there are zero degrees of freedom for assessing the piece-wise fit of the model using M_{ij}, and therefore

TABLE 5
Overall Goodness-of-Fit Tests for the Fit of a 1-parameter Logistic Model to the Chilean Mathematical Proficiency Data

stat	*value*	*df*	*p*	*RMSEA*
M_2	121.32	104	0.12	0.007
R_2	133.40	113	0.09	0.008
$\overline{Y}_2$	22.88	23.52	0.50	0
$\overline{L}_2$	121.47	104.04	0.12	0.007
$\overline{\overline{Y}}_2$	102.64	104	0.52	0
$\overline{\overline{L}}_2$	121.43	104	0.12	0.007

Notes: $N = 3{,}000$; $n = 15$; $K = 2$.

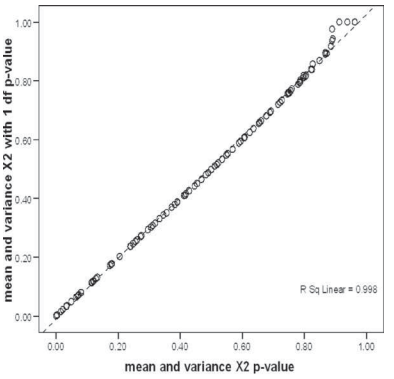
$\overline{\overline{X}}_{ij}^2$. Piece-wise assessment can be performed using z_{ij} as its asymptotic distribution is standard normal. It can also be assessed using $\overline{X}_{ij}^2$ and R_{ij} as for these statistics degrees of freedom are estimated as a real number and as an integer, respectively. In this application, the estimated degrees of freedom for $\overline{X}_{ij}^2$ ranged between 1.02 and 1.06 with an average of 1.03. The estimated degrees of freedom for R_{ij} were 2 for 6 item pairs, and 3 for the remaining 99 item-pairs. None of the z_{ij}, $\overline{X}_{ij}^2$ or R_{ij} statistics is statistically significant at the 5% level when using a Bonferroni adjustment. We have failed to reject the hypothesis that the 1PL model is the generating model for these data, and we have failed to detect any parts of the model whose fit may be improved. This is a remarkable result.

Note that the $\overline{X}_{ij}^2$ statistics are not amenable to being displayed in table form since they are on different scales (their degrees of freedom). To avoid this problem, and if mean-and-variance adjusted X^2 statistics are to be reported, I advocate reporting $\overline{\overline{X}}_{ij}^2$ with 1 degree of freedom when fitting binary data. This is a reasonable approach in as much as the p-values obtained using $\overline{X}_{ij}^2$ and $\overline{\overline{X}}_{ij}^2$ are very close. To illustrate this point, I have plotted both sets of p-values for this example in Figure 1a. As we can see in this figure, in this example the relationship is almost perfectly linear (except for very high p-values) with an intercept very close to zero. For completeness, in Figure 1b I also provide the relationship between the p-values obtained using $\overline{X}_{ij}^2$ and R_{ij}. In this figure we see that except for 6 item pairs, there is a very strong curvilinear relationship between the 2 sets of p-values. Only for 2 of these outlying pairs are there 2 degrees of freedom for testing R_{ij}. Therefore, there is no relationship between being an outlying item pair in Figure 1b and the number of degrees of freedom of R_{ij}. Finally, in Figure 1c the p-values for $\overline{X}_{ij}^2$ and z_{ij} are displayed. We see that there is a strong linear relationship between the 2 sets of p-values. However, the relationship is clearly heteroscedastic.

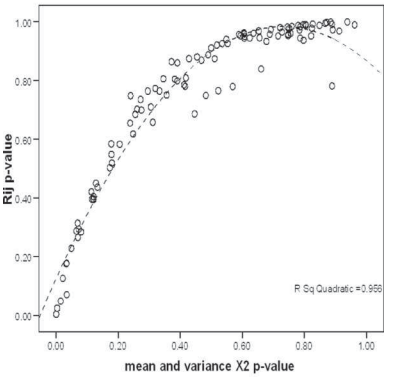
GOODNESS-OF-FIT METHODS FOR THE ORDINAL FACTOR ANALYSIS MODEL ESTIMATED VIA POLYCHORIC CORRELATIONS

A logistic or a normal ogive link function can be used in the graded-response model, and differences in fit when one or the other is used are generally small. When a normal ogive link function is used, the model is formally equivalent to a factor analysis model for multivariate normal responses that have been categorized (i.e., an ordinal factor analysis model). That is, if (a) a normally distributed psychological value is assumed to underlie the response to each item, (b) a common factor model is assumed to hold for the unobserved psychological values, and (c) the psychological values are discretized according to a set of thresholds, then Samejima's normal ogive graded-response model and the factor analysis model for ordinal responses are equivalent.

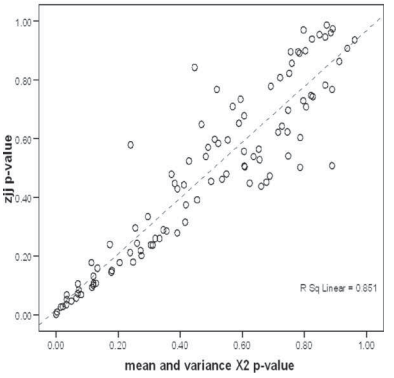
Also, the graded-response model can be specified using unstandardized parameters (intercepts and slopes) or standardized parameters (standardized thresholds and standardized factor loadings) (Forero & Maydeu-Olivares, 2009). Up to this point we have focused on the logistic version of the model because it is the most widely used when ML estimation of the model is used. Also, when ML estimation is used, unstandardized parameters are used. But, if the normal ogive version of the model is used and the model is specified using standardized parameters, there is a separability



a) $\bar{X}^2_{ij}$ with $\bar{\bar{X}}^2_{ij}$ with one degree of freedom p-values



b) $\bar{X}^2_{ij}$ vs. R_{ij} p-values



c) $\bar{X}^2_{ij}$ vs. z_{ij} p-values

FIGURE 1 Relationship between p-values for different statistics for bivariate piecewise assessment of the 1-parameter logistic model fitted to the Chilean mathematical proficiency data.

of parameters that does not take place in the logistic version of the model, even when standardized parameters are used. More specifically, when the normal ogive version is used with standardized parameters, then (a) the univariate margins depend only on the standardized thresholds and (b) the bivariate margins only depend on the standardized thresholds and the polychoric correlations. A polychoric correlation is the correlation between 2 unobserved psychological values (if they had been observed), and when both items are binary, it is referred to as tetrachoric correlation. This separability of parameters in the normal ogive version of the model has implications for model estimation, and also for the goodness-of-fit testing of the model.

Because of this parameter separability, the normal ogive model can be estimated sequentially. First, the standardized thresholds are estimated separately for each item using maximum likelihood. Second, each polychoric correlation is estimated separately by maximum likelihood, holding the thresholds at the values estimated in the second stage. This is pseudo maximum likelihood in the terminology of Gong and Samaniego (1981). Third, the standardized factor loadings (and other model parameters, such as the interfactor correlations) are estimated from the thresholds and polychoric correlations by minimizing a weighted least squares (LS) function.

Fully weighted LS, diagonally weighted LS, or unweighted LS may be employed in the third stage of the estimation method, and the latter may be the best performing method (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Muthén, 1993). The procedure just described is implemented in popular structural equation modeling programs such as Lisrel (Jöreskog & Sörbom, 2007) or Mplus (Muthén & Muthén, 2011). The main difference between the 2 implementations is how they estimate the asymptotic covariance matrix of the parameter estimates (Jöreskog, 1994; Maydeu-Olivares, 2006; Muthén, 1984; Muthén, 1993). A very similar procedure is implemented in Eqs (Bentler, 2004; see also Lee, Poon, & Bentler, 1995).

The parameter separability present in the model also has implications for goodness-of-fit testing: The overall discrepancy between the model and the data can be decomposed into a distributional discrepancy (the extent to which the data arises from a discretized multivariate normal distribution, and a structural discrepancy (the extent to which the constraints imposed on the thresholds and polychoric correlations are correctly specified).

Current implementations of polychoric estimation methods provide an array of methods to test the structural assumptions (i.e., whether a 1-factor model adequately reproduces the estimated polychoric correlations): (a) an overall goodness-of-fit test, (b) z statistics for residual polychoric correlations, and (c) score tests (aka modification indices), for instance, for correlations among the unique errors.

Although these tests are informative, and therefore may be useful, they do not assess model-data misfits. Rather, they assess solely the structural discrepancy and they rely on the distributional assumptions being met. It is not clear how robust they are to violations of the discretized multivariate normality assumption. Furthermore, the distributional assumption of underlying multivariate normality is currently only assessed in a piece-wise fashion, computing X_{ij}^2 for each pair of variables after the thresholds and polychoric correlations are estimated and before a structural model is fitted. In this case, X_{ij}^2 is not asymptotically chi-square, because the 2-stage procedure used to estimate the thresholds and polychoric correlations is not asymptotically efficient. In principle, M_{ij} should be used instead as this statistic is asymptotically correct when polychoric correlations are computed in 2 stages. However, the simulation results by Maydeu-Olivares, García-Forero, Gallardo-Pujol, and Renom (2009) reveal that p-values of X_{ij}^2 are just as

accurate as those of M_{ij} in this case due to the high efficiency of the 2-stage polychoric correlation estimator. Yet, it is not clear what to conclude if the assumption of underlying normality is rejected for some pairs but not for others. If assessing separately the distributional and structural restrictions of this model are of interest, then an overall test of the distributional assumptions is needed, and M_2 (10) and M_{ord} (21) may be used to this end. To test the distributional assumptions using these statistics, the parameters estimated in the first 2 stages of the sequential estimation procedure are used (i.e., unrestricted thresholds and polychoric correlations).

In contrast, the methods described in the previous sections of this article assess the overall discrepancy between the model and the data. For instance, to assess the overall discrepancy M_2 and M_{ord} may be applied, using in this case the restricted polychoric correlations and thresholds (as implied for instance by a 1-factor model). Of the 2, in principle, the statistic M_{ord} is better suited for ordinal factor analysis problems as it is computationally less intensive, and extant theory and simulation studies suggest that it has higher power (e.g., Cai & Hansen, 2012). However, when there are no degrees of freedom for testing using M_{ord}, M_2 must be used. Also, using the estimated statistics one can compute the RMSEA_2 and RMSEA_{ord} overall goodness-of-approximation statistics as well as confidence intervals for the population parameters.

To assess the overall source of misfit, M_{ij} (38) can also be applied directly. The difference is that when testing the overall misfit one uses the model-implied polychoric correlation; whereas, when testing only the distributional restriction one uses the unrestricted polychoric correlation. Maydeu-Olivares and Liu (2012) report that empirical rejection rates of M_{ij} when applied to assess the overall discrepancy in an ordinal factor analysis are right on target even with as few as 100 observations. They also observe that the empirical rejection rates of the asymptotically incorrect X^2_{ij} also fare well in this case. In other words, the correction introduced by M_{ij} on X^2_{ij} for this model is very small.

In closing this section, the other overall statistics presented earlier, as well as the other piecewise fit statistics, can also be computed for the ordinal factor analysis model. All that is needed is to use the correct expression for the asymptotic covariance matrices in lieu of expressions (4) and (40) given for the MLE. The asymptotic covariance matrix of the residual univariate and bivariate moments for polychoric estimation methods is given in Maydeu-Olivares and Joe (2006, eq. 2.6). An expression that is easier to program is given in Maydeu-Olivares (2006, eq. 31).

DISCUSSION

In this article, I have presented an overview of limited information test statistics for assessing the goodness-of-fit of IRT models. The limited information methods presented here are quadratic-form statistics in univariate and bivariate residuals.

Three general strategies have been proposed in the literature to construct overall limited information test statistics. In one strategy, these residuals are weighted by a consistent estimate of their asymptotic covariance matrix. The second strategy weights the residuals such that a consistent estimate of their asymptotic covariance matrix is a generalized inverse of the weight matrix. The third strategy simply chooses a computationally convenient weight matrix. When the first and second strategies are used, a statistic with an asymptotic chi-square distribution is obtained. With the third strategy, the asymptotic distribution of the statistic is a mixture of independent chi-square variables, and p-values are generally obtained by computing a mean-and-variance adjustment

to the statistic. In any case, because the asymptotic distribution of these residuals depends on 4-way probabilities, the distribution of these quadratic-form statistics can be well approximated by asymptotic methods even in small samples.

When either the first or third strategy is pursued, the computation of the p-values involves an estimate of the asymptotic covariance matrix of the item parameters. There are, in turn, 3 widely used approaches to obtain this matrix when ML estimation is applied: the expected information matrix, the observed information matrix, and the cross-products information matrix. The expected information matrix yields good results for goodness-of-fit testing purposes, but it can only be computed when the model does not involve too many possible response patterns. Specifically, it can seldom be computed when the items are polytomous. As a result, either the observed matrix or the cross-products matrix must be used in most applications. In our limited experience, the cross-products approximation can only be used for goodness-of-fit purposes in fairly large samples; whereas, the observed information matrix yields good results in small samples. More research is needed to investigate which of the 3 strategies yields the best results in terms of Type I errors and power, and how to best approximate the covariance matrix of the item parameter estimates.

The same 3 strategies can be used to construct statistics to assess the source of misfit in poorly fitting models. An additional approach can be used in this case, namely, the use of z statistics that are asymptotically standard normal. In the case of binary data, it is relatively natural to obtain a z statistic for the residual means and cross-product. This same approach can also be used in the case of polytomous ordinal variables. Residual cross-products are directly related to residual (product moment) correlations. Hence, the use of residual cross-products leads back, in many ways, to the methods used to assess the goodness-of-fit of classical factor analysis models. However, as the anxiety example illustrates, different piece-wise test statistics may suggest different ways to modify the model. For this reason, I suggest that 2 different piece-wise procedures be used (such as z statistics and mean-and-variance corrected X^2_{ij}).

The methods currently implemented in ordinal factor analysis models estimated using polychorics are closely related to some of the methods described in this article. However, they only assess the structural restrictions imposed by the model (how well the model fits the thresholds and polychoric correlations). In contrast, the methods described in the first sections of the article assess the overall restrictions imposed by the model (how well the model fits the data).

There is an important class of statistics for assessing the source of misfit that I have not covered in this review: score tests (Glas, 1999; Glas & Suárez-Falcón, 2003; Liu & Thissen, 2012). Score tests (aka Lagrange multiplier tests) have a known asymptotic distribution and are an attractive alternative to the methods described in this article. They differ from the statistics presented here in that they are directional. Thus, whereas the alternative hypothesis in the tests presented here is unrestricted (it simply states that the fitted model does not hold), in score tests the alternative hypothesis is fully specified and the fitted model is nested within the model specified by the null hypothesis. For instance, if the true model is a bifactor model and a unidimensional model is fitted, then a score test that uses a bifactor model as an alternative hypothesis has maximum power, but if the alternative model is not a bifactor model, then the power of the test needs to be investigated. For some fitted models and some true models, Type I errors and power rates of score tests appear to be relatively unaffected by the choice of alternative model used in the score test (Liu & Maydeu-Olivares, 2012b). However, in other cases the choice of alternative model does

make a difference (Liu & Thissen, 2013). In any case, more research is needed to compare the performance of the statistics presented here with that of the score tests.

CONCLUDING REMARKS

In applications, researchers should always assess how well IRT models fit their data. Whenever the model does not fit the data, item parameter estimates are biased and, therefore, individuals' scores are biased. The magnitude of the bias is unknown but it depends on the magnitude of misfit of the model. Depending on the research question being addressed, the bias incurred may be negligible from a substantive viewpoint if the model is only slightly misspecified (Reise, Scheines, Widaman, & Haviland, 2012). In other applications, however, the use of misfitting models may lead to seriously invalid conclusions. For policy analysis, invalid conclusions erroneously backed up by data modeling are more dangerous than theoretical conjectures because we invariably assign more confidence to conclusions supported by data.

Goodness-of-fit should be distinguished from model selection. Model selection examines model-model fit. As such, it need not provide us with information as to whether we should avoid using any of the models under consideration. Goodness-of-fit examines model-data fit. Goodness-of-fit and model selection should be used in tandem. If we find several competing models that fit the data well, then we need to consider which model to use. Alternatively, if we first select the best model from among a set of competing models using some model selection criterion, we need to address the question of whether this best fitting model should be used or an alternative model should be sought.

Assessing the overall goodness-of-fit has been outside the research agenda of IRT modeling until very recently due to the lack of reliable statistics. The use of limited information test statistics has changed this situation and in this article I have presented a general overview of these methods. See also Glas and Verhelst (1995) for a review of methods specifically designed for Rasch-type models.

Because limited information statistics concentrate the information available in the data, they are generally more powerful than full information statistics. Of course, they have little power to detect the misfit located in places in which they are not looking. Thus, if only bivariate information is used for testing and the model misfit is located in the 3-way or higher associations between the items, the statistics will have very low power. However, in our experience, it is hard not to reject IRT models using the limited information statistics described in this article. On the other hand, as the Rasch application shows, it is possible to find well-fitting IRT models for carefully constructed sets of items, even in large samples. Yet, in my experience, such well-fitting applications are rare, and they are more common when binary items are used and when educational contents are measured. If this is indeed the case, why is it so? Is it because it is easier to model binary items than polytomous items? Is it because it is easier to model educational aspects than personality, attitudes, or patient reported outcomes? Is it because our models for polytomous data are inappropriate? Or is it simply because the test statistics are more powerful for polytomous than for binary data? These are all very interesting questions, and it is clear that much more research will be needed before we can answer them.

More research is also needed on a number of other topics. First, a great deal of research on goodness-of-fit tests has focused on the graded-response model and Rasch-type models. There

is some evidence (Maydeu-Olivares, 2005) that the graded-response model may be the best fitting parametric model for rating data. However, more research is needed to investigate how to assess the fit of alternative, more highly parameterized, models, such as the 3-parameter logistic (3PL) or Bock's (1972) nominal model. Can their fit be assessed using bivariate information? Do we need to use trivariate information? Second, simulation studies are needed to determine which statistic should be used to assess the overall goodness-of-fit, and the source of misfit. Importantly, simulation results obtained with overall statistics should not be extrapolated to piece-wise statistics or vice versa. For instance, in a recent study (Maydeu-Olivares & Liu, 2012) we found that in conditions where the overall M_2 statistic has maximum power (power = 1), the power of the piece-wise M_2 statistic (i.e. M_{ij}) was only slightly above the nominal level. Third, and perhaps most importantly, more research is needed to investigate the robustness of substantive conclusions for modeling misspecification of varying degrees. For instance, if our model yields an $RMSEA_2$ of 0.05, what does this information tell us about the accuracy of our latent trait estimates (i.e. individual scores)? However, this suggestion for future work should not distract us from the fact that a large set of methods is now available for applied researchers to assess the fit of their IRT models. I look forward to seeing many applications of these methods in the near future.

ACKNOWLEDGEMENTS

This research was supported by an ICREA-Academia Award and grant SGR 2009 74 from the Catalan government and grant PSI2009-07726 from the Spanish Ministry of Education. The author is indebted to Yang Liu for computational assistance and for countless discussions that have greatly improved the article.

REFERENCES

Asparouhov, T., & Muthén, B. O. (2010). Simple second order chi-square correction scaled chi-square statistics. Technical Report. Los Angeles, CA: Muthén and Muthén.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*, 1–15. doi: 10.1348/000711002159617

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods and Research*, *27*, 525–546. doi: 10.1177/0049124199027004003

Bentler, P. M. (2004). EQS 6 [Computer program]. Encino, CA: Multivariate Software.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. doi: 10.1007/BF02291411

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. doi: 10.1007/BF02293801

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing made fit. In K.A. Bollen & J.S. Long (Eds), *Testing Structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *The British Journal of Mathematical and Statistical Psychology*, *66*, 245–76. doi: 10.1111/j.2044-8317.2012.02050.x

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 tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173–194. doi: 10.1348/000711005X66419

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*(3), 275–299. doi: 10.1037/a0015825

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*(4), 625–641. doi: 10.1080/10705510903203573

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*(3), 273–294. doi: 10.1007/BF02294296

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*(2), 87–106. doi: 10.1177/0146621602250530

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

Gong, G., & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *Annals of Statistics*, *9*, 861–869.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, *48*, 419–426. doi: 10.2307/2332763

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*(3), 393–419. doi: 10.1007/s11336-010-9165-5

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*(3), 381–389. doi: 10.1007/BF02296131

Jöreskog, K. G., & Sörbom, D. (2007). LISREL (Version 8.8) [Computer program]. Chicago, IL: Scientific Software.

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. doi: 10.1016/j.csda.2012.04.010

Koehler, K. J., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, *75*(370), 336–344. doi: 10.1080/01621459.1980.10477473

Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, *48*, 339–58.

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

Liu, Y., & Maydeu-Olivares, A. (2012b). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*, 254–274. doi: 10.1177/0013164412453841

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. (2013). Local dependence score tests for the graded response model. *Under review*.

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research*, *40*(2), 261–279. doi: 10.1207/s15327906mbr4002_5

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*, *71*(1), 57–77. doi: 10.1007/s11336-005-0773-4

Maydeu-Olivares, A., García-Forero, C., Gallardo-Pujol, D., & Renom, J. (2009). Testing categorized bivariate normality with two-stage polychoric correlation estimates. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *5*(4), 131–136. doi: 10.1027/1614-2241.5.4.131

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2 n contingency tables. *Journal of the American Statistical Association*, *100*(471), 1009–1020. doi: 10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. doi: 10.1007/s11336-005-1295-9

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. (in press). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*.

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

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551–560. doi: 10.1007/BF02293813

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. doi: 10.1007/BF02294210

Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.

Muthén, L. K., & Muthén, B. (2011). MPLUS 6 [Computer program].

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*(3), 263–83. doi: 10.1177/1073191111411667

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2012). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*(1), 5–26. doi: 10.1177/0013164412449831

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

Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology*, *61*(Part 2), 331–360.

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

Satorra, A., & Bentler, P. M. (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.

Steiger, J. H., & Lind, J. C. (1980, June). *Statistically-based tests for the number of common factors*. Paper presented at the Annual Meeting of the Psychometric Iowa.

Tollenaar, N., & Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 271–288.

Von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research Online*, *2*(2), 29–48. doi: 10.1017/S1041610212001652