Multivariate Behavioral Research, 49:305–328, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2014.911075

Routledge
Taylor & Francis Group

Assessing Approximate Fit in Categorical Data Analysis

Alberto Maydeu-Olivares
Faculty of Psychology, University of Barcelona

Harry Joe
Department of Statistics, University of British Columbia

A family of Root Mean Square Error of Approximation (RMSEA) statistics is proposed for assessing the goodness of approximation in discrete multivariate analysis with applications to item response theory (IRT) models. The family includes RMSEAs to assess the approximation up to any level of association of the discrete variables. Two members of this family are RMSEA$_2$, which uses up to bivariate moments, and the full information RMSEA$_n$. The RMSEA$_2$ is estimated using the M_2 statistic of Maydeu-Olivares and Joe (2005, 2006), whereas for maximum likelihood estimation, RMSEA$_n$ is estimated using Pearson's X^2 statistic. Using IRT models, we provide cutoff criteria of adequate, good, and excellent fit using the RMSEA$_2$. When the data are ordinal, we find a strong linear relationship between the RMSEA$_2$ and the Standardized Root Mean Squared Residual goodness-of-fit index. We are unable to offer cutoff criteria for the RMSEA$_n$ as its population values decrease as the number of variables and categories increase.

Parametric models are fitted to categorical data in an attempt to capture the underlying process that may have generated the data. Yet, in applications one should expect discrepancies between the postulated parametric model and the population probabilities that, given a sufficiently large sample size, will not be attributed to chance and will lead to rejecting the fitted model. As the number of variables being modeled increases, good parametric approximations to the population probabilities become increasingly difficult and much smaller sample sizes will suffice to reveal discrepancies between the population probabilities and the model specified under the null hypothesis. In this context, and paraphrasing Steiger (1990), the question of interest is how well our model approximates the unknown population probabilities. Ultimately, however, researchers need to decide whether the approximation provided by the fitted model is good enough. This can be accomplished by testing whether the discrepancy between the population probabilities and the fitted model is less than or equal to some arbitrary value (Browne & Cudeck, 1993).

Although well developed in the structural equation modeling (SEM) literature where they arose, the notions of assessing goodness of approximation and testing for close fit are yet to be developed in categorical data analysis. This article aims to fill this gap with an eye on applications to item response modeling (IRT). Interestingly, assessing the overall exact model fit in categorical data analysis had proved so difficult, except for very small models that are usually not of interest in applications, that the issue of goodness-of-fit has been outside the IRT research agenda for many years. Recently, there has been a growing interest in goodness-of-fit assessment in IRT (Bartholomew & Leung, 2001; Bartholomew & Tzamourani, 1999; Cai & Hansen, 2013; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Glas, 1988; Glas & Verhelst, 1989, 1995; Joe & Maydeu-Olivares, 2006, 2010; Langeheine, Pannekoek, & van de Pol, 1996; Maydeu-Olivares, 2006; Maydeu-Olivares & Cai, 2006; Maydeu-Olivares & Joe, 2005, 2006; Reiser, 1996, 2008; Reiser & VandenBerg, 1994; Tollenaar & Mooijaart, 2003; Von Davier, 1997) and recent reviews (Mavridis, Moustaki, & Knott, 2007; Maydeu-Olivares, 2013; Maydeu-Olivares & Joe, 2008) suggest that the issue of assessing whether IRT models fit exactly is well under way to being solved.

Correspondence concerning this article should be addressed to Alberto Maydeu-Olivares, Faculty of Psychology, University of Barcelona, P. Valle de Hebrón, 171, 08035 Barcelona, Spain. E-mail: amaydeu@ub.edu

In this article we provide a general framework for assessing the goodness of approximation of a model in categorical data analysis. The developments presented here closely parallel the existing ones in SEM. For instance, we make use of the Root Mean Square Error of Approximation (RMSEA) first introduced by Steiger and Lind (1980) in the SEM literature. The RMSEAs introduced here are simply a transformation of the discrepancy between the fitted model and the population probabilities that adjusts for model complexity and expresses such discrepancy in the metric of the summary statistics used to assess model fit. For ease of exposition we concentrate on models for multivariate data obtained under multinomial sampling although the framework can be easily extended to models obtained under other sampling schemes. Also, for concreteness, our presentation focuses on applications to IRT modeling although applications to other statistical models for multivariate discrete data are straightforward. The presentation necessarily is somewhat technical although most technical material is relegated to an Appendix.

The remainder of the article is organized as follows: We begin by reviewing tests of exact fit in multinomial models, such as Pearson's X^2, and discussing why asymptotic p values for these statistics are inaccurate unless model size is very small. We also discuss in this section how to obtain accurate asymptotic p values for tests of exact fit, namely, by using test statistics that only use bivariate information. In the second section we propose using a Mahalanobis distance to measure the discrepancy between the population probabilities and the fitted model in categorical data analysis. Using this distance, an RMSEA can be formed, which can be estimated using Pearson's X^2 statistic. Unfortunately, the sampling distribution of this RMSEA estimate will only be well approximated using asymptotic methods in very small models. To overcome this problem we propose using the same strategy used to overcome the exact goodness-of-fit testing problem, using an RMSEA that uses only bivariate information. We describe the relationship between the full information RMSEA (referred as RMSEA$_n$) and the bivariate RMSEA (referred as RMSEA$_2$) and we examine using simulation studies the extent to which their empirical sampling distribution can be approximated in finite samples. We also examine the issues of what cutoff value to use for the RMSEA$_n$ and RMSEA$_2$ and how to assess the goodness of approximation in models that are so large that the RMSEA$_2$ cannot be computed.

TESTING FOR EXACT FIT IN MULTINOMIAL MODELS

Consider a set of n multinomial variables, such as n test items, Y_1 to Y_n, each with K response alternatives. We assume that all variables have the same number of categories simply to ease the notation and simplify the exposition. Responses to these variables can be placed in a K^n contingency table. The cells of this contingency table will be indexed by $c = 1, \ldots, C = K^n$. The C dimensional column vector of population probabilities will be denoted by $\boldsymbol{\pi}$. Cell proportions (the sample counterpart of $\boldsymbol{\pi}$) based on a sample of size N will be denoted by $\mathbf{p}$. We consider a parametric model for the probability vector that depends on a q-dimensional vector of parameters $\boldsymbol{\theta}$ to be estimated from the data. Generically, we denote such a model as $\boldsymbol{\pi}(\boldsymbol{\theta})$. For example, in the case of the two-parameter logistic model with a standard normal distributed trait (2PLM) widely used in IRT, $q = 2n$ and $\boldsymbol{\theta}$ is the vector of intercepts and slopes, one for each item, and $\boldsymbol{\pi}(\boldsymbol{\theta})$ are the restrictions imposed by the 2PLM on the set of probabilities, one for each possible pattern.

Full Information Statistics

In a test of exact fit we assess the null hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ against the alternative $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$. That is, we assess whether the population probability vector arises exactly from the parametric model $\boldsymbol{\pi}(\boldsymbol{\theta})$ against the alternative that the model is incorrect. The two classical goodness-of-fit statistics for testing this null hypothesis are the likelihood ratio statistic, G^2, and Pearson's X^2 statistic. In scalar form these statistics may be written as $G^2 = 2N\sum_{c=1}^{C} p_c \ln(p_c/\hat{\pi}_c)$,where p_c and $\hat{p}_c = p_c(\hat{\theta})$ denote the observed proportion and estimated probability for cell c, and $X^2 = N\sum_{c=1}^{C}(p_c - \hat{\pi}_c)^2/\hat{p}_c$. In matrix form, Pearson's X^2 statistic can be written as

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

where $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$, $\mathbf{p} - \hat{\boldsymbol{\pi}}$ are the cell residuals, and $\hat{\mathbf{D}} = diag(\hat{\boldsymbol{\pi}})$ is a diagonal matrix of estimated probabilities. If the model parameters have been estimated using the maximum likelihood (ML) method and if the fitted model holds exactly in the population, the empirical distribution of the G^2 and X^2 statistics can be approximated in large samples using a chi-square distribution with $C - q - 1$ degrees of freedom.

Unfortunately it is well known that the p values obtained using this large sample approximation are grossly incorrect except in small models. A useful rule of thumb to determine whether the large sample p values for G^2 and X^2 are reliable is to compare them. If the p values are similar, they are likely to be accurate. If they are slightly different, the p value for X^2 is likely to be the most accurate (Koehler & Larntz, 1980). If they differ widely, it is likely that both p values are incorrect. Regrettably, it is common to find in IRT applications that G^2 yields a p value of 1 and that X^2 yields a p value of 0, which clearly suggests that both p values are grossly incorrect.

In SEM, under correct model specification, asymptotic p values for overall goodness-of-fit statistics may be incorrect if the sample size is not large enough. But most often the asymptotic p values for G^2 and X^2 fail regardless of sample size. To understand why, consider the more accurate of these two statistics, X^2. The empirical variance of X^2 and its variance under its reference asymptotic distribution differ by a term that depends on the inverse of the cell probabilities (Cochran, 1952). As a result, when the cell probabilities become small the discrepancy between the empirical and asymptotic variances of X^2 can be large. The empirical variance of X^2 is larger than the expected variance under the reference chi-square distribution and *p* values for X^2 computed using the reference asymptotic distribution are too small, leading to reject the model. But as the number of cells, *C*, increases, the probabilities must be small as they must add up to one. Thus, for large *C*, small cell probabilities *must* be encountered and *p* values obtained using the reference asymptotic distribution *must* be too small (Bartholomew & Tzamourani, 1999). How large must *C* be for the asymptotic *p* values for X^2 to be useless? As Thissen and Steinberg (1997) put it, when the number of categories is five or more, the approximation becomes invalid for any model as soon as the number of variables is greater than six "with any conceivable sample size" (p. 61). This is what has been referred to in the categorical data literature as the problem of sparse expected tables (or sparseness, for short) and some guidelines on what the expected counts, Np_c, should be for the asymptotic approximation to X^2 to be accurate have been offered. This is the reason in typical IRT applications the asymptotic approximation to the empirical distribution of X^2 will fail—regardless of sample size. Of course, in nonsparse conditions, for any model $\boldsymbol{\pi}(\boldsymbol{\theta})$ of a given size, *C*, the accuracy of the asymptotic *p* values for X^2 will also depend on sample size, *N*, just as in SEM.

G^2 and X^2 can be called full information statistics in the sense that they use all the information available in the data to test the model. How can we obtain a statistic whose distribution can be well approximated by asymptotic methods in sparse situations, that is, when the number of cells *C* is large? By using limited information test statistics, that is, by using statistics that only use a limited amount of the information available in the data. These statistics pool, with overlap, the cells of the contingency table using the multidimensional structure of the data. The distribution of limited information statistics is better approximated by asymptotic methods in large models than for full information statistics because pooled cells must have higher expected probabilities. Also, because limited information statistics concentrate the available information, they can be more powerful, perhaps surprisingly, than full information statistics. Maydeu-Olivares and Joe (2005, 2006) provided a framework that unifies limited and full information testing in multivariate discrete data to which we now turn.

Limited Information Statistics

Perhaps the best way to understand limited information testing is by realizing that a model for a population contingency table admits at least two representations. One of them uses cell probabilities. The other representation uses moments.

Consider the smallest multivariate categorical data problem, a 2 × 2 table arising from two binary variables each coded as $\{0, 1\}$. The cell representation uses four cell probabilities that must add up to one. The alternative representation uses three moments: the two means, $\pi_1^1 = \Pr(Y_1 = 1)$ and $\pi_2^1 = \Pr(Y_2 = 1)$, and the cross product $\pi_{12}^{11} = \Pr(Y_1 = 1, Y_2 = 1)$. Both representations are depicted here.

cells

Y_1 \ Y_2	0	1
0	π_{00}	π_{01}
1	π_{10}	π_{11}

margins

Y_1 \ Y_2	0	1	
0			
1		π_{12}^{11}	π_1^1
		π_2^1	

It is obvious that the relationship is one-to-one and invertible. One can always go from one representation to the other regardless of the number of binary variables involved.

The same is true for contingency tables involving polytomous variables where not necessarily all variables consist of the same number of alternatives. This is shown here, again for the simplest case, a 2 × 3 table. In this case there are six cell probabilities, which must add up to one. The alternative representation uses five moments: three univariate moments, $\pi_1^1 = \Pr(Y_1 = 1)$, $\pi_2^1 = \Pr(Y_2 = 1)$, and $\pi_2^2 = \Pr(Y_2 = 2)$, and two bivariate moments, $\pi_{12}^{11} = \Pr(Y_1 = 1, Y_2 = 1)$ and $\pi_{12}^{12} = \Pr(Y_1 = 1, Y_2 = 2)$. Note that the moments are simply the marginal probabilities that do not involve category 0.

cells

Y_1 \ Y_2	0	1	2
0	π_{00}	π_{01}	π_{02}
1	π_{10}	π_{11}	π_{12}

margins

Y_1 \ Y_2	0	1	2	
0				
1		π_{12}^{11}	π_{12}^{12}	π_1^1
		π_2^1	π_2^2	

We use $\dot{\boldsymbol{\pi}}_1$ to denote the set of all univariate moments, $\dot{\boldsymbol{\pi}}_2$ to denote the set of bivariate moments, and so forth. Also, we use $\boldsymbol{\pi}_r$ to denote the column vector of all population moments up to order *r*; that is, $\boldsymbol{\pi}'_r = (\dot{\boldsymbol{\pi}}'_1, \dot{\boldsymbol{\pi}}'_2, \ldots, \dot{\boldsymbol{\pi}}'_r)$. Its sample counterpart (marginal proportions) is denoted by $\mathbf{p}_r$. Later on we describe in more detail why these quantities are moments.

Maydeu-Olivares and Joe (2005, 2006) proposed the family of test statistics M_r that are simply quadratic forms in residual moments. More specifically, the family of test statistics M_r is

$$M_r = N\,(\mathbf{p}_r - \hat{\boldsymbol{\pi}}_r)'\,\hat{\mathbf{C}}_r\,(\mathbf{p}_r - \hat{\boldsymbol{\pi}}_r)\,. \tag{2}$$

The weight matrix in Equation (2) is

$$\begin{aligned}\mathbf{C}_r &= \boldsymbol{\Xi}_r^{-1} - \boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r(\boldsymbol{\Delta}'_r\boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r)^{-1}\boldsymbol{\Delta}'_r\boldsymbol{\Xi}_r^{-1}\\ &= \boldsymbol{\Delta}_r^{(c)}(\boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r^{(c)})^{-1}\boldsymbol{\Delta}_r^{(c)\prime},\end{aligned} \tag{3}$$

evaluated at the parameter estimates. In Equation (3), $\boldsymbol{\Delta}_r$ denotes the matrix of derivatives of the moments up to order *r* with respect to the parameter vector $\boldsymbol{\theta}$, and $\boldsymbol{\Xi}_r$ denotes *N* times the asymptotic covariance matrix of the sample moments up to order *r* evaluated at the parameter estimates $\hat{\boldsymbol{\theta}}$.

$\mathbf{\Delta}_r^{(c)}$ is simply a matrix such that $\mathbf{\Delta}_r^{(c)\prime}\mathbf{\Delta}_r = \mathbf{0}$. See the Appendix for further details.

M_r is a family of test statistics: $M_1, M_2, \ldots,$ up to M_n. In M_1 only means are used. In M_2, only means and cross products are used, and so forth, up to M_n where up to n-way moments are used. When all moments are used (i.e., when M_n is used) a statistic that uses all the information available in the data (i.e., a full information statistic) is obtained. Furthermore, for the ML estimator, M_n equals Pearson's X^2 algebraically. On the other hand, when $r < n$, the statistics M_r are limited information statistics.

If the model is identified from moments up to order r (i.e., if it could be estimated using only the sample moments up to order r), and if the model holds exactly in the population, the empirical distribution of M_r can be approximated asymptotically using a chi-square distribution with $df_r = s_r - q$ degrees of freedom, where when all items consist of the same number of categories, K, $s_r = \sum_{i=1}^{r} \binom{n}{i}(K-1)^i$. For routine applications, Maydeu-Olivares and Joe (2005, 2006) proposed using M_2. This is simply a quadratic form in residual means and cross products with $df_2 = n(K-1) + \frac{n(n-1)}{2}(K-1)^2 - q$ degrees of freedom.

Why use only bivariate information for testing? Maydeu-Olivares and Joe (2005) showed that the empirical variance of M_r and its variance under its reference asymptotic distribution differ by a term that depends on the inverse of the marginal probabilities of order $2r$ (or n, should $n < 2r$). Thus, if for instance M_2 is employed, the accuracy of the asymptotic approximation depends at most on four-way marginal probabilities. In contrast, the accuracy of the asymptotic approximation to M_3 depends on up to six-way marginal probabilities. This means that the accuracy of the asymptotic approximation to M_r improves with decreasing r (because, for instance, four-way probabilities are larger than six-way probabilities). Consequently, they recommended testing using the lowest amount of information possible. As most IRT models are identified using only univariate and bivariate information, they recommended using M_2 for general testing in IRT. Consistent with asymptotic theory, Maydeu-Olivares and Joe (2005, 2006) showed that the accuracy of the asymptotic approximation to the sampling distribution of M_r statistics worsened as model size increased, sample size decreased, and r increased. Yet, M_2 yielded accurate p values even for the largest model considered, a graded response model (Samejima, 1969) for 10 variables each with five response categories, and the smallest sample considered, $N =$ 300 observations. Despite that this is a small model for IRT standards, the number of cells of the contingency table, C, is close to 10 million, and there are almost as many degrees of freedom available for full information testing.

Clearly, it is unlikely that any restricted model will fit exactly such large contingency tables in applications. What is needed, paraphrasing Browne and Cudeck, (1993, pp. 137–138) is a procedure for assessing how well a model with unknown, but optimally chosen, parameter values approximates the population probability vector if it were available. Also, following Steiger (1990), we may wish to take into account model complexity, as models with more parameters are likely to approximate the population probability vector better than less parameterized models. Finally, we need to determine how precisely we have determined the goodness of approximation from our sample data (Steiger, 1990).

FULL INFORMATION GOODNESS OF APPROXIMATION

As before, we consider fitting a model for a C-dimensional probability vector expressed as a function of q parameters $\boldsymbol{\pi}_0 = \boldsymbol{\pi}(\boldsymbol{\theta})$. We refer to $\boldsymbol{\pi}_0$ as the fitted model and also as the null model because it is the model specified in the null hypothesis of exact fit. Now, suppose that the null model is not the data-generating model. Rather, the population probability vector that generated the data is $\boldsymbol{\pi}_T$. One way to assess the discrepancy between the population probability vector and the null model is by using the Mahalanobis distance, D_n, between them, where

$$D_n = (\boldsymbol{\pi}_T - \boldsymbol{\pi}_0)' \mathbf{D}_0^{-1} (\boldsymbol{\pi}_T - \boldsymbol{\pi}_0). \tag{4}$$

D_n is a discrepancy due to approximation between a population probability vector and the null model. It is also the distance in Pearson's X^2 metric between them. Notice that it is a population quantity as there are no data involved in the expression.

The Mahalanobis distance D_n will generally decrease when parameters are added to the null model. As a result, if model parsimony is of concern, we may wish to use instead, following Steiger (1990) and Browne and Cudeck (1993), a Root Mean Square Error of Approximation, RMSEA$_n$,

$$\varepsilon_n = \sqrt{\frac{D_n}{df}} \tag{5}$$

as a measure of the discrepancy due to approximation per degree of freedom.

Like for any other parameter, estimates of the RMSEA$_n$ are subject to sampling fluctuations. We can convey the precision with which the RMSEA$_n$ is estimated by providing a confidence interval for its population value (with say 90% confidence). Also, we may wish to test whether the null model is a good enough approximation to the population probability vector. That is, we could use the following test of close fit,

$$H_0^* : \varepsilon_n \leq c_n \text{ vs. } H_1^* : \varepsilon_n > c_n, \tag{6}$$

where c_n is some cutoff value. Notice that if there is no error of approximation, that is, if $D_n = 0$, then $\varepsilon_n = 0$. Thus, testing $H_0^* : \varepsilon_n = 0$ vs. $H_1^* : \varepsilon_n > 0$ is equivalent to the usual test of exact model fit $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ vs. $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}_0$.

Now, how can we estimate the RMSEA$_n$ from data? How can we obtain a confidence interval for it? And, if we wish to do so, how can we obtain a p value for the test of close fit in Equation (6)? As a special case of the results in the next section, if the model parameters $\boldsymbol{\theta}$ have been estimated by ML, and under an assumption of a sequence of local alternatives, the RMSEA$_n$ given by Equation (5) can be estimated using Pearson's X^2 and its degrees of freedom (df) using

$$\hat{\varepsilon}_n = \sqrt{\text{Max}\left(\frac{\hat{X}^2 - df}{N \times df}, 0\right)}, \tag{7}$$

where $\hat{X}^2$ is the observed value of the X^2 statistic for the data set, and we have taken into account that $\hat{X}^2 - df$, with $df = C - q - 1$, may be negative in applications. Also, a 90% confidence interval for ε_n is given by

$$\left(\sqrt{\frac{\hat{L}}{N \times df}}; \sqrt{\frac{\hat{U}}{N \times df}}\right), \tag{8}$$

where $\hat{L}$ and $\hat{U}$ are the solution to

$$F_{\chi^2}(\hat{X}^2; df, \hat{L}) = 0.95, \quad \text{and} \quad F_{\chi^2}(\hat{X}^2; df, \hat{U}) = 0.05, \tag{9}$$

respectively, where $F_{\chi^2}(\cdot\,; df, \lambda)$ is the noncentral chi-square distribution function with df degrees of freedom and noncentrality parameter λ. Note that $F_{\chi^2}(\cdot\,; df, \lambda)$ is nondecreasing as λ increases, and $F_{\chi^2}(\hat{X}^2; df, 0) > 0.95$ if the p value of the chi-square test statistic is less than .05. In this case, the asymptotic p value for the test of close fit in Equation (6) is

$$p = 1 - F_{\chi^2}\left(\hat{X}^2; df, N \times df \times c_n^2\right). \tag{10}$$

Now, the accuracy of the confidence intervals for the RMSEA$_n$ and the accuracy of the p value in Equation (10) for the test of close fit depend on the accuracy of the noncentral chi-square approximation to the distribution of X^2 under a sequence of local alternatives assumption. Unfortunately, this approximation suffers from the same problems discussed previously for the asymptotic approximation of X^2 under the null hypothesis of exact fit. If the sampling variability of X^2 when the model holds exactly is underestimated by its reference (central) chi-square distribution, its variability under a sequence of local alternatives will be underestimated under its reference noncentral chi-square distribution as well. This implies that the except in nonsparse tables, the confidence intervals obtained using asymptotic methods will underestimate the true variability of the RMSEA estimate in Equation (7).

What is needed is a statistic for assessing goodness of approximation whose precision can be well estimated in practice. For testing exact fit, a solution to this problem was obtained by considering the family of limited information statistics M_r, and using the statistic within this family that is best approximated by its reference chi-square distribution, M_2. The same approach can be used to obtain a test of approximate fit, and more generally, to assess the goodness of approximation of categorical data models. This leads us to consider a family of limited information population discrepancies between the population probability vector and the null model as well as a family of limited information RMSEAs. Also, unlike the results presented in this section, which apply to models estimated by ML, results in the next section are applicable to any consistent and asymptotically normal estimator. This includes, among others, the ML estimator, estimators based on polychorics (e.g., Muthén, 1993), and the pairwise likelihood estimators of Katsikatsou, Moustaki, Yang-Wallentin, and Jöreskog (2012).

A FAMILY OF LIMITED INFORMATION RMSEAs

Consider the family of discrepancies between the population probabilities and null model given by

$$D_r = \left(\boldsymbol{\pi}_r^T - \boldsymbol{\pi}_r^0\right)' \mathbf{C}_r^0 \left(\boldsymbol{\pi}_r^T - \boldsymbol{\pi}_r^0\right) \tag{11}$$

with $\boldsymbol{\pi}_r^0$ being the moments up to order r under the null model, $\boldsymbol{\pi}_r^T$ being the moments implied by the population probability vector, and $\mathbf{C}_r^0$ being Equation (3) based on the null model. D_1 is the population discrepancy between the univariate moments under the population probability vector and the null model; D_2 is the population discrepancy between univariate and bivariate moments; and so forth up to D_n, a population discrepancy involving all moments.

The null model $\boldsymbol{\pi}_0$ here corresponds to the value of $\boldsymbol{\theta}_0$ that minimizes the Kullback-Leibler (KL) discrepancy between the population probabilities and the model specified under the null hypothesis. That is, the vector $\boldsymbol{\pi}(\boldsymbol{\theta}_0)$ minimizes

$$\begin{aligned} D_{KL}(\boldsymbol{\pi}_T, \boldsymbol{\pi}(\boldsymbol{\theta}_0)) &= \boldsymbol{\pi}_T' \ln(\boldsymbol{\pi}_T / \boldsymbol{\pi}(\boldsymbol{\theta}_0)) \\ &= \boldsymbol{\pi}_T' \left[\ln(\boldsymbol{\pi}_T) - \ln(\boldsymbol{\pi}(\boldsymbol{\theta}_0))\right]. \end{aligned} \tag{12}$$

Then this special case of Equation (11) when all moments are used is algebraically equal to the expression for D_n given in Equation (4).

Taking into account model parsimony, Equation (11) leads to a family of Root Mean Square Error of Approximation RMSEA$_r$'s given by

$$\varepsilon_r = \sqrt{\frac{D_r}{df_r}}, \tag{13}$$

where $df_r = s_r - q$ denotes the degrees of freedom available for testing when only up to rth-way moments used.

For any consistent and asymptotically normal estimator, $\hat{\boldsymbol{\theta}}$, we show in the Appendix that under a sequence of local alternatives, an estimate of the RMSEA$_r$ is

$$\hat{\varepsilon}_r = \sqrt{\text{Max}\left(\frac{\hat{M}_r - df_r}{N \times df_r}, 0\right)}, \tag{14}$$

where $\hat{M}_r$ is the observed value of the M_r statistic for the data set. Also, a 90% confidence interval for e_r is given by

$$\left(\sqrt{\frac{\hat{L}_r}{N \times df_r}}; \qquad \sqrt{\frac{\hat{U}_r}{N \times df_r}}\right), \tag{15}$$

where $\hat{L}_r$ and $\hat{U}_r$ are the solution to

$$F_{\chi^2}(\hat{M}_r; df_r, \hat{L}_r) = 0.95, \quad \text{and} \quad F_{\chi^2}(\hat{M}_r; df_r, \hat{U}_r) = 0.05, \tag{16}$$

respectively, assuming $F_{\chi^2}(\hat{M}_r; df_r, 0) > 0.95$.

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

$$H_0^* : \varepsilon_r \leq c_r \text{ vs. } H_1^* : \varepsilon_r > c_r, \tag{17}$$

where c_r is an arbitrary cutoff value that depends on r, the highest level of association used. p values for Equation (17) are obtained using

$$p = 1 - F_{\chi^2}\left(\hat{M}_r; df_r, N \times df_r \times c_r^2\right). \tag{18}$$

Equation (17) defines a family of tests of close fit. Which member of this family should be used? We advocate using the member of this family for which we can obtain a more accurate p value. Consequently, from extant theory we advocate testing using the smallest r at which the model is identified, generally two. That is, we recommend using RMSEA$_2$ in applications, the RMSEA statistic obtained from the M_2 statistic using Equation (14).

THE SAMPLING DISTRIBUTION OF THE SAMPLE RMSEA$_2$ AND RMSEA$_n$

To show that the distribution of the RMSEA$_2$ can be well approximated using asymptotic methods even in large models and small samples, whereas the distribution of the RMSEA$_n$ can only be well approximated in small models, we report the results of a small simulation study using IRT models estimated by (marginal) ML.

Binary data were generated using a two-parameter logistic model (2PLM)

$$\Pr(Y_i = 1|\eta) = \frac{1}{1 + \exp\left[\alpha_i + \beta_{\mathrm{i}}\eta\right]}, \quad i = 1, \ldots n \tag{19}$$

with a standard normal latent trait η and a one-parameter logistic model (1PLM) was fitted. The 1PLM is obtained by setting all slopes β_i equal. Six conditions were investigated. The six conditions were obtained by crossing two model sizes (n = 5, 10) and three sample sizes (N = 100, 500, 3,000). One thousand replications per condition were used. The intercepts and slopes for the five item condition were

$$\boldsymbol{\alpha}' = (-1, -0.5, 0, 0.5, 1), \quad \text{and} \quad \boldsymbol{\beta}' = (0.6, 1, 1.7, 1, 0.6), \tag{20}$$

respectively. In the 10-item condition these values were simply duplicated. For each replication, two tests of close fit in Equation (17) were performed: $H_0^* : \varepsilon_2 \leq c_2$ vs. $H_1^* : \varepsilon_2 > c_2$ and $H_0^* : \varepsilon_n \leq c_n$ vs. $H_1^* : \varepsilon_n > c_n$. The cutoff criteria c_2 and c_n were set equal to the population RMSEA$_2$ and RMSEA$_n$ for convenience, as with this choice the expected rejection rates need not be computed. Thus, for each replication, data were generated using a 2PLM with parameter values in Equation (20) and a 1PLM was fitted by ML. The sample RMSEA$_2$ and RMSEA$_n$ were computed using Equations (7) and (14), respectively, and p values for the test of close fit with values c_2 and c_n were computed using Equations (10) and (18). For testing the RMSEA$_n$ there are 25 *df* when $n = 5$ and 1,012 *df* when $n = 10$. In contrast, for testing the RMSEA$_2$ there are 9 *df* when $n = 5$ and 44 *df* when $n = 10$.

The population RMSEAs were computed by choosing the 1PLM parameter vector that minimized the KL function in Equation (12). Minimizing Equation (12) is equivalent to using a multinomial ML discrepancy function between the population probability vector and the 1PLM specified under the null hypothesis (Maydeu-Olivares & Montaño, 2013; Reiser, 2008; see also Jöreskog, 1994). More specifically, the procedure used to compute the population RMSEAs was as follows: First, cell probabilities under the 2PLM with parameters in Equation (20) were computed. This is the probability vector $\boldsymbol{\pi}_T$. These probabilities were then treated as if they were sample proportions and a 1PLM was fitted by ML. This yields the 1PLM parameter vector $\boldsymbol{\theta}_0$ closest to the population 2PLM in the KL metric in Equation (12). Using these 1PLM parameters the fitted probability vector $\boldsymbol{\pi}_0 = \boldsymbol{\pi}(\boldsymbol{\theta}_0)$ and the population RMSEA$_2$ and RMSEA$_n$ parameters are computed. Using this procedure, we found that for $n = 5$, the population RMSEA$_n$ = 0. 0306 and RMSEA$_2$ = 0.0509; for $n = 10$, RMSEA$_n$ = 0. 0098 and RMSEA$_2$ = 0.04654.

Table 1 reports the results of the simulation: expected rates under the null hypothesis that the RMSEAs equal their population values and empirical rejection rates across 1,000 replications. The results are as expected from asymptotic theory: The sampling distribution of the full information RMSEA$_n$ based on Pearson's X^2 is well approximated by asymptotic methods when the model is small ($n = 5$) but not when the model is so large than some probabilities under the fitted model become too small ($n = 10$). In contrast, the sampling distribution of the bivariate information RMSEA$_2$ based on M_2 is well approximated by asymptotic methods under all conditions, even for a sample size of 100. A sample size of 100 also suffices to approximate the distribution of the RMSEA$_n$ (based on X^2) with $n = 5$.

For our choice of population values and fitted models and for these particular parameter values, (a) the values for the population bivariate RMSEA are larger than the values of the full information RMSEA, and (b) the values of the population RMSEA$_2$ and RMSEA$_n$ decrease as the number of variables increases. Are these findings typical? To address

TABLE 1
Empirical Rejection Rates for Tests of RMSEA_2 and RMSEA_n Equal to Their Population Values

Stat	n	N	1%	5%	10%	20%
RMSEA_n	5	100	2.5	8.1	13.2	22.5
	5	500	1.0	6.1	11.3	22.1
	5	3000	1.7	6.0	12.9	22.8
	10	100	30.8	37.6	40.8	47.1
	10	500	26.8	39.4	48.5	56.0
	10	3,000	16.9	34.8	44.7	57.3
RMSEA_2	5	100	1.8	6.8	11.7	23.0
	5	500	0.9	5.2	9.6	19.4
	5	3,000	1.1	5.8	10.8	21.5
	10	100	1.3	5.4	10.0	19.5
	10	500	1.5	5.8	10.0	20.1
	10	3,000	1.1	4.9	10.3	20.8

Note. Data were generated according to a two parameter logistic model with a normally distributed latent trait and a one parameter logistic model was fitted by maximum likelihood (ML). 1,000 replications were used. The population RMSEA_n for $n = 5$, 10 are 0.0306 and 0.0098. The population RMSEA_2 for $n =$ 5, 10 are 0.0509 and 0.0465. Degrees of freedom for RMSEA_n are 25 and 1012 for $n = 5$, 10. Degrees of freedom for RMSEA_2 are 9 and 44 for $n = 5$, 10.

this issue we examine in the next section how the population RMSEA_2 and RMSEA_n change for different configurations of model misspecification in IRT models as well as how they change as the number of variables increases. We also address in this section the issue of what cutoff values could be used for RMSEA_2 and RMSEA_n, that is, what criteria could be used to determine that the fit of a model is "close."

CHOICE OF RMSEA CUTOFF VALUES IN IRT MODELS

Binary Data

We used the procedure described in the previous section to compute the population RMSEA_2 and RMSEA_n values when the population probabilities arise from a bidimensional three-parameter logistic model (3PLM) with standard normal latent traits and unidimensional 1PLM and 2PLM were used as null models. The item response function for this 3PLM is

$$\Pr(Y_i = 1|\eta_1, \eta_2) = c_i + \frac{1 - c_i}{1 + \exp\left[-\left(\alpha_i + \beta_{i1}\eta_1 + \beta_{i2}\eta_2\right)\right]}, \quad i = 1, \ldots n, \tag{21}$$

where c_i denotes the "guessing" parameter. Seventy-two conditions were obtained by crossing (a) two null models (unidimensional 1PLM and 2PLM), (b) four levels for the correlation between the latent traits ($\rho = 0$, 0.3, 0.6, 1), (c) three levels of the "guessing" parameter ($c = 0$, 0.1, 0.2), and (d) three levels of model size ($n = 6, 8, 10$). An independent clusters configuration was used for the bidimensional 3PLM: $\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\beta} \end{pmatrix}$. The population parameters were as follows: for $n = 6$, $\boldsymbol{\alpha} = (1.19, 0, –1.19)'$—duplicated twice, and $\boldsymbol{\beta} = (1.67, 2.27, 1.67)'$ for $n = 8$, $\boldsymbol{\alpha} = (2.38, 1.42, –1.42, –2.38)'$—duplicated twice, and $\boldsymbol{\beta} = (1.67, 2.27, 2.27, 1.67)'$ and for $n = 10$, $\boldsymbol{\alpha} = (2.38\ 1.42, 0, –1.42, –2.38)'$—duplicated twice, and $\boldsymbol{\beta} = (1.67, 2.27, 1.28, 2.27, 1.67)'$. The c parameters were set equal for all items. In the metric of the normal ogive IRT model (also known as ordinal factor analysis), these parameter values correspond to the following thresholds and factor loadings (e.g., Flora & Curran, 2004; Forero & Maydeu-Olivares, 2009): for $n = 6$, $\boldsymbol{\tau} = (–0.5, 0, 0.5)'$, $\boldsymbol{\lambda} = (0.7, 0.8, 0.7)'$; for $n = 8$, $\boldsymbol{\tau} = (–1, –0.5, 0.5, 1)'$, $\boldsymbol{\lambda} = (0.7, 0.8, 0.8, 0.7)'$; and for $n = 10$, $\boldsymbol{\tau} = (–1, –0.5, 0', 0.5, 1)'$, $\boldsymbol{\lambda} = (0.7, 0.8, 0.6, 0.8, 0.7)'$.[1]

When the null model is a 1PLM, population RMSEA_n values ranged from 0.008 to 0.105 with a median of 0.026, whereas population RMSEA_2 values ranged from 0.026 to 0.191 with a median of 0.073. When the null model is a 2PLM, population RMSEA_n values ranged from 0 to 0.110 with a median of 0.025, whereas population RMSEA_2 values ranged[2] from 0 to 0.238 with a median of 0.085. Of the 72 conditions investigated only in 2 of them the RMSEA_n was larger than the RMSEA_2 and the difference was less than 0.01. These were unidimensional models with $n = 6$. In all other conditions the RMSEA_2 was larger than the RMSEA_n and the largest difference was 0.13 (when $n = 10$, $c = 0$, and $\rho = 0$).[3] Furthermore, because the 1PLM is a special case of the 2PLM, the distance between the population probability

[1] Ten additional sets of 72 RMSEA_2 and RMSEA_n where obtained drawing the model parameter values at random from a uniform distribution. Intercepts were drawn between –2.38 and 2.38, slopes were drawn between 1.28 and 2.27, correlations were drawn between 0 and 1, and guessing parameters between 0 and 0.2. Results similar to the ones reported here for a single fixed set of parameters were found.

[2] Because the data-generating model is a 3PLM involving two dimensions and the matrix of slopes has an independent clusters structure, when $c = 0$ and $\rho = 1$, the 2PL is correctly specified and the population RMSEA_n and RMSEA_2 equal 0.

[3] The largest differences between RMSEA_2 and RMSEA_n were found whenever $c = 0$, $\rho = 0$, and a 2PL was fitted; the difference in all three cases ($n = 6, 8, 10$) was at least 0.12.

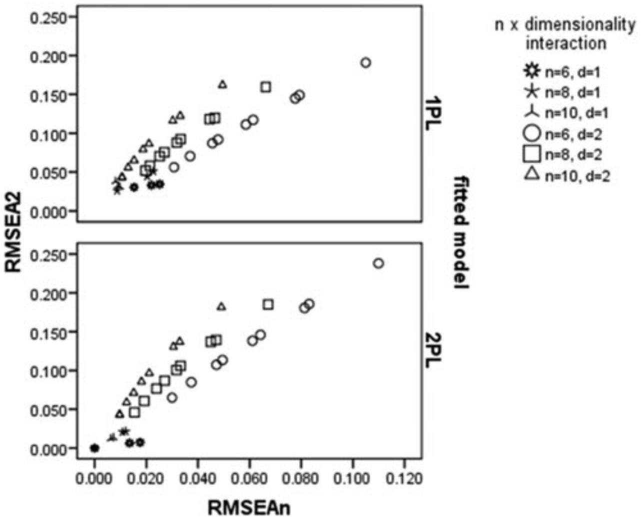

FIGURE 1 Plot of RMSEA$_2$ values as a function RMSEA$_n$, number of variables, latent trait dimensionality, and fitted model.

vector and the model implied by the null hypothesis must be smaller for the 2PLM than for the 1PLM, but also degrees of freedom must be smaller for the 2PLM. As a result, because in the RMSEAs this distance is divided by the degrees of freedom, the RMSEAs for the 2PLM need not be smaller than for the 1PLM.

Figure 1 displays the population (RMSEA$_n$, RMSEA$_2$) pairs for each of the 72 conditions, separately for each null model (1PLM or 2PLM). We see several interesting patterns in Figure 1. First, the relationship between RMSEA$_n$ and RMSEA$_2$ values depends on dimensionality and model size (n). Second, for the 2PLM, the values of the RMSEAs are lower for unidimensional models than for bidimensional models. However, for the 1PLM some bidimensional models yield RMSEA values lower than some unidimensional models. Third, for bidimensional models, RMSEA$_2$ values can be well predicted from RMSEA$_n$ values and n with R^2 of over 99% when a quadratic regression model is used. The curvature of the quadratic relationship increases as n increases; for $n = 6$ a linear model yields an R^2 of one.

The relationship between RMSEA values and model size is more clearly seen in Figure 2. In this figure, RMSEA$_n$ and RMSEA$_2$ values are displayed separately as a function of number of variables (n), correlation between the traits (ρ), guessing parameter (c), latent trait dimensionality (d), and null model (1PLM and 2PLM). Each line in this figure joins three values of RMSEA: for n = 6, 8, and 10 and a particular $\rho \times c$ combination. Figure 2 reveals that RMSEA$_n$ decreases as n increases. We also see in this figure that the RMSEA$_2$ decreases as n increases but that it appears to asymptote at $n = 8$ for both the 1PLM and 2PLM. The reason for this behavior is as follows: The value of the full information noncentrality parameter D_n increases as n increases, but the full information degrees of freedom $(2^n - q - 1)$ increase at a faster rate than D_n as n increases. As a result, RMSEA$_n$ decreases as n increases. In contrast, bivariate information degrees of freedom $n(n + 1)/2 - q$ increase at a slower rate than full information degrees of freedom as n increases. Therefore, for a given model for multidimensional multinomial data, RMSEA$_2$ may asymptote as n increases.

The results obtained reveal the difficulty of specifying cutoff criteria for multinomial RMSEAs. First, a different criterion must be employed for RMSEA$_n$ and RMSEA$_2$. Second, the cutoff may depend on model size (n). Finally, the cutoff may depend on the population probabilities and the null model. Yet, it can be easily argued that latent trait dimensionality is the most important substantive consideration when determining the fit of an IRT model. That is, we may be

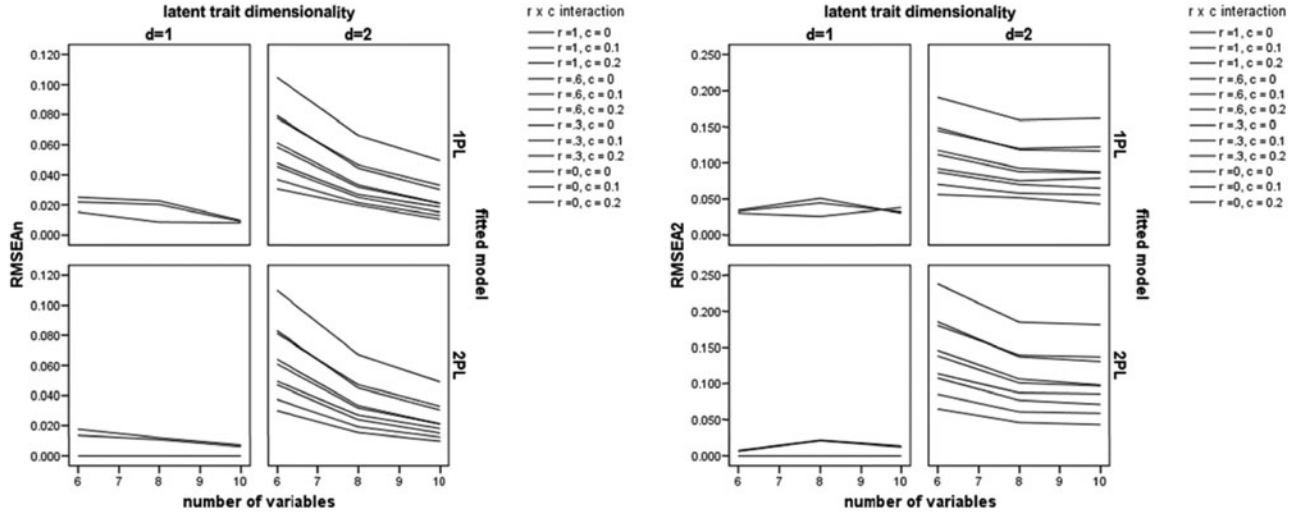

FIGURE 2 Plot of RMSEA$_n$ and RMSEA$_2$ values as a function of number of variables, correlation between the traits (ρ), guessing parameter (c), latent trait dimensionality, and fitted model. Each line joints three values.

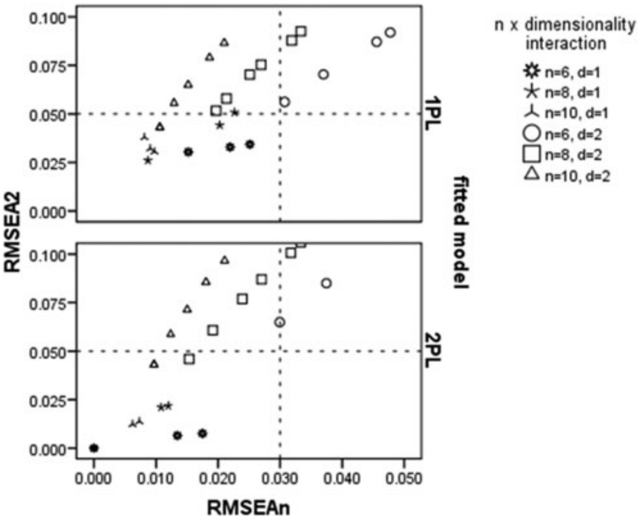

FIGURE 3 Plot of RMSEA$_2$ values as a function RMSEA$_n$, number of variables, latent trait dimensionality, and fitted model in the range RMSEA$_2$ < 0.10 and RMSEA$_n$ < 0.05.

willing to retain a misspecified IRT model if the latent trait dimensionality is correctly specified whereas we are unlikely to retain it if the latent trait dimensionality is misspecified. Interestingly, the results shown in Figure 1 reveal that for the population probabilities and null models considered (IRT models for binary data) the main driver of the value of the RMSEAs is the latent trait dimensionality in the population probabilities. Thus, it appears to be possible to establish cutoff criteria for the RMSEAs to separate incorrectly specified binary IRT models with a single latent trait from models involving more than one latent trait. To do so, in Figure 3 we zoom in the results presented in Figure 1 to examine in detail the area where the RMSEAs approach 0. We see in this figure that a cutoff criterion of close fit of RMSEA$_2$ $\leq$ 0.05 will retain most misspecified unidimensional 1PLM and 2PLM while rejecting most misspecified bidimensional models. Furthermore, this cutoff does not depend on model size as the largest RMSEA$_2$ values for unidimensional models are found when $n = 8$ and not for the largest model. Thus, a value of 0.05 for the bivariate RMSEA seems a reasonable cutoff criterion for close fit in binary IRT models: models with a RMSEA$_2$ above this cutoff are likely to have the wrong latent trait dimensionality, whereas models below this cutoff are likely to have the correct latent trait dimensionality. Furthermore, note than when a 2PLM is fitted all unidimensional IRT models yield a population RMSEA$_2$ values less than 0.03. A smaller cutoff value should be used for the full information RMSEA as RMSEA$_n$ is most often smaller than the corresponding bivariate RMSEA. The results shown in Figure 3 suggest that if the criterion is to retain misspecified unidimensional IRT models, a value of 0.03 for RMSEA$_n$ seems a reasonable cutoff criterion for close fit when full information testing is used.

Polytomous Data

To investigate what population values may be expected in the polytomous case, we computed population RMSEA$_2$ and RMSEA$_n$ values when the population probabilities conform to a bidimensional logistic graded response model (GRM) with standard normal latent traits. This is a suitable IRT model for ordinal responses (Maydeu-Olivares, 2005). The item response function for this model is

$$\Pr(Y_i = k \,|\eta_1,\eta_{\mathbf{2}}) = \begin{cases} 1 - \Psi_{i,1} & \text{if} \quad k = 0 \\ \Psi_{i,k} - \Psi_{i,k+1} & \text{if } 0 < k < K - 1 \\ \Psi_{i,K-1} & \text{if} \quad k = K - 1 \end{cases}, \tag{22}$$

$$\Psi_{i,k} = \frac{1}{1 + \exp[-(\alpha_{i,k} + \beta_{i1}\eta_1 + \beta_{i2}\eta_2)]}.\ i = 1, \ldots n. \tag{23}$$

This model reduces to the 2PLM in the binary case. Forty-eight conditions were obtained by crossing (a) four levels for the correlation between the latent traits (ρ = 0.6, 0.7, 0.8, 0.9), (b) four levels of model size (n = 6, 8, 10, 12), and (c) three levels of number of categories (K = 2, 3, 4). The same configuration of slopes used in the binary case was used. For each condition, the population values of the intercepts were set equal across items. The intercepts used were for $K = 4$, $\boldsymbol{\alpha}_i$ = (1.42, 0, –1.42); for $K = 3$, $\boldsymbol{\alpha}_i$ = (1.42, –1.42); and for $K = 2$, $\boldsymbol{\alpha}_i$ = (0). In all conditions, the null model was a unidimensional GRM (Samejima, 1969).

Figure 4 displays the population RMSEA$_n$ and RMSEA$_2$ values as a function of the number of variables, the correlation between the traits, and the number of categories. RMSEA$_n$ values were only computed for up to 10 variables due to the size of the model involved when $K = 4$. We see in this figure that in addition to depending on the intertrait correlation, the population RMSEA$_n$ depends strongly on the number of variables (the larger, the smaller the RMSEA$_n$) but even more strongly on the number of categories (the larger, the smaller the RMSEA$_n$). In contrast, we see that the RMSEA$_2$ is relatively unaffected by the number of variables when the amount of model misspecification is small (i.e., large ρ). However, the value of the population RMSEA$_2$ depends on the number of categories although less so than the RMSEA$_n$. If an RMSEA$_2$ less than or equal to 0.05 were used as cutoff value for close fit, all models with an intertrait correlation larger than 0.6 would be judged to be a close fit for $K = 4$; but for $K = 3$, only models with a correlation larger than 0.7 would be retained; and for K = 2, only models with a correlation larger than 0.9 would be judged to be a close fit.

However, because for small levels of model misspecification (e.g., $\rho = 0.9$) the population RMSEA$_2$ appears to be robust to the effect of the number of variables, it may be feasible to provide a common cutoff for a RMSEA$_2$ adjusted by number of categories. More specifically, let the adjusted RMSEA$_2$ be (K – 1) × RMSEA$_2$. Values of this adjusted

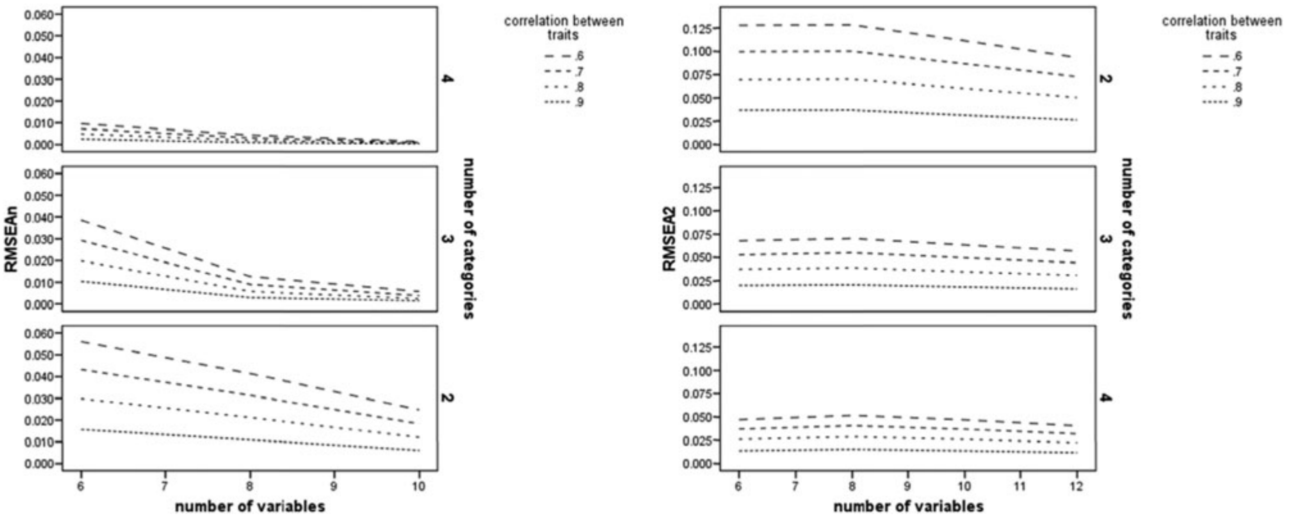

FIGURE 4 Plot of RMSEA$_n$ and RMSEA$_2$ values as a function of number of variables, correlation between the traits, and number of categories.

RMSEA$_2$ are plotted in Figure 5 as a function of the number of variables, the correlation between the traits, and the number of categories. We see that the values of this adjusted RMSEA$_2$ are relatively stable across number of categories and items for small levels of model misspecification. Thus, if c_2 is the cutoff value for close fit when the RMSEA$_2$ is used with binary data, we suggest using as cutoff of excellent fit $c_2/(K-1)$ for $K \geq 2$. It does not appear possible to offer a similar set of cutoffs for the RMSEA$_n$ as its values also depend on the number of variables.

All in all, the results illustrate the difficulty of choosing a cutoff criterion of close fit. Even if we circumscribe ourselves to a set of IRT models, population RMSEA$_2$ values depend on the number of categories of the data: the more categories, the smaller the RMSEA$_2$ population value. This is because the RMSEAs do not adjust for model size. As model size increases (either because the number of categories increases or because the number of variables increases) the noncentrality parameter increases. The RMSEAs adjust the noncentrality parameter by degrees of freedom to penalize models with too many parameters, but they do not adjust for model size. As a result, RMSEAs will tend to decrease as model size increases. This is simply a reflection of the fact that keeping all other factors constant, the noncentrality parameter increases less rapidly than degrees of freedom as model size increases. Fortunately, our results indicate that the RMSEA$_2$ is relatively robust to the effect of number of variables. In contrast, the values of the population RMSEA$_n$ decrease not only as the number of categories increases but also as the number of variables increases. This fact, coupled with the fact that the sampling distribution of the RMSEA$_2$ is much better approximated using asymptotic methods than the sampling distribution of RMSEA$_n$ makes the RMSEA$_2$ a much better candidate than RMSEA$_n$ to assess the degree of approximation of categorical data models.

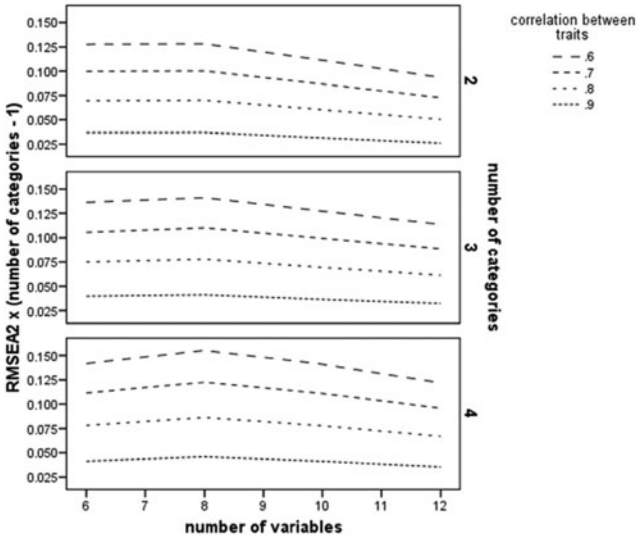

FIGURE 5 Plot of adjusted RMSEA$_2$ values as a function of number of variables, correlation between the traits, and number of categories.

Increasing the number of categories has a further effect on goodness-of-fit assessment of categorical data models, namely, as the number of categories increases, model size increases so rapidly that M_2, and hence the RMSEA$_2$, may not be computed due to memory limitations. For instance, when $K = 5$ and $n = 10$ the number of univariate and bivariate moments that enter in the computation of M_2 and the RMSEA$_2$ is $s = n(K-1) + \frac{n(n-1)}{2}(K-1)^2 = 760$, but when $K = 7$ and $n = 20$, $s = 6{,}960$. We consider in the next section how to assess the goodness-of-fit of a model for ordinal data when M_2 cannot be computed because the model is too large.

ASSESSING GOODNESS-OF-FIT IN LARGE MODELS FOR ORDINAL DATA: M_{ord} AND RMSEA$_{ord}$

Multivariate discrete data can be summarized using C sample proportions. For n variables with K categories, $C = K^n$. Limited information goodness-of-fit testing is based on using a smaller set of statistics that summarize the information contained in the data as much as possible, but not too much, in such a way that potential models can be discriminated by the summaries (Joe & Maydeu-Olivares, 2010, p. 413). Generally, the summaries make use only of univariate and bivariate information so that their distribution may be well approximated using asymptotic methods and higher power may be obtained. When M_2 is used, the summary statistics used are the means and cross products of indicator (dummy) variables used to denote each of the categories in the multinomial variables involved.

For example, let Y_{i} and Y_j be two multinomial variables each with three categories, $k = \{0, 1, 2\}$. Y_{i} and Y_j can be characterized using the indicator variables $I_{i,1}$, $I_{i,2}$ and $I_{j,1}$, $I_{j,2}$, respectively, where

Y_{i}	$I_{i,1}$	$I_{i,2}$
0	0	0
1	1	0
2	0	1

Then, the summary statistics used in M_2 are the sample means of these indicator variables and the sample cross products involving indicator variables from different variables. That is, the summary statistics used in M_2 are the sample counterparts of

$$\begin{array}{ll} E[I_{i,1}] = \Pr(Y_i = 1) & E[I_{i,1}I_{j,1}] = \Pr(Y_i = 1, Y_j = 1) \\ E[I_{i,2}] = \Pr(Y_i = 2) & E[I_{i,1}I_{j,2}] = \Pr(Y_i = 1, Y_j = 2) \\ E[I_{j,1}] = \Pr\left(Y_j = 1\right) & E[I_{i,2}I_{j,1}] = \Pr(Y_i = 2, Y_j = 1) \\ E[I_{j,2}] = \Pr\left(Y_j = 2\right) & E[I_{i,2}I_{j,2}] = \Pr(Y_i = 2, Y_j = 2) \end{array} \cdot \tag{24}$$

Therefore the summary statistics used in M_2 are just the univariate and bivariate proportions that exclude category 0.

As n and particularly K increase, the number of summary statistics used in M_2 increase very rapidly, to the point that for large values of n and K computing M_2 may no longer be feasible. In such large models, it is necessary to further reduce the information used for testing. The means and cross products of the multinomial variables ignoring the multivariate nature of the multinomial variables is a natural choice of statistics in this case. That is, one can use as summary statistics the sample counterparts of

$$\begin{aligned} \kappa_i &= E[Y_i] = 0 \times \Pr(Y_i = 0) + \ldots + (K_i - 1) \\ &\quad \times \Pr(Y_i = K_i - 1), \end{aligned} \tag{25}$$

$$\begin{aligned} \kappa_{ij} &= E[Y_iY_j] = 0 \times 0 \times \Pr(Y_i = 0, Y_j = 0) \\ &\quad + \ldots + (K_i - 1) \times (K_j - 1) \\ &\quad \times \Pr\left(Y_i{=}K_i - 1, Y_j{=}K_j{-}1\right). \end{aligned} \tag{26}$$

For our previous example, these simplify to

$$\begin{aligned} \kappa_i &= E[Y_i] = 1\Pr(Y_i = 1) + 2\Pr(Y_i = 2) \\ \kappa_j &= E[Y_j] = 1\Pr(Y_j = 1) + 2\Pr(Y_j = 2) \\ \kappa_{ij} &= E[Y_iY_j] = 1 \times 1\Pr(Y_i = 1, Y_j = 1) + 1 \\ &\quad \times 2\Pr(Y_i = 1, Y_j = 2) + 2 \times 1\Pr(Y_i = 2, Y_j = 1) \\ &\quad + 2 \times 2\Pr(Y_i = 2, Y_j = 2). \end{aligned} \tag{27}$$

Comparing Equation (24) with Equation (27), we see that quantities in Equation (27) are simply a linear function of those in Equation (24). Therefore, the sample counterparts of Equation (27)—means and cross products of variables coded as $\{0, 1, \ldots, K_i\}$—are a further reduction of the data than the sample counterparts of Equation (24)—univariate and bivariate proportions.

Let $\hat{\boldsymbol{\kappa}} = \boldsymbol{\kappa}(\hat{\boldsymbol{\theta}})$ be the statistics in Equation (27), which depend on the model parameters and are evaluated at their estimates, and let $\mathbf{m}$ be the sample counterpart of Equation (27). Using these statistics, from theory in Joe and Maydeu-Olivares (2010), a quadratic-form statistic can be formed similar to M_2:

$$\begin{aligned} M_{ord} &= N\left(\mathbf{m} - \hat{\kappa}\right)' \hat{\mathbf{C}}_{ord}\left(\mathbf{m} - \hat{\kappa}\right), \\ C_{ord} &= \boldsymbol{\Xi}_{ord}^{-1} - \boldsymbol{\Xi}_{ord}^{-1}\boldsymbol{\Delta}_{ord}(\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_{ord}^{-1}\boldsymbol{\Delta}_{ord})^{-1}\boldsymbol{\Delta}_{ord}'\boldsymbol{\Xi}_{ord}^{-1}. \end{aligned} \tag{28}$$

M_{ord} differs from M_2 in that the statistics used for testing are different, and so are their asymptotic covariance matrix $\boldsymbol{\Xi}$, and the matrix of derivatives involved, $\boldsymbol{\Delta}$, both of which are to be evaluated at the parameter estimates. The sample statistics used in M_{ord} are $\mathbf{m} = \left(\bar{\mathbf{y}}', \mathbf{c}'\right)'$, the n sample means $\bar{\mathbf{y}}$, and the $n(n\text{-}1)/2$ cross products $\mathbf{c} = vecr\left(\mathbf{Y}'\mathbf{Y}/N\right)$. Here $\mathbf{Y}$ denotes the $N \times n$ data matrix and *vecr*() denotes an operator that takes the lower diagonal of a matrix (excluding the diagonal) and stacks it on a column vector. Also, when all variables are binary, M_{ord} reduces to M_2. We provide in the Appendix details on $\boldsymbol{\Delta}_{ord}$ and $\boldsymbol{\Xi}_{ord}$ for the graded IRT response model. Also, Cai and Hansen (2013) provided details[4] on how to compute these for bifactor IRT graded response models.

From theory in Joe and Maydeu-Olivares (2010), M_{ord} follows an asymptotic chi-square distribution with $df_{ord} = n(n{+}1)/2 - q$ degrees of freedom for any consistent and

[4] Cai and Hansen (2013) used M_2^* to refer to the statistic we refer to as M_{ord} to emphasize that it should only be used with ordinal data.

asymptotically normal estimator. Also, under a sequence of local alternatives the noncentrality parameter of M_{ord} divided by sample size is

$$D_{ord} = (\boldsymbol{\kappa}^T - \boldsymbol{\kappa}^0)'\, \mathbf{C}^0_{ord}(\boldsymbol{\kappa}^T - \boldsymbol{\kappa}^0) \tag{29}$$

with $\boldsymbol{\kappa}^0$ and $\boldsymbol{\kappa}^T$ being the means and cross products in Equation (27) under the fitted (i.e., null) model and population probabilities, respectively, and $\mathbf{C}^0_{ord}$ is given in Equation (28) and it is computed based on the fitted (null) model. An RMSEA can be constructed using Equation (29):

$$\varepsilon_{ord} = \sqrt{\frac{D_{ord}}{df_{ord}}}, \tag{30}$$

and an estimate of this RMSEA is

$$\hat{\varepsilon}_{ord} = \sqrt{\mathrm{Max}\left(\frac{\hat{M}_{ord} - df_{ord}}{N \times df_{ord}}, 0\right)}. \tag{31}$$

As we did for the previous RMSEAs discussed in this article we can construct confidence intervals for this ordinal RMSEA and test whether the ordinal RMSEA is smaller than some cutoff value.

Now, because M_{ord} is a quadratic form of statistics that are a further reduction of the data than the statistics used in M_2, from theory in Joe and Maydeu-Olivares (2010), (a) the empirical sample distribution of M_{ord} is likely to be better approximated in small samples than the distribution of M_2; and (b) if $(D_{ord}/D_2) > 0.9$, that is, if the ratio of noncentrality parameters for M_{ord} and M_2 is sufficiently large, M_{ord} will be more powerful than M_2 over a variety of alternative directions because there are fewer degrees of freedom associated with M_{ord} than with M_2. Cai and Hansen (2013) investigated the small sample distribution of M_{ord} and M_2 in bifactor logistic models for polytomous data and reported that the sampling distribution of M_{ord} is better approximated than that of M_2 when there are small expected counts in the bivariate tables. They also reported that M_{ord} has higher power than M_2 to detect misspecified bifactor models.

The use of M_{ord} is not without limitations. First, the computation of the population means and cross products in Equation (27) and their sample counterparts must be meaningful. These are simply weighted combinations of univariate and bivariate probabilities. Such linear combinations are meaningless when the categorical data are nominal. Hence, the use of M_{ord} is only justified when data are ordinal—hence its name. Second, the number of items must be large enough for the degrees of freedom of M_{ord} to be positive. For instance, for the GRM with a single latent trait, the minimum number of items needed for the degrees of freedom of M_{ord} to be positive is $n > K + 2$. A larger number of items is needed in the case of multidimensional models. Thus, there is an interesting trade-off between M_{ord} and M_2: When the number of items and categories is large M_2 cannot be used for computational reasons, but when the number of items is small and the number of categories is large, M_{ord} cannot be used due to lack of degrees of freedom.

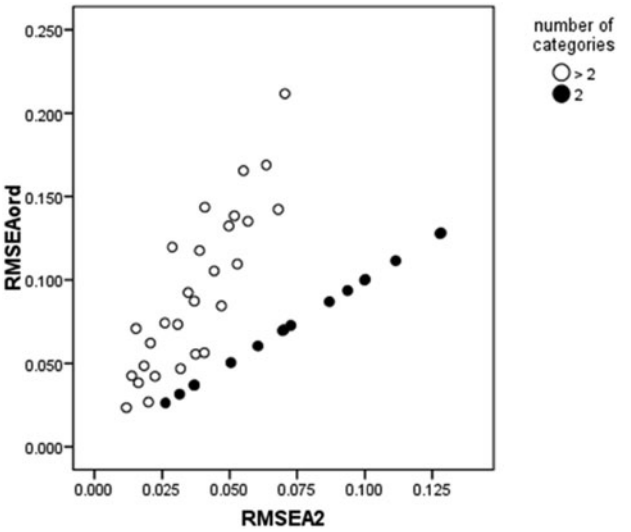

FIGURE 6 Plot of RMSEA$_{ord}$ values as a function of RMSEA$_2$ and number of categories (2 or higher).

Choice of RMSEA$_{ord}$ Cutoff Values in IRT Models

What is the relationship between the RMSEA$_2$ and RMSEA$_{ord}$ population values? What cutoff values should be used when using RMSEA$_{ord}$? To address these questions we used the 48 conditions of the previous subsection. Due to the lack of degrees of freedom, RMSEA$_{ord}$ cannot be computed when $n = 6$ and $K = 4$. As a result, the effective number of conditions in this case is 44. RMSEA$_2$ versus RMSEA$_{ord}$ population values are plotted in Figure 6. When $K = 2$, RMSEA$_2$ equals RMSEA$_{ord}$. For $K > 2$ we see that in all cases RMSEA$_{ord}$ is greater than RMSEA$_2$, reflecting that the noncentrality parameter of M_{ord} is greater than that of M_2. As a result, M_{ord} has more power than M_2 to reject this particular type of model misspecification. Furthermore, we see in Figure 6 that the relationship between both RMSEAs when $K > 2$ is approximately linear ($R^2 = 75\%$).

RMSEA$_{ord}$ values are plotted in Figure 7 as a function of number of variables and number of categories. We see in this figure that for the parameter values chosen RMSEA$_{ord}$ values are larger when $K = 3$ than when $K = 2$ except for $n = 6$ and $\rho = .8$ or $.9$. We also see in this figure that for $K = 4$ RMSEA$_{ord}$ values decrease as the number of variables increases. In fact, for $n = 12$, RMSEA$_{ord}$ values are lower when $K = 4$ than for $K = 2{,}3$. All in all, it appears very difficult to offer a cutoff value for the RMSEA$_{ord}$ across different values of n and K.

The Standardized Root Mean Squared Residual (SRMSR)

A solution to the problem of how to assess the approximate fit of large models when the data are ordinal may lie in

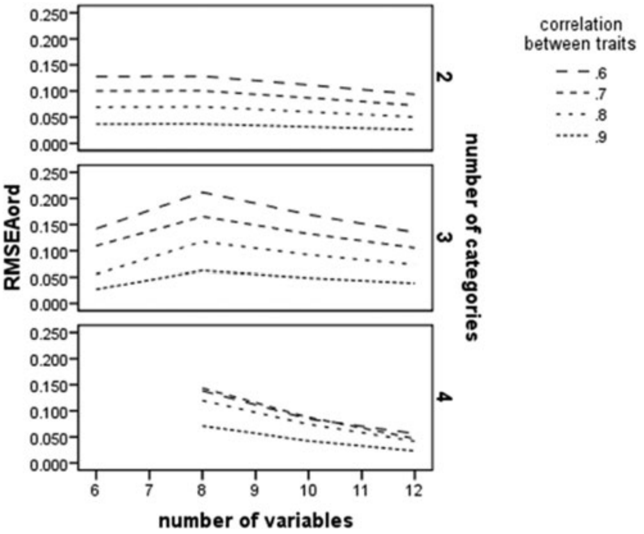

FIGURE 7 Plot of RMSEA$_{ord}$ values as a function of number of variables, correlation between the traits, and number of categories.

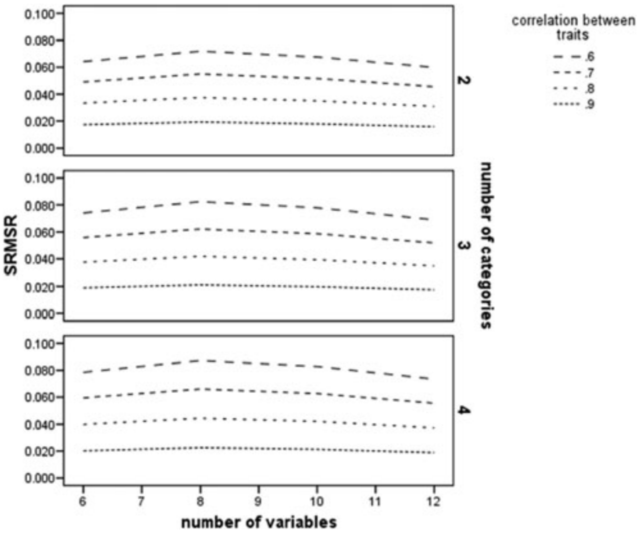

FIGURE 8 Plot of Standardized Root Mean Squared Residual (SRMSR) values as a function of number of variables, correlation between the traits, and number of categories.

the use of the Standardized Root Mean Squared Residual (SRMSR) borrowed from the factor analysis literature. For a pair of items i and j, the residual correlation is 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. Thus, the residual correlation is

$$r_{ij} - \hat{\rho}_{ij} = r_{ij} - \frac{\hat{\kappa}_{ij} - \hat{\kappa}_i \hat{\kappa}_i}{\sqrt{\hat{\kappa}_{ii} - \hat{\kappa}_i^2}\sqrt{\hat{\kappa}_{jj} - \hat{\kappa}_j^2}}, \qquad (32)$$

where the means (κ_i and κ_j) and the cross product κ_{ij} were given in Equations (25) and (26), and κ_{ii} is

$$\begin{aligned} k_{ii} &= E[Y_i^2] = 0^2 \times \Pr(Y_i = 0) + \ldots (K_i - 1)^2 \\ &\quad \times \Pr(Y_i = K_i - 1)\,. \end{aligned} \qquad (33)$$

The sample SRMSR is simply the square root of the average of these squared residual correlations[5]

$$\widehat{SRMSR} = \sqrt{\sum_{i<j} \frac{(r_{ij} - \hat{\rho}_{ij})^2}{n(n-1)/2}}. \qquad (34)$$

Being an average of residual correlations, the SRMSR should not be affected by the number of items, all other factors being held constant. To investigate whether this is indeed the case, we computed the population SRMSR for the previous 48 conditions. The population SRMSR is defined as

$$SRMSR = \sqrt{\sum_{i<j} \frac{\left(\hat{\rho}_{ij}^{T} - \hat{\rho}_{ij}^{0}\right)^2}{n(n-1)/2}}. \qquad (35)$$

This is the squared root of the mean of the squared differences between the correlations implied by the population probabilities and fitted model. The values of these population SRMSRs are plotted in Figure 8 as a function of the number of variables, number of categories, and correlation.[6] As we can see in this figure, for the conditions investigated, the values of the SRMSR are relatively stable across the number of variables and categories for small amounts of model misspecification.

What is the relationship between the SRMSR and the bivariate RMSEA? To address this question, we plot in Figure 9 the values of the population SRMSR and RMSEA$_2$ across the 48 conditions as a function of the number of categories. We see in this figure that their relationship, for the models investigated, is increasingly linear as the number of categories increases: for $K = 2$, 3, and 4, the R^2 between SRMSR and RMSEA$_2$ are 95%, 98%, and 99%. In fact, as Figure 10 shows, when the RMSEA$_2$ is adjusted by $(K - 1)$ its relationship to the SRMSR is quite linear ($R^2 = 97\%$). Thus, it is possible to relate cutoff values for SRMSR and RMSEA$_2$. In so doing, we find that a population SRMSR value of 0.05 corresponds roughly to an adjusted RMSEA$_2$ of 0.09 and that a population RMSEA$_2$ value of 0.05 corresponds roughly to a value of SRMSR of 0.03.

[5] A more appropriate label for the statistic in Equation (34) is Root Mean Squared Residual Correlation (RMSRC) but in the factor analysis literature where it originated this statistic is commonly referred to as RMSR (correlation) or SMSR. The latter is used here to avoid introducing additional terminology.

[6] The SRMSR can be computed even when there are no degrees of freedom available for testing using M_{ord} . As a result, 48 conditions are displayed in Figure 8 but only 44 in Figure 7.

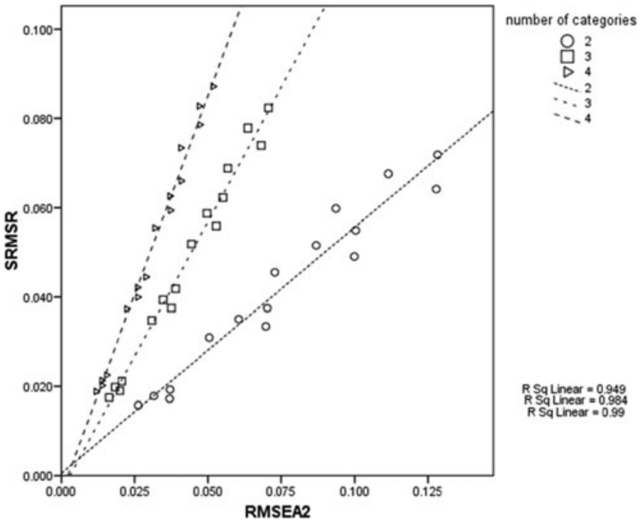

FIGURE 9 Plot of SRMSR values as a function of RMSEA$_2$ values. Their linear association increases as the number of categories increases.

DISCUSSION

Assessing the goodness-of-fit in multivariate data analysis is more complicated for categorical than for continuous variables: As the number of variables increases, the asymptotic p values for the usual test statistics for categorical data analysis such as Pearson's X^2 become inaccurate regardless of sample size. This problem can be solved by using limited information test statistics such as Maydeu-Olivares and Joe's (2005, 2006) M_2 statistic as asymptotic p values for this statistic are accurate even when the data are extraordinarily sparse. A similar approach can be used to reliably assess the goodness of approximation in multivariate categorical data analysis.

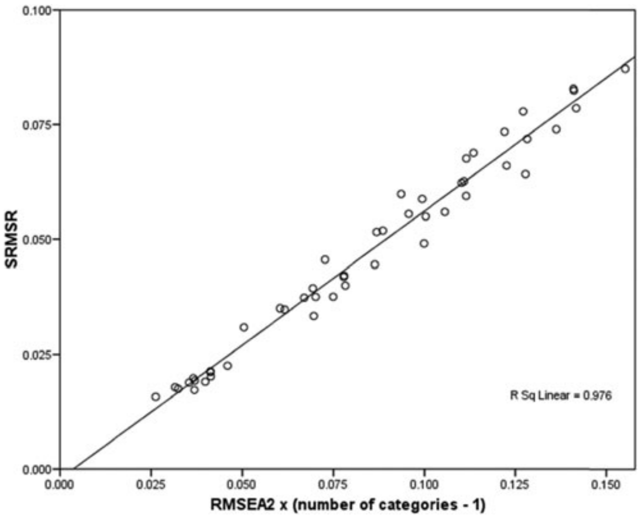

FIGURE 10 Plot of population SRMSR values as a function of adjusted RMSEA$_2$ values, that is, $(K - 1)$ RMSEA$_2$.

The RMSEA as first proposed in the context of factor analysis by Steiger and Lind (1980) assesses model to data fit while imposing a penalty for model complexity. As a result, it can be used to assess goodness of approximation but also for model selection. Furthermore, emphasis is given to computing confidence intervals on the population parameter. This requires that the test statistic used to estimate the population RMSEA has a known sampling distribution. In this article we have proposed a family of RMSEA parameters for multivariate categorical data analysis. Two members of this family are the full information RMSEA$_n$ and the bivariate RMSEA$_2$. For any consistent and asymptotically normal estimator, these RMSEAs can be estimated using M_n and M_2, respectively. In the special case where the ML estimator is used, the RMSEA$_n$ can be estimated using Pearson's X^2 statistic. Because these statistics have known asymptotic sampling distributions, it is possible to construct confidence intervals and to perform tests of close fit based on them. For the ML estimator, the RMSEA$_n$ may be used when the sampling distribution of X^2 is well approximated (i.e., in models with not too many possible response patterns), whereas the RMSEA$_2$ may be used in very large models. However, in applications we may find categorical models that involve so many response patterns that M_2 simply cannot be computed. This will occur if both the number of items and of response alternatives per item is large. In this case, an RMSEA suitable for ordinal data can be computed and a confidence interval for it constructed using the asymptotic sampling distribution of M_{ord}.

However, by construction, these categorical RMSEAs present two features that may be perceived as undesirable for the purpose of assessing goodness of approximation proper (regardless of model complexity). First, the RMSEAs are hard to interpret substantively. Because confidence intervals for the RMSEAs are of interest, the RMSEAs are constructed using test statistics with known sampling distributions. But test statistics with known sampling distributions are usually weighted averages, which are inherently difficult to interpret, rendering the substantive interpretation of the RMSEAs difficult. Second, because by construction the RMSEAs adjust for model parsimony by dividing a test statistic by its degrees of freedom, the population RMSEAs will generally not be invariant as model size increases keeping all determinants of model misfit constant. As a result, it may be difficult to offer cutoff values for the RMSEAs that are independent of the number of variables and categories used.

We have seen that this is indeed the case for the RMSEA$_n$ and RMSEA$_{ord}$ but not for the RMSEA$_2$. That is, for the models investigated, both the population RMSEA$_n$ and RMSEA$_{ord}$ decrease as the number of variables and categories increase, but the population RMSEA$_2$ appeared to be relatively stable for small levels of misspecification as the number of variables increased and it could be easily adjusted by the number of categories. As a result, cutoff population values for the RMSEA$_2$ but not for the RMSEA$_n$ and

RMSEA$_{ord}$ can be offered. For IRT models, we argued that distinguishing between misspecified models with the correct and incorrect latent dimensionality may be the most important consideration in applications. Consequently, we searched for an RMSEA$_2$ population value that enabled researchers to distinguish between these two sets of models. The resulting population value, RMSEA$_2$ ≤ 0.05, is our criterion for close fit. Because population values of the RMSEA$_2$ decrease as the number of categories increases but values of the RMSEA$_2$ adjusted by the number of categories using RMSEA$_2$/ $(K - 1)$ remain relatively stable for small degrees of model misfit, our criterion for excellent fit is adjusted RMSEA$_2$ ≤ 0.05. Notice that the adjusted RMSEA$_2$ equals the raw RMSEA$_2$ when the data are binary. Therefore, our criteria for close and excellent fit are equal in the case of binary data.

The use of an RMSEA$_2$ adjusted by the number of response categories enables us to address the second of the concerns we put forth regarding the use of RMSEAs, namely, invariance to the number of items and categories, but it does not fully address the first. What can we say about the magnitude of an estimated RMSEA$_2$ equal to say 0.12? To address this issue, we examined the relationship between population RMSEA$_2$ values and Standardized Root Mean Squared Residual (SRMSR) values in IRT models. The SRMR can be described simply as the squared root of the average of the residual product-moment correlations squared and therefore should only be computed with binary or ordinal data. For the IRT models investigated there is a strong linear relationship between the population adjusted RMSEA$_2$ and SRMR values, which enables associating adjusted RMSEA$_2$ values to SRMR values. If we are willing to consider a model for ordinal data that yields an SRMR ≤ 0.05 as an acceptable approximation to the data, we can use this value as cutoff for acceptable fit and link it to the RMSEA$_2$ values as summarized in Table 2.

When the data are ordinal or binary, we recommend using the RMSEA$_2$ and SRMR in tandem to assess the goodness of approximation of the fitted models. The SRMR provides a normed effect size of the model misfit and therefore its magnitude can be easily judged. In contrast, the RMSEA$_2$ gives us a measure of model misfit adjusted by degrees of freedom and it should be used when selecting among competing models fitted to the same data. It is straightforward to obtain a confidence interval for the RMSEA$_2$ and to perform a test of close fit. In contrast, it is possible but cumbersome to obtain a confidence interval for the SRMR for categorical data we propose. Thus, it is best to use the SRMR as a goodness-of-fit index. As a goodness-of-fit index, the SRMR can be easily computed for models of any size. In particular, it can be computed for models with so many response patterns that the RMSEA$_2$ cannot be computed. For such large models for ordinal data we prefer the SRMR to the RMSEA$_{ord}$ as we cannot offer cutoff criteria for the latter.

TABLE 2
Suggested Cutoff Criteria for Approximate Fit in Categorical Data Analysis

Criterion	RMSEA$_2$	SRMR
Adequate fit	0.089	0.05
Close fit	0.05	0.027
Excellent fit	0.05 / $(K - 1)$	0.027 / $(K - 1)$

Note. The Squared Root Mean Residual (SRMR) should only be computed for ordinal and binary data.

Interestingly, our cutoff criteria for close and acceptable fit using the RMSEA$_2$ are very similar to those put forth by Browne and Cudeck (1993) in the context of factor analysis. What is the relationship between the RMSEAs and SRMSR proposed here and those currently in use in structural equation modeling? We address this issue in the next subsection.

Relationship Between RMSEAs and SRMSR for Categorical and Continuous Data

The IRT models used in the previous sections are equivalent to the ordinal factor analysis model used when fitting a factor analysis model using polychoric correlations except for the choice of link function—logistic for the former, normal for the latter (Takane & de Leeuw, 1987). However, the RMSEA and SRMSR obtained when fitting an ordinal factor analysis using polychoric correlations are different from the ones introduced here. In ordinal factor analysis, the SRMSR can be interpreted approximately as the average residual polychoric correlation, whereas the SRMSR introduced can be interpreted approximately as the average residual product-moment correlation. Therefore, they are different statistics and they take different values in applications, particularly when the number of categories is small and, as a result, estimated polychoric correlations differ from product-moment correlations.

The bivariate RMSEA$_2$ introduced here is conceptually different from the RMSEA in use when fitting an ordinal factor analysis to polychoric correlations. The latter reflects how well the model reproduces the polychoric correlations, whereas the former reflects how well the model reproduces the bivariate tables. When fitting an ordinal factor model via polychoric correlations the overall discrepancy between the model and the data can be decomposed (Maydeu-Olivares, 2006; Muthén, 1993) into a distributional discrepancy (the extent to which the assumption of discretized multivariate normality underlying the use of polychorics is tenable) and a structural discrepancy (the extent to which the model reproduces the polychoric correlations). The polychoric RMSEA only assesses the latter, whereas the RMSEA$_2$ assesses the overall discrepancy. Generally, when fitting an ordinal factor analysis the distributional discrepancy is much larger than the structural discrepancy (Maydeu-Olivares, 2006) and therefore we believe that the RMSEA$_2$ introduced here should be used instead of the polychoric RMSEA when assessing the goodness of approximation of ordinal factor analysis models. However, because the RMSEAs adjust the discrepancies (overall or structural) by their respective degrees of freedom and the degrees of freedom associated to the overall bivariate discrepancy are much larger than the degrees of freedom from fitting the hypothesized structural model to polychoric correlations, the overall RMSEA$_2$ may be smaller than the polychoric RMSEA in applications. Further research is needed on the relationship between the polychoric RMSEA and the RMSEA$_2$ introduced in this article.

The SRMSR given in Equation (34) is the same used in linear factor analysis for continuous data. However, the SRMSR reported by software programs for factor analysis or structural equation modeling may differ from the SRMSR defined in Equation (34). In Equation (34) the expected correlations are obtained by dividing the expected covariances by the expected standard deviations. Existing implementations may compute the expected correlations using sample standard deviations instead. In our experience, when Equation (34) using Equation (32) is used to compute the SRMSR for an IRT model fitted to ordinal data and the same expression is used to compute the SRMSR for a comparable linear factor analysis model fitted to the same data (treating them as if they were continuous), similar results are obtained. However, the models are based on different assumptions and are estimated differently. Strictly speaking, the linear factor analysis is misspecified when applied to ordinal data because its predicted values cannot take integer values (Maydeu-Olivares, Cai, & Hernández, 2011; McDonald, 1999). An analogy is applying linear regression to predict a dichotomous dependent variable. In contrast, an IRT model may be the correctly specified data-generating model. A linear factor analysis attempts to account for the observed bivariate associations present in the data (covariances or correlations). Furthermore, when a linear factor analysis is fitted to ordinal data using ML, a misspecified density is assumed (the distribution of the data is multinomial, but a normal density is assumed when estimating the model) although corrections are used (e.g., Satorra & Bentler, 1994) to ensure that the test statistic is robust to density misspecification. In contrast, IRT models attempt to account for the observed frequencies of the response patterns (or equivalently for the univariate, bivariate, trivariate, . . . , up to n-way associations present in the data), and when ML is used to fit an IRT model to ordinal data, a correctly specified density, multinomial, is employed. When a linear factor analysis is fitted to ordinal data the SRMSR should be low if the model provides a reasonable approximation to the data as a discrepancy function between the sample and expected bivariate moments is minimized. When an IRT model is fitted to ordinal data, the SRMSR need not be low as a discrepancy function between all sample and expected moments is being minimized. In this context, the SRMSR should be taken simply as a computationally convenient, substantively easy to interpret, goodness-of-fit index to gauge the magnitude of the misfit to the low order margins of the contingency table because assessing the magnitude of the misfit to the full table is impractical. From this point of view, it is reassuring to find that IRT and linear factor analysis yield similar SRMSR indices.

SOME DATA EXAMPLES

In this section, we provide some examples to illustrate the theory set forth in the previous sections. In all cases we used maximum likelihood (ML) to estimate the IRT models. This is generally called marginal ML in the IRT literature (see Bock & Aitkin, 1981). The software flexMIRT (Cai, 2012) with default settings was used in all cases. This software provides estimates of X^2 and the full information RMSEA$_n$ (provided there are not too many response patterns), M_2 and the RMSEA$_2$, and M_{ord} and the RMSEA$_{ord}$. IRTPRO (Cai, du Toit, & Thissen, 2011) also computes X^2 and RMSEA$_n$ and M_2 and RMSEA$_2$.

Fitting Logistic IRT Models to the LSAT7 Data

This data set (Bock & Lieberman, 1970) consists of the responses of 1,000 individuals to five selected items of the Law Scholastic Aptitude Test. The answers to these items have been coded dichotomously (correct, incorrect). This is a very small model, as there are only $2^5 = 32$ possible response patterns. As a consequence, the asymptotic approximation to the distribution of the full information RMSEA can be trusted. We fitted a 1PLM and a 2PLM to these data using ML. Table 3 provides the results for the test statistics for assessing exact fit. Two statistics are reported in this table, M_2, which uses only univariate and bivariate information, and X^2. X^2 assesses how well the model reproduces the univariate, bivariate, trivariate, four-variate, and five-variate moments. As can be seen in this table, when the p values of X^2 can be trusted, M_2 and X^2 provide fairly similar results.

The goodness of approximation results are also shown in this table. The RMSEA$_2$ and the RMSEA$_n$ agree in that both the 1PLM and 2PLM provide good fits to these data. However, we do not observe quite as close an agreement between the RMSEA$_2$ and RMSEA$_n$ as we did for the p values of the test of exact fit. Indeed, for both models considered, the limited information RMSEAs are larger. The M_2 and X^2 test statistics are members of a general family of test statistics described in Joe and Maydeu-Olivares (2010). For two test statistics within this family they show that if one statistic is obtained by concentrating the information used in the other statistic, the statistic that further concentrates the information will be more powerful to detect many alternatives of interest. Furthermore, they show that power will generally be related to the ratio of the value of the statistic divided by its degrees of freedom. The M_2 statistic "concentrates" the information used in X^2 into fewer degrees of freedom. Because the

TABLE 3
Goodness-of-Fit Assessment for Two IRT Models Fitted to the LSAT 7 Data

Tests of Exact Fit						
	1PLM			2PLM		
	Value	*df*	*p*	Value	*df*	*p*
M_2	23.17	9	.01	11.94	5	.04
X^2	44.15	25	.01	32.48	21	.05

Goodness of Approximation								
	1PLM				2PLM			
	Value	90% CI	*df*	*p*	Value	90% CI	*df*	*p*
RMSEA$_2$	0.040	(0.020; 0.060)	9	.78	0.037	(0.009; 0.065)	5	.75
RMSEA$_n$	0.028	(0.013; 0.041)	25	.58	0.023	(0; 0.038)	21	.74

Note. $n = 5$, $K = 2$, $N = 1{,}000$. For the RMSEA$_2$ we test H_0: $\varepsilon_2 \leq 0.05$; for the RMSEA$_n$ we test H_0: $\varepsilon_n \leq 0.03$.

RMSEAs are a function of the statistics divided by their degrees of freedom, as power increases so does the RMSEA estimate. Thus, one should expect the estimated RMSEA$_2$ to be larger than the RMSEA$_n$ reflecting the fact that the test statistic has more power to detect that the fitted model does not approximate well the population probabilities.

Because the RMSEA$_2$ is generally larger than the RMSEA$_n$, different cutoff values should be used, otherwise models that would be rejected using RMSEA$_2$ would be accepted using RMSEA$_n$. Our results suggest that if a cutoff of 0.05 is used for the RMSEA$_2$, a cutoff of 0.03 should be used for the RMSEA$_n$. These are the cutoff criteria that we use for testing in Table 3. Using these criteria, a similar *p* value for the tests of close fit is obtained for the 2PLM but not so much for the 1PLM. The use of the same cutoff criteria for the RMSEA$_n$ as for the RMSEA$_2$, 0.05, leads to a *p* value of 1 when testing for close fit using the RMSEA$_n$.

Next, we consider a realistic example.

Fitting an IRT Model to Beck's Hopelessness Scale

Chang, D'Zurilla, and Maydeu-Olivares (1994) modeled the responses of 393 individuals to Beck's Hopelessness Scale (Beck, Weissman, Lester, & Trexler, 1974). This is a set of $n = 20$ true-or-false questions used to predict depression, suicidal ideation, and suicidal intent. We fitted a 2PLM to these data using ML and we report in Table 4 tests of exact fit using M_2 and X^2 as well as the goodness of approximation of the model using RMSEA$_n$ and RMSEA$_2$.

There are 2^{20} (>1 million) cells in the contingency table. As a result, the degrees of freedom obtained after fitting any IRT model to these data will be over a million. It is questionable whether we are interested in testing how well the model reproduces the joint moments of the data up to 20th order, which is what X^2 does. Furthermore, the *p* value for X^2 is useless due to data sparseness. In contrast, the *p* value for M_2 can be trusted, even with the small sample size of this example, and it can be argued that M_2 performs a more meaningful assessment, how well the model reproduces the bivariate margins of the table, and in so doing it may be more powerful than X^2 to reject the model. However, it is questionable to expect that any model for these data will fail to be rejected by a test of exact fit because recall that the model is trying to fit all 2^{20} possible response patterns. With such large models, assessing the goodness of approximation of the fitted model is a more sensible endeavor.

The use of the RMSEA$_n$ based on X^2 to assess the goodness of approximation in this example is questionable for several reasons. First, the RMSEA$_n$ assesses how well we approximate all moments of the data. We do not feel that such an assessment is of substantive interest. Second, due to data sparseness, we cannot assess the precision of a RMSEA$_n$ estimate (we cannot obtain reliable confidence intervals or a test of close fit). Third, we cannot even qualitatively assess the magnitude of the estimated RMSEA$_n$ because the RMSEA$_n$ decreases as the number of variables increases. Thus, for this example we obtained an RMSEA$_n$ point estimate of 0.055 and we provide a 90% confidence interval in Table 4. However, this confidence interval is incorrect; it overestimates

TABLE 4
Goodness-of-Fit Assessment for a One-Dimensional Two-Parameter Logistic Model Fitted to Beck's Hopelessness Scale

Tests of Exact Fit			
	Value	*df*	*p*
M_2	231.50	170	.001
X^2	2,299,697.064	1,048,535	0
Goodness of Approximation			
	Value	90% CI	*df*
RMSEA$_2$	0.030	(0.020; 0.040)	170
RMSEA$_n$	0.055	(0.055; 0.055)	1,048,535

Note. $n = 20$, $K = 2$, $N = 393$.

the precision with which we estimate the RMSEA$_n$. Qualitatively, the RMSEA$_n$ appears small given the number of degrees of freedom involved. However, we know that the population RMSEA$_n$ decreases as the number of variables increases, so we really do not know if an RMSEA$_n$ $= 0.055$ is "small" or not given 20 binary variables.

In contrast, in our view, the RMSEA$_2$ performs a more meaningful assessment, how closely the model approximates the bivariate margins. Also, reliable confidence intervals for its population value can be obtained. Finally, because the population RMSEA$_2$ appears to be relatively robust to the number of variables involved we can offer cutoff values and a test of close fit can be performed.

It is interesting to see that in this example the RMSEA$_n$ estimate, 0.055, is larger than the RMSEA$_2$ estimate, 0.030. Why is the RMSEA$_2$ smaller than the RMSEA$_n$ in this case? We believe that it is because X^2 and hence, the RMSEA$_n$, have such a large sampling variability in this case that they cannot be trusted.

In closing this example we point out that if there is interest in assessing how well the model approximates higher order moments, then RMSEA$_3$ (involving trivariate margins), RMSEA$_4$ (involving four-way margins), and so forth, can be estimated and its precision assessed using the theory provided in this article. However, increasingly larger sample sizes are needed to estimate accurately the precision of these higher order RMSEAs. We now turn to an example involving an IRT model for polytomous ordinal data.

Fitting an IRT Model to the PROMIS Depression Items

We fitted a unidimensional GRM with a normally distributed latent trait to the $n = 28$ PROMIS depression items (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*. The responses were coded from 0 to 4 for the analyses. We used the $N = 768$ complete responses to these data kindly provided by the authors. Preliminary analyses revealed quite large slope estimates and we used 100 rectangular quadrature points between –8 and 8 to ensure the accuracy of the results reported here.

There are over 37 trillion possible response patterns and Pearson's X^2 cannot be computed in this case. Yet, the model only involves $q = 5 \times 28 = 140$ parameters. Testing the exact fit of the model does not make much sense. It would be wonderful if we failed to reject any model to these data, but it is not realistic to expect such an outcome. M_2 can barely be computed in this example and its computation takes considerably longer than the estimation of the model itself. This is because in this example there are $s = 6{,}160$ moments to be computed, along with their asymptotic covariance matrix. Degrees of freedom for M_2 are therefore 6,160 – 140 = 6,020. The M_2 estimate is 8,543.56 and consequently the RMSEA$_2$ estimate is 0.023. Using our suggested cutoff criteria, we conclude that, overall, the fitted model provides a close fit to the PROMIS depression data, but it falls short of our criteria for excellent fit (0.05/4 = 0.0125) as the 90% confidence interval for the population bivariate RMSEA is (0.022; 0.024). Because the data are ordinal, we can compute the residual correlations implied by the model and the SRMSR. Its estimate is 0.037. Certainly, the overall magnitude of the misfit of the model is small. However, even if a model provides a good overall approximation as in this case, it is necessary to investigate whether there are some parts of the model whose fit can be improved. If the data are ordinal as in this case, this can be accomplished by examining z statistics for the residual means in Equation (25) and cross products in Equation (26) (Maydeu-Olivares & Liu, 2012). The residual correlations provide us with an estimate of the size of the misfit for each cross product z statistic. Examining the residual correlations, we find that there are nine residual correlations larger in absolute value than 0.10. We conclude that although the GRM provides a good overall approximation to these data, the model can be fine-tuned to provide a better approximation as there are associations between these items that are not well captured by the model.

This model is about the largest model for which the RMSEA$_2$ can be computed. Yet, the RMSEA$_{ord}$ can be computed effortlessly in this example and it can be computed for much larger models for ordinal data. A 90% confidence interval for the population RMSEA$_{ord}$ in this example is (0.065; 0.073). Hence, the value obtained is larger than for the RMSEA$_2$, reflecting that M_{ord} is more powerful than M_2 (Cai & Hansen, 2013). As a consequence, different cutoff criteria should be used for the RMSEA$_2$ and RMSEA$_{ord}$. The RMSEA$_{ord}$ can be used to compare competing models (adjusting for parsimony) fitted to the same data, but we cannot offer a cutoff criterion of close fit as this parameter decreases as the number of variables and categories fitted increase.

CONCLUDING REMARKS

We have introduced a family of RMSEAs that enables researchers to assess the goodness of approximation of their models for multivariate categorical data. The family consists of the statistics RMSEA$_1$, RMSEA$_2$, . . ., to RMSEA$_n$. The RMSEA$_1$ describes the goodness of approximation of the model to the univariate margins of the contingency table, the RMSEA$_2$ to the bivariate margins, and so forth up to RMSEA$_{n,}$ which describes the goodness of approximation to the full table. These RMSEAs can be conveniently estimated using Maydeu-Olivares and Joe's M_r statistics: M_1 can be used to estimate the RMSEA$_1$; M_2 to estimate the RMSEA$_2$;and so forth up to $M_{n,}$ which can be used to estimate the RMSEA$_n$. For ML estimation, Pearson's X^2 equals

M_n (Maydeu-Olivares & Joe, 2005). Consequently, X^2 can be used to estimate the full information RMSEA in this case.

At what level of association shall we assess the goodness of approximation of our models? At least at the level at which the model is identified, otherwise the RMSEA cannot be computed. Thus, if a categorical data model can be estimated using only bivariate information, RMSEA$_2$ to RMSEA$_n$ can be used. The smallest the level of association used, the better we can determine the precision of the sample RMSEA. Thus, we recommend assessing the goodness of approximation at the smallest level of association at which a model is identified. Many models for categorical data, such as the IRT models we have used in this article, are identified using bivariate information. Thus, for routine applications, we recommend using the bivariate RMSEA$_2$. In some applications it may be of interest to examine higher order RMSEAs. But assessing fit at higher order associations requires increasingly larger sample sizes. Also, population RMSEAs of different orders (e.g., RMSEA$_2$ vs. RMSEA$_n$) are on different scales and therefore different cutoff criteria of close fit should be used. For any given RMSEA, cutoff criteria of good fit can only be meaningfully given if population values remain relatively stable for increasing number of variables and number of categories. Fortunately, for the models investigated here the RMSEA$_2$ meets this requirement and we have been able to offer cutoff values of adequate, good, and excellent fit for the RMSEA$_2$. In contrast, population RMSEA$_n$ values decrease as the number of categories and the number of variables increase and therefore any fixed cutoff value would favor large models.

As soon as the number of cells is larger than about 300, confidence intervals for the RMSEA$_n$ may be inaccurate. For models with over a million cells, the RMSEA$_n$ can no longer be computed. The bivariate RMSEA can still be computed and its precision assessed in models of this size, but in models with over a trillion cells it can no longer be computed. An RMSEA$_{ord}$ can be computed in this case, and hence the goodness-of-fit of approximation assessed, but only if the data are ordinal. Unfortunately, the population RMSEA$_{ord}$ decreases as the number of variables and categories increases and as a result we are unable to offer cutoff criteria of close fit for it. However, the RMSEA$_{ord}$ can (and should) be used in selecting among competing models fitted to the same data. For assessing the goodness of approximation in large models for ordinal data we suggest using the SRMSR borrowed from the factor analysis literature. Population values of this goodness-of-fit index are relatively stable across number of variables and categories and we have been able to offer cutoff criteria of adequate, good, and excellent fit for the SRMSR as well. Furthermore, for the IRT models investigated, population values of SRMSR and RMSEA$_2$ show a strong linear relationship.

Finally, in dealing with models for multivariate data it does not suffice to inspect summary measures of fit such as the RMSEA$_2$ or the SRMSR. Rather, the fit of the model to *all* the variables must be assessed to check if there is an obvious model deviation that can be remedied with a slightly more complex model. Only when no apparent trend is apparent in the residual diagnostics, the RMSEA$_2$ and/or the SRMSR can be considered as a measure of the goodness of approximation of the model. z statistics for residual means and cross products provide suitable residual diagnostics for binary and ordinal data (Maydeu-Olivares & Liu, 2012), and residual correlations provide an assessment of the magnitude of the misfit identified by z statistics for residual cross products. For nominal polytomous data, other residual diagnostics for each item and pair of items must be employed.

In closing, for ease of exposition, in our presentation we have reported just the results obtained with a single set of parameter values, although different sets of parameters were used to investigate the robustness of the cutoff values offered. Future research should thoroughly examine the effects of potential determinants of the results, such as item skewness. Also, although the presentation here has focused on applications to IRT modeling, the framework presented here is completely general and can be applied to any model for discrete data under multivariate multinomial assumptions. We expect that the cutoff values offered here will be useful to applied researchers using the IRT models considered in this article. However, more research is needed to investigate whether the cutoffs offered (using dimensionality as criterion) are useful when other criteria are of interest. For instance, researchers may be interested in retaining misspecified models where the correlation between the true latent traits and the estimated latent traits is above some value, say 0.99, but not if the correlation is smaller. The procedures described in this article can be used to check the usefulness of the cutoffs offered in distinguishing IRT models based on this or other criterion. A final remark: Assessing the goodness of approximation of a model does not provide us with information about its usefulness. On the other hand, assessing the goodness of approximation of a model that has been judged to be useful tells us how much room there is for improvement.

FUNDING

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

REFERENCES

Bartholomew, D. J., & Leung, S. O. (2001). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*, 1–16.

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

Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology*, *42*, 861–865.

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

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197.

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

Cai, L. (2012). *flexMIRT: A numerical engine for multilevel item factor analysis and test scoring [Computer software]*. Seattle, WA: Vector Psychometric Group.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modelling [Computer software]*. Chicago, IL: Scientific Software International.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276. 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 2p tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173–194.

Chang, E. C., D'Zurilla, T. J., & Maydeu-Olivares, A. (1994). Assessing the dimensionality of optimism and pessimism using a multimeasure approach. *Cognitive Therapy and Research*, *18*, 143–160.

Cochran, W. G. (1952). The X^2 test of goodness of fit. *Annals of Mathematical Statistics*, *23*, 315–345.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491. doi:10.1037/1082-989X.9.4.466

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

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

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

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

Joe, H., & Maydeu-Olivares, A. (2006). On the asymptotic distribution of Pearson's X^2 in cross-validation samples. *Psychometrika*, *71*, 587–592.

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

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

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*, 4243–4258. doi:10.1016/j.csda.2012.04.010

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

Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, *24*, 492–516.

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

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

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*, *71*, 57–77.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement*, *11*, 71–101.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2(dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*, 55–64.

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling*, *18*, 333–356.

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

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

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

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

Maydeu-Olivares, A., & Montaño, R. (2013). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika*, *78*, 116–133. doi:10.1007/s11336-012-9293-1

McDonald, R. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

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.

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), 263–283. doi:10.1177/1073191111411667

Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika*, *61*, 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*, 331–360.

Reiser, M., & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, *47*, 85–107.

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

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

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the Psychometrika Society meeting, Iowa City, IA.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51–66). New York, NY: Springer-Verlag.

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*, 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*, 29–48.

APPENDIX

Derivation of the Asymptotic Distribution of the Sample RMSEA$_r$ in Equation (14) Under a Sequence of Local Alternatives Assumption for Any Consistent and Asymptotically Normal Estimator

The derivations presented here are very similar to those of Browne and Cudeck (1993) for structural equation models for continuous data. Consider n multinomial variables each with K categories and two models for the resulting $C = K^n$ contingency table: the fitted (null) model $\boldsymbol{\pi}_0 = \boldsymbol{\pi}\,(\boldsymbol{\theta})$, a parametric model that depends on q parameters to be estimated from the data, and the population probabilities $\boldsymbol{\pi}_T$. We assume that the population probabilities are related to the null model by the standard assumption of a sequence of local alternatives, $\boldsymbol{\pi}_T = \boldsymbol{\pi}_{T,N} = \boldsymbol{\pi}_0 + \boldsymbol{\delta}/\sqrt{N}$, where $\boldsymbol{\pi}_0 = \boldsymbol{\pi}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\theta}_0$ is in the interior of the parameter space. This assumption is also known as the parameter drift assumption. It implies that there is a sequence of probabilities converging to a point where the null model is satisfied. That is, as $N \to \infty$, $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_T$.

We assume that $\hat{\boldsymbol{\theta}}$, based on the null parametric model, is asymptotically normal under the sequence of local alternatives assumption (and consistent if $\boldsymbol{\delta} = 0$). Then, the asymptotic distribution of M_r is noncentral chi-square with $df_r = s_r$ – q degrees of freedom and noncentrality parameter (Joe & Maydeu-Olivares, 2010; Maydeu-Olivares & Joe, 2005):

$$\lambda_r = \lim_{N\to\infty} N\left(\boldsymbol{\pi}_r^T - \boldsymbol{\pi}_r^0\right)' \mathbf{C}_r^0 \left(\boldsymbol{\pi}_r^T - \pi_r^0\right) = \lim_{N\to\infty} ND_r, \tag{36}$$

where $\mathbf{C}_r$ given in Equation (3) is computed under the null model, and $\boldsymbol{\pi}_r^T$ depends on $\boldsymbol{\delta}/\sqrt{N}$. For ease of exposition, we assume $\boldsymbol{\delta}/\sqrt{N}$ and $D_r = \lambda_r/N$ for a large sample of size N.

Let $\hat{M}_r$ be the observed value of the M_r statistic for a data set. From the properties of the noncentral chi-square distribution, asymptotically,

$$\Pr(\hat{M}_r > M_r) = 1 - F_{\chi^2}(\hat{M}_r; df_r, \lambda_r), \tag{37}$$

and this is increasing as λ_r increases. Under $H_0^* : \varepsilon_r = \sqrt{D_r/df_r} \le c_r$, the largest value of $\lambda_r = ND_r$ is $N \times df_r \times c_r^2$ so that the p value for this H_0^* of close fit is

$$\max_{H_0^*}\left\{1 - F_{\chi^2}\left(\hat{M}_r; df_r, \lambda_r\right)\right\} = 1 - F_{\chi^2}\left(\hat{M}_r; df_r,\right.$$
$$\left. N \times df_r \times c_r^2\right). \tag{38}$$

The asymptotic 95% confidence interval for λ_r is $(\hat{L}_r, \hat{U}_r)$, where

$$1 - F_{\chi^2}(\hat{M}_r; df_r, \hat{L}_r) = 0.05,$$
$$1 - F_{\chi^2}(\hat{M}_r; df_r, \hat{U}_r) = 0.95, \tag{39}$$

provided $0.05 > 1 - F_{\chi^2}(\hat{M}_r; df_r, 0)$ (the right-hand side is the p value for H_0 of exact fit of the null parametric model). Because by Equation (13) and the aforementioned, $\varepsilon_r = \sqrt{D_r/df_r} = \sqrt{\lambda_r/(N \times df_r)}$, an asymptotic 95% confidence interval for ε_r is Equation (15).

Now, the mean of a noncentral chi-square distribution is its noncentrality parameter plus the degrees of freedom. Therefore, asymptotically, $E[M_r] = \lambda_r + df_r = \left(\lim\limits_{N\to\infty} ND_r\right) + s_r - q$. Using the method of moments we can estimate D_r using

$$\hat{D}_r = \frac{\hat{M}_r - df_r}{N}. \tag{40}$$

Finally, the point estimate of $\varepsilon_r = \sqrt{D_r/df_r}$ is Equation (14).

Computation of M$_2$. Consider n items coded as $Y_i = \{0, 1, \ldots, K-1\}$. That is, for ease of exposition we assume all items consist of the same number of categories, K. Let

$$\pi_i^a = \Pr(Y_i = a), \quad \pi_{ij}^{ab} = \Pr(Y_i = a, Y_j = b). \tag{41}$$

Then, $\dot{\boldsymbol{\pi}}_1$ denotes the set of all univariate population moments. Its dimension is $n(K-1)$, and its elements are $\pi_1^1, \pi_1^2, \ldots, \pi_1^{K-1}, \pi_2^1, \pi_2^2, \ldots, \pi_2^{K-1}, \ldots, \pi_n^1, \pi_n^2, \ldots, \pi_n^{K-1}$. Similarly, $\dot{\boldsymbol{\pi}}_2$ denotes the set of bivariate population moments. Its dimension is $\binom{n}{2}(K-1)^2$, and its elements are $\pi_{12}^{11}, \pi_{12}^{12}, \ldots, \pi_{12}^{K-1,K-1}, \ldots, \pi_{13}^{11}, \pi_{13}^{12}, \ldots, \pi_{13}^{K-1,K-1}, \ldots, \pi_{K-1,K}^{11}, \pi_{K-1,K}^{12}, \ldots, \pi_{K-1,K}^{K-1,K-1}$.

Then, $\boldsymbol{\pi}_2' = (\dot{\boldsymbol{\pi}}_1', \dot{\boldsymbol{\pi}}_2')$ is the set of univariate and bivariate moments, with sample counterpart $\mathbf{p}_2' = (\dot{\mathbf{p}}_1', \dot{\mathbf{p}}_2')$. The asymptotic covariance matrix of $\sqrt{N}\,(\mathbf{p}_2 - \boldsymbol{\pi}_2)$ is denoted by $\boldsymbol{\Xi}_2$. It can be partitioned according to the partitioning of p$_2$ into $\boldsymbol{\Xi}_{11} = \sqrt{N}\text{Acov}\,(\dot{\mathbf{p}}_1)$, $\boldsymbol{\Xi}_{21} = \sqrt{N}\text{Acov}\,(\dot{\mathbf{p}}_2, \dot{\mathbf{p}}_1)$, and $\boldsymbol{\Xi}_{22} = \sqrt{N}\text{Acov}\,(\dot{\mathbf{p}}_2, \dot{\mathbf{p}}_2)$, where Acov() denotes asymptotic covariance matrix. $\boldsymbol{\Xi}_{11}$, $\boldsymbol{\Xi}_{21}$, and $\boldsymbol{\Xi}_{22}$ have elements

$$\sqrt{N}\text{Acov}\left(p_i^a, p_j^b\right) = \pi_{ij}^{ab} - \pi_i^a \pi_j^b, \tag{42}$$

$$\sqrt{N}\text{Acov}\left(p_{ij}^{ab}, p_k^c\right) = \pi_{ijk}^{abc} - \pi_{ij}^{ab}\pi_k^c, i < j, \tag{43}$$

$$\sqrt{N}\text{Acov}\left(p_{ij}^{ab}, p_{kl}^{cd}\right) = \pi_{ijkl}^{abcd} - \pi_{ij}^{ab}\pi_{kl}^{cd},\ \text{i} < j, k < l, \tag{44}$$

respectively, where

$$\pi_{ijk}^{abc} = \Pr(Y_i = a, Y_j = b, Y_k = c),$$
$$\pi_{ijkl}^{abcd} = \Pr(Y_i = a, Y_j = b, Y_k = c, Y_l = d). \tag{45}$$

Now, in Equation (42)

$$\pi_{ij}^{ab} = \begin{cases} 0, & \text{if} \quad i = j, a \neq b, \\ \pi_i^a, & \text{if} \quad i = j, a = b, \\ \pi_{ij}^{ab}, & \text{otherwise.} \end{cases} \tag{46}$$

In Equation (43),

$$\pi_{ijk}^{abc} = \begin{cases} 0, & \text{if} \quad (i = k, a \neq c) \vee (j = k, b \neq c), \\ \pi_{ij}^{ab}, & \text{if} \quad (i = k, a = c) \vee (j = k, b = c), \\ \pi_{ijk}^{abc}, & \text{otherwise,} \end{cases} \tag{47}$$

whereas in Equation (44)

$$\pi_{ijkl}^{abcd} = \begin{cases} 0, & \text{if} \quad (i = k, a \neq c) \vee (i = l, a \neq d) \\ & \quad \vee (j = k, b \neq c) \vee (j = l, b \neq d), \\ \pi_{ij}^{ab}, & \text{if} \quad \{i, j\} = \{k, l\}, \{a, b\} = \{c, d\}, \\ \pi_{ijl}^{abc}, & \text{if} \quad (i = k, a = c) \vee (j = k, b = c), \\ \pi_{ijk}^{abd}, & \text{if} \quad (i = l, a = d) \vee (j = l, b = d), \\ \pi_{ijkl}^{abcd}, & \text{otherwise.} \end{cases} \tag{48}$$

Consider now a parametric model for the vector of population probabilities, $\boldsymbol{\pi}(\boldsymbol{\theta})$, where there is a q-parameter vector. The null model used throughout this article is the graded response model (GRM) with a standard normal latent trait. Under this model,

$$\pi_i^a = \int_{-\infty}^{\infty} \Pr(Y_i = a \,|\eta)\, \phi(\eta)\, d\eta,$$

$$\pi_{ij}^{ab} = \int_{-\infty}^{\infty} \Pr(Y_i = a \,|\eta) \Pr(Y_j = b \,|\eta) \phi(\eta)\, d\eta, i < j, \tag{49}$$

and analogous expressions hold for trivariate π_{ijk}^{abc} and four-way moments π_{ijkl}^{abcd}. In Equation (49) $\phi(\eta)$ denotes a standard normal density, and

$$\Pr(Y_i = a \,|\eta) = \begin{cases} 1 - \Psi_{i,1} & \text{if} \quad a = 0 \\ \Psi_{i,a} - \Psi_{i,a+1} & \text{if } 0 < a < K - 1 \\ \Psi_{i,K-1} & \text{if} \quad a = K - 1 \end{cases}, \tag{50}$$

$$\Psi_{i,a} = \frac{1}{1 + \exp[-(\alpha_{i,a} + \beta_i \eta)]}, \quad i = 1, \dots n, \tag{51}$$

with $\alpha_{i,a}$ decreasing in a for all i. To compute the test statistic M_2, the matrix $\boldsymbol{\Delta}_2 = \frac{\partial \boldsymbol{\pi}_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$ is needed. With $\boldsymbol{\theta}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$ we have

$$\boldsymbol{\Delta}_2 = \begin{pmatrix} \boldsymbol{\Delta}_{11} & \boldsymbol{\Delta}_{12} \\ \boldsymbol{\Delta}_{21} & \boldsymbol{\Delta}_{22} \end{pmatrix} = \begin{pmatrix} \frac{\partial \dot{\boldsymbol{\pi}}_1(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'} & \frac{\partial \dot{\boldsymbol{\pi}}_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \\ \frac{\partial \dot{\boldsymbol{\pi}}_2(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'} & \frac{\partial \dot{\boldsymbol{\pi}}_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \end{pmatrix}.$$

For this null model $\boldsymbol{\Delta}_{11}$ has elements

$$\frac{\partial \pi_i^a}{\partial \alpha_{\pi,g}} = \begin{cases} \displaystyle\int_{-\infty}^{\infty} \frac{\partial \Psi_{i,a}}{\partial \alpha_{i,a}} \phi(\eta)\, d\eta, & \text{if} \quad i = p, a = g, \\ \displaystyle -\int_{-\infty}^{\infty} \frac{\partial \Psi_{i,a}}{\partial \alpha_{i,a}} \phi(\eta)\, d\eta, & \text{if} \quad i = p, a + 1 = g, \\ 0, & \text{otherwise,} \end{cases} \tag{52}$$

where

$$\frac{\partial \Psi_{i,a}}{\partial \alpha_{i,a}} = \Psi_{i,a}(1 - \Psi_{i,a}), \tag{53}$$

$\boldsymbol{\Delta}_{21}$ has elements

$$\frac{\partial \pi_{ij}^{ab}}{\partial \alpha_{p,g}} = \begin{cases} \displaystyle\int_{-\infty}^{\infty} \frac{\partial \Psi_{i,a}}{\partial \alpha_{i,a}} \Pr(Y_j = b \,|\eta)\, \phi(\eta)\, d\eta & \text{if } i = p, a = g, \\ \displaystyle -\int_{-\infty}^{\infty} \frac{\partial \Psi_{i,a}}{\partial \alpha_{i,a}} \Pr(Y_j = b \,|\eta)\, \phi(\eta)\, d\eta, & \text{if } i = p, a + 1 = g, \\ \displaystyle\int_{-\infty}^{\infty} \frac{\partial \Psi_{j,b}}{\partial \alpha_{j,b}} \Pr(Y_i = a \,|\eta)\, \phi(\eta)\, d\eta, & \text{if } j = p, b = g, \\ \displaystyle -\int_{-\infty}^{\infty} \frac{\partial \Psi_{j,b}}{\partial \alpha_{j,b}} \Pr(Y_i = a \,|\eta)\, \phi(\eta)\, d\eta, & \text{if } j = p, b + 1 = g, \\ 0, & \text{otherwise.} \end{cases} \tag{54}$$

$\boldsymbol{\Delta}_{12}$ has elements

$$\frac{\partial \pi_i^a}{\partial \beta_p} = \begin{cases} \displaystyle\int_{-\infty}^{\infty} \frac{\partial \Psi_{i,a}}{\partial \beta_i} \phi(\eta)\, d\eta, & \text{if } i = p, a = K - 1, \\ \displaystyle -\int_{-\infty}^{\infty} \left(\frac{\partial \Psi_{i,a}}{\partial \beta_i} - \frac{\partial \Psi_{i,a+1}}{\partial \beta_i} \right) \phi(\eta)\, d\eta, & \text{if } i = p, a < K - 1, \\ 0, & \text{otherwise,} \end{cases} \tag{55}$$

where

$$\frac{\partial \Psi_{i,a}}{\partial \beta_i} = \eta \Psi_{i,a}(1 - \Psi_{i,a}). \tag{56}$$

Finally, $\mathbf{\Delta}_{22}$ has elements

$$\frac{\partial \Psi_{ij}^{ab}}{\partial \beta_p} = \begin{cases} \int\limits_{-\infty}^{\infty} \frac{\partial \Psi_{i,a}}{\partial \beta_i} \Pr\left(Y_j = b\,|\eta\right) \phi\left(\eta\right) d\eta, \\ \quad \text{if} \quad i = p, a = K-1, \\ -\int\limits_{-\infty}^{\infty} \left(\frac{\partial \Psi_{i,a}}{\partial \beta_i} - \frac{\partial Y_{i,a+1}}{\partial \beta_i}\right) \Pr\left(Y_j = b\,|\eta\right) \phi\left(\eta\right) d\eta, \\ \quad \text{if} \quad i = p, a < K-1, \\ \int\limits_{-\infty}^{\infty} \frac{\partial \Psi_{j,b}}{\partial \beta_j} \Pr\left(Y_i = a\,|\eta\right) \phi\left(\eta\right) d\phi, \\ \quad \text{if} \quad j = p, b = K-1, \\ -\int\limits_{-\infty}^{\infty} \left(\frac{\partial \Psi_{j,b}}{\partial \beta_j} - \frac{\partial Y_{j,b+1}}{\partial b_j}\right) \Pr\left(Y_i = a\,|\eta\right) \Psi\left(\eta\right) d\eta, \\ \quad \text{if} \quad j = p, b < K-1, \\ 0, \quad \text{otherwise.} \end{cases} \tag{57}$$

Computation of $\mathrm{M}_{\mathrm{ord}}$. The computation of M_{ord} involves $\mathbf{\Xi}_{ord}$, the asymptotic covariance matrix of $\sqrt{N}(\mathbf{m} - \kappa)$, and the matrix $\mathbf{\Delta}_{ord} = \frac{\partial \boldsymbol{\kappa}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$. $\mathbf{\Xi}_{ord}$ can be partitioned according to the partitioning of m into $\mathbf{\Xi}_{11} = \sqrt{N}\mathrm{Acov}\left(\mathbf{m}_1\right)$, $\mathbf{\Xi}_{21} = \sqrt{N}\mathrm{Acov}\left(\mathbf{m}_2, \mathbf{m}_1\right)$, and $\mathbf{\Xi}_{22} = \sqrt{N}\mathrm{Acov}\left(\mathbf{m}_2, \mathbf{m}_2\right)$. $\mathbf{\Xi}_{11}$, $\mathbf{\Xi}_{21}$, and $\mathbf{\Xi}_{22}$ have elements

$$\sqrt{N}\mathrm{Acov}(m_i, m_j) = \mathrm{E}[Y_i Y_j] - \mathrm{E}[Y_i]\mathrm{E}[Y_j], \tag{58}$$

$$\sqrt{N}\mathrm{Acov}(m_{ij}, m_k) = \mathrm{E}[Y_i Y_j Y_k] - \mathrm{E}[Y_i Y_j]\mathrm{E}[Y_k],\ i<j, \tag{59}$$

$$\sqrt{N}\mathrm{Acov}(m_{ij}, m_{kl}) = \mathrm{E}[Y_i Y_j Y_k Y_l] - \mathrm{E}[Y_i Y_j]\mathrm{E}[Y_k Y_l], \quad i < j\ ,\ k < l, \tag{60}$$

respectively. In Equation (58),

$$\mathrm{E}[Y_i Y_j] = \begin{cases} \mathrm{E}[Y_i^2], & \text{if } i = j, \\ \mathrm{E}[Y_i Y_j], & \text{otherwise.} \end{cases} \tag{61}$$

In Equation (59),

$$\mathrm{E}[Y_i Y_j Y_k] = \begin{cases} \mathrm{E}[Y_i^2 Y_j], & \text{if } i = k, \\ \mathrm{E}[Y_j^2 Y_i], & \text{if } j = k, \\ \mathrm{E}[Y_i Y_j], & \text{otherwise,} \end{cases} \tag{62}$$

and a similar expression is obtained for $\mathrm{E}[Y_i Y_j Y_k Y_l]$ in Equation (60).

For the null model employed in this article, the GRM with a normally distributed latent trait, these expressions can be computed as follows:

$$E\left[Y_i\right] = \int\limits_{-\infty}^{+\infty} \left[\sum_{a=1}^{K-1} \Psi_{i,a}\right] f(\eta) d\eta, \tag{63}$$

$$\mathrm{E}\left[Y_i Y_j\right] = \begin{cases} \int\limits_{-\infty}^{+\infty} \left[\sum_{a=1}^{K-1} (2a-1)\Psi_{i,a}\right] \phi(\eta) d\eta, & \text{if } i = j, \\ \int\limits_{-\infty}^{+\infty} \left[\sum_{a=1}^{K-1} \Psi_{i,a}\right]\left[\sum_{b=1}^{K-1} \Psi_{j,b}\right] \phi(\eta) d\eta, & \text{otherwise,} \end{cases} \tag{64}$$

$$\mathrm{E}\left[Y_i Y_j Y_k\right] = \begin{cases} \int\limits_{-\infty}^{+\infty} \left[\sum_{a=1}^{K-1} (2a-1)\Psi_{i,a}\right]\left[\sum_{b=1}^{K-1} Y_{j,b}\right] \phi(\eta) d\eta & \text{if } i = k, \\ \int\limits_{-\infty}^{+\infty} \left[\sum_{b=1}^{K-1} (2b-1)\Psi_{i,b}\right]\left[\sum_{a=1}^{K-1} \Psi_{i,a}\right] \phi(\eta) d\eta & \text{if } j = k, \\ \int\limits_{-\infty}^{+\infty} \left[\sum_{a=1}^{K-1} \Psi_{i,a}\right]\left[\sum_{b=1}^{K-1} \Psi_{j,b}\right]\left[\sum_{c=1}^{K-1} \Psi_{k,c}\right] \phi(\eta) d\eta & \text{otherwise,} \end{cases} \tag{65}$$

$$\mathrm{E}\left[Y_i Y_j Y_k Y_l\right] = \begin{cases} \int\limits_{-\infty}^{+\infty} \left[\sum_{a=1}^{K-1} (2a-1)\Psi_{i,a}\right]\left[\sum_{b=1}^{K-1} (2b-1)\Psi_{j,b}\right] \phi(\eta) d\eta \\ \quad \text{if } i = k,\ j = l, \\ \int\limits_{-\infty}^{+\infty} \left[\sum_{a=1}^{K-1} (2a-1)\Psi_{i,a}\right]\left[\sum_{b=1}^{K-1} \Psi_{j,b}\right]\left[\sum_{d=1}^{K-1} \Psi_{l,d}\right] \phi(\eta) d\eta \\ \quad \text{if } i = k, \\ \int\limits_{-\infty}^{+\infty} \left[\sum_{a=1}^{K-1} (2a-1)\Psi_{i,a}\right]\left[\sum_{b=1}^{K-1} \Psi_{j,b}\right]\left[\sum_{c=1}^{K-1} \Psi_{k,c}\right] \phi(\eta) d\eta \\ \quad \text{if } j = l, \\ \int\limits_{-\infty}^{+\infty} \left[\sum_{b=1}^{K-1} (2b-1)\Psi_{j,b}\right]\left[\sum_{a=1}^{K-1} \Psi_{i,a}\right]\left[\sum_{d=1}^{K-1} \Psi_{l,d}\right] \phi(\eta) d\eta \\ \quad \text{if } j = k, \\ \int\limits_{-\infty}^{+\infty} \left[\sum_{b=1}^{K-1} (2b-1)\Psi_{j,b}\right]\left[\sum_{a=1}^{K-1} \Psi_{i,a}\right]\left[\sum_{c=1}^{K-1} \Psi_{k,c}\right] \phi(\eta) d\eta \\ \quad \text{if } j = l, \\ \int\limits_{-\infty}^{+\infty} \left[\sum_{a=1}^{K-1} \Psi_{i,a}\right]\left[\sum_{b=1}^{K-1} \Psi_{j,b}\right]\left[\sum_{c=1}^{K-1} \Psi_{k,c}\right]\left[\sum_{d=1}^{K-1} \Psi_{l,d}\right] \phi(\eta) d\eta \\ \quad \text{otherwise.} \end{cases} \tag{66}$$

Similarly, $\mathbf{\Delta}_{ord}$ can be partitioned as

$$\mathbf{\Delta}_{ord} = \begin{pmatrix} \mathbf{\Delta}_{11} & \mathbf{\Delta}_{12} \\ \mathbf{\Delta}_{21} & \mathbf{\Delta}_{22} \end{pmatrix} = \begin{pmatrix} \frac{\partial \boldsymbol{\kappa}_1(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'} & \frac{\partial \boldsymbol{\kappa}_1(\beta)}{\partial \boldsymbol{\beta}'} \\ \frac{\partial \boldsymbol{\kappa}_2(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'} & \frac{\partial \boldsymbol{\kappa}_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \end{pmatrix}$$

and for this model $\mathbf{\Delta}_{11}$, $\mathbf{\Delta}_{12}$, $\mathbf{\Delta}_{21}$, and $\mathbf{\Delta}_{22}$ have elements

$$\frac{\partial E\left[Y_i\right]}{\partial \alpha_{p,g}} = \begin{cases} 0, & \text{if } j \neq i, \\ \int\limits_{-\infty}^{+\infty} \left[(1 - \Psi_{p,g})\Psi_{p,g}\right] \phi(\eta) d\eta, & \text{if } j = i, \end{cases} \tag{67}$$

$$\frac{\partial E\left[Y_i\right]}{\partial \beta_p} = \begin{cases} 0, & \text{if } j \neq i, \\ \int\limits_{-\infty}^{+\infty} \left\{ \sum_{g=1}^{K-1} \left[\left(1 - \Psi_{p,g}\right) \Psi_{p,g}\right] \right\} \eta\phi(\eta) d\phi, & \text{if } j = i, \end{cases} \tag{68}$$

$$\frac{\partial \mathrm{E}\left[Y_i Y_j\right]}{\partial \alpha_{p,g}} = \begin{cases} 0, & \text{if } p \neq i, j, \\ \int\limits_{-\infty}^{+\infty} \left(\sum_{b=1}^{K-1} \Psi_{j,b} \right) \left[1 - \Psi_{p,g}\right] \Psi_{p,g} \phi(\eta) d\eta, & \text{if } p = i, \\ \int\limits_{-\infty}^{+\infty} \left(\sum_{a=1}^{K-1} \Psi_{i,a} \right) \left[1 - \Psi_{p,g}\right] \Psi_{p,g} \phi(\eta) d\eta, & \text{if } p = j, \end{cases} \tag{69}$$

$$\frac{\partial \mathrm{E}\left[Y_i Y_j\right]}{\partial \beta_p} = \begin{cases} 0, & \text{if } p \neq i, j, \\ \int\limits_{-\infty}^{+\infty} \left(\sum_{b=1}^{K-1} \Psi_{j,b} \right) \left[\sum_{g=1}^{K-1} \left(\left[1 - \Psi_{p,g}\right] \Psi_{p,g} \right) \right] \eta\phi(\eta) d\eta, & \text{if } p = i, \\ \int\limits_{-\infty}^{+\infty} \left(\sum_{a=1}^{K-1} \Psi_{i,a} \right) \left[\sum_{g=1}^{K-1} \left(\left[1 - \Psi_{p,g}\right] \Psi_{p,g} \right) \right] \eta\phi(\eta) d\eta, & \text{if } p = j, \end{cases} \tag{70}$$

respectively.