

Published in final edited form as:

Br J Math Stat Psychol. 2013 May ; 66(2): 245–276. doi:10.1111/j.2044-8317.2012.02050.x.

Limited-information Goodness-of-fit Testing of Hierarchical Item Factor Models

Li Cai and Mark Hansen

University of California, Los Angeles

Abstract

In applications of item response theory, assessment of model fit is a critical issue. Recently, limited-information goodness-of-fit testing has received increased attention in the psychometrics literature. In contrast to full-information test statistics such as Pearson's X^2 or the likelihood ratio G^2 , these limited-information tests utilise lower order marginal tables rather than the full contingency table. A notable example is Maydeu-Olivares and colleagues' M_2 family of statistics based on univariate and bivariate margins. When the contingency table is sparse, tests based on M_2 retain better Type I error rate control than the full-information tests and can be more powerful. While in principle the M_2 statistic can be extended to test hierarchical multidimensional item factor models (e.g., bifactor and testlet models), the computation is non-trivial. To obtain M_2 , a researcher often has to obtain (many thousands of) marginal probabilities, derivatives, and weights. Each of these must be approximated with high-dimensional numerical integration. We propose a dimension reduction method that can take advantage of the hierarchical factor structure so that the integrals can be approximated far more efficiently. We also propose a new test statistic that can be substantially better calibrated and more powerful than the original M_2 statistic when the test is long and the items are polytomous. We use simulations to demonstrate the performance of our new methods and illustrate their effectiveness with applications to real data.

Keywords

model fit; multidimensional item response theory; hierarchical item factor analysis; bifactor model; testlet response model; two-tier model; dimension reduction

1 Introduction

In this paper we consider the application of limited-information tests (e.g., Bartholomew & Leung, 2002; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2005) to the case of hierarchical multidimensional item response theory (IRT) models (e.g., Cai, 2010b; Cai, Yang, & Hansen, 2011; Gibbons & Hedeker, 1992; Wainer, Bradlow, & Wang, 2007). In the context of IRT, limited-information goodness-of-fit test statistics have been suggested as a potential solution to the problem of extreme sparseness of the underlying multinomial table upon which the IRT model is defined (see, e.g., Bartholomew & Tzamourani, 1999). In contrast to the usual full-information test statistics (e.g., Pearson's statistic) that depend on probabilities of the full set of cells in the table (hence full-information), the limited-information statistics use only the lower-order marginal probabilities. While the idea of limited-information estimation is not new (e.g., Christofferson, 1975; Muthén, 1978; Jöreskog & Moustaki, 2001) and the seeds of limited-

information testing have been planted earlier in the psychometric literature (e.g., Reiser, 1996), the theory of limited-information goodness-of-fit tests has seen major new development in recent years largely thanks to a theoretical breakthrough made by Maydeu-Olivares and Joe (2005). We now have a comprehensive set of test statistics that have far better calibrated Type I error rates than their full-information counterparts under sparseness. We also understand that limited-information tests can be more powerful than full-information tests. They work for a variety of IRT models, full- and sub-tables alike (Maydeu-Olivares & Joe, 2006), and statistics not restricted to marginal tables have also been developed (Joe & Maydeu-Olivares, 2010).

At the same time, large-scale data collection efforts in both health outcomes research (e.g., PROMIS; Reeve et al., 2007) and educational surveys (e.g., PISA; Adams & Wu, 2002) have resulted in item response data for which hierarchical item factor models prove to be natural. In these contexts, unidimensionality is often violated either because of the complexity of the domain under consideration or because of the assembly of the measurement instruments. A prime example of the former is described in detail by Gibbons et al. (2007), who considered a quality-of-life scale for mental health patients that is made up of questions from a multitude of subdomains, such as family, health, finance, social, leisure, etc. These subdomains contribute to the measurement of quality-of-life, but they also introduce additional nuisance dimensions above and beyond the general quality-of-life dimension. Gibbons et al. (2007) fitted a bifactor model to the data to not only account for the local dependence due to the subdomains but also to reflect the essential goal of the measurement endeavour, which is to derive an estimate of the patient's quality-of-life. In educational testing, it is routine to organise several test items around a common stimulus, such as the questions following a passage in a reading comprehension test. The existence of these so-called "testlets" significantly alters the latent dimensionality of the tests. To obtain correct estimates of ability as well as standard errors of measurement, testlet response theory models have been proposed (Wainer et al., 2007). As an alternative, Cai, Yang, and Hansen (2011) also considered bifactor models for large-scale educational survey data.

In both bifactor and testlet response theory models, there is a general dimension on which all items load. In addition, there may be S group-specific dimensions that are mutually orthogonal, and orthogonal to the general dimension. As a critical feature, an item is allowed to load on at most one group-specific dimension. This is in contrast to standard item factor models (e.g., Reckase, 2009) wherein the factors may be correlated and items may load on more than two factors. Historically, Holzinger and Swineford's (1937) bifactor analysis and Thurstone's (1947) multiple factor analysis reflect two somewhat different intellectual leanings in factor analysis, with the former being a derivative of Spearman's (1904) one-factor model. In an attempt to merge the two types of multidimensional IRT models, Cai (2010b) proposed a two-tier model in which an item can load on P potentially correlated primary dimensions, as in Reckase's (2009) models, and at the same time, is only permitted to load on a single group-specific dimension, as in the bifactor model.

From a computational perspective, these hierarchical item factor models are highly attractive because maximum marginal likelihood estimation for these models requires at most $P + 1$ dimensional numerical integration as opposed to the full $P + S$ dimensional integration (Cai, 2010b; Gibbons & Hedeker, 1992; Rijmen, 2009). When P is (most often) equal to 1, the computational burden of obtaining the maximum likelihood solution does not increase exponentially as the number of group-specific dimensions increases. The standard (correlated factors) multidimensional models do not enjoy such freedom from the "curse of dimensionality." In some applications, the number of group-specific dimensions in the model may be quite large (Cai, 2010b, for example, presented an analysis of educational test data in which the fitted model included 12 group-specific factors), but dimension reduction

guarantees that the item parameters can still be estimated efficiently using maximum likelihood.

It would seem from the growing awareness and importance of limited-information goodness-of-fit testing and hierarchical item factor models that it is only natural to assess the fit of bifactor or testlet or two-tier models using limited-information test statistics – the best of both worlds. Yet to our knowledge, no researcher has attempted the limited-information testing of hierarchical item factor models.

Theoretically, existing results on limited-information goodness-of-fit testing can be straightforwardly extended to the case of multidimensional IRT. However, a major computational stumbling block remains. The limited-information statistics for a realistic psychological or educational test depend on the calculation of hundreds or sometimes thousands of model-implied probabilities and derivatives, all of which must be approximated by numerical integration. If $P+S$ dimensional integration is required, when S is 10 or more, the computation burden may render limited-information goodness-of-fit testing practically useless for hierarchical item factor models. This dilemma is similar to the state of affairs in IRT parameter estimation prior to Gibbons and Hedeker's (1992) discovery of the bifactor dimension reduction method.

Our main purpose in this paper is to extend the dimension reduction technique, already used in parameter estimation, to limited-information goodness-of-fit testing. We show that the dimensionality of integration problem can be directly tackled in the same manner as in item parameter estimation. For example, for a bifactor model, where P is equal to 1, the probabilities and derivatives involved in limited-information goodness-of-fit test statistics require at most 2 dimensional numerical integration, regardless of the number of factors in the model. As such, we are now able to test much larger models with much larger latent dimensionality than what was practically available previously.

As a secondary goal, we consider a new quadratic form test statistic, which we call M_2^* , that is based on the general limited-information testing principles proposed by Joe and Maydeu-Olivares (2010). The statistic is best understood as a further reduction (or concentration) of the univariate and bivariate marginal tables. The residuals used in the quadratic form are linear functions of the multinomial cell residuals, but they are not marginal moments. Maydeu-Olivares, Cai, and Hernandez (2011) proposed a similar (but not identical) statistic in the context of unidimensional graded IRT models.

We propose this new test statistic in response to a phenomenon well-known in limited-information estimation, but relatively obscure in limited-information goodness-of-fit testing. That is, when the items are polytomous, even the bivariate marginal tables can become too sparse. Consider the following example. For the quality-of-life of life scale analysed by Gibbons et al. (2007) and described above, respondents rate their satisfaction with respect to various aspects of their lives using a seven-point scale. Accordingly, each bivariate marginal table has 49 cells. Because all the items are intended to measure the same primary construct, we expect that the responses for any two items would generally have a positive relationship. Moreover, some response combinations (e.g., a high rating for one item, low for the other) would be rather uncommon or unexpected, especially for items belonging to the same subdomain.

To illustrate the situation, we fit a unidimensional logistic graded IRT model to data from 586 respondents to the quality-of-life scale and obtained marginal maximum likelihood item parameter estimates. We focus our attention on items 4 and 5, which belong to the “family” subdomain and are fairly strongly related to the underlying dimension (with standardised

factor loadings of .65 and .68, respectively). The parameter estimates for each item were used to generate univariate category response curves. For each of the 49 possible rating combinations, we took the product of the category response curve for item 4, the response curve for item 5, and the assumed (standard normal, in this example) population distribution. These distributions were then appropriately normalised to obtain the posterior of the particular rating combination. The posteriors of all combinations (i.e., all cells in the bivariate marginal table for this item pair) and the corresponding model-implied cell probabilities are shown in Figure 1. Not surprisingly, the cells with the highest model-implied probabilities and greatest posterior densities are along the main diagonal, where ratings for the two items are the same or within one or two points of each other. Cells in the bottom-left and upper-right corners, where the differences between the two ratings are most different, have rather low expected probabilities. In other words, we observe sparseness even in the second order marginal table.

Recall that the primary reason why limited-information test statistics have better chi-square approximations than the full-information test statistics is that the lower-order marginal tables are better filled. However, as we have seen, this central tenet in support of limited-information goodness-of-fit testing can be invalid when models for polytomous items are used. Again, as in full-information testing, the sparseness in the marginal tables leads to inadequate chi-square approximations. We will subsequently show that the distribution of limited-information test statistics as originally described by Maydeu-Olivares and Joe (2006) are often stochastically smaller than their theoretical limiting chi-square distributions for tests made up of polytomous items. This leads to severely mis-calibrated Type I error rates and significant loss of power to detect misspecification. On the other hand, by further condensing the marginal tables, the new test statistic has a much improved chi-square approximation and higher power than the original M_2 . The new test statistic reduces to the original M_2 when all items are dichotomous.

The remainder of this paper will be organised in the following manner. In Section 2, we present some technical details related to the computation of the original M_2 statistic and the proposed alternative. We describe in detail the dimension reduction strategy that may be utilised in the case of hierarchical item factor models to improve the efficiency of computing M_2 . In Section 3, we apply the limited-information test statistics to simulated data. First, we illustrate the improvement in computational efficiency achieved through dimension reduction. Then, we examine the distribution of the two M_2 statistics across a wide range of conditions, with and without model misspecification. Section 4 presents an analysis of an empirical dataset. We conclude in Section 5 with a brief discussion of this work, summarising our findings and highlighting directions for further research.

2 Technical Details

2.1 An Item Bifactor Analysis Model

Let there be n polytomously scored items so that the dichotomous responses become a special case. Let the i th response pattern be $\mathbf{y}_i = (y_{i1}, \dots, y_{in})$, where $y_{ij} = 0, \dots, K_j - 1$, where the number of categories K_j for item j can vary. The total number of possible response patterns is $C = \prod_{j=1}^n K_j$. Let there be a simple random (calibration) sample of size N . Without loss of generality, let us consider the logistic version of the graded item bifactor model (Cai, Yang, & Hansen, 2011) as a representative hierarchical item factor analysis model. Other models such as the two-tier model (Cai, 2010b) follow the derivations naturally, but we do not address them here due to space constraints. In the graded bifactor model, the cumulative response probabilities (i.e., the probability of a response in category k and above) are expressed as

$$\begin{aligned}
\Pr_j^+(0|\eta_0, \dots, \eta_s, \theta) &= 1, \\
\Pr_j^+(1|\eta_0, \dots, \eta_s, \theta) &= \frac{1}{1 + \exp(-\alpha_{j1} - \beta_{j0}\eta_0 - \beta_{js}\eta_s)}, \\
&\vdots \\
\Pr_j^+(k|\eta_0, \dots, \eta_s, \theta) &= \frac{1}{1 + \exp(-\alpha_{jk} - \beta_{j0}\eta_0 - \beta_{js}\eta_s)}, \\
&\vdots \\
\Pr_j^+(K_j-1|\eta_0, \dots, \eta_s, \theta) &= \frac{1}{1 + \exp(-\alpha_{jK_j-1} - \beta_{j0}\eta_0 - \beta_{js}\eta_s)},
\end{aligned}$$

where the latent variable η_0 is the general dimension with slope β_{j0} , and η_s is the s th group-specific dimension with slope β_{js} . Note that item j is permitted to load on at most one group-specific dimension s . For an item with K_j categories, there are also $K_j - 1$ intercepts (the α 's). The item parameters are components of the parameter vector $\theta \in \Theta$, where Θ is the parameter space. Denote the dimensionality of Θ as ν . This is equal to the number of unconstrained parameters in the model.

Taken together, the category response probability can be written as the difference:

$$\Pr_j(k|\eta, \theta) = \Pr_j^+(k|\eta_0, \dots, \eta_s, \theta) - \Pr_j^+(k+1|\eta_0, \dots, \eta_s, \theta), \quad (1)$$

where $\eta = (\eta_0, \eta_1, \dots, \eta_s, \dots, \eta_S)'$. For any generic item j , we may use

$$\pi(y_{ij}|\eta, \theta) = \prod_{k=0}^{K_j-1} [\Pr_j(k|\eta, \theta)]^{\chi_k(y_{ij})} \quad (2)$$

to denote the conditional density of the observed item response y_{ij} on the latent variables η , where $\chi_k(y_{ij})$ is an indicator function that takes on a value of 1 if and only if $y_{ij} = k$, and 0 otherwise. The formulation in Equation (2) is especially flexible because if y_{ij} is not among the legitimate response categories (e.g., missing data), it does not affect the estimation of θ .

The item bifactor model has $S + 1$ latent variables. As with any IRT model, the response pattern probability $\pi_i(\theta) = \pi(\theta|\mathbf{y}_i)$ is obtained by integrating over the latent variables,

$$\pi_i(\theta) = \int \prod_{j=1}^n \pi(y_{ij}|\eta, \theta) h(\eta) d\eta, \quad (3)$$

where $h(\eta)$ is the assumed (prior) distribution of the latent variables, typically taken as standard multivariate normal (see Cai, Yang, & Hansen, 2011).

2.2 Some Asymptotic Distribution Theory

As with any other IRT model, the item bifactor model is defined on an underlying n -way contingency table made up of the full list of C cross-classifications of the response patterns. Let the C (unstructured) multinomial cell probabilities be $\boldsymbol{\pi} = (\pi_1, \dots, \pi_i, \dots, \pi_C)'$. The corresponding observed cell proportions based on a sample of size N are $\mathbf{p} = (p_1, \dots, p_i, \dots, p_C)'$. The restrictions $\sum_{i=1}^C \pi_i = 1$ and $\sum_{i=1}^C p_i = 1$ are implicit. Following standard results in

Bishop, Fienberg, and Holland (1975), the asymptotic distribution of $\mathbf{p} - \boldsymbol{\pi}$ is C -variate normal with zero means and limiting covariance matrix $\boldsymbol{\Xi} = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$, i.e.,

$$\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{D} \mathcal{N}_C(\mathbf{0}, \boldsymbol{\Xi}). \quad (4)$$

The bifactor model imposes the following structure on the multinomial cell probabilities: $\boldsymbol{\pi}(\boldsymbol{\theta}) = [\pi_1(\boldsymbol{\theta}), \dots, \pi_A(\boldsymbol{\theta}), \dots, \pi_C(\boldsymbol{\theta})]'$. The log-likelihood is

$$\log L(\boldsymbol{\theta}) \propto \sum_{i=1}^C p_i \log \pi_i(\boldsymbol{\theta}). \quad (5)$$

Fitting this model to the observed proportions by maximum likelihood yields the solution $\hat{\boldsymbol{\theta}}$. Standard estimation algorithms such as the Bock-Aitkin EM algorithm (Bock & Aitkin, 1981) can be used to optimise the log-likelihood.

It is once again a standard discrete multivariate analysis result (see, e.g., Maydeu-Olivares & Joe, 2005 for a derivation) that the cell residual vector $\mathbf{e} = \mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ is also asymptotically normally distributed with zero means and limiting covariance matrix $\boldsymbol{\Omega}(\boldsymbol{\theta}) = \boldsymbol{\Xi}(\boldsymbol{\theta}) - \boldsymbol{\Delta}(\boldsymbol{\theta})[\mathcal{F}(\boldsymbol{\theta})]^{-1}\boldsymbol{\Delta}(\boldsymbol{\theta})'$, i.e.,

$$\sqrt{N}[\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})] \xrightarrow{D} \mathcal{N}_C[\mathbf{0}, \boldsymbol{\Omega}(\boldsymbol{\theta})], \quad (6)$$

where $\boldsymbol{\Delta}(\boldsymbol{\theta}) = \partial \boldsymbol{\pi}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is the $C \times \nu$ Jacobian matrix containing all first order partial derivatives of the model-implied probabilities with respect to the parameters, and $\mathcal{F}(\boldsymbol{\theta})$ is the Fisher information matrix $\mathcal{F}(\boldsymbol{\theta}) = \boldsymbol{\Delta}(\boldsymbol{\theta})' \text{diag}[\boldsymbol{\pi}(\boldsymbol{\theta})]^{-1} \boldsymbol{\Delta}(\boldsymbol{\theta})$.

Testing the goodness-of-fit of the model under maximum likelihood estimation involves the following composite hypothesis H_0 : there exists some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$. The null is tested against the alternative H_A : $\boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. While algorithms developed by Gibbons and Hedeker (1992) and extended by Cai, Yang, and Hansen (2011) have made parameter estimation for item bifactor analysis practical in routine applications, testing the goodness-of-fit of the models is not straightforward when n is relatively large. This issue is well-known in categorical data analysis. The number of cells increases exponentially as the number of items increases, which will quickly exceed any conceivable sample size in psychological and educational measurement. For more than a dozen or so items, the underlying contingency table is extremely sparse, with many cells having near zero expected

probabilities. Under sparseness, the Pearson test statistic $X^2 = N \sum_{i=1}^C [p_i - \pi_i(\hat{\boldsymbol{\theta}})]^2 / \pi_i(\hat{\boldsymbol{\theta}})$, or its large-sample equivalent, the likelihood ratio test statistic $G^2 = 2N \sum_{i=1}^C p_i \log [p_i / \pi_i(\hat{\boldsymbol{\theta}})]$, can no longer achieve the limiting central chi-square distribution with $C - \nu - 1$ degrees-of-freedom under the null hypothesis. As documented extensively in the literature, both Type I error rates and power are adversely affected when the contingency table is sparse (see, e.g., Bartholomew & Leung, 2002).

2.3 Limited-information Goodness-of-fit Testing

To resolve the sparseness issue, limited-information tests have been suggested, most notably by Maydeu-Olivares and Joe (2005). In limited-information goodness-of-fit tests, marginal sub-tables are used. A number of authors, including Maydeu-Olivares and Joe (2005) and

Cai et al. (2006), consider marginal residuals up to order 2, i.e., first and second order margins. The marginal residuals are obtained via operator matrices, e.g.,

$$\mathbf{e}_2 = \mathbf{L}_2 \mathbf{e} = \mathbf{L}_2 \mathbf{p} - \mathbf{L}_2 \pi(\hat{\theta}) = \mathbf{p}_2 - \pi_2(\hat{\theta}), \quad (7)$$

where \mathbf{L}_2 is a particular $d \times C$ fixed matrix of 0s and 1s that can collapse the cell residuals into first and second order marginal residuals (see, e.g., Maydeu-Olivares & Joe, 2006). Let us elaborate a little more on the dimensionality of the various matrices.

As Maydeu-Olivares and Joe (2006) explained, the number of linearly independent first order marginal residuals is equal to $d_1 = \sum_{j=1}^n (K_j - 1)$. The number of linearly independent second order marginal residuals is equal to $d_2 = \sum_{l=2}^n \sum_{m=1}^{l-1} (K_l - 1)(K_m - 1)$. Thus, taken together, $d = d_1 + d_2$. In partitioned form, we may also write

$$\mathbf{e}_2 = \begin{pmatrix} \dot{\mathbf{e}}_1 \\ \dot{\mathbf{e}}_2 \end{pmatrix} = \begin{pmatrix} \dot{\mathbf{L}}_1 \\ \dot{\mathbf{L}}_2 \end{pmatrix} \mathbf{e} = \begin{pmatrix} \dot{\mathbf{L}}_1 \\ \dot{\mathbf{L}}_2 \end{pmatrix} \mathbf{p} - \begin{pmatrix} \dot{\mathbf{L}}_1 \\ \dot{\mathbf{L}}_2 \end{pmatrix} \pi(\hat{\theta}) = \begin{pmatrix} \dot{\mathbf{p}}_1 - \dot{\pi}_1(\hat{\theta}) \\ \dot{\mathbf{p}}_2 - \dot{\pi}_2(\hat{\theta}) \end{pmatrix}, \quad (8)$$

where $\dot{\mathbf{L}}_1$ is a $d_1 \times C$ operator matrix that collapses the cell proportions and model-implied cell probabilities into d_1 first order marginal proportions $\dot{\mathbf{p}}_1 = \dot{\mathbf{L}}_1 \mathbf{p}$ and marginal probabilities $\dot{\pi}_1(\hat{\theta}) = \dot{\mathbf{L}}_1 \pi(\hat{\theta})$, and $\dot{\mathbf{L}}_2$ is a $d_2 \times C$ operator matrix that collapses the cell proportions and probabilities into d_2 second order marginal proportions $\dot{\mathbf{p}}_2 = \dot{\mathbf{L}}_2 \mathbf{p}$ and marginal probabilities $\dot{\pi}_2(\hat{\theta}) = \dot{\mathbf{L}}_2 \pi(\hat{\theta})$. It is also clear that \mathbf{e}_2 is made up of the first and second order marginal residuals $\dot{\mathbf{e}}_1$ and $\dot{\mathbf{e}}_2$.

For convenience, let us order the first order marginal probabilities by item number, i.e., $j = 1, \dots, n$. Similarly, let us order the second order marginal probabilities by unique item pairs, i.e., $(l, m) = (2, 1), (3, 1), (3, 2), \dots, (n, n-1)$. There are $n(n-1)/2$ such unique pairs. As mentioned above, for any item j , there are K_j cells in the first order marginal table, but there are only $K_j - 1$ linearly independent marginal probabilities because the K_j cells must sum to 1. These can be conveniently obtained by removing cells that correspond to the first category, the 0 category. By similar reasoning, while there are $K_l K_m$ cells in each bivariate table, there are only $(K_l - 1)(K_m - 1)$ linearly independent marginal probabilities. Removing the cells that correspond to the 0 category for both item l and item m would leave the linearly independent probabilities. Let us denote a generic first order marginal probability as $\pi_{1,j,k}(\hat{\theta})$. It can be interpreted as the model-implied probability of responses in category k for item j . Let a generic second order marginal probability be $\pi_{2,l,m,k_l k_m}(\hat{\theta})$. It is interpreted as the model-implied joint probability of responses in category k_l for item l , and in category k_m for item m .

The following is an example of obtaining first order marginal probabilities for a test with 3 items. The first two items have 3 categories, and the 3rd one has 2 categories:

$$\begin{aligned}
 \dot{\pi}_1(\widehat{\theta}) &= \begin{pmatrix} \dot{\pi}_{1,1,1}(\widehat{\theta}) \\ \dot{\pi}_{1,1,2}(\widehat{\theta}) \\ \dot{\pi}_{1,2,1}(\widehat{\theta}) \\ \dot{\pi}_{1,2,2}(\widehat{\theta}) \\ \dot{\pi}_{1,3,1}(\widehat{\theta}) \end{pmatrix} = \dot{\mathbf{L}}_1 \pi(\widehat{\theta}) \\
 &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_{(000)}(\widehat{\theta}) \\ \pi_{(001)}(\widehat{\theta}) \\ \pi_{(010)}(\widehat{\theta}) \\ \pi_{(011)}(\widehat{\theta}) \\ \pi_{(020)}(\widehat{\theta}) \\ \pi_{(021)}(\widehat{\theta}) \\ \pi_{(100)}(\widehat{\theta}) \\ \pi_{(101)}(\widehat{\theta}) \\ \pi_{(110)}(\widehat{\theta}) \\ \pi_{(111)}(\widehat{\theta}) \\ \pi_{(120)}(\widehat{\theta}) \\ \pi_{(121)}(\widehat{\theta}) \\ \pi_{(200)}(\widehat{\theta}) \\ \pi_{(201)}(\widehat{\theta}) \\ \pi_{(210)}(\widehat{\theta}) \\ \pi_{(211)}(\widehat{\theta}) \\ \pi_{(220)}(\widehat{\theta}) \\ \pi_{(221)}(\widehat{\theta}) \end{pmatrix}, \quad (9)
 \end{aligned}$$

where the subscripts in parentheses for the 18×1 vector above indicate a particular (reverse lexicographical) ordering of the $C = 18$ response patterns created by the 3 items. There are $2 + 2 + 1 = 5$ linearly independent first order marginal probabilities. In particular, this may be verified from the fact that $\dot{\mathbf{L}}_1$ has full row rank. To obtain the $(3 - 1)(3 - 1) + (3 - 1)(2 - 1) + (3 - 1)(2 - 1) = 8$ second order marginal probabilities, consider the following:

$$\begin{aligned}
 \dot{\pi}_2(\hat{\theta}) &= \begin{pmatrix} \dot{\pi}_{2,2,1,1,1}(\hat{\theta}) \\ \dot{\pi}_{2,2,1,1,2}(\hat{\theta}) \\ \dot{\pi}_{2,2,1,2,1}(\hat{\theta}) \\ \dot{\pi}_{2,2,1,2,2}(\hat{\theta}) \\ \dot{\pi}_{2,3,1,1,1}(\hat{\theta}) \\ \dot{\pi}_{2,3,1,1,2}(\hat{\theta}) \\ \dot{\pi}_{2,3,2,1,1}(\hat{\theta}) \\ \dot{\pi}_{2,3,2,1,2}(\hat{\theta}) \end{pmatrix} = \mathbf{L}_2 \dot{\pi}(\hat{\theta}) \\
 &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \pi_{(000)}(\hat{\theta}) \\ \pi_{(001)}(\hat{\theta}) \\ \pi_{(010)}(\hat{\theta}) \\ \pi_{(011)}(\hat{\theta}) \\ \pi_{(020)}(\hat{\theta}) \\ \pi_{(021)}(\hat{\theta}) \\ \pi_{(100)}(\hat{\theta}) \\ \pi_{(101)}(\hat{\theta}) \\ \pi_{(110)}(\hat{\theta}) \\ \pi_{(111)}(\hat{\theta}) \\ \pi_{(120)}(\hat{\theta}) \\ \pi_{(121)}(\hat{\theta}) \\ \pi_{(200)}(\hat{\theta}) \\ \pi_{(201)}(\hat{\theta}) \\ \pi_{(210)}(\hat{\theta}) \\ \pi_{(211)}(\hat{\theta}) \\ \pi_{(220)}(\hat{\theta}) \\ \pi_{(221)}(\hat{\theta}) \end{pmatrix}. \quad (10)
 \end{aligned}$$

The observed marginal proportions can be obtained similarly. We shall denote $\hat{p}_{1,j,k}$ as the observed first order marginal proportion for item j in category k . The observed second order marginal proportion for item l in category k_l and item m in category k_m is \hat{p}_{2,l,m,k_l,k_m} .

From the above derivations, it is clear that $\mathbf{e}_2 = \mathbf{p}_2 - \pi_2(\hat{\theta})$ is a linear combination of \mathbf{e} . The rank of \mathbf{L}_2 is equal to d . As such, the marginal residual moments are asymptotically d -variate normally distributed with zero means and limiting covariance matrix $\mathbf{\Omega}_2(\theta) = \mathbf{\Xi}_2(\theta) - \mathbf{\Delta}_2(\theta) [\mathcal{F}(\theta)]^{-1} \mathbf{\Delta}_2(\theta)'$,

$$\sqrt{N}[\mathbf{p}_2 - \pi_2(\hat{\theta})] \xrightarrow{D} \mathcal{N}_d[\mathbf{0}, \mathbf{\Omega}_2(\theta)], \quad (11)$$

where $\mathbf{\Xi}_2(\theta) = \mathbf{L}_2 \mathbf{\Xi}(\theta) \mathbf{L}_2'$ is $d \times d$, and the Jacobian $\mathbf{\Delta}_2(\theta) = \mathbf{L}_2 \mathbf{\Delta}(\theta) = \partial \pi_2(\theta) / \partial \theta$ is $d \times \nu$ containing all first order partial derivatives of the marginal probabilities with respect to the parameters of the model.

Maydeu-Olivares and Joe (2005) noted that if the model is locally identified from the marginal moments, i.e., when $\mathbf{\Delta}_2(\theta)$ has full column rank, then there exists a $d \times (d - \nu)$ orthogonal complement $\bar{\mathbf{\Delta}}_2(\theta)$ such that $\bar{\mathbf{\Delta}}_2(\theta)' \mathbf{\Delta}_2(\theta)$ is zero. By implication, $\bar{\mathbf{\Delta}}_2(\hat{\theta})' \mathbf{e}_2$ is $(d - \nu)$ -variate normal with zero means and limiting covariance matrix $\bar{\mathbf{\Delta}}_2(\theta)' \mathbf{\Xi}_2(\theta) \bar{\mathbf{\Delta}}_2(\theta)$, i.e.,

$$\sqrt{N} \bar{\mathbf{\Delta}}_2(\hat{\theta})' \mathbf{e}_2 = \bar{\mathbf{\Delta}}_2(\hat{\theta})' \sqrt{N}[\mathbf{p}_2 - \pi_2(\hat{\theta})] \xrightarrow{D} \mathcal{N}_{d-\nu}(\mathbf{0}, \bar{\mathbf{\Delta}}_2(\theta)' \mathbf{\Xi}_2(\theta) \bar{\mathbf{\Delta}}_2(\theta)). \quad (12)$$

Consequently, by arguments in Maydeu-Olivares and Joe (2005), the statistic

$$M_2 = N \mathbf{e}_2' \bar{\mathbf{\Delta}}_2(\hat{\theta}) [\bar{\mathbf{\Delta}}_2(\hat{\theta})' \mathbf{\Xi}_2(\hat{\theta}) \bar{\mathbf{\Delta}}_2(\hat{\theta})]^{-1} \bar{\mathbf{\Delta}}_2(\hat{\theta})' \mathbf{e}_2 \quad (13)$$

is asymptotically chi-square distributed with $d - \nu$ degrees-of-freedom under the null hypothesis that the model fits exactly.

2.4 A New Test Statistic

As discussed in Section 1, a potential problem with the original M_2 statistic as given in Equation (13) when the test items are polytomous is the sparseness of the bivariate marginal tables. We adopt the strategies endorsed by Joe and Maydeu-Olivares (2010) and Maydeu-Olivares et al. (2011) in a new test statistic that is based on a further reduction or concentration of the cells in the marginal tables. Specifically, let $\dot{e}_{1,j,k} = \dot{p}_{1,j,k} - \dot{\pi}_{1,j,k}(\hat{\theta})$ denote a first order marginal residual for item j in category k . We sum across the categories such that

$$\dot{e}_{1,j} = \sum_{k=1}^{K_j-1} k \dot{e}_{1,j,k} = \sum_{k=1}^{K_j-1} k [\dot{p}_{1,j,k} - \dot{\pi}_{1,j,k}(\hat{\theta})] = \begin{pmatrix} 1 & \cdots & K_j-1 \end{pmatrix} \begin{pmatrix} \dot{e}_{1,j,1} \\ \vdots \\ \dot{e}_{1,j,K_j-1} \end{pmatrix} = \dot{\mathbf{T}}_{1,j} \dot{\mathbf{e}}_{1,j} \quad (14)$$

is the *reduced* first order marginal residual for item j , where $\dot{\mathbf{T}}_{1,j}$ is a $1 \times (K_j - 1)$ vector and $\dot{\mathbf{e}}_{1,j} = (\dot{e}_{1,j,1}, \dots, \dot{e}_{1,j,K_j-1})'$ is the vector of all linearly independent first order marginal residuals for item j . Let $\dot{\mathbf{e}}_1 = (\dot{e}_{1,1}, \dots, \dot{e}_{1,n})'$ be the $n \times 1$ vector of reduced first order marginal residuals for all n items. We may represent $\dot{\mathbf{e}}_1$ using matrices

$$\dot{\mathbf{e}}_1 = \begin{pmatrix} \dot{e}_{1,1} \\ \vdots \\ \dot{e}_{1,j} \\ \vdots \\ \dot{e}_{1,n} \end{pmatrix} = \begin{pmatrix} \dot{\mathbf{T}}_{1,1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & & & \vdots \\ \vdots & & \dot{\mathbf{T}}_{1,j} & & \vdots \\ \vdots & & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \dot{\mathbf{T}}_{1,n} \end{pmatrix} \begin{pmatrix} \dot{e}_{1,1} \\ \vdots \\ \dot{e}_{1,j} \\ \vdots \\ \dot{e}_{1,n} \end{pmatrix} = \dot{\mathbf{T}}_1 \dot{\mathbf{e}}_1. \quad (15)$$

From Equation (15) and Equation (8), it is clear that $\dot{\mathbf{e}}_1 = \dot{\mathbf{T}}_1 \dot{\mathbf{e}}_1 = \dot{\mathbf{T}}_1 \dot{\mathbf{L}}_1 \mathbf{e}$, which shows that the reduced first order marginal residual vector $\dot{\mathbf{e}}_1$ is a linear function of the multinomial cell residuals. Furthermore, it is clear that $\dot{\mathbf{T}}_1$ is an $n \times d_1$ matrix that has full row rank. Therefore, the product $\dot{\mathbf{T}}_1 \dot{\mathbf{L}}_1$ has full row rank.

Moving on to the second order residuals. Let $\dot{e}_{2,l,m,k_l k_m} = \dot{p}_{2,l,m,k_l k_m} - \dot{\pi}_{2,l,m,k_l k_m}(\hat{\theta})$ be a second order marginal residual for item pair l and m in categories k_l and k_m , respectively. We define the *reduced* marginal residual for item pair l and m as a product moment:

$$\begin{aligned} \dot{\epsilon}_{2,l,m} &= \sum_{k_l=1}^{K_l-1} \sum_{k_m=1}^{K_m-1} k_l k_m \dot{\epsilon}_{2,l,m,k_l,k_m} = \sum_{k_l=1}^{K_l-1} \sum_{k_m=1}^{K_m-1} k_l k_m [\dot{p}_{2,l,m,k_l,k_m} - \pi_{2,l,m,k_l,k_m}(\hat{\theta})] \\ &= (\dot{\mathbf{T}}_{1,l} \otimes \dot{\mathbf{T}}_{1,m}) \begin{pmatrix} \dot{\epsilon}_{2,l,m,1,1} \\ \vdots \\ \dot{\epsilon}_{2,l,m,1,K_m-1} \\ \dot{\epsilon}_{2,l,m,2,1} \\ \vdots \\ \dot{\epsilon}_{2,l,m,2,K_m-1} \\ \vdots \\ \dot{\epsilon}_{2,l,m,K_l-1,1} \\ \vdots \\ \dot{\epsilon}_{2,l,m,K_l-1,K_m-1} \end{pmatrix} = \dot{\mathbf{T}}_{2,l,m} \dot{\mathbf{e}}_{2,l,m}, \end{aligned} \quad (16)$$

where $\dot{\mathbf{T}}_{2,l,m}$ is $1 \times (K_l - 1)(K_m - 1)$, formed by a direct product of $\dot{\mathbf{T}}_{1,l}$ with $\dot{\mathbf{T}}_{1,m}$, and $\dot{\mathbf{e}}_{2,l,m}$ is the vector of all linearly independent second order marginal residuals for item pair (l, m) . Let $\dot{\mathbf{e}}_2 = (\dot{\epsilon}_{2,2,1}, \dots, \dot{\epsilon}_{2,n,n-1})'$ be the $n(n-1)/2 \times 1$ vector of reduced second order marginal residuals for all $n(n-1)/2$ unique item pairs. We may represent $\dot{\mathbf{e}}_2$ using matrices

$$\dot{\mathbf{e}}_2 = \begin{pmatrix} \dot{\epsilon}_{2,2,1} \\ \vdots \\ \dot{\epsilon}_{2,i,m} \\ \vdots \\ \dot{\epsilon}_{2,n,n-1} \end{pmatrix} = \begin{pmatrix} \dot{\mathbf{T}}_{2,2,1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & & & \vdots \\ \vdots & & \dot{\mathbf{T}}_{2,i,m} & & \vdots \\ \vdots & & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \dot{\mathbf{T}}_{2,n,n-1} \end{pmatrix} \begin{pmatrix} \dot{\epsilon}_{2,2,1} \\ \vdots \\ \dot{\epsilon}_{2,i,m} \\ \vdots \\ \dot{\epsilon}_{2,n,n-1} \end{pmatrix} = \dot{\mathbf{T}}_2 \dot{\mathbf{e}}_2 = \dot{\mathbf{T}}_2 \dot{\mathbf{L}}_2 \mathbf{e}. \quad (17)$$

It is clear that the reduced second order marginal residual vector $\dot{\mathbf{e}}_2$ is a linear function of the multinomial cell residuals. Again, because $\dot{\mathbf{T}}_2$ is an $n(n-1)/2 \times d_2$ matrix that has full row rank, $\dot{\mathbf{T}}_2 \dot{\mathbf{L}}_2$ has full row rank as well.

Given the foregoing development, let the $n + n(n-1)/2 = n(n+1)/2 = \kappa$ reduced first and second order marginal residuals be

$$\mathbf{e}_2 = \begin{pmatrix} \dot{\mathbf{e}}_1 \\ \dot{\mathbf{e}}_2 \end{pmatrix} = \begin{pmatrix} \dot{\mathbf{T}}_1 & \mathbf{0} \\ \mathbf{0} & \dot{\mathbf{T}}_2 \end{pmatrix} \begin{pmatrix} \dot{\mathbf{L}}_1 \\ \dot{\mathbf{L}}_2 \end{pmatrix} \mathbf{e} = \mathbf{T}_2 \mathbf{e}_2. \quad (18)$$

From Equation (18), we have

$$\sqrt{N} \mathbf{e}_2 = \sqrt{N} \mathbf{T}_2 \mathbf{e}_2 \xrightarrow{D} \mathcal{N}_\kappa[\mathbf{0}, \Phi(\theta)], \quad (19)$$

where $\Phi(\theta) = \mathbf{T}_2 \Xi_2(\theta) \mathbf{T}_2' = \mathbf{T}_2 \Omega_2(\theta) \mathbf{T}_2' - \mathbf{T}_2 \Lambda_2(\theta) [\mathcal{F}(\theta)]^{-1} \Lambda_2'(\theta) \mathbf{T}_2'$. Rewrite $\mathbf{T}_2 \Omega_2(\theta) \mathbf{T}_2'$ as $\Psi_2(\theta)$ and $\mathbf{T}_2 \Lambda_2(\theta)$ as $\Gamma_2(\theta)$. We see that $\Phi(\theta) = \Psi_2(\theta) - \Gamma_2(\theta) [\mathcal{F}(\theta)]^{-1} \Gamma_2'(\theta)$ has a similar structure as $\Omega_2(\theta)$ in Equation (19). In particular, if the $\kappa \times \nu$ reduced marginal Jacobian $\Gamma_2(\theta)$ has full column rank, i.e., when the model is locally identified from the reduced first and second order marginal residuals, there exists a $\kappa \times (\kappa - \nu)$ orthogonal complement $\bar{\Gamma}_2(\theta)$ such that $\bar{\Gamma}_2(\theta)' \Gamma_2(\theta)$ is zero. Therefore, $\bar{\Gamma}_2(\hat{\theta})' \mathbf{e}_2$ is $(\kappa - \nu)$ -variate normal with zero means and limiting covariance matrix $\bar{\Gamma}_2(\theta)' \Psi_2(\theta) \bar{\Gamma}_2(\theta)$, i.e.,

$$\sqrt{N}\bar{\Gamma}_2(\hat{\theta})' \varepsilon_2 \xrightarrow{D} \mathcal{N}_{k-\nu}(\mathbf{0}, \bar{\Gamma}_2(\theta)' \Psi_2(\theta) \bar{\Gamma}_2(\theta)), \quad (20)$$

and we now present the new test statistic:

$$M_2^* = N \varepsilon_2' \bar{\Gamma}_2(\hat{\theta}) [\bar{\Gamma}_2(\theta)' \Psi_2(\theta) \bar{\Gamma}_2(\theta)^{-1} \bar{\Gamma}_2(\hat{\theta})] \varepsilon_2, \quad (21)$$

which is asymptotically chi-square distributed with $k - \nu$ degrees-of-freedom under the null hypothesis that the model fits exactly. Again, M_2 and M_2^* are numerically equivalent when all items are dichotomous because then the matrix \mathbf{T}_2 is going to be an identity matrix.

The further reduction in M_2^* is akin to using means and product moments for model fit testing. We will show using simulations that the new statistic has well calibrated Type I error rates and significantly higher power than the original M_2 when the test is made up of polytomous items. When the test items are dichotomous, \mathbf{T}_2 becomes an identity matrix, implying that M_2^* is equivalent to M_2 . A critical requirement of M_2^* is that the model must be locally identified from the set of reduced marginals so that the degrees-of-freedom is positive. For graded response bifactor models, the minimum number of items to achieve local identification is typically not excessive. If all test items are dichotomous, at least 6 items are needed. If the number of categories is equal to 5 for all items, the minimum is 12 items.

2.5 Computing the Marginals with Dimension Reduction

The observed marginal proportions are easy to tabulate from the observed data univariate and bivariate frequencies. On the other hand, while the model-implied marginal probabilities and elements of Δ_2 are defined as linear combinations of elements of π and Δ , their computation need not involve the \mathbf{L} matrices at all. The \mathbf{L} matrices are mostly useful as theoretical devices for studying the asymptotic distributions of the marginal residuals. The number of columns in \mathbf{L} is same as C , the total number of response patterns. Multiplications of dense matrices with dimension C quickly become infeasible as the number of items n increases.

2.5.1 The First and Second Order Marginal Probabilities—In reality, the marginal probabilities are computed directly, as integrals over the latent variable distribution. The integrals nominally have dimensionality equal to $S + 1$, and they must be numerically approximated using quadrature. For instance, the first order marginal probability $\pi_{1,j,k}(\hat{\theta})$ can be computed as

$$\begin{aligned} \pi_{1,j,k}(\hat{\theta}) &= \int_j \Pr(k|\eta, \hat{\theta}) h(\eta) d\eta \\ &\approx \underbrace{\sum_{q_S=1}^Q \cdots \sum_{q_0=1}^Q}_{(S+1) \text{ fold}} \Pr_j(k|X_{q_0}, \dots, X_{q_S}, \hat{\theta}) W(X_{q_0}) \cdots W(X_{q_S}), \quad (22) \end{aligned}$$

where X_{q_0}, \dots, X_{q_S} denotes a set of (direct product) quadrature nodes, with Q points per dimension, and $W(\cdot)$ denotes the corresponding quadrature weights, e.g., Gauss-Hermite nodes and weights (Abramowitz & Stegun, 1964). Similarly the second order marginal probability $\pi_{2,l,m,k|k_m}(\hat{\theta})$ is equal to

$$\begin{aligned} \dot{\pi}_{2,l,m,k_l,k_m}(\widehat{\theta}) &= \int \Pr(k_l|\eta, \widehat{\theta}) \Pr(k_m|\eta, \widehat{\theta}) h(\eta) d\eta \\ &\approx \underbrace{\sum_{q_s=1}^Q \cdots \sum_{q_0=1}^Q}_{(S+1) \text{ fold}} \Pr(k_l|X_{q_0}, \dots, X_{q_s}, \widehat{\theta}) \Pr(k_m|X_{q_0}, \dots, X_{q_s}, \widehat{\theta}) W(X_{q_0}) \cdots W(X_{q_s}). \end{aligned} \quad (23)$$

Clearly, when S is large, numerical approximation suffers from the curse of dimensionality since there must be Q^{S+1} function evaluations per marginal probability.

Just as in parameter estimation, we may use the item bifactor model's special feature (that each item loads on at most one group-specific dimension) to analytically reduce the dimensionality of the integrals. For the first order marginal probability, this is simple to accomplish. Assume for the moment that item j loads on group-specific dimension s . We have:

$$\begin{aligned} \dot{\pi}_{1,j,k}(\widehat{\theta}) &= \int \Pr(k|\eta, \widehat{\theta}) h(\eta) d\eta \\ &= \int \int \Pr(k|\eta_0, \eta_s, \widehat{\theta}) h(\eta_s) h(\eta_0) d\eta_s d\eta_0 \\ &= \int \left[\int \Pr(k|\eta_0, \eta_s, \widehat{\theta}) h(\eta_s) d\eta_s \right] h(\eta_0) d\eta_0 \quad (24) \\ &\approx \sum_{q_0=1}^Q \left[\sum_{q_s=1}^Q \Pr(k|X_{q_0}, X_{q_s}, \widehat{\theta}) W(X_{q_s}) \right] W(X_{q_0}). \end{aligned}$$

To go from the first to the second line in (24), we simply need to realise that besides the general dimension η_0 , out of all the group-specific dimensions $\eta_1, \dots, \eta_s, \dots, \eta_S$, item j is only influenced by η_s . For the second order marginal probability, we have to consider two distinct possibilities. When items l and m both load on the same group-specific dimension s , we have:

$$\begin{aligned} \dot{\pi}_{2,l,m,k_l,k_m}(\widehat{\theta}) &= \int \Pr(k_l|\eta, \widehat{\theta}) \Pr(k_m|\eta, \widehat{\theta}) h(\eta) d\eta \\ &= \int \int \Pr(k_l|\eta_0, \eta_s, \widehat{\theta}) \Pr(k_m|\eta_0, \eta_s, \widehat{\theta}) h(\eta_s) h(\eta_0) d\eta_s d\eta_0 \\ &= \int \left[\int \Pr(k_l|\eta_0, \eta_s, \widehat{\theta}) \Pr(k_m|\eta_0, \eta_s, \widehat{\theta}) h(\eta_s) d\eta_s \right] h(\eta_0) d\eta_0 \quad (25) \\ &\approx \sum_{q_0=1}^Q \left[\sum_{q_s=1}^Q \Pr(k_l|X_{q_0}, X_{q_s}, \widehat{\theta}) \Pr(k_m|X_{q_0}, X_{q_s}, \widehat{\theta}) W(X_{q_s}) \right] W(X_{q_0}). \end{aligned}$$

When item l loads on group-specific dimension s and item m loads on group-specific dimension t , we have instead of Equation (25):

$$\begin{aligned} \dot{\pi}_{2,l,m,k_l,k_m}(\widehat{\theta}) &= \int \Pr(k_l|\eta, \widehat{\theta}) \Pr(k_m|\eta, \widehat{\theta}) h(\eta) d\eta \\ &= \int \int \Pr(k_l|\eta_0, \eta_s, \widehat{\theta}) \Pr(k_m|\eta_0, \eta_t, \widehat{\theta}) h(\eta_s) h(\eta_t) h(\eta_0) d\eta_s d\eta_t d\eta_0 \\ &= \int \left[\int \Pr(k_l|\eta_0, \eta_s, \widehat{\theta}) h(\eta_s) d\eta_s \right] \left[\int \Pr(k_m|\eta_0, \eta_t, \widehat{\theta}) h(\eta_t) d\eta_t \right] h(\eta_0) d\eta_0 \quad (26) \\ &\approx \sum_{q_0=1}^Q \left[\sum_{q_s=1}^Q \Pr(k_l|X_{q_0}, X_{q_s}, \widehat{\theta}) W(X_{q_s}) \right] \left[\sum_{q_t=1}^Q \Pr(k_m|X_{q_0}, X_{q_t}, \widehat{\theta}) W(X_{q_t}) \right] W(X_{q_0}). \end{aligned}$$

In either case, the dimensionality of integration is equal to 2. More generally, regardless of the number of factors in the model, we can compute the marginal probabilities with $(P+1)$ -

dimensional quadrature, as long as the model is a bifactor ($P = 1$), testlet ($P = 1$), or two-tier ($P \geq 1$) type of hierarchical item factor model. The resulting computational savings in the calculation of goodness-of-fit has been previously unavailable to researchers.

2.5.2 The Elements of the Jacobian—As to the elements of the Jacobian, they are also integrals that can be approximated by the same quadrature rule as for the marginal probabilities. The dimensionality of integration can be reduced by the same line of reasoning. Consider a generic element θ in $\boldsymbol{\theta}$. Assuming that item j loads on group-specific dimension s , the derivative of the first order marginal probability $\pi_{1,j,k}(\boldsymbol{\theta})$ with respect to θ , when evaluated at $\hat{\boldsymbol{\theta}}$, can be numerically approximated as the following 2-dimensional integral:

$$\begin{aligned} \frac{\partial \pi_{1,j,k}(\hat{\boldsymbol{\theta}})}{\partial \theta} &= \int_j \Pr(k|\eta, \hat{\boldsymbol{\theta}}) \frac{\partial \log \Pr_j(k|\eta, \hat{\boldsymbol{\theta}})}{\partial \theta} h(\eta) d\eta \\ &= \int \left[\int_j \Pr(k|\eta_0, \eta_s, \hat{\boldsymbol{\theta}}) \frac{\partial \log \Pr_j(k|\eta_0, \eta_s, \hat{\boldsymbol{\theta}})}{\partial \theta} h(\eta_s) d\eta_s \right] h(\eta_0) d\eta_0 \quad (27) \\ &\approx \sum_{q_0=1}^Q \left[\sum_{q_s=1}^Q \Pr(k|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}}) \frac{\partial \log \Pr_j(k|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}})}{\partial \theta} W(X_{q_s}) \right] W(X_{q_0}). \end{aligned}$$

For the derivative of second order marginal probabilities, we find after some algebra

$$\begin{aligned} \frac{\partial \pi_{2,l,m,k_l,k_m}(\hat{\boldsymbol{\theta}})}{\partial \theta} &= \int_l \Pr(k_l|\eta, \hat{\boldsymbol{\theta}}) \Pr(k_m|\eta, \hat{\boldsymbol{\theta}}) \left[\frac{\partial \log \Pr_l(k_l|\eta, \hat{\boldsymbol{\theta}})}{\partial \theta} + \frac{\partial \log \Pr_m(k_m|\eta, \hat{\boldsymbol{\theta}})}{\partial \theta} \right] h(\eta) d\eta \\ &\approx \sum_{q_0=1}^Q \left[\sum_{q_s=1}^Q \Pr(k_l|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}}) \Pr(k_m|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}}) \frac{\partial \log \Pr_l(k_l|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}})}{\partial \theta} W(X_{q_s}) \right] + \\ &\quad \left[\sum_{q_s=1}^Q \Pr(k_l|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}}) \Pr(k_m|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}}) \frac{\partial \log \Pr_m(k_m|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}})}{\partial \theta} W(X_{q_s}) \right] W(X_{q_0}), \end{aligned}$$

provided that both item l and item m load on the same group-specific dimension s . As in the case in Equation (26), when item l loads on group-specific dimension s and item m loads on group-specific dimension t , the derivative can be numerically approximated as:

$$\begin{aligned} \frac{\partial \pi_{2,l,m,k_l,k_m}(\hat{\boldsymbol{\theta}})}{\partial \theta} &= \int_l \Pr(k_l|\eta, \hat{\boldsymbol{\theta}}) \Pr(k_m|\eta, \hat{\boldsymbol{\theta}}) \left[\frac{\partial \log \Pr_l(k_l|\eta, \hat{\boldsymbol{\theta}})}{\partial \theta} + \frac{\partial \log \Pr_m(k_m|\eta, \hat{\boldsymbol{\theta}})}{\partial \theta} \right] h(\eta) d\eta \\ &= \int \left[\int_l \Pr(k_l|\eta_0, \eta_s, \hat{\boldsymbol{\theta}}) \frac{\partial \log \Pr_l(k_l|\eta_0, \eta_s, \hat{\boldsymbol{\theta}})}{\partial \theta} h(\eta_s) d\eta_s \right] \left[\int_m \Pr(k_m|\eta_0, \eta_t, \hat{\boldsymbol{\theta}}) h(\eta_t) d\eta_t \right] h(\eta_0) d\eta_0 \\ &\quad \left[\int_l \Pr(k_l|\eta_0, \eta_s, \hat{\boldsymbol{\theta}}) h(\eta_s) d\eta_s \right] \left[\int_m \Pr(k_m|\eta_0, \eta_t, \hat{\boldsymbol{\theta}}) \frac{\partial \log \Pr_m(k_m|\eta_0, \eta_t, \hat{\boldsymbol{\theta}})}{\partial \theta} h(\eta_t) d\eta_t \right] h(\eta_0) d\eta_0 \\ &\approx \sum_{q_0=1}^Q \left[\sum_{q_s=1}^Q \Pr(k_l|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}}) \frac{\partial \log \Pr_l(k_l|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}})}{\partial \theta} W(X_{q_s}) \right] \left[\sum_{q_t=1}^Q \Pr(k_m|X_{q_0}, X_{q_t}, \hat{\boldsymbol{\theta}}) W(X_{q_t}) \right] + \\ &\quad \left[\sum_{q_s=1}^Q \Pr(k_l|X_{q_0}, X_{q_s}, \hat{\boldsymbol{\theta}}) W(X_{q_s}) \right] \left[\sum_{q_t=1}^Q \Pr(k_m|X_{q_0}, X_{q_t}, \hat{\boldsymbol{\theta}}) \frac{\partial \log \Pr_m(k_m|X_{q_0}, X_{q_t}, \hat{\boldsymbol{\theta}})}{\partial \theta} W(X_{q_t}) \right] W(X_{q_0}). \end{aligned}$$

We can see that the amount of computation required for the Jacobian elements is equal to a constant multiple of Q^2 function evaluations, after dimension reduction. For two-tier models, the number of function evaluations is equal to Q^{P+1} , when P is relatively small, as opposed to the full Q^{P+S} function evaluations without dimension reduction.

2.5.3 The Elements of the Weight Matrix—Dropping reference to $\boldsymbol{\theta}$ for the moment, the weight matrix $\mathbf{\Xi}_2$ can be written as

$$\Xi_2 = \mathbf{L}_2 \Xi \mathbf{L}_2' = \mathbf{L}_2 \text{diag}(\pi) \mathbf{L}_2' - \mathbf{L}_2 \pi \pi' \mathbf{L}_2' = \mathbf{L}_2 \text{diag}(\pi) \mathbf{L}_2' - \pi_2 \pi_2' = \sum -\pi_2 \pi_2'. \quad (28)$$

In the last expression, the second part involves the first and second order marginal probabilities in π_2 , which are already computed following results in Section 2.5.1. The first part $\sum = \mathbf{L}_2 \text{diag}(\pi) \mathbf{L}_2'$ involves second, third, and fourth order marginal probabilities. We may gain a clearer understanding of what it takes to compute the elements of Σ from a partitioning of \mathbf{L}_2 , as in Equation (8):

$$\sum = \mathbf{L}_2 \text{diag}(\pi) \mathbf{L}_2' = \begin{pmatrix} \dot{\mathbf{L}}_1 \text{diag}(\pi) \dot{\mathbf{L}}_1' & \\ \dot{\mathbf{L}}_2 \text{diag}(\pi) \dot{\mathbf{L}}_1' & \dot{\mathbf{L}}_2 \text{diag}(\pi) \dot{\mathbf{L}}_2' \end{pmatrix} = \begin{pmatrix} \sum_{11} & \\ \sum_{21} & \sum_{22} \end{pmatrix}.$$

Consider the upper-left block $\sum_{11} = \dot{\mathbf{L}}_1 \text{diag}(\pi) \dot{\mathbf{L}}_1'$. As is clear from Equation (9), each row of $\dot{\mathbf{L}}_1$ can be understood as a set of binary logical conditions that whether each of the C response patterns contributes to a particular first order marginal. Suppose row r of $\dot{\mathbf{L}}_1$ corresponds to the first-order marginal for item l in category k_l . And further let row c of $\dot{\mathbf{L}}_1$ corresponds to the first-order marginal for item m in category k_m . We realise that the (r, c) element of \sum_{11} is equal to the second order marginal probability $\pi_{2,l,m,k_l k_m}$ if $l \neq m$. If $l = m$, the entry is zero.

Now consider the lower-left block $\sum_{21} = \dot{\mathbf{L}}_2 \text{diag}(\pi) \dot{\mathbf{L}}_1'$. Suppose row r of $\dot{\mathbf{L}}_2$ corresponds to the second-order marginal wherein item l is in category k_l and item m is in category k_m . Let row c of $\dot{\mathbf{L}}_1$ correspond to the first-order marginal for item l' in category $k_{l'}$. Then the (r, c) element of \sum_{21} is equal to a third order marginal probability $\pi_{3,l,m,l',k_l,k_m,k_{l'}}$ i.e., indicating the joint probability of item l in category k_l , item m in category k_m , and item l' in category $k_{l'}$ if l, m , and l' are distinct. Recall that l and m are always distinct due to the construction of $\dot{\mathbf{L}}_2$ (see Equation 10). If l' is equal to either l or m and at the same time if $k_{l'}$ is equal to either k_l or k_m , then the (r, c) element of \sum_{21} is equal to the second order marginal probability $\pi_{2,l,m,k_l k_m}$. Otherwise the (r, c) element of \sum_{21} is equal to zero. Thus, we must be able to compute $\pi_{3,l,m,j,k_l k_m,k_{j'}}$ in addition to the pre-computed second order marginals.

Finally, consider the lower-right block $\sum_{22} = \dot{\mathbf{L}}_2 \text{diag}(\pi) \dot{\mathbf{L}}_2'$. Let row r of $\dot{\mathbf{L}}_2$ correspond to the second-order marginal wherein item l is in category k_l and item m is in category k_m . And let row c of $\dot{\mathbf{L}}_2$ correspond to the second-order marginal wherein item l' is in category $k_{l'}$ and item m' is in category $k_{m'}$. If l, m, l' , and k' are distinct, the (r, c) element of \sum_{22} is equal to the fourth order marginal $\pi_{4,l,m,l',m',k_l,k_m,k_{l'},k_{m'}}$. If l' is equal to either l or m and at the same time if $k_{l'}$ is equal to either k_l or k_m , the (r, c) element of \sum_{22} is equal to the third order marginal $\pi_{3,l,m,m',k_l,k_m,k_{m'}}$ or zero otherwise. If m' is equal to either l or m and at the same time if $k_{m'}$ is equal to either k_l or k_m , the (r, c) element of \sum_{22} is equal to the third order marginal $\pi_{3,l,m,l',k_l,k_m,k_{l'}}$ or zero otherwise. If $l = l'$ and $m = m'$ and at the same time $k_l = k_{l'}$ and $k_m = k_{m'}$, the (r, c) element of \sum_{22} is equal to the second order marginal $\pi_{2,l,m,k_l k_m}$ or zero otherwise.

We note here that the third and fourth order marginal probabilities can be computed with dimension reduction as well, so that the actual dimension of integration is equal to 2 for bifactor/testlet models and $P + 1$ for two-tier models with P primary dimensions. The computational details are lengthy and are relegated to the Appendix. The numeric engine of IRTPRO (Cai, Thissen, & du Toit, 2011) contains an implementation of M_2 and M_2^* used in the remainder of this paper.

3 Applications to Simulated Data

3.1 The Efficiency of Dimension Reduction

In order to illustrate the computational efficiency achieved through the strategy described above, we conducted a simple timing experiment. Data were simulated from a graded item bifactor model with 15 items in 2, 3, 4, or 5 categories. Three items loaded on each group-specific factor (in addition to the general dimension), and the number of group-specific factors (S) ranged from 2 to 5. The total number of dimensions ($S + 1$) thus ranged from 3 to 6. For each condition, the generating model was fit to the simulated data, and we recorded the time (in seconds) required to calculate M_2 with and without dimension reduction. Nine quadrature points spread evenly from -4 to 4 standard deviation units were used in approximating the marginal probabilities. This relatively small number is necessary to allow us to obtain results in the high-dimensional cases where we do not utilise dimension reduction. However, with dimension reduction, a larger number of quadrature points can and should be used, in order to obtain the best possible approximations of the marginal probabilities. The results for Maydeu-Olivares and Joe's (2005) M_2 test statistic based on the full first- and second-order marginal tables (Equation 13) are presented in Table 1. Without implementing dimension reduction, the time required to obtain the test statistic increases with both the number of categories and the number of dimensions. In contrast, when dimension reduction is used, the computing time remains relatively constant for models with a given number of response categories, regardless of the number of group-specific dimensions. Though not presented here, a similar pattern of results was observed for the alternative M_2^* statistic based on further reduced residuals.

3.2 Monte Carlo Simulations

A second simulation study was conducted to evaluate the performance of the test statistics under repeated sampling. Both Maydeu-Olivares and Joe's (2005) M_2 statistic based on the full first- and second-order marginal tables (Equation 13) and the alternative version M_2^* (Equation 21) were considered. We generated data in two sample sizes ($N = 750$ or 1500) for 15 items from models that varied in their factor structure and the number of response categories ($K = 2, 3, 4$, or 5).

The factor structures of the generating models in this study can be broadly characterised according to their "major" and "minor" factor domains (see, e.g., Tucker, Koopman, & Linn, 1969; MacCallum & Tucker, 1991). For each condition, the major domain consisted of either a single dimension or 4 dimensions with a bifactor structure (1 general dimension and 3 orthogonal group-specific dimensions). In some conditions, a minor factor domain was added to the generating model, utilizing a variation of the Tucker, Koopman, and Linn (1969) procedure (TKL). This minor domain consisted of 50 additional common factors that (1) were orthogonal to one another and to the dimension(s) in the major domain and (2) collectively accounted for some fixed proportion of variance not accounted for by the common factor(s) in the major domain.

The TKL procedure has been utilised in numerous simulation studies of principal components and factor analytic methods (see, e.g., MacCallum & Tucker, 1991; Hong,

1999; Briggs & MacCallum, 2003; Timmerman & Lorenzo-Seva, 2011). To our knowledge, this may be the first application to simulation studies in IRT, though Davey, Nering, and Thompson (1997) implemented an approach based on Tucker et al.'s (1969) logic. The key benefit of the TKL procedure, as articulated by Tucker et al. (1969), lies in its realism. The minor factors cannot, in principle, be represented parsimoniously in any item factor model. As such, the minor dimensions result in a kind of model misspecification that is likely to be quite pervasive in analyses of real data. Such misspecification is not tied to the omission of just one or a few parameters from the generating model to the fitted model, as in most other simulation studies.

In all, there were 6 different factor structures obtained by crossing the major and minor domain structures. Specifically, the generating models consisted of, first, either a unidimensional or 4-dimensional bifactor structure as the major domain and, second, additional minor common factors accounting for 0% (i.e., no additional minor factors), 10% (TKL10), or 30% (TKL30) of variance. For the conditions with no minor factors, an exactly correctly specified model does exist; it is either the unidimensional model or the bifactor model (depending on the major domain factor structure). These represent the null conditions. For TKL10 and TKL30, the unidimensional or bifactor models will be mildly misspecified (accounting for most influential dimensions while ignoring the minor ones).

The simulation procedure started with the major domain. The generating item slope parameters for the major domain factors were randomly drawn from a log-normal distribution with a mean of 0 and standard deviation of 0.2. A total of 30 slope parameters were drawn, 15 for the general dimension (also used in the unidimensional conditions) and 15 for the group-specific factors. Threshold values were randomly drawn from uniform distributions with ranges that varied according to the number of response categories. For each item, these values were then multiplied by the negative of the corresponding general dimension slope parameters; these products were used as intercept parameters. For simplicity, the same intercepts were used for all conditions with a given number of response categories, regardless of the factor structure. The generating item parameters for the unidimensional and bifactor model conditions (with no additional minor domain factors) are summarised in Table 2.

In order to obtain slopes for the minor domain factors, the slope parameters for the major domain factors in Table 2 were first transformed to standardised factor loadings. Next, factor loadings were randomly drawn using the TKL procedure for the 50 minor factors such that (1) successive minor factors would have diminishing influence on the item responses, and (2) the total variance explained by the minor factors would equal the intended value (either 10% or 30%). The loadings were then transformed back to the metric of logistic slopes. The item response data for the TKL10 and TKL30 conditions were then simulated from this (highly) multidimensional IRT model, with 51 total factors when the major domain consisted of a single factor and 54 total factors when the major domain had a bifactor structure.

For each condition, 1000 datasets were generated. After generating data, item parameters were estimated by fitting either a unidimensional or bifactor model using the Bock-Aitkin EM algorithm (Bock & Aitkin, 1981). Details concerning item bifactor model estimation are provided in Cai, Yang, and Hansen (2011). The number of quadrature points was $Q = 19$, and convergence was declared if maximum absolute inter-cycle parameter change dropped below 10^{-4} . In all cases, the fitted model corresponded to the structure of the “major” factor domain of the data generating model; minor domain factors were ignored, if present. This allowed us to test goodness-of-fit in both the null case and under what we consider to be rather mild model misspecification. Given the data and item parameter estimates, we

computed both M_2 and M_2^* test statistics for each replication. The dimension reduction strategy described in Section 2.5 was used for all bifactor conditions. Results from unconverged replications or in cases where numerical instability was evident were identified and omitted from subsequent analyses. Across all conditions, the minimum number of replications reported is 922.

3.2.1 Type I Error Rates—In Tables 3 and 4, we present simulation study results for the unidimensional and item bi-factor models, respectively, under the null condition, in which the fitted and generating models were the same. Here, we are concerned with whether the test statistics follow their asymptotic distribution and the extent to which the empirical model rejection rates match their nominal alpha levels. The results for the unidimensional and bifactor models are quite similar. For both models, it appears that the original M_2 test statistic is well-calibrated for dichotomous ($K = 2$) response data. However, as the number of response categories increases ($K > 2$), it appears that this M_2 statistic becomes stochastically smaller than the expected chi-square distribution. This is seen in the smaller mean value and in the empirical rejection rates, which fall well below the corresponding alpha levels, starting with 3 categories for the unidimensional model and 4 categories for the item bifactor model. In contrast, the M_2^* statistic based on the reduced marginal tables appears to be fairly well-calibrated for the conditions we tested. As Maydeu-Olivares et al. (2011), we calculated Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993) values from the M_2 and M_2^* statistics. Mean RMSEA and a 90% confidence interval (based on the the observed 5th and 95th percentiles across replications) are reported. The means ranged from .001 to .007. For all conditions, the lower limit of the 90% confidence interval was 0; the highest upper limit was .024.

Observed and expected quantiles of the M_2 and M_2^* statistics are presented in Figure 2. Two-sided Kolmogorov-Smirnov tests were performed to compare the observed and expected distributions; the resulting p -values are shown within each panel of Figure 2. Consistent with the results in Tables 3 and 4, the observed statistics are smaller than expected for the original M_2 statistic for items with 4 or more categories (3 or more for the unidimensional model), while the alternative M_2^* statistic seems to better follow the expected distribution.

3.2.2 Power to Detect Misspecification—We examined the sensitivity of the test statistics to model misspecification by fitting models which ignored the common factors in the minor domain of the data generating model. As described above, those minor domain factors collectively accounted for 10 or 30% of unique item variance in the TKL10 and TKL30 conditions, respectively. Table 5 presents results for the unidimensional model; results for the item bifactor model are shown in Table 6. Null condition results are included for comparison. The empirical rejection rates for all conditions are presented in a set of power curves in Figure 3.

Consistent with the results for the null case, the original M_2 statistic appears to have less power than the statistic based on the reduced marginal tables. These differences are most evident for the TKL10 conditions. It is interesting to note that the RMSEA values are generally quite small for the conditions examined in the study. The mean RMSEA based on the alternative M_2^* statistic was roughly .02 for TKL10 conditions and .05–.06 for TKL30; for the bifactor model, the corresponding means were .01 and .03–.04. These relatively small values are consistent with our intention of simulating data with rather mild misspecification.

4 An Application to Empirical Data

Having examined the performance of the test statistics in a range of simulated conditions, we now present an analysis of real data. A sample of 1042 daily cigarette smokers responded to 18 items regarding possible consequences of quitting smoking. Each item was rated on a 5-point ordinal scale. The items were administered within a larger questionnaire dealing with various attitudes, beliefs, and behaviors related to smoking (see Shadel, Edelen, & Tucker, 2011). The study was itself part of the development of the National Institute of Health's Patient Reported Outcomes Measurement Information System (PROMIS; see, e.g., Reeve et al., 2007).

All 18 items dealt with negative consequences of quitting, and it seems plausible that the items would be influenced by a common dimension. At the same time, there were small groups of items that dealt with rather similar consequences. Two items asked about potential difficulty concentrating, three items related to weight control, two items dealt with stress management, and so on. In total, we identified 7 possible groups among the 18 items. In similar testing contexts, such item groups with shared content create dependencies that cannot be fully accounted for by a single common dimension. The bifactor model provides a way to explicitly model these dependencies, while still assuming a common dimension. Thus, we consider the unidimensional and bifactor models as competing alternatives. We fit these models to the data and then used the M_2 and M_2^* statistics, as well as likelihood-based criteria, to compare them. The bifactor model used here allows all items to load on a general dimension. In addition, groups of two or three items load onto one of 7 group-specific factors, based on shared content. Thus, this model has a total of 8 dimensions. The results of this model comparison are presented in Table 7.

Based on the log-likelihood and information criteria, it appears that the bifactor model should be preferred. Because the unidimensional model is nested within the bifactor model, a likelihood ratio test may be performed. Under appropriate conditions, -2 times the difference in the log-likelihoods is asymptotically distributed as a central chi-square variable and provides a test of whether the constraints imposed in the nested model are correct (Haberman, 1977). In this case, the value of the likelihood ratio test statistic is over 3000, while the degrees of freedom (equal to the difference in the number of parameters estimated) is 13. The implication is that the restrictions imposed in going from the bifactor to the unidimensional model (which amount to fixing all group-specific factor slopes to 0) are not plausible. Such an assessment of relative fit has been a common approach to IRT model selection. Among the "appropriate conditions" for using the likelihood ratio test, however, is that the less constrained model be correctly specified (Maydeu-Olivares & Cai, 2006), which is a matter of absolute model fit. Here, we use the fit statistics to evaluate how well this condition is met (i.e., the extent to which the bifactor model fits the data).

We present results for both the original M_2 and alternative M_2^* statistics. Both the unidimensional and bifactor models are rejected by either statistic. This is not surprising, given the sensitivity of these statistics demonstrated in the simulation study. The RMSEA values provide an assessment of the severity of the model misspecification. Using the RMSEA computed from the original M_2 statistic, both the unidimensional and bifactor model appear to provide reasonably good fit, at least by standards suggested for linear factor analysis and structural equation models (e.g., Browne & Cudeck, 1993). Of course, the applicability of such suggestions to IRT models has not been fully investigated. Perhaps more importantly, however, we observed in the simulation studies that for polytomous data, the original M_2 statistic was often stochastically smaller than its expected distribution, and the discrepancy between the observed and expected distributions increased as the number of response categories increased. For data scored in 5 categories, as we have in this analysis,

the alternative M_2^* statistic, based on the reduction of the marginal tables, was found to have much better calibration. The RMSEA values based on M_2^* – in contrast to the original M_2 – reveal a substantial difference between the fit of the uni-dimensional and bifactor models. The RMSEA for the unidimensional model is .156 – a value that we consider to be indicative of rather poor fit (unlike the value of .049 obtained from the original M_2). Meanwhile, the RMSEA for the bifactor model is about .043, which is similar to the values observed in the simulation study for the TKL30 bifactor model conditions. We conclude that the 8-dimensional bifactor model fits the smoking data reasonably well. Had we used the original M_2 statistic and the associated RMSEA, we would have perhaps come to the wrong conclusion that the unidimensional model is reasonable for this data set.

5 Discussion

It is a well-known problem in IRT modelling that full-information tests such as the likelihood ratio G^2 and Pearson's X^2 do not follow their asymptotic distributions due to sparseness in the full contingency table. As a consequence, tests of absolute model fit based on these statistics may not be trustworthy. Work by Maydeu-Olivares and colleagues has demonstrated that limited-information statistics based on lower-order margins are a promising alternative to full-information statistics. The M_2 family of statistics is a notable example. M_2 is based on the one- and two-way margins of the contingency table, thus less susceptible to sparseness. In previous studies (see, e.g., Maydeu-Olivares & Joe, 2006), M_2 has been shown to have better calibration than the full-information tests. Moreover, it can be more powerful to detect model misspecification.

In this research, we considered the application of limited-information goodness-of-fit testing to hierarchical multidimensional models, which include item bifactor models (Gibbons et al., 2007; Cai, Yang, & Hansen, 2011), testlet response models (Wainer et al., 2007), and two-tier models (Cai, 2010b). Extension of M_2 to these conditions is non-trivial; approximating marginal response pattern probabilities, derivatives, and weights can involve high-dimensional integration. However, for hierarchical IRT models, some of this complexity can be alleviated. We presented an approach to computing M_2 that takes advantage of the constraints that characterise these models in order to greatly improve the computational efficiency. This analytical dimension reduction is closely related to the strategies for estimating hierarchical models (e.g., Gibbons & Hedeker, 1992) that have contributed to the popularity of these models. After deriving the strategy, we demonstrated the improved efficiency through a timing experiment with simulated data.

We have also considered the use of M_2 with polytomous data. Although limited-information testing is based on the premise that lower-order margins are better filled than the full contingency table, we showed that there are circumstances in which even the lower-order margins can become sparse. This is not to suggest that the limited-information statistics would be inferior to the full-information statistics in such cases. After all, sparseness at the lower-order margins guarantees sparseness in higher-order margins and in the full contingency table. However, the sparseness undermines the ability of the limited-information statistics to overcome the very problem they were intended to address. This problem was anticipated by Joe and Maydeu-Olivares (2010) and Maydeu-Olivares et al. (2011), who suggested that the lower-order margins might be further reduced or summarised. Following this suggestion, we have proposed a new M_2^* statistic for polytomous data and hierarchical item factor models that is based on a collapsing of the first- and second-order marginal tables.

In simulation studies, we demonstrated that the original M_2 statistic performs well with dichotomous data for both the unidimensional and bifactor models. However, as the number

of categories increased, it was shown that the original test statistic was stochastically smaller than expected, resulting in Type I error rates that were well below the nominal alpha levels. In contrast, the alternative M_2^* demonstrated rejection rates very close to the alpha level, even as the number of response categories (and, thus, the size of each marginal contingency table) increases. In addition, the proposed M_2^* statistic was more sensitive to model misspecification, demonstrating higher empirical rejection rates for the conditions of model misspecification tested. We used the M_2 and M_2^* statistics to evaluate the fit of two models to an empirical data set, showing how the ability to assess absolute model fit provides a useful complement to tests of relative fit.

In both simulation study and our empirical example, RMSEA values were calculated. In the simulation study, these were found to be rather small by traditional standards (e.g., Browne & Cudeck, 1993) for the conditions tested. This may be indicative of the fact that ignoring “minor” factors (present in the data generating models but absent from the fitted models) constituted only a rather mild level of model misspecification. On the other hand, it is possible that the suggested guidelines for interpreting RMSEA, based on experience fitting models for continuous data, may not apply to IRT models. Given the apparent power of the test statistics to reject misspecified models, further research will be needed in order to develop guidelines for interpreting RMSEA values (and any other indices we might derive from the M_2 statistics). Such guidelines could suggest what RMSEA values might be expected for a “good”-fitting model, as well as the values that would suggest only “acceptable” or even “poor” fit. The analogous guidelines for factor analysis and structural equation models were based largely on experience – developed over time – of fitting many models and observing the values of the fit indices. Our hope is that by developing test statistics for IRT models and by making them available in software, practitioners may begin to accumulate that sort of experience.

Although we have demonstrated the computation of M_2 for hierarchical IRT models, we have not addressed the challenge of obtaining the test statistic for high-dimensional models that do not conform to this structure. The strategy we implemented (analytical dimension reduction) to achieve improved efficiency followed an approach already applied in model estimation. Accordingly, methods for obtaining fit statistics for general multidimensional IRT models could perhaps draw upon the strategies developed for estimating those models (e.g., Cai, 2010a).

Acknowledgments

Part of this research is made possible by a pre-doctoral training grant (R305B080016) and a statistical methodology grant (R305D100039) from the Institute of Education Sciences. Li Cai’s research is also supported by grants from the National Institute on Drug Abuse (R01DA026943 and R01DA030466).

References

- Abramowitz, M.; Stegun, IA. Handbook of mathematical functions with formulas, graphs, and mathematical tables. New York, NY: Dover; 1964.
- Adams, R.; Wu, M. PISA 2000 technical report. Paris: Organization for Economic Cooperation and Development; 2002.
- Bartholomew DJ, Leung SO. A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*. 2002; 55:1–15. [PubMed: 12034008]
- Bartholomew DJ, Tzamourani P. The goodness-of-fit of latent trait models in attitude measurement. *Sociological Methods and Research*. 1999; 27:525–546.
- Bishop, YMM.; Fienberg, SE.; Holland, PW. Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press; 1975.

- Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*. 1981; 46:443–459.
- Briggs N, MacCallum RC. Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*. 2003; 38:25–56.
- Browne, MW.; Cudeck, R. Alternative ways of assessing model fit. In: Bollen, K.; Long, J., editors. *Testing structural equation models*. Newbury Park, CA: Sage; 1993. p. 136–162.
- Cai L. High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*. 2010a; 75:33–57.
- Cai L. A two-tier full-information item factor analysis model with applications. *Psychometrika*. 2010b; 75:581–612.
- Cai L, Maydeu-Olivares A, Coffman DL, Thissen D. Limited-information goodness-of-fit testing of item response theory models for sparse 2^P tables. *British Journal of Mathematical and Statistical Psychology*. 2006; 59:173–194. [PubMed: 16709285]
- Cai, L.; Thissen, D.; du Toit, SHC. IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International, Inc; 2011.
- Cai L, Yang JS, Hansen M. Generalized full-information item bifactor analysis. *Psychological Methods*. 2011; 16:221–248. [PubMed: 21534682]
- Christofferson A. Factor analysis of dichotomized variables. *Psychometrika*. 1975; 40:5–32.
- Davey, T.; Nering, ML.; Thompson, T. Realistic simulation of item response data (ACT Research Report Series No 97–4). Iowa City, IA: ACT; 1997.
- Gibbons RD, Bock RD, Hedeker D, Weiss DJ, Segawa E, Bhaumik DK, Grochocinski VJ. Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*. 2007; 31:4–19.
- Gibbons RD, Hedeker D. Full-information item bifactor analysis. *Psychometrika*. 1992; 57:423–436.
- Haberman SJ. Log-linear models and frequency tables with small expected cell counts. *The Annals of Statistics*. 1977; 5:1148–1169.
- Holzinger KJ, Swineford F. The bi-factor method. *Psychometrika*. 1937; 2:41–54.
- Hong S. Generating correlation matrices with model error for simulation studies in factor analysis: A combination of the Tucker-Koopman-Linn model and Wijsman's algorithm. *Behavior Research Methods, Instruments, & Computers*. 1999; 31:727–730.
- Joe H, Maydeu-Olivares A. A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*. 2010; 75:393–419.
- Jöreskog KG, Moustaki I. Factor analysis of ordinal variables: a comparison of three approaches. *Multivariate Behavioral Research*. 2001; 36:347–387.
- MacCallum RC, Tucker LR. Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*. 1991; 109:502–511.
- Maydeu-Olivares A, Cai L. A cautionary note on using $G^2(\text{dif})$ to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*. 2006; 41:55–64.
- Maydeu-Olivares A, Cai L, Hernandez A. Comparing the fit of IRT and factor analysis models. *Structural Equation Modeling*. 2011; 18:333–356.
- Maydeu-Olivares A, Joe H. Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*. 2005; 100:1009–1020.
- Maydeu-Olivares A, Joe H. Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*. 2006; 71:713–732.
- Muthén B. Contributions of factor analysis to dichotomous variables. *Psychometrika*. 1978; 43:551–560.
- Reckase, MD. *Multidimensional item response theory*. New York, NY: Springer; 2009.
- Reeve BB, Hays RD, Björner JB, Cook KF, Crane PK, Teresi JA, Cella D. Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Medical Care*. 2007; 45:S22–31. [PubMed: 17443115]

- Reiser M. Analysis of residuals for the multinomial item response theory model. *Psychometrika*. 1996; 61:509–528.
- Rijmen, F. Efficient full information maximum likelihood estimation for multidimensional IRT models (Tech Rep No RR-09-03). Educational Testing Service; 2009.
- Shadel WG, Edelen M, Tucker JS. A unified framework for smoking assessment: The PROMIS Smoking Initiative. *Nicotine & Tobacco Research*. 2011; 13:399–400. [PubMed: 21330279]
- Spearman C. General intelligence objectively determined and measured. *American Journal of Psychology*. 1904; 15:201–293.
- Thurstone, LL. Multiple factor analysis. Chicago, IL: The University of Chicago Press; 1947.
- Timmerman ME, Lorenzo-Seva R. Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*. 2011; 16:209–220. [PubMed: 21500916]
- Tucker LR, Koopman RF, Linn RL. Evaluation of factor analytic procedures by means of simulated correlation matrices. *Psychometrika*. 1969; 34:421–459.
- Wainer, H.; Bradlow, ET.; Wang, X. Testlet response theory and its applications. New York, NY: Cambridge University Press; 2007.

Appendix: Computation of Elements of the Weight Matrix

The computation of third and fourth order moments continues to benefit from dimension reduction. Let us consider the third order moment $\pi_{3,l,m,j,k_l,k_m,k_j}$ as a case in point. When items l , m , and j all load on the same group-specific dimension s , the marginal probability is simply

$$\begin{aligned}\dot{\pi}_{3,l,m,j,k_l,k_m,k_j} &= \int \int \Pr(k_l|\eta_0, \eta_s, \hat{\theta}) \Pr(k_m|\eta_0, \eta_s, \hat{\theta}) \Pr(k_j|\eta_0, \eta_s, \hat{\theta}) h(\eta_s) h(\eta_0) d\eta_s d\eta_0 \\ &= \int \left[\int \Pr(k_l|\eta_0, \eta_s, \hat{\theta}) \Pr(k_m|\eta_0, \eta_s, \hat{\theta}) \Pr(k_j|\eta_0, \eta_s, \hat{\theta}) h(\eta_s) d\eta_s \right] h(\eta_0) d\eta_0 \\ &\approx \sum_{q_0=1}^Q \left[\sum_{l=1}^Q \Pr(k_l|X_{q_0}, X_{q_s}, \hat{\theta}) \Pr(k_m|X_{q_0}, X_{q_s}, \hat{\theta}) \Pr(k_j|X_{q_0}, X_{q_s}, \hat{\theta}) W(X_{q_s}) \right] W(X_{q_0}).\end{aligned}$$

When items l , m load on group-specific dimension s and j on group-specific dimension t , the marginal probability becomes

$$\begin{aligned}\dot{\pi}_{3,l,m,j,k_l,k_m,k_j} &= \int \int \Pr(k_l|\eta_0, \eta_s, \hat{\theta}) \Pr(k_m|\eta_0, \eta_s, \hat{\theta}) \Pr(k_j|\eta_0, \eta_t, \hat{\theta}) h(\eta_s) h(\eta_t) h(\eta_0) d\eta_s d\eta_t d\eta_0 \\ &= \int \left[\int \Pr(k_l|\eta_0, \eta_s, \hat{\theta}) \Pr(k_m|\eta_0, \eta_s, \hat{\theta}) h(\eta_s) d\eta_s \right] \left[\int \Pr(k_j|\eta_0, \eta_t, \hat{\theta}) h(\eta_t) d\eta_t \right] h(\eta_0) d\eta_0 \\ &\approx \sum_{q_0=1}^Q \left[\sum_{l=1}^Q \Pr(k_l|X_{q_0}, X_{q_s}, \hat{\theta}) \Pr(k_m|X_{q_0}, X_{q_s}, \hat{\theta}) W(X_{q_s}) \right] \left[\sum_{j=1}^Q \Pr(k_j|X_{q_0}, X_{q_t}, \hat{\theta}) W(X_{q_t}) \right] W(X_{q_0}).\end{aligned}$$

Finally, when items l , m , and j each loads on a different group-specific dimension, say, s , t , and u , the marginal probability $\pi_{3,l,m,j,k_l,k_m,k_j}$ becomes

$$\begin{aligned}
& \int \left[\int \Pr(k_l | \eta_0, \eta_s, \widehat{\theta}) h(\eta_s) d\eta_s \right] \left[\int \Pr(k_m | \eta_0, \eta_t, \widehat{\theta}) h(\eta_t) d\eta_t \right] \left[\int \Pr(k_j | \eta_0, \eta_u, \widehat{\theta}) h(\eta_u) d\eta_u \right] h(\eta_0) d\eta_0 \\
& \approx \sum_{q_0=1}^Q \left[\sum_{q_s=1}^Q \Pr(k_l | X_{q_0}, X_{q_s}, \widehat{\theta}) W(X_{q_s}) \right] \left[\sum_{q_t=1}^Q \Pr(k_m | X_{q_0}, X_{q_t}, \widehat{\theta}) W(X_{q_t}) \right] \times \\
& \quad \left[\sum_{q_u=1}^Q \Pr(k_j | X_{q_0}, X_{q_u}, \widehat{\theta}) W(X_{q_u}) \right] W(X_{q_0}).
\end{aligned}$$

The case of fourth order moment is completely analogous.

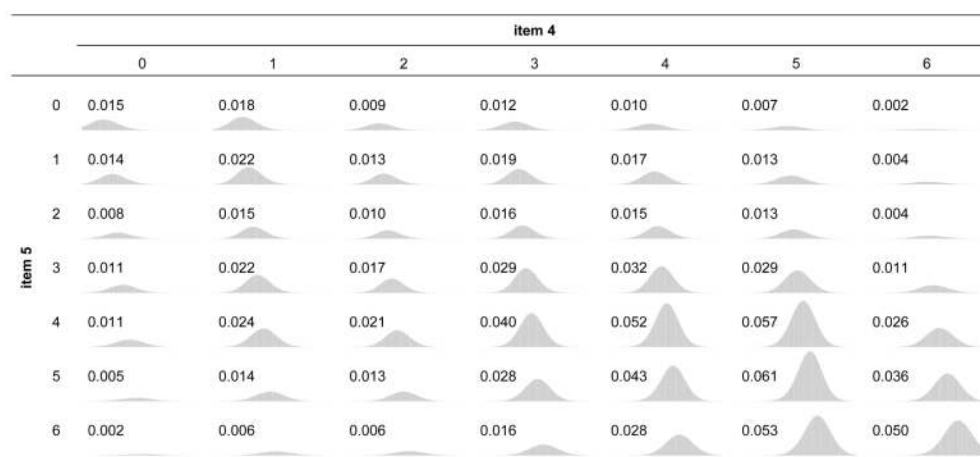
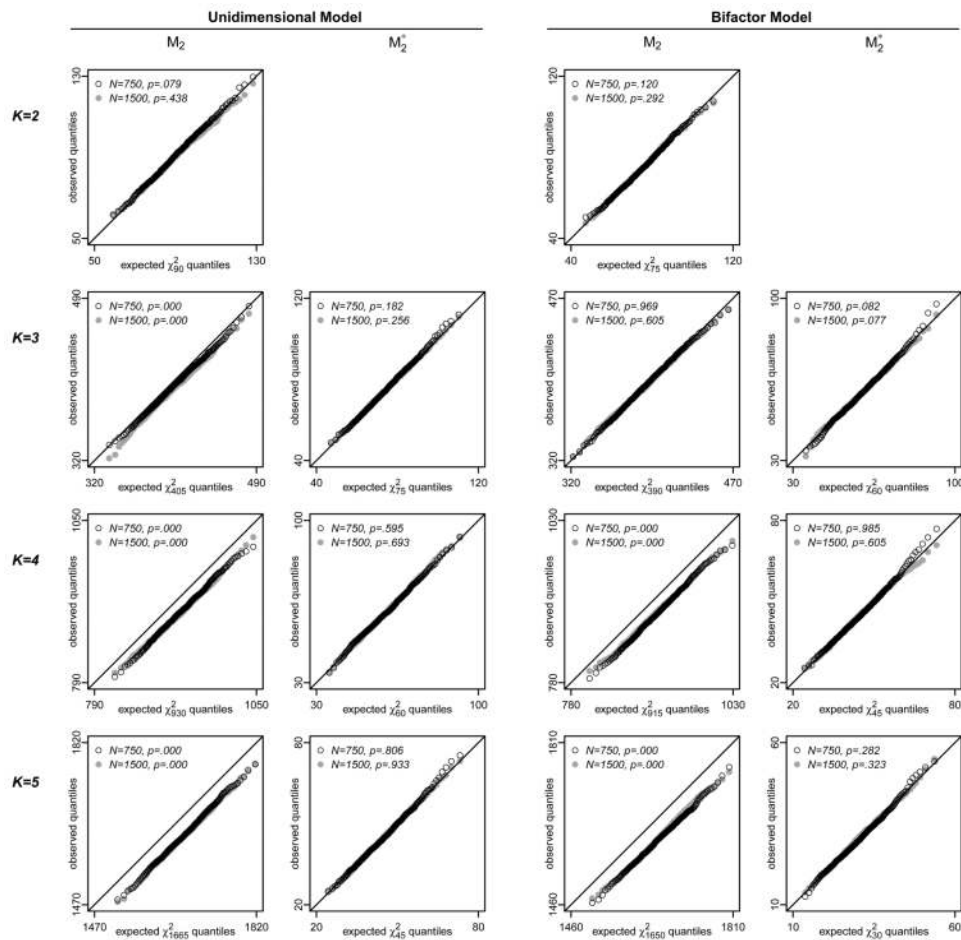


Figure 1. Bivariate posterior distributions and expected cell probabilities for items 4 and 5 of a quality-of-life scale.

**Figure 2.**

Q-Q plots of observed M_2 and M_2^* values and their reference chi-square distributions (degrees-of-freedom shown in the subscripts of the x-axis labels). Closed grey circles indicate conditions with sample size $N = 1500$; open black circles indicate $N = 750$. Reported p -values are for a two-tailed Komogorov-Smirnov test.

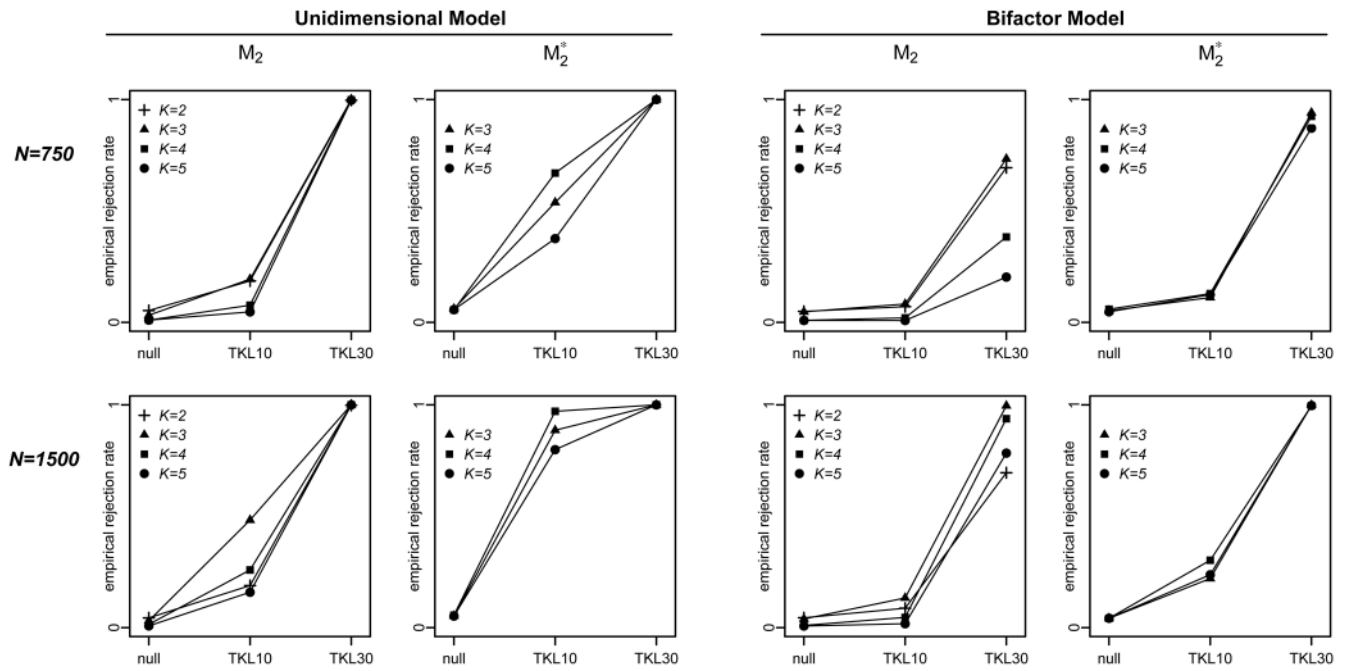


Figure 3.

Empirical rejection rates at $\alpha = .05$ for M_2 and M_2^* across all simulated conditions. K is the number of categories.

Table 1

Timing Comparison Results

#Group-Specific Dimensions	Number of Response Categories			
	2	3	4	5
<i>without dimension reduction</i>				
2	0.4	3.1	26.5	147.0
3	3.5	16.7	70.8	262.4
4	32.3	145.3	489.4	1369.5
5	335.2	1459.8	5305.7	13437.5
<i>with dimension reduction</i>				
2	0.2	2.0	22.4	141.4
3	0.2	2.2	22.8	145.3
4	0.2	2.2	23.7	146.7
5	0.2	2.2	22.4	142.9

Notes. The entries are time (in seconds) required to compute M_2 for a graded item bifactor model with 15 items.

Table 2

Generating Item Parameters with No Additional Minor Domain Factors

Item	General	Group-Specific			$K = 2$			$K = 3$			$K = 4$			$K = 5$		
		β_0	β_1	β_2	β_3	$\alpha_{j,1}$	$\alpha_{j,2}$	$\alpha_{j,1}$	$\alpha_{j,2}$	$\alpha_{j,1}$	$\alpha_{j,2}$	$\alpha_{j,3}$	$\alpha_{j,1}$	$\alpha_{j,2}$	$\alpha_{j,3}$	$\alpha_{j,4}$
1	0.65	0.99				1.44	0.83	0.83	-0.98	1.46	0.49	-1.68	1.48	0.71	-0.84	-1.59
2	0.87	2.16				0.21	1.82	1.82	-0.25	1.10	0.71	-2.23	1.80	0.16	-0.38	-2.30
3	0.84	0.70				2.45	2.51	2.51	-0.22	1.18	0.21	-1.95	1.66	0.15	-0.32	-1.43
4	1.13	1.76				-1.53	2.45	2.45	-1.09	1.95	1.05	-2.13	2.13	0.57	-1.50	-3.02
5	0.57	0.97				1.58	0.01	0.01	-1.04	1.53	0.00	-0.82	1.65	0.78	-0.21	-1.26
6	1.41		1.47			-2.13	3.21	3.21	-2.45	1.84	-0.82	-1.95	2.72	0.98	-1.85	-2.87
7	0.78		2.00			2.06	2.07	2.07	-2.10	1.02	-0.57	-1.34	1.49	1.16	-0.51	-1.81
8	0.73		1.18			-1.03	1.59	1.59	-1.09	1.32	0.61	-1.34	2.09	0.69	-0.22	-1.59
9	1.03		0.91			-1.00	2.95	2.95	-1.71	1.88	-0.62	-1.69	2.07	1.16	-1.33	-2.08
10	0.47		0.61			-0.62	0.48	0.48	-0.10	0.56	0.04	-0.98	1.15	0.29	-0.16	-0.91
11	1.00				0.82	-2.26	2.91	2.91	-1.30	1.82	0.99	-2.07	2.26	1.27	-0.36	-1.82
12	0.38				1.41	0.64	1.03	1.03	-0.89	0.89	-0.02	-0.57	0.87	0.30	-0.04	-0.71
13	1.07				1.59	2.41	1.01	1.01	-0.48	1.66	0.86	-2.24	2.86	1.40	-1.50	-3.20
14	1.36				1.86	-1.58	0.56	0.56	-1.90	2.63	-0.23	-3.36	2.10	0.15	-0.75	-2.75
15	0.77				0.29	-1.53	0.09	0.09	-2.17	1.86	-0.19	-0.82	2.24	0.75	-0.14	-1.90

Table 3

Simulation Results: Unidimensional Model, Null Condition

K	N	reps	M ₂		RMSEA		rejection rates				
			df	Mean	Var	Mean	(90% CI)	.150	.100	.050	.010
test statistic based on full one- and two-way marginal tables (M ₂)											
2	750	1000	90	90.9	183.4	.006	(.000,.019)	.164	.104	.053	.016
	1500	1000	90	89.9	162.0	.004	(.000,.013)	.133	.075	.044	.008
3	750	1000	405	399.8	781.8	.003	(.000,.011)	.099	.062	.032	.009
	1500	1000	405	396.2	794.8	.002	(.000,.008)	.082	.063	.031	.005
4	750	1000	930	904.8	1857.9	.002	(.000,.008)	.055	.033	.010	.002
	1500	1000	930	903.5	1716.1	.001	(.000,.006)	.054	.034	.015	.001
5	750	1000	1665	1622.1	3380.4	.001	(.000,.007)	.042	.020	.010	.001
	1500	1000	1665	1619.6	3320.3	.001	(.000,.004)	.033	.022	.008	.000
test statistic based on reduced one- and two-way marginal tables (M ₂ [*])											
3	750	1000	75	75.7	158.5	.006	(.000,.020)	.160	.106	.061	.020
	1500	1000	75	75.5	149.8	.004	(.000,.014)	.155	.103	.053	.012
4	750	1000	60	60.2	120.2	.006	(.000,.021)	.159	.106	.056	.010
	1500	1000	60	60.4	129.4	.005	(.000,.015)	.165	.117	.056	.012
5	750	1000	45	45.0	95.5	.007	(.000,.023)	.150	.105	.057	.019
	1500	1000	45	45.1	91.7	.005	(.000,.016)	.158	.106	.052	.012

Notes. K is the number of categories. N is the sample size. The number of valid (converged and stable) solutions are reported under the column heading “reps.” The reference distribution of M_2 and M_2^* statistics is central chi-square with df degrees of freedom. Note that M_2 is equivalent to M_2^* when K is 2.

Table 4

Simulation Results: Bifactor model, Null Condition

K	N	reps	df	M ₂		RMSEA		rejection rates				
				Mean	Var	Mean	(90% CI)	.150	.100	.050	.010	
test statistic based on full one- and two-way marginal tables (M ₂)												
2	750	935	75	74.4	147.1	.006	(.000,.019)	.144	.106	.049	.006	
	1500	922	75	74.7	147.8	.004	(.000,.013)	.153	.104	.044	.007	
3	750	995	390	390.1	754.1	.004	(.000,.012)	.153	.098	.047	.009	
	1500	998	390	389.8	706.4	.003	(.000,.009)	.140	.093	.039	.007	
4	750	998	915	888.7	1838.3	.002	(.000,.008)	.057	.030	.009	.002	
	1500	1000	915	891.2	1703.6	.001	(.000,.006)	.059	.035	.010	.003	
5	750	989	1650	1601.5	3149.9	.001	(.000,.006)	.029	.015	.008	.001	
	1500	989	1650	1604.8	3318.4	.001	(.000,.004)	.033	.014	.007	.001	
test statistic based on reduced one- and two-way marginal tables (M ₂ [*])												
3	750	995	60	59.4	133.5	.006	(.000,.021)	.140	.096	.052	.014	
	1500	998	60	59.4	116.5	.004	(.000,.014)	.128	.091	.042	.009	
4	750	998	45	45.2	96.6	.007	(.000,.023)	.154	.106	.059	.017	
	1500	1000	45	45.1	86.3	.005	(.000,.015)	.157	.097	.042	.007	
5	750	989	30	29.8	62.3	.006	(.000,.024)	.140	.092	.055	.017	
	1500	989	30	30.1	60.2	.004	(.000,.017)	.151	.095	.047	.013	

Notes. K is the number of categories. N is the sample size. The number of valid (converged and stable) solutions are reported under the column heading “reps.” The reference distribution of M₂ and M₂^{*} statistics is central chi-square with df/degrees of freedom. Note that M₂ is equivalent to M₂^{*} when K is 2.

Table 5

Simulation Results: Unidimensional model, Power at .05 Alpha Level

<i>K</i>	<i>N</i>	<i>df</i>	null model			TKL10			TKL30		
			<i>M</i>	RMSEA (90% CI)	rej	<i>M</i>	RMSEA (90% CI)	rej	<i>M</i>	RMSEA (90% CI)	rej
<i>test statistic based on full one- and two-way marginal tables (<i>M</i>₂)</i>											
2	750	90	90.9	.006 (.000,.019)	.053	100.1	.011 (.000,.023)	.186	180.7	.036 (.028,.044)	.997
	1500	90	89.9	.004 (.000,.013)	.044	100.2	.008 (.000,.016)	.189	180.1	.026 (.019,.032)	.998
3	750	405	399.8	.003 (.000,.011)	.032	425.5	.007 (.000,.015)	.194	620.1	.026 (.022,.031)	1.000
	1500	405	396.2	.002 (.000,.008)	.031	451.4	.008 (.000,.013)	.483	841.5	.027 (.024,.030)	1.000
4	750	930	904.8	.002 (.000,.008)	.010	937.2	.004 (.000,.011)	.077	1180.6	.019 (.015,.022)	.999
	1500	930	903.5	.001 (.000,.006)	.015	973.1	.005 (.000,.009)	.259	1462.2	.019 (.017,.022)	1.000
5	750	1665	1622.1	.001 (.000,.007)	.010	1663.9	.003 (.000,.009)	.047	1933.1	.015 (.011,.018)	.997
	1500	1665	1619.6	.001 (.000,.004)	.008	1699.9	.003 (.000,.007)	.159	2246.4	.015 (.013,.017)	1.000
<i>test statistic based on reduced one- and two-way marginal tables (<i>M</i>₂[*])</i>											
3	750	75	75.7	.006 (.000,.020)	.061	98.3	.019 (.000,.030)	.538	246.0	.055 (.046,.064)	1.000
	1500	75	75.5	.004 (.000,.014)	.053	119.4	.019 (.011,.026)	.886	413.0	.055 (.049,.061)	1.000
4	750	60	60.2	.006 (.000,.021)	.056	87.0	.023 (.007,.035)	.670	206.2	.057 (.045,.070)	1.000
	1500	60	60.4	.005 (.000,.015)	.056	113.9	.024 (.016,.031)	.971	337.8	.055 (.047,.064)	1.000
5	750	45	45.0	.007 (.000,.023)	.057	59.2	.018 (.000,.033)	.376	138.7	.052 (.041,.064)	1.000
	1500	45	45.1	.005 (.000,.016)	.052	73.8	.020 (.010,.028)	.799	232.7	.053 (.045,.060)	1.000

Notes. *K* is the number of categories. *N* is the sample size. The reference distribution of *M*₂ and *M*₂^{*} statistics is central chi-square with *df* degrees of freedom. Note that *M*₂ is equivalent to *M*₂^{*} when *K* is 2. Means are reported under *M* and the empirical rejection rates at alpha level of .05 are shown under column heading "rej."

Table 6

Simulation Results: Bifactor model, Power at .05 Alpha Level

K	N	df	null model			TKL10			TKL30		
			M	RMSEA (90% CI)	rej	M	RMSEA (90% CI)	rej	M	RMSEA (90% CI)	rej
test statistic based on full one- and two-way marginal tables (M_2)											
2	750	75	74.4	.006 (.000,.019)	.049	77.6	.007 (.000,.020)	.070	104.6	.022 (.010,.032)	.694
	1500	75	74.7	.004 (.000,.013)	.044	77.7	.005 (.000,.015)	.087	105.1	.016 (.007,.023)	.695
3	750	390	390.1	.004 (.000,.012)	.047	397.5	.005 (.000,.014)	.082	458.1	.015 (.007,.021)	.734
	1500	390	389.8	.003 (.000,.009)	.039	406.5	.005 (.000,.011)	.133	529.0	.015 (.012,.019)	.996
4	750	915	888.7	.002 (.000,.008)	.009	901.1	.002 (.000,.009)	.020	974.0	.009 (.000,.014)	.383
	1500	915	891.2	.001 (.000,.006)	.010	908.6	.002 (.000,.007)	.046	1061.2	.010 (.007,.013)	.938
5	750	1650	1601.5	.001 (.000,.006)	.008	1614.3	.001 (.000,.007)	.008	1698.1	.006 (.000,.011)	.203
	1500	1650	1604.8	.001 (.000,.004)	.007	1625.1	.001 (.000,.006)	.018	1796.8	.007 (.004,.010)	.783
test statistic based on reduced one- and two-way marginal tables (M_2^*)											
3	750	60	59.4	.006 (.000,.021)	.052	64.3	.009 (.000,.024)	.111	106.8	.032 (.020,.042)	.941
	1500	60	59.4	.004 (.000,.014)	.042	70.3	.009 (.000,.019)	.220	153.8	.032 (.025,.038)	.999
4	750	45	45.2	.007 (.000,.023)	.059	50.2	.011 (.000,.026)	.129	84.3	.033 (.020,.045)	.925
	1500	45	45.1	.005 (.000,.015)	.042	56.2	.011 (.000,.022)	.303	121.6	.033 (.025,.041)	.996
5	750	30	29.8	.006 (.000,.024)	.055	33.3	.011 (.000,.030)	.125	60.4	.035 (.018,.050)	.872
	1500	30	30.1	.004 (.000,.017)	.047	37.1	.011 (.000,.024)	.237	93.0	.037 (.027,.047)	.997

Notes. *K* is the number of categories. *N* is the sample size. The reference distribution of *M₂* and *M₂^{*}* statistics is central chi-square with *df* degrees of freedom. Note that *M₂* is equivalent to *M₂^{*}* when *K* is 2. Means are reported under *M* and the empirical rejection rates at alpha level of .05 are shown under column heading "rej."

Table 7

Application to Empirical Data

	Unidimensional Model	Bifactor Model
number of free parameters, ν	90	103
$-2 \times \log\text{-likelihood}$	47374.5	44361.0
Akaike Information Criterion (AIC)	47554.5	44567.0
Bayesian Information Criterion (BIC)	47999.9	45076.7
<i>test statistic based on full one- and two-way marginal tables (M_2)</i>		
M_2	8489.4	6541.6
degrees-of-freedom	2430	2417
p	0.0001	0.0001
RMSEA (90% Confidence Interval)	0.049 (0.048,0.050)	0.032 (0.031,0.034)
<i>test statistic based on reduced one- and two-way marginal tables (M_2^*)</i>		
M_2	2131.8	199.3
degrees-of-freedom	81	68
p	0.0001	0.0001
RMSEA (90% Confidence Interval)	0.156 (0.150,0.162)	0.043 (0.036,0.050)