PSYCHOMETRIKA—VOL. 71, NO. 4, 713–732
DECEMBER 2006
DOI: 10.1007/s11336-005-1295-9

LIMITED INFORMATION GOODNESS-OF-FIT TESTING IN MULTIDIMENSIONAL CONTINGENCY TABLES

ALBERT MAYDEU-OLIVARES

UNIVERSITY OF BARCELONA AND INSTITUTO DE EMPRESA BUSINESS SCHOOL

HARRY JOE

UNIVERSITY OF BRITISH COLUMBIA

We introduce a family of goodness-of-fit statistics for testing composite null hypotheses in multidimensional contingency tables. These statistics are quadratic forms in marginal residuals up to order r. They are asymptotically chi-square under the null hypothesis when parameters are estimated using any asymptotically normal consistent estimator. For a widely used item response model, when r is small and multidimensional tables are sparse, the proposed statistics have accurate empirical Type I errors, unlike Pearson's X^2. For this model in nonsparse situations, the proposed statistics are also more powerful than X^2. In addition, the proposed statistics are asymptotically chi-square when applied to subtables, and can be used for a piecewise goodness-of-fit assessment to determine the source of misfit in poorly fitting models.

Key words: multivariate discrete data, categorical data analysis, multivariate multinomial distribution, composite likelihood, item response theory, Lisrel.

1. Introduction

Consider the problem of modeling N independent and identically distributed observations on n discrete random variables consisting, respectively, of $K_1, \ldots, K_n$ categories. This type of data arises, for ingfstance, in surveys, educational tests, or social science questionnaires when the number of choices is not constant over items. The observed data can be gathered in an n-dimensional contingency table with $C = \prod_i^n K_i$ cells.

Now, consider a parametric model, $\boldsymbol{\pi}(\boldsymbol{\theta})$, where $\boldsymbol{\pi}$ is the C-dimensional vector of cell probabilities, which depends on a q-dimensional parameter vector $\boldsymbol{\theta}$ which is typically estimated from the data. For assessing the fit of the model, consider a composite null hypothesis H_0 : $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$ versus $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$. Researchers confronted with testing such a composite hypothesis face two problems. First, how to assess the overall goodness of fit of the hypothesized model and, second, how to determine the source of the misfit in poorly fitting models.

The two most commonly used goodness-of-fit statistics for testing the overall goodness of fit of a parametric model in multivariate categorical data analysis are Pearson's $X^2 = 2N \sum_{c=1}^{C} (p_c - \pi_c)^2/\pi_c$, and the likelihood ratio statistic $G^2 = 2N \sum_{c=1}^{C} p_c \ln(p_c/\pi_c)$. When the model holds, the two statistics are asymptotically equivalent. Under H_0, they are asymptotically distributed as chi-square with $C - q - 1$ degrees of freedom. However, it is well known that in sparse tables

This research has been supported by the Department of Universities, Research, and Information Society (DURSI) of the Catalan Government, by grant BSO2003-08507 of the Spanish Ministry of Science and Technology, and an NSERC Canada grant. We are grateful to the referees for comments leading to improvements.

Requests for reprints should be sent to Albert Maydeu-Olivares. Faculty of Psychology. University of Barcelona. P. Valle de Hebrón, 171, 08035 Barcelona (Spain). E-mail: amaydeu@ub.edu

the empirical Type I error rates of the X^2 and G^2 test statistics do not match their expected rates under their asymptotic distribution. Of the two statistics, X^2 is less adversely affected by the sparseness of the contingency table than G^2 (Koehler & Larntz, 1980).

One reason for the poor empirical performance of X^2 is that the empirical variance of X^2 and its variance under its reference asymptotic distribution differ by a term that depends on the inverse of the cell probabilities (Cochran, 1952). When the cell probabilities become small the discrepancy between the empirical and asymptotic variances of X^2 can be large and the Type I error for X^2 will be larger than the α level based on its asymptotic critical value. Thus, the accuracy of the Type I errors will depend on the model being fitted to the table (as it determines the cell probabilities), but also on the size of the contingency table. This is because when the size of the contingency table is large, the cell probabilities must be small (Bartholomew & Tzamourani, 1999). However, for C and $\boldsymbol{\pi}(\boldsymbol{\theta})$ fixed the accuracy of the asymptotic p-values for X^2 also depends on sample size, N. As N becomes smaller some of the cell proportions increasingly become more poorly estimated (their estimates will be zero) and the empirical Type I errors of X^2 will become inaccurate. The degree of sparseness N/C summarizes the relationship between sample size and model size. Thus, the accuracy of the asymptotic p-values for X^2 depends on the model and the degree of sparseness of the contingency table.

Three alternative strategies have been proposed for obtaining Type I errors when the accuracy of the asymptotic p-values of X^2 is suspect: (a) pooling cells; (b) resampling methods; and (c) limited information methods. Our new statistical procedures are in category (c); below we point out the advantages of (c) over (a) and (b).

Regarding (a), pooling cells before the model is fitted is a useful approach as it reduces the size of the contingency table, and thus the degree of sparseness. However, there is a limit in the amount of pooling that can be performed without distorting the purpose of the analysis. Also, pooling cells ad hoc after the model has been fitted may result in a test statistic with an unknown asymptotic null sampling distribution. Regarding (b), generating the empirical sampling distribution of the goodness-of-fit statistic using a resampling method such as the parametric bootstrap method (e.g., Collins, Fidler, Wugalter, & Long, 1993; Bartholomew & Tzamourani, 1999) may result in trustworthy p-values (but see Tollenaar & Mooijaart, 2003). However, resampling methods may be very time-consuming if the researcher is interested in comparing the fit of several models. On the other hand, limited information methods use only the information contained in the low-order marginals of the contingency table to assess the model, and amounts to pooling cells a priori. The cells are pooled in a systematic way, so that the resulting statistics have a known asymptotic null distribution. These procedures are computationally much more efficient than resampling methods.

There have been several proposals in Psychometrics to use low-order marginals in goodness-of-fit assessment of *binary* contingency tables, most notably Christoffersson (1975), Reiser (1996), Bartholomew and Leung (2002); see also Cai, Maydeu-Olivares, Coffman, and Thissen (2006), Maydeu-Olivares (2001a, 2001b), and Maydeu-Olivares and Joe (2005). Limited information statistics appear as a viable framework to assess the overall goodness of fit of models for multidimensional contingency tables as they have more accurate empirical Type I errors and may be asymptotically more powerful than full information statistics such as X^2 (Maydeu-Olivares & Joe, 2005).

A second challenge a researcher must confront when modeling multivariate categorical data is to identify the source of the misfit when the overall test suggests significant misfit. The inspection of cell residuals is often not very useful to this aim. It is difficult to find trends in inspecting these residuals, and even for moderate n the number of residuals to be inspected is too large. Perhaps, most importantly, Bartholomew and Tzamourani (1999) point out that because the cell frequencies are integers and the expected frequencies in large tables must be very small, the resulting residuals will be either very small or very large. To overcome this challenge, numerous

authors have advocated examining residuals from the two- and three-way margins to assess the goodness of fit in binary contingency tables. Some key references in this literature are Reiser (1996), Reiser and Lin (1999), Reiser and VandenBerg (1994), Bartholomew and Tzamourani (1999), Bartholomew and Leung (2002), and Maydeu-Olivares and Joe (2005). However, when the observed variables are not binary, the number of marginal residuals grows very rapidly as the number of categories and variables increases, and it may be difficult to draw useful information by inspecting individual marginal residuals. To overcome this problem, it has been suggested (Drasgow, Levine, Tsien, Williams, & Mead, 1995) to compute X^2 for single variables, pairs, and triplets. However, X^2 applied to subtables is not asymptotically chi-square under the null hypothesis even for the maximum likelihood estimator (MLE).

In this paper, the main ideas and results of Maydeu-Olivares and Joe (2005) for the binary case ($K_i = 2$ for all i) are extended in two directions. First, we provide goodness-of-fit test statistics for multidimensional contingency tables of arbitrary dimensions. The statistics are quadratic forms in the residuals of marginal tables up to order r, for small r. These test statistics are asymptotically chi-square for any $\sqrt{N}$ -consistent and asymptotically normal estimator. The extension is straightforward but the computational implementation is more cumbersome. Second, we provide statistics for assessing the goodness of fit in r-dimensional subtables. These statistics are also asymptotically chi-square under the same conditions than the statistics to assess the overall goodness of fit and they can be useful to identify the source of the misfit in poorly fitting models.

The remainder of the paper is organized as follows. In Section 2 we provide a convenient representation of multivariate categorical data which are a random sample from a multivariate multinomial (MVM) distribution, and we also provide the asymptotic distribution of multivariate marginal residuals for different estimators. In Section 3 we consider extensions of the family of limited information statistics M_r proposed by Maydeu-Olivares and Joe (2005). These statistics can be used with nominal categorical variables as they are invariant to arbitrary relabeling of the categories. Section 3 also includes a small simulation study to illustrate the small sample distributions of M_r (for small r) and X^2. In Section 4 we consider the use of marginal residuals and M_r statistics on r-dimensional subtables to identify the source of the misfit. Section 5 contains two examples to illustrate our results. Finally, Section 6 has a discussion of the different limited information approaches that have been proposed, as well as directions for further research.

For completeness, we also discuss in an Appendix goodness-of-fit testing of simple null hypotheses under MVM assumptions as a straightforward extension of the results of Maydeu-Olivares and Joe (2005) for multivariate Bernoulli assumptions. Computational details for estimation, evaluation of M_r, and simulations are also given in the Appendix.

2. Multivariate Multinomial Distributions and Asymptotic Distribution of Marginal Residuals

In this section we define the notation used in the remainder of this paper and we give two representations of the MVM distribution. One of them uses the cell probabilities, while the other uses a set of multivariate marginal probabilities. There is a one-to-one linear map between the two representations. We also provide the asymptotic distribution of cell residuals and of marginal residuals for MVM models where the parameters have been estimated using a $\sqrt{N}$ -consistent and asymptotically normal estimator (including limited information estimators).

2.1. Representation of the MVM Distribution

By an MVM distribution, we mean a multivariate distribution with univariate margins that are multinomial. If the ith ($1 \leq i \leq n$) variable consists of $K_i \geq 2$ categories labeled as $0, 1, \ldots, K_i - 1$, with respective probabilities $p_{i0}, \ldots, p_{i,K_i-1}$, then one observation of the ith

variable Y_i has a Multinomial(1; $p_{i0}, \ldots, p_{i,K_i-1}$) distribution. Using indicator functions, we give a representation of the MVM distribution. In the case where each $K_i = 2$, the representation is the same as that of Teugels (1990).

With the notation $Y_i = j$ meaning that Y_i has category j, we define the following indicator variables for $Y_1, \ldots, Y_n$:

$$I_{ij} = I(Y_i = j), \qquad j = 1, \ldots, K_i - 1, \quad i = 1, \ldots, n. \tag{2.1}$$

The univariate moments are $\mathrm{E}(I_{ij})$, $j = 1, \ldots, K_i - 1$, $i = 1, \ldots, n$; the bivariate moments are $\mathrm{E}[I_{i_1 j_1} I_{i_2 j_2}] = \Pr(Y_{i_1} = j_1, Y_{i_2} = j_2)$, $j_1 = 1, \ldots, K_{i_1} - 1$, $j_2 = 1, \ldots, K_{i_2} - 1$, $1 \le i_1 < i_2 \le n$. The trivariate up to n-dimensional moments can be defined in a similar way. Note that these moments consist of all joint and marginal probabilities of $Y_1, \ldots, Y_n$ that do not involve category 0 for any variables.

The distribution is characterized by all the moments involving the I_{ij} up to the nth moments, since all joint probabilities, including those involving the 0 categories, can be deduced from these moments. This follows by letting $I_{i0} = 1 - I_{i1} - \cdots - I_{iK_i}$; then

$$\Pr(Y_1 = j_1, \ldots, Y_n = j_n) = \mathrm{E}[I_{1j_1} \cdots I_{nj_p}],$$

and after expanding out any term with $j_i = 0$, this is a linear combination of the moments not involving any category of 0.

Consider the set $\mathcal{A}_r$ of expectations or moments that come from products of 1 to r indicators in (2.1). Then all probabilities up to the rth-dimensional margins can be obtained from the set $\mathcal{A}_r$. There are no redundant moments in $\mathcal{A}_r$ in that no moment can be obtained as a linear combination of other moments in $\mathcal{A}_r$. The cardinality of $\mathcal{A}_r$ is equal to

$$s(r) \stackrel{\text{def}}{=} \sum_{j=1}^{r} \sum_{1 \le i_1 < \cdots < i_j \le n} \prod_{\ell=1}^{j} (K_{i_\ell} - 1), \tag{2.2}$$

where the middle sum is over the combinations of size j from n indices; this is smaller than the number $\sum_{1 \le i_1 < \cdots < i_r \le n} \prod_{\ell=1}^{r} K_{i_\ell}$ of rth-order marginal probabilities. For example, if $K_i = K$ for $i = 1, \ldots, n$, then the number of rth-order marginal probabilities is $\binom{n}{r} K^r$ which is larger than $s(r) = \sum_{j=1}^{r} \binom{n}{j} (K-1)^j$, the cardinality of $\mathcal{A}_r$. Note that $s(n) = C - 1$, with $C = \prod_{i=1}^{n} K_i$.

In the next section we will be constructing quadratic form statistics based on residuals corresponding to the moments in $\mathcal{A}_r$. Because of the relationships mentioned above, the quadratic form statistics can also be expressed in terms of the residuals associated with all rth-order marginal probabilities. It is an advantage computationally to work with the set $\mathcal{A}_r$ so that we can deal with smaller matrices in the quadratic form statistics. Note that even the cardinality of $\mathcal{A}_r$ increases rapidly as K_i and n increase. Also, for any goodness-of-fit statistic defined based on the moments up to order r, it is necessary to check/prove that the statistic is invariant to the labeling of the categories, since it is generally arbitrary which category is labeled as category 0.

Further insight into the relationship between the multivariate moment and the cell representation is obtained by using a notation analogous to that employed in Maydeu-Olivares and Joe (2005). In what follows we assume for notational ease that $K_i = K$ for all i. Consider an n-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ of K-category random variables, with $\pi_i(j) = \Pr(Y_i = j)$, $i = 1, \ldots, n$, and joint distribution:

$$\pi_{\mathbf{y}} = \Pr(Y_i = y_i, i = 1, \ldots, n), \qquad \mathbf{y} = (y_1, \ldots, y_n), \quad y_i \in \{0, \ldots, K-1\}.$$

When we consider a parametric model with parameter vector $\boldsymbol{\theta}$, we write $\pi_{\mathbf{y}}(\boldsymbol{\theta})$ for an individual probability and $\boldsymbol{\pi}(\boldsymbol{\theta})$ for the vector of K^n joint probabilities. Also, we write $\boldsymbol{\pi}_1$ for the $n(K-1)$ vector of univariate marginal probabilities. Similarly, we write $\boldsymbol{\pi}_2$ for the $\binom{n}{2}(K-1)^2$ vector

of bivariate marginal probabilities, and so forth up to $\dot{\boldsymbol{\pi}}_n$, the $\binom{n}{n}(K-1)^n$ vector of nth way probabilities. Finally, let $\dot{\boldsymbol{\pi}}' = (\dot{\boldsymbol{\pi}}_1', \dot{\boldsymbol{\pi}}_2', \ldots, \dot{\boldsymbol{\pi}}_n')'$. Then, we can write $\dot{\boldsymbol{\pi}} = \mathbf{T}\boldsymbol{\pi}$, where $\mathbf{T}$ is a $(K^n - 1) \times K^n$ matrix of 1s and 0s, of full row rank (if K_i is not constant, then $\mathbf{T}$ is $(C-1) \times C$).

$\mathbf{T}$ can be partitioned according to the partitioning of $\dot{\boldsymbol{\pi}}$,

$$\begin{pmatrix} \dot{\boldsymbol{\pi}}_1 \\ \dot{\boldsymbol{\pi}}_2 \\ \vdots \\ \dot{\boldsymbol{\pi}}_n \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{n1} \\ \mathbf{T}_{n2} \\ \vdots \\ \mathbf{T}_{nn} \end{pmatrix} \boldsymbol{\pi}.$$

The vector of multivariate moments up to order r ($r \leq n$), denoted by $\boldsymbol{\pi}_r = (\dot{\boldsymbol{\pi}}_1', \ldots, \dot{\boldsymbol{\pi}}_r')'$, can be written as

$$\boldsymbol{\pi}_r = \mathbf{T}_r \boldsymbol{\pi},$$

where $\mathbf{T}_r = (\mathbf{T}_{n1}', \ldots, \mathbf{T}_{nr}')'$. Note that by definition $\boldsymbol{\pi}_n = \dot{\boldsymbol{\pi}}$. That is, $\mathbf{T}_r$ is the mapping of the C-dimensional vector of cell probabilities to the moments in $\mathcal{A}_r$.

2.2. *Asymptotic Distribution of Marginal Residuals*

In this section we state the main results on the asymptotic distribution of marginal residuals which are needed in deriving the statistics in the next section. The results are essentially those given in Maydeu-Olivares and Joe (2005) for the binary case.

First, consider an MVM parametric model $\boldsymbol{\pi}(\boldsymbol{\theta})$ for a fixed a priori vector $\boldsymbol{\theta}$ of dimension q. For a random sample of size N from this model, let $\mathbf{p}$ and $\dot{\mathbf{p}}$ denote the C-dimensional vector of cell proportions, and the $(C-1)$-dimensional vector of sample joint moments, respectively. Then,

$$\sqrt{N}(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}(\boldsymbol{\theta})) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Xi}(\boldsymbol{\theta})), \qquad \boldsymbol{\Xi}(\boldsymbol{\theta}) = \mathbf{T}\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{T}',$$

where $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \mathbf{D}(\boldsymbol{\theta}) - \boldsymbol{\pi}(\boldsymbol{\theta})\boldsymbol{\pi}'(\boldsymbol{\theta})$, and $\mathbf{D}(\boldsymbol{\theta}) = \operatorname{diag}(\boldsymbol{\pi}(\boldsymbol{\theta}))$. Hence,

$$\sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta})) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{\Xi}_r(\boldsymbol{\theta})\big), \qquad \boldsymbol{\Xi}_r(\boldsymbol{\theta}) = \mathbf{T}_r \boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{T}_r', \tag{2.3}$$

where $\mathbf{p}_r$ be the vector of sample moments up to order r; it has dimension $s(r)$ as given in (2.2).

In practice, in most applications for multivariate categorical data, one is interested in comparing one or more MVM models where $\boldsymbol{\theta}$ is estimated from the data. Let $\hat{\boldsymbol{\theta}}$ be a $\sqrt{N}$ -consistent and asymptotically normal estimator. We assume that the usual regularity conditions on the model are satisfied so as to fulfill the consistency and asymptotic normality of the estimates. In particular, we assume that $\hat{\boldsymbol{\theta}}$ satisfies

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{H}\sqrt{N}\,(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1) \tag{2.4}$$

for some $q \times C$ matrix $\mathbf{H}$. This includes minimum variance (or best asymptotically normal (BAN)) estimators such as the MLE or the minimum chi-square estimator. For BAN estimators $\mathbf{H} = \boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1}$, where $\boldsymbol{\mathcal{I}} = \boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta}$ is the Fisher information matrix, and $\boldsymbol{\Delta} = \partial\boldsymbol{\pi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$. Also, the limited information estimators considered by Christoffersson (1975), Jöreskog (1994); see also Maydeu-Olivares (2006), Jöreskog and Moustaki (2001), Lee, Poon, and Bentler (1995), Maydeu-Olivares (2001b), and Muthén (1978, 1984, 1993) are special cases of this framework.

Using (2.4), the asymptotic distribution of the vector of cell residuals $\hat{\mathbf{e}} = \mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ is $\sqrt{N}\,\hat{\mathbf{e}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$ with asymptotic covariance matrix

$$\boldsymbol{\Sigma} = (\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})\boldsymbol{\Gamma}(\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})'. \tag{2.5}$$

For the marginal residuals, $\hat{\mathbf{e}}_r = \mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}) = \mathbf{T}_r\hat{\mathbf{e}}$, $\sqrt{N}\,\hat{\mathbf{e}}_r \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_r)$, where

$$\begin{aligned}\boldsymbol{\Sigma}_r &= \mathbf{T}_r\boldsymbol{\Sigma}\mathbf{T}_r' = (\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{H})\boldsymbol{\Gamma}(\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{H})' \\ &= \boldsymbol{\Xi}_r - \boldsymbol{\Delta}_r\mathbf{H}\boldsymbol{\Gamma}\mathbf{T}_r' - \mathbf{T}_r\boldsymbol{\Gamma}\mathbf{H}'\boldsymbol{\Delta}_r' + \boldsymbol{\Delta}_r[\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}']\boldsymbol{\Delta}_r', \end{aligned} \tag{2.6}$$

where $\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}'$ is the asymptotic covariance matrix of $\sqrt{N}\,\hat{\boldsymbol{\theta}}$, and $\boldsymbol{\Delta}_r = \partial\boldsymbol{\pi}_r(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ is an $s(r) \times q$ matrix. In the special case of BAN estimators such as the MLE, equations (2.5) and (2.6) reduce to $\boldsymbol{\Sigma} = \boldsymbol{\Gamma} - \boldsymbol{\Delta}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}'$, and $\boldsymbol{\Sigma}_r = \boldsymbol{\Xi}_r - \boldsymbol{\Delta}_r\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}_r'$, respectively.

3. Overall Goodness-of-Fit Testing Using Marginal Residuals

In this section we consider testing a composite null hypothesis using quadratic forms in the marginal residuals. That is, we consider the hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$ versus $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$, when parameters are estimated using a method that yields $\sqrt{N}$-consistent and asymptotically normal estimates. Let r_0 be the smallest integer r such that the model is (locally) identified from the marginal residuals up to order r. Then, for $r \geq r_0$, the matrix $\boldsymbol{\Delta}_r$ is of full column rank q. Also, we assume that $s(r) > q$ so as to exclude the case $s(r) = q$. Finally, we assume that $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ does not imply linear dependencies among the marginal moments in $\mathcal{A}_r$, so that $\boldsymbol{\Sigma}_r$ has full rank $s(r)$.

3.1. The Family of Test Statistics M_r

In the special case $K_i = 2$, Maydeu-Olivares and Joe (2005) introduced the family of statistics M_r for testing composite null hypotheses for multivariate binary models. Their results readily extend to MVM models for contingency tables of arbitrary dimensions. The notation is basically the same but the dimension of the matrices is larger, and numerical computations are harder.

Consider an $s(r) \times (s(r) - q)$ orthogonal complement to $\boldsymbol{\Delta}_r$, say $\boldsymbol{\Delta}_r^{(c)}$, such that $[\boldsymbol{\Delta}_r^{(c)}]'\boldsymbol{\Delta}_r = \mathbf{0}$. Let

$$\mathbf{C}_r = \mathbf{C}_r(\boldsymbol{\theta}) = \boldsymbol{\Delta}_r^{(c)}\big([\boldsymbol{\Delta}_r^{(c)}]'\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r^{(c)}\big)^{-1}[\boldsymbol{\Delta}_r^{(c)}]' = \boldsymbol{\Xi}_r^{-1} - \boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r\big(\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r\big)^{-1}\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}. \tag{3.1}$$

Note that $\mathbf{C}_r$ is invariant to the choice of orthogonal complement (if $\boldsymbol{\Delta}_r^{(c)}$ is a full rank orthogonal complement, then so is $\boldsymbol{\Delta}_r^{(c)}\mathbf{A}$ for a nonsingular matrix $\mathbf{A}$), and the last equality in (3.1) follows from a result in Rao (1973, p. 77).

The limited information statistic M_r of order r is given by

$$M_r = M_r(\hat{\boldsymbol{\theta}}) = N\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})\big)'\widehat{\mathbf{C}}_r\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})\big). \tag{3.2}$$

In (3.2), $\widehat{\mathbf{C}}_r$ denotes $\mathbf{C}_r(\hat{\boldsymbol{\theta}})$ and other matrices are also evaluated at $\hat{\boldsymbol{\theta}}$. It is straightforward to verify that $\mathbf{C}_r = \mathbf{C}_r\boldsymbol{\Sigma}_r\mathbf{C}_r$, that is, $\boldsymbol{\Sigma}_r$ is a generalized inverse of $\mathbf{C}_r$. Now, $M_r \xrightarrow{d} \chi^2_{s(r)-q}$, where the degrees of freedom are obtained from a result in Rao (1973, p. 30) using the fact that $\boldsymbol{\Delta}_r^{(c)}$ is of full column rank $s(r) - q$ and hence $\mathbf{C}_r$ is also of rank $s(r) - q$. Note that (3.2) does not use the generalized inverse of $\boldsymbol{\Sigma}_r$ because this may be numerically unstable with a small singular value. Also computation of $\mathbf{C}_r$, which depends on $\boldsymbol{\Delta}_r$ and $\boldsymbol{\Xi}_r$, is much easier than that of $\boldsymbol{\Sigma}_r$, which depends also on $\mathbf{H}\boldsymbol{\Gamma}\mathbf{T}_r'$ and $\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}'$ (or $\boldsymbol{\mathcal{I}}$ in the case of the MLE).

$\{M_r\}$ is a family of test statistics based on residuals up to r-variate margins whose members are $\{M_1, \ldots, M_n\}$. M_1 is defined only if $s(1) > q$, that is, for models that do not have many parameters; for example, it is not defined for the item response model that we use later in this paper. M_1 is a quadratic form in univariate residuals, whereas M_2 is a quadratic form in univariate and bivariate residuals, and so forth, up to M_n which is a full information test statistic. M_n can

be written as

$$M_n = N\big(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}})\big)'\widehat{\mathbf{C}}_n\big(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}})\big) \tag{3.3}$$

with $\widehat{\mathbf{C}}_n = \mathbf{C}_n(\hat{\boldsymbol{\theta}})$. Since $C = s(n) + 1$ is the number of possible cells in the contingency table, the full information test statistic $M_n = M_n(\hat{\boldsymbol{\theta}})$ is asymptotically χ^2_{C-1-q} for this large class of consistent estimators. Maydeu-Olivares and Joe (2005) show that this statistic can be alternatively written as a quadratic form in the cell residuals as

$$M_n = N\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big)'\widehat{\mathbf{U}}\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big)$$

with $\widehat{\mathbf{U}} = \mathbf{U}(\hat{\boldsymbol{\theta}})$, where $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{D}^{-1} - \mathbf{D}^{-1}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1}$. Note that with $X^2(\hat{\boldsymbol{\theta}})$ representing the X^2 statistic based on $\hat{\boldsymbol{\theta}}$, the results in the Appendix of Maydeu-Olivares and Joe (2005) imply that $M_n(\hat{\boldsymbol{\theta}}) \leq X^2(\hat{\boldsymbol{\theta}})$. That is, for a consistent estimator that is not the MLE, the asymptotic null distribution of $X^2(\hat{\boldsymbol{\theta}})$ is stochastically larger than χ^2_{C-1-q}. Also, $M_n = X^2$ when $\hat{\boldsymbol{\theta}}$ is the MLE. But for other minimum variance asymptotically normal estimators, $M_n \leq X^2$ and M_n and X^2 are equivalent only asymptotically.

Also with a proof very similar to that in the Appendix of Maydeu-Olivares and Joe (2005), M_r is invariant to the labeling of the categories, assuming that with permuted categories $\boldsymbol{\theta} \to \boldsymbol{\theta}_\Lambda$, a permuted vector, and $\hat{\boldsymbol{\theta}}$ is an equivariant estimator.

Maydeu-Olivares and Joe (2005) pointed out that the asymptotic variance of M_r is influenced by the smallest marginal probability of dimension $\min\{2r, n\}$. Therefore, the asymptotic null distribution of M_r can be acceptable if the rth-order margins are not sparse, and larger sample sizes are needed as r increases for the null asymptotics to be valid. This was illustrated using a simulation study where a two-parameter logistic model (Lord & Novick, 1968) was estimated by the MLE. For the less sparse situations, the small sample behavior of $M_n = X^2$ was close to its asymptotic reference distribution. But as sparseness increased the empirical Type I errors of X^2 first—and with increased sparseness M_3 as well—departed from its expected rates. Only the empirical Type I errors of M_2 remained accurate throughout the different sparseness conditions considered in their study. In the next subsection we extend their simulation results by: (a) considering an item response theory (IRT) model for variables where $K_i > 2$; (b) considering much larger contingency tables; and (c) investigating the behavior of the test statistics for a limited information estimator.

3.2. Small Sample Performance of M_r

For an illustration of the small sample performance of M_r consider a unidimensional *item response model* (e.g., van der Linden & Hambleton, 1997)

$$\Pr\left[\bigcap_{i=1}^{n}\{Y_i = y_i\}\right] = \int_{-\infty}^{\infty}\prod_{i=1}^{n}\Pr(Y_i = y_i \mid \eta)f(\eta)\,d\eta, \qquad y_i \in \{0, \ldots, K-1\}, \tag{3.4}$$

where $f(\eta)$ denotes the density of a continuous unobserved variable (i.e, a latent trait). Note that under this family of models, the probabilities conditional on the latent trait are assumed to be independent. For ordered categorical variables, Samejima (1969) proposed letting $f(\eta)$ be a standard normal density function and

$$\Pr(Y_i = j \mid \eta) = \begin{cases} 1 - G(\alpha_{i,1} + \beta_i\eta) & \text{if } j = 0, \\ G(\alpha_{i,j} + \beta_i\eta) - G(\alpha_{i,j+1} + \beta_i\eta) & \text{if } 0 < j < K-1, \\ G(\alpha_{i,K-1} + \beta_i\eta) & \text{if } j = K-1, \end{cases} \tag{3.5}$$

where $G(z)$ equals either the standard logistic distribution function

$$\Psi(z) = [1 + \exp\{-z\}]^{-1} \tag{3.6}$$

or the standard normal distribution function

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}\, dx. \tag{3.7}$$

Thus, in this model, for each item there is one slope parameter β_i and $K-1$ intercept parameters $\alpha_{i,j}$; $\alpha_{i,j}$ is decreasing in j for each i. Samejima (1969) referred to the model specified by equations (3.4)–(3.7) as the (logistic or normal) graded model. Bartholomew and Knott (1999) refer to these models as the logit-normit and normit-normit models, respectively. The family of models (3.4) are random effects members of the larger *generalized linear mixed model* (GLMM) family (see Agresti (2002, Chap. 12)).

Note that for model (3.4)–(3.6), the number of parameters is $q = nK$ so that M_r in (3.2) is defined only for $r \geq 2$ since $s(1) = n(K-1) < q$. M_3 would be useful to compute only if $s(3)/N$ is large enough. For most IRT applications M_2 is the statistic of choice in the M_r family.

To illustrate the small sample behavior of M_2 for maximum likelihood (ML) estimation, we generated data according to Samejima's logistic model for many different parameter vectors. We summarize some representative results in Table 1 , which has three cases of (K, n) : (3, 5) with $C = 243$ cells; (5, 5) with $C = 3125$ cells; and (5, 10) with $C = 9765625 \approx 10^7$ cells. The sample sizes are N = 300, 1000, and 3000. The procedure used to generate the data is explained in the Appendix subsection on computing notes. For $K = 3$, $\boldsymbol{\alpha} = (-1, 1)$ for all items, and for $K = 5$, $\boldsymbol{\alpha} = (-1, -0.5, 0.5, 1)$ for all items. For $n = 5$, $\boldsymbol{\beta} = (1, 1.5, 2, 1.5, 1)$, whereas for $n = 10$, $\boldsymbol{\beta} = (1, 1.5, 2, 1.5, 1, 1, 1.5, 2, 1.5, 1)$.

A small model with $K = 3$ and $n = 5$ was chosen to show that the empirical rejection rates of M_2 are similar to those of X^2 when the latter are accurate. The other cases with larger C were chosen to show that the empirical rejection rates of M_2 remain accurate, unlike those of X^2, even

TABLE 1.
Small sample distribution for X^2 and M_2 with ML estimation. Mean, variance, and exceedances of asymptotic upper 0.2, 0.1, 0.05, 0.01 quantiles.

n	K	N	Statistic	df	Mean	Var.	$\alpha = .2$	$\alpha = .1$	$\alpha = .05$	$\alpha = .01$
5	3	300	X^2	227	227.4	460.8	0.21	0.11	0.058	0.011
			M_2	35	35.1	67.1	0.20	0.10	0.051	0.009
5	3	1000	X^2	227	227.5	487.4	0.21	0.12	0.066	0.015
			M_2	35	34.9	71.8	0.21	0.11	0.045	0.006
5	3	3000	X^2	227	227.1	470.5	0.19	0.10	0.056	0.011
			M_2	35	35.1	70.4	0.22	0.11	0.050	0.005
5	5	300	X^2	3099	3094	75200	0.35	0.31	0.27	0.21
			M_2	155	155	300	0.20	0.10	0.042	0.005
5	5	1000	X^2	3099	3097	23675	0.30	0.24	0.19	0.11
			M_2	155	155	311	0.20	0.10	0.053	0.010
5	5	3000	X^2	3099	3097	12108	0.26	0.16	0.10	0.04
			M_2	155	155	301	0.20	0.09	0.042	0.012
10	5	300	X^2	9765574	9.68×10^6	6.07×10^{12}	0.37	0.37	0.37	0.36
			M_2	710	711	1482	0.21	0.11	0.064	0.011
10	5	1000	X^2	9765574	9.73×10^6	1.13×10^{12}	0.42	0.42	0.42	0.41
			M_2	710	710	1339	0.18	0.10	0.055	0.008
10	5	3000	X^2	9765574	9.77×10^6	3.60×10^{11}	0.48	0.48	0.48	0.48
			M_2	710	708	1315	0.19	0.08	0.039	0.008

Note: 1000 replications. MVM model given in (3.4)–(3.6).

TABLE 2.
Small sample distribution for M_2 with BCL estimator. Mean, variance, and exceedances of asymptotic upper 0.2, 0.1, 0.05, 0.01 quantiles.

n	K	N	Statistic	df	Mean	Var.	$\alpha = .2$	$\alpha = .1$	$\alpha = .05$	$\alpha = .01$
5	3	300	M_2	35	35.1	67.0	0.20	0.10	0.050	0.008
5	3	1000	M_2	35	34.9	71.7	0.21	0.10	0.045	0.006
5	3	3000	M_2	35	35.1	70.4	0.22	0.11	0.049	0.005
5	5	300	M_2	155	155	299	0.20	0.10	0.041	0.005
5	5	1000	M_2	155	155	309	0.19	0.10	0.050	0.009
5	5	3000	M_2	155	155	301	0.20	0.09	0.041	0.013
10	5	300	M_2	710	711	1435	0.20	0.11	0.056	0.009
10	5	1000	M_2	710	710	1327	0.18	0.10	0.051	0.007
10	5	3000	M_2	710	708	1310	0.19	0.09	0.036	0.009

Note: 1000 replications. MVM model given in (3.4–3.6).

for extremely sparse tables. As can be seen in Table 1, the empirical Type I errors for M_2 remain close to their nominal levels even at the highest degree of sparseness considered, whereas those of X^2 are only accurate in the small model with $K = 3$, $n = 5$.

Also, we use a bivariate composite likelihood (BCL) estimator (Zhao & Joe, 2005) under the same conditions as above to illustrate the behavior of M_2 for estimators that are not BAN. The results are shown in Table 2. The BCL estimator is the maximum of the sum of the $\binom{n}{2}$ bivariate marginal log-likelihood, rather than the maximum of the joint n-dimensional log-likelihood. In one special setting, Jöreskog and Moustaki (2001) refer to this as the underlying bivariate normal (UBN) approach. If the trivariate margins are not sparse, one could consider the trivariate composite likelihood (TCL) estimator. The asymptotic analysis of this estimator can be done using the theory of estimating equations (Godambe, 1991) and the asymptotic covariance matrix of the BCL estimator is an inverse Godambe information matrix, which can be compared with the inverse Fisher information matrix. We were able to compute both of these for different parameter vectors for (3.5), and look at ratios of the diagonals of these two matrices. For all cases that we computed for $n \leq 10$ and $2 \leq K \leq 5$, the asymptotic relative efficiency of any component of the BCL estimator is over 0.98; the average efficiency tends to slowly decrease as n increases.

As can be seen in Table 2, the finite sample null distribution of the M_2 statistic with the BCL estimator behaves very similarly to M_2 with the MLE. Although we have only studied the small sample performance for one (commonly used) model for item response categorical data, we expect the behavior to be similar for other models, for the MLE, and for other $\sqrt{N}$ -consistent estimators.

3.3. Power Comparison of X^2 and M_r when Data Are Not Sparse

For the binary case, Maydeu-Olivares and Joe (2005) have an asymptotic power comparison under a sequence of local alternatives for model (3.4)–(3.6). They report that M_2 and M_3 typically had more power asymptotically than X^2 for the null hypothesis of a common slope parameter.

For $K_i = K > 2$, we have done some simulations that show a similar behavior for M_2 for finite sample sizes. A finite sample power comparison of X^2 and M_r is meaningful only in the nonsparse cases where X^2 can be used with chi-square critical values. Consequently, Table 3 has some summaries for representative cases for small sample power comparison for $n = 5$, $K = 3$, using model (3.4)–(3.6) with constant $\beta_i = \beta$ as the null nested model.

TABLE 3.
Small sample power for M_2 vs. X^2 with MLE estimator; MVM model given in (3.4)–(3.6) with a common slope parameter for the null hypothesis. Exceedances of asymptotic upper 0.2, 0.1, 0.05, 0.01 quantiles based on 1000 replications.

$\boldsymbol{\alpha}$	β (altern.)	N	Stat.	$\alpha = .2$	$\alpha = .1$	$\alpha = .05$	$\alpha = .01$
$1-1, 1-1, 1-1, 1-1, 1-1$	1 1 1 .5 .5	500	X^2	0.37	0.21	0.12	0.04
			M_2	0.60	0.42	0.31	0.14
$1-1, 1-1, 1-1, 1-1, 1-1$	1 1 1 1 .5	500	X^2	0.36	0.20	0.13	0.03
			M_2	0.56	0.40	0.29	0.11
$1-1, 1-1, 1-1, 1-1, 1-1$	1 .9 .8 .9 .8	1200	X^2	0.24	0.11	0.05	0.03
			M_2	0.26	0.14	0.09	0.02
$1-1, 1-1, 1-1, .5-.5, .5-.5$	1 1 1 1 .5	500	X^2	0.31	0.18	0.09	0.03
			M_2	0.54	0.37	0.26	0.11

3.4. Remarks on Hypotheses of Specified Marginal Distributions

In this paper we focus on testing the null hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$ versus the alternative $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$. These are full information hypotheses. The asymptotic distribution of the family $\{M_r : 2 \leq r \leq n\}$ is derived under this null hypothesis. It is not derived under a marginal null of the type $H_0' : \boldsymbol{\pi}_r = \boldsymbol{\pi}_r(\boldsymbol{\theta})$. In fact, under the latter, M_r is not defined, because the matrix of the quadratic form involves marginal probabilities of order up to $r_2 = \min\{2r, n\}$. For $H_0' : \boldsymbol{\pi}_r = \boldsymbol{\pi}_r(\boldsymbol{\theta})$, only r-dimensional marginal distributions are assumed, higher-order margins are left unspecified, and a statistic other than M_r would have to be used. A suitable statistic for H_0' is M_r' introduced by Maydeu-Olivares and Joe (2005). M_r' differs from M_r in that the marginal probabilities up to r_2 are consistently evaluated using sample proportions rather than estimated probabilities as in M_r.

In applications one typically has a completely specified probability model in mind. Also, one cannot arbitrarily come up with $\boldsymbol{\pi}_r(\boldsymbol{\theta})$. The only sure way of specifying a well-defined $\boldsymbol{\pi}_r(\boldsymbol{\theta})$ or $\boldsymbol{\pi}_{r_2}(\boldsymbol{\theta})$ is from marginalizing an n-dimensional distribution $\boldsymbol{\pi}(\boldsymbol{\theta})$.

We use limited information statistics to assess full information hypotheses. M_r can be used when sample r-dimensional margins are not sparse and higher-dimensional margins are sparse. However, limited information tests based on marginal moments up to order r have no power to distinguish among models with the same margins up to order r but different higher-order margins. Also, hypothesis testing based on limited information is somewhat unbalanced, as if we fail to reject the full information null hypothesis using margins up to order r, we are unable to detect if the model does not fit well higher-order margins.

4. Using Marginal Residuals To Assess the Source of Misfit

When the M_r statistic suggests a model misfit, the vector of standardized marginal residuals can be inspected. This is $N[\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})]$, the differences of observed and model expected counts or moments for margins, divided by the square root of $\text{diag}\,(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))$. Note that this vector includes only those categories for which no category index is 0. The remaining residuals can be obtained based on zero sum constraints or by computing the residuals from inverse coded categories (the M_r statistic is invariant to the inverse coding). In large models, particularly when the number of categories for some variables is large, there will be a large number of marginal residuals involved and it may be difficult to draw useful information. Furthermore, the standardized residuals may be difficult to compute in large models.

A more fruitful avenue to assess the source of misfit might be to examine the rth-dimensional marginal tables. Note that this is like multiple testing after a jointly significant result. An analogy is Fisher's least significant difference following a significant F-ratio in ANOVA. In other words, we recommend assessing the source of misfit by computing $M_r^{(b)}(\hat{\boldsymbol{\theta}})$ for each subset b of $\{1, \ldots, n\}$ with cardinality r. For a submodel for r-dimensional margins, with $C_r(b) = \prod_{i \in b} K_i$ cells depending on $q_r(b)$ parameters, $M_r^{(b)}(\hat{\boldsymbol{\theta}})$ has an asymptotic null chi-square distribution with $C_r(b) - q_r(b) - 1$ degrees of freedom, provided the submodel is identified, the estimator is consistent and asymptotically normal, and $C_r(b) - 1 > q_r(b)$. When $r = 2$, we write $M_2^{(ij)}$ for $1 \leq i < j \leq n$. Also if $K_i = K$ for all i, then $C_2(b) = K^2$.

Consider M_r applied to the r-variate subset b. Let the vector of sample and model moments for this subset be denoted as $\dot{\mathbf{p}}_{rb}$ and $\dot{\boldsymbol{\pi}}_{rb}(\hat{\boldsymbol{\theta}}_b)$, respectively, both of dimension $C_r(b) - 1$. Typically, $\boldsymbol{\theta}_b$ is a subset of the vector $\boldsymbol{\theta}$. Let $q_r(b)$ be the dimension of $\boldsymbol{\theta}_b$. Using (3.3), we can write M_r in this case as

$$M_r^{(b)}(\hat{\boldsymbol{\theta}}) = M_r^{(b)}(\hat{\boldsymbol{\theta}}_b) = N(\dot{\mathbf{p}}_{rb} - \dot{\boldsymbol{\pi}}_{rb}(\hat{\boldsymbol{\theta}}_b))' \widehat{\mathbf{C}}_{rb} (\dot{\mathbf{p}}_{rb} - \dot{\boldsymbol{\pi}}_{rb}(\hat{\boldsymbol{\theta}}_b))$$

for some $\sqrt{N}$ -consistent and asymptotically normal estimator $\hat{\boldsymbol{\theta}}$. We assume that $\boldsymbol{\Delta}_{rb} = \partial \dot{\boldsymbol{\pi}}_{rb}(\boldsymbol{\theta}_b)/\partial \boldsymbol{\theta}_b'$ is of full rank $q_r(b)$, so that the submodel is (locally) identified. Also, we assume that $C_r(b) - 1 - q_r(b) > 0$. The matrix of the above quadratic form is

$$\widehat{\mathbf{C}}_{rb} = \mathbf{C}_{rb}(\hat{\boldsymbol{\theta}}_b) = \boldsymbol{\Delta}_{rb}^{(c)} \left([\boldsymbol{\Delta}_{rb}^{(c)}]' \boldsymbol{\Xi}_{rb} \boldsymbol{\Delta}_{rb}^{(c)} \right)^{-1} [\boldsymbol{\Delta}_{rb}^{(c)}]'$$

evaluated at $\hat{\boldsymbol{\theta}}_b$, where $\boldsymbol{\Delta}_{rb}^{(c)}$ is an orthogonal complement to $\boldsymbol{\Delta}_{rb}$, and $\boldsymbol{\Xi}_{rb}$ is N times the asymptotic covariance matrix of $\mathbf{p}_{rb} - \boldsymbol{\pi}_{rb}(\boldsymbol{\theta}_b)$. Now, $(\mathbf{p}_{rb} - \boldsymbol{\pi}_{rb}(\hat{\boldsymbol{\theta}}_b)) = \mathbf{T}_{rb}(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}))$ for some $(C_r(b) - 1) \times C$ matrix $\mathbf{T}_{rb}$. Thus, using (2.5), the asymptotic covariance matrix of $\sqrt{N}\,(\mathbf{p}_{rb} - \boldsymbol{\pi}_{rb}(\hat{\boldsymbol{\theta}}_b))$ is $\boldsymbol{\Sigma}_{rb} = \mathbf{T}_{rb}(\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})\boldsymbol{\Gamma}(\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})'\mathbf{T}_{rb}' = (\mathbf{T}_{rb} - \boldsymbol{\Delta}_{rb}\mathbf{H})\boldsymbol{\Gamma}(\mathbf{T}_{rb} - \boldsymbol{\Delta}_{rb}\mathbf{H})'$.

A necessary and sufficient condition for $M_r^{(b)}$ to be asymptotically distributed (under H_0) as a chi-square with ν degrees of freedom in this setup is (Schott, 1997, Theorem 9.10)

$$\boldsymbol{\Sigma}_{rb}\mathbf{C}_{rb}\boldsymbol{\Sigma}_{rb}\mathbf{C}_{rb}\boldsymbol{\Sigma}_{rb} = \boldsymbol{\Sigma}_{rb}\mathbf{C}_{rb}\boldsymbol{\Sigma}_{rb} \qquad \text{for any } \boldsymbol{\theta}, \tag{3.8}$$

where $\nu = \operatorname{tr}(\mathbf{C}_{rb}\boldsymbol{\Sigma}_{rb})$. Since $\boldsymbol{\Xi}_{rb} = \mathbf{T}_{rb}\boldsymbol{\Gamma}\mathbf{T}_{rb}'$, it can be readily verified that $\mathbf{C}_{rb} = \mathbf{C}_{rb}\boldsymbol{\Sigma}_{rb}\mathbf{C}_{rb}$. That is, $\boldsymbol{\Sigma}_{rb}$ is a generalized inverse for $\mathbf{C}_{rb}$. So, (4.1) is satisfied. Also, the degrees of freedom are obtained using the fact that $\boldsymbol{\Delta}_{rb}^{(c)}$ is of full column rank $C_r(b) - 1 - q_r(b) > 0$ and hence $\mathbf{C}_{rb}$ is also of rank $C_r(b) - 1 - q_r(b)$. Thus, the null distribution of $M_r^{(b)}(\hat{\boldsymbol{\theta}}_b)$ is asymptotically chi-square with degrees of freedom $C_r(b) - 1 - q_r(b)$.

On the other hand, Pearson's X^2 is not asymptotically chi-square under H_0 when applied to subsets of variables even for BAN estimators. To see this, from the Appendix of Maydeu-Olivares and Joe (2005), X^2 applied to the r-variate subset b can be written as $X_b^2 = N(\dot{\mathbf{p}}_b - \dot{\boldsymbol{\pi}}_b(\hat{\boldsymbol{\theta}}_b))' \boldsymbol{\Xi}_{rb}^{-1}(\dot{\mathbf{p}}_b - \dot{\boldsymbol{\pi}}_b(\hat{\boldsymbol{\theta}}_b))$. Now, using (2.5), the asymptotic covariance matrix of $\sqrt{N}\,(\dot{\mathbf{p}}_b - \dot{\boldsymbol{\pi}}_b(\hat{\boldsymbol{\theta}}_b))$ for BAN estimators such as the MLE is $\boldsymbol{\Sigma}_{rb} = \mathbf{T}_{rb}(\boldsymbol{\Gamma} - \boldsymbol{\Delta}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}')\mathbf{T}_{rb}' = \boldsymbol{\Xi}_{rb} - \boldsymbol{\Delta}_{rb}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}_{rb}' = \boldsymbol{\Xi}_{rb} - \mathbf{A}$, where $\mathbf{A} = \boldsymbol{\Delta}_{rb}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}_{rb}'$ is symmetric. For this $\boldsymbol{\Sigma}_{rb}$, it can be readily verified that $\boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb} = \boldsymbol{\Xi}_{rb} - 2\mathbf{A} + \mathbf{A}\boldsymbol{\Xi}_{rb}^{-1}\mathbf{A}$ and $\boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb} = \boldsymbol{\Xi}_{rb} - 3\mathbf{A} + 3\mathbf{A}\boldsymbol{\Xi}_{rb}^{-1}\mathbf{A} - \mathbf{A}\boldsymbol{\Xi}_{rb}^{-1}\,\mathbf{A}\boldsymbol{\Xi}_{rb}^{-1}\mathbf{A}$, so that $\boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb} \neq \boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb}$ in general. To get a null asymptotic distribution that is chi-square, a BAN estimator based on the variables in the subset b must be used.

In closing this section note that from the results in Section 3.1, (incorrectly) using X^2 instead of $M_r^{(b)}(\hat{\boldsymbol{\theta}}_b)$ with a chi-square with degrees of freedom $C_r(b) - 1 - q_r(b)$ reference distribution would result in an undue impression of poor fit as X^2 is stochastically larger than $M_r^{(b)}(\hat{\boldsymbol{\theta}}_b)$.

5. Data Examples

In this section we provide two numerical data examples to illustrate our results. In these examples we used Samejima's (1969) graded logistic model to fit questionnaire data using the MLE. In the first example a small model is considered, $C = 3^5 = 243$, where the contingency table is not very sparse. In the second example we fit a larger model, $C = 5^{10} \approx 10^7$, to illustrate a highly sparse situation.

5.1. The Satisfaction with Life Scale Data

The Satisfaction with Life Scale (SWLS) (Diener, Emmons, Larsen, & Griffin, 1985) is a widely used questionnaire consisting of $n = 5$ statements intended to obtain a global cognitive judgment of one's life. The SWLS is usually completed using a 7-point rating scale. However, Kramp (2006) investigated experimentally the effects of varying the number of response options in several rating scales, among them the SWLS. Here we shall fit Samejima's graded logistic model to an experimental version of the SWLS where respondents were asked to employ the following three-point scale: 0 = disagree, 1 = neither agree nor disagree, and 2 = agree. The sample size is $N = 429$, so the contingency table is not very sparse ($N/C = 1.77$). However, even in this situation, 141 cells have zero counts. As a consequence of these zero observed counts the full information test statistics X^2 and G^2 yield very different conclusions: $X^2 = 310$, $p = 0.0002$ and $G^2 = 199$, $p = 0.91$, both on 227 degrees of freedom. The M_2 statistic, on the other hand, suggests that the model does not fit well, but not as poorly as X^2 : $M_2 = 57.05$ on 35 degrees of freedom, $p = 0.01$. Notice that, in this case, since we are using ML estimation, $X^2 = M_5$.

As the model does not fit well, we proceed to investigate the source of the misfit. Large standardized cell residuals were obtained for the patterns (01222), (20122), (00210), (11211), (02000), (22120), (02200), (00122), (21000), (22102), (10202), (00021), (22010). We cannot meaningfully extract any trend in these patterns. As an alternative, we computed goodness-of-fit statistics for bivariate subtables.

Each bivariate table depends on $2(K-1)$ intercepts and two slopes. Thus, there are $(K^2 - 1) - 2(K-1) - 2 = 2$ degrees of freedom when M_2 is applied to bivariate subtables. We cannot assess how well this model fits each item separately using M_1 as the univariate submodels are not identified. There are $K-1$ mathematically independent probabilities in each univariate table. But each univariate table depends on $K-1$ intercepts α and one slope β.

We provide in Table 4 the bivariate statistics computed for every pair of variables. As can be seen in this table, the pairwise $M_2^{(ij)}$ statistics suggest that the model does not fit well for item 2. To verify this conjecture we reestimated the model to each subset of $n - 1 = 4$ items. The results are shown in Table 5.

TABLE 4.
$M_2^{(ij)}$ statistics applied to bivariate subtables for the SWLS data.

Items	1	2	3	4	5
1	—	0.03	6.00	6.19	0.08
2	0.03	—	1.41	9.13*	8.71*
3	6.00	1.41	—	0.83	0.05
4	6.19	9.13*	0.83	—	1.52
5	0.08	8.71*	0.05	1.52	—

Note: Statistics significant at the $\alpha = 0.05$ significance level are marked with $*$.

TABLE 5.
Overall goodness-of-fit results for subsets of $n - 1 = 4$ items for the SWLS data.

Dropping	X^2	p	M_2	p
1	86.98	0.06	24.97	0.20
2	64.59	0.59	22.91	0.29
3	91.93	0.03	35.70	0.02
4	82.20	0.12	36.37	0.01
5	105.36	<0.01	30.06	0.07

Note: $X^2 = M_4^{(-i)}$; there are 68 degrees of freedom for X^2 and 20 for $M_2^{(-i)}$ (ith item deleted).

The results of this table strongly suggest that the fit of Samejima's graded logistic model to these data can be improved by removing item 2, as suggested by the bivariate $M_2^{(ij)}$ statistics. Also notice that in this table there are some discrepancies between $X^2 = M_4^{(-i)}$ and $M_2^{(-i)}$ with the ith item deleted, which suggests, given our simulation results in Section 3.2, that sparseness can still have some adverse effects on the small sample behavior of X^2 even in such small tables.

5.2. *The Negative Problem Orientation Data*

Following Drasgow et al. (1995), Maydeu-Olivares (2005) used X^2 statistics for single items, item pairs, and item triplets to compare in a descriptive fashion the fit of several unidimensional IRT models to each of the five scales of the Social Problem Solving Inventory-Revised (SPSI-R) (D'Zurilla, Nezu, & Maydeu-Olivares, 2002). The models considered were Samejima's graded logistic model, Masters's (1982) partial credit model, Thissen and Steinberg's (1986) extension of the latter, and Bock's (1972) nominal model. In all scales Samejima's graded logistic model yielded the best fit. However, since the statistics employed to compare the models had an unknown sampling distribution, nothing could be concluded about the absolute fit of the models. In this example, we shall reanalyze data in Maydeu-Olivares (2005) from one of the SPSI-R scales, the Negative Problem Orientation (NPO) scale, to investigate whether the best fitting model, Samejima's graded logistic model, indeed provides an adequate fit to the data.

The NPO scale consists of 10 items intended to measure individual differences in: (a) viewing a problem as a significant threat to well-being; (b) doubting one's personal ability to solve problems successfully; and (c) easily becoming frustrated and upset when confronted with problems in everyday living. Individuals are asked to respond to each item using one of five categories: "0 = Not at all true of me"; "1 = Slightly true of me"; "2 = Moderately true of me"; "3 = Very true of me"; "4 = Extremely true of me." The sample size is $N = 1053$.

Samejima's graded logistic model was estimated by ML. The parameter estimates and standard errors are reported in Table 6. The number of degrees of freedom available for testing using X^2 and G^2 is very large, $df = 9765574$, and each statistic offers a very different picture: $X^2 \approx 6 \times 10^7 \gg df$, $G^2 \approx 13000 \ll df$. Given the extremely large degree of sparseness of the data, neither statistic can be trusted and we resort to M_2 to assess the overall fit of the model. With 710 degrees of freedom we obtained $M_2 \approx 1500$, $p \ll .001$. Thus, the model fits very poorly.

To assess the source of misfit we used, as in the previous example, pairwise $M_2^{(ij)}$ statistics, each with 14 degrees of freedom. These statistics are shown in Table 7. In this table we used a Bonferroni adjustment for the $M_2^{(ij)}$ statistics. Thus, those statistics that exceed 35.82, the upper $0.05/45 = .0011$ quantile of the χ_{14}^2 distribution, are indicated with an asterisk. Even with this correction, Table 7 reveals that the misfit of the model cannot be attributed to any particular item.

TABLE 6.
ML estimates and standard errors for the NPO data.

Parameters	Estimates	Standard errors
$\alpha_{11}, \ldots, \alpha_{14}$	1.97, 0.14, −1.52, −3.16	0.11, 0.09, 0.10, 0.15
$\alpha_{21}, \ldots, \alpha_{24}$	2.05, −0.06, −1.94, −3.99	0.12, 0.10, 0.12, 0.19
$\alpha_{31}, \ldots, \alpha_{34}$	2.29, 0.15, −1.39, −3.28	0.12, 0.09, 0.10, 0.15
$\alpha_{41}, \ldots, \alpha_{44}$	2.15, −0.02, −1.63, −3.52	0.12, 0.09, 0.11, 0.16
$\alpha_{51}, \ldots, \alpha_{54}$	0.89, −0.92, −2.46, −4.22	0.10, 0.10, 0.13, 0.19
$\alpha_{61}, \ldots, \alpha_{64}$	2.92, 0.73, −0.90, −3.05	0.14, 0.10, 0.10, 0.15
$\alpha_{71}, \ldots, \alpha_{74}$	1.65, −0.68, −2.52, −4.51	0.13, 0.11, 0.15, 0.22
$\alpha_{81}, \ldots, \alpha_{84}$	1.63, −0.18, −1.34, −2.76	0.10, 0.09, 0.10, 0.13
$\alpha_{91}, \ldots, \alpha_{94}$	1.12, −0.86, −2.52, −4.65	0.12, 0.12, 0.15, 0.22
$\alpha_{10,1}, \ldots, \alpha_{10,4}$	2.33, −0.37, −2.28, −4.63	0.14, 0.12, 0.15, 0.22
$\beta_1, \ldots, \beta_5$	1.57, 2.06, 1.78, 1.71, 1.74	0.09, 0.11, 0.10, 0.10, 0.10
$\beta_6, \ldots, \beta_{10}$	2.02, 2.43, 1.51, 2.47, 2.57	0.11, 0.13, 0.09, 0.14, 0.14

Rather, it is widespread. Thus, we conclude that although results in Maydeu-Olivares (2005) suggest that Samejima's logistic graded model was the best fitting model for these data among a set of parametric IRT models, this model does not provide a satisfactory fit to this questionnaire. An alternative model is needed.

6. Discussion and Conclusions

Applied researchers confronted with the problem of modeling sparse multidimensional contingency tables are faced with the problem of how to assess the overall goodness of fit of the model and, should the overall fit be poor, how to identify the source of the misfit. Much attention has been devoted in recent years to limited information procedures to overcome these problems.

With regard to overall goodness-of-fit assessment, most of the statistics proposed are quadratic forms in low-order marginal residuals. The statistics differ in that they are based on different marginal residuals, and/or on the choice of weight matrix. Some of the statistics, such as the overall *GFfit* statistic proposed by Jöreskog and Moustaki (2001), are used

TABLE 7.
Bivariate subtable $M_2^{(ij)}$ for the NPO data.

Items	1	2	3	4	5	6	7	8	9	10
1	—	48.37*	23.35	31.37	12.84	39.73*	34.63*	26.67	51.11*	19.87
2	48.37*	—	55.31*	28.64	19.49	30.13	15.43	21.35	18.31	33.08
3	23.35	55.31*	—	28.40	24.46	36.47*	31.14	27.08	37.92*	35.73
4	31.37	28.64	28.40	—	31.67	57.42*	29.69	32.32	34.99	49.28*
5	12.84	19.49	24.46	31.67	—	30.51	32.70	46.82*	38.77*	27.45
6	39.73*	30.13	36.47*	57.42*	30.51	—	31.43	50.55*	50.97*	55.49*
7	34.63	15.43	31.14	29.69	32.70	31.43	—	45.23*	58.72*	92.67*
8	26.67	21.35	27.08	32.32	46.82*	50.55*	45.23*	—	25.75	29.34
9	51.11*	18.31	37.92*	34.99	38.77*	50.97*	58.72*	25.75	—	35.89*
10	19.87	33.08	35.73	49.28*	27.45	55.49*	92.67*	29.34	35.89*	—

Note: Statistics significant at the $\alpha = 0.05/45 = .0011$ level are marked with $*$.

heuristically. Other limited information statistics can be referenced to a chi-square distribution. To obtain asymptotic p-values for limited information statistics, two procedures have been used. One approach is to construct a quadratic form statistic that is asymptotically chi-square. This is the approach taken by Christoffersson (1975), Reiser (1996), and Maydeu-Olivares and Joe (2005). Another approach is to construct a quadratic form statistic that is asymptotically distributed as a mixture of independent chi-square variates and approximate its distribution using a chi-square distribution by matching moments. This is the approach used in Maydeu-Olivares (2001a, 2001b), Bartholomew and Leung (2002), and Cai et al. (2006).

Christoffersson (1975) proposed a statistic for binary data models that is asymptotically chi-square for asymptotically efficient estimators based on univariate and bivariate margins. For other estimators the statistic is not asymptotically chi-square (Maydeu-Olivares & Joe, 2005). Reiser (1996) proposed another statistic for binary data models which is asymptotically chi-square for BAN estimators such as the MLE. This statistic uses as weight matrix a generalized inverse of the asymptotic covariance matrix of the marginal residuals ($\mathbf{\Sigma}_r$ in our notation), which requires computing the asymptotic covariance matrix of the parameter estimates.

Statistics based on moment corrections also require computing the covariance matrix of the parameter estimates. Thus, in two separate reports, Maydeu-Olivares (2001a, 2001b) gave formulas for obtaining p-values for the same statistic for two different estimators for binary data models. In one case, for models estimated using the three-stage estimation procedure proposed by Muthén (1993) and implemented in Mplus (Muthén & Muthén, 2001) and Lisrel (Jöreskog & Sörbom, 2001); and in, the second case, for models estimated using the two-stage estimation procedure implemented in NOHARM (Fraser & McDonald, 1988). Bartholomew and Leung (2002) proposed another limited information statistic for binary data models estimated with the MLE using moment corrections without computing the covariance matrix of the ML parameter estimates. Cai et al. (2006) showed that the effect of parameter estimation in Bartholomew and Leung's statistic was too large to be ignored and that the computation of the covariance matrix of the ML parameter estimates is necessary to obtain accurate p-values for that statistic.

In contrast, the statistics in the M_r family proposed by Maydeu-Olivares and Joe (2005) for binary data models do not require computing the asymptotic covariance matrix of the parameter estimates. Furthermore, unlike previous statistics, which are associated with particular estimators, M_r is asymptotically chi-squared distributed for *all* members of the class of $\sqrt{N}$ -consistent and asymptotically normal estimators.

In this paper we have extended Maydeu-Olivares and Joe's (2005) work on limited information goodness-of-fit testing of composite hypotheses in multidimensional binary contingency tables to multidimensional contingency tables of arbitrary dimensions. We have shown that their M_r family of overall goodness-of-fit statistics extends readily to the general case.

To date only one limited information test statistic had been proposed for multidimensional contingency tables of arbitrary dimensions (Maydeu-Olivares, 2006). In this statistic, asymptotic p-values are again obtained by matching moments, and it requires the computation of the asymptotic covariance matrix of the parameter estimates. The p-values are only valid when parameters are estimated using the sequential estimator described in Jöreskog (1994) and implemented in Lisrel.

An interesting alternative line of research on limited information testing with second-order moments is that of Glas and co-workers, most notably Glas (1988, 1999) and Glas and Verhelst (1989). A direction of future research is to compare their approach with ours, for statistical power and detection of various misfits, in the context of item response models.

With regard to the assessment of the source of misfit, Drasgow et al. (1995) suggested computing Pearson's X^2 statistic (adjusted for sample size) to univariate, bivariate, and trivariate subtables, particularly in cross-validation samples. Also, Jöreskog and Moustaki (2001)

heuristically proposed computing X^2 statistics for univariate and bivariate subtables (their *GFfit* statistic). Here, we have suggested employing the M_r statistic for r-dimensional subtables. Provided the subtable's model is identified, the $M_r^{(b)}$ statistics are asymptotically chi-square with degrees of freedom equal to the number of cells in the subtable minus the number of parameters involved in the subtable minus one. This result holds for all $\sqrt{N}$-consistent and asymptotically normal estimators. Furthermore, we have shown that X^2 =*GFfit* applied to subtables is not asymptotically chi-square even for the MLE. These $M_r^{(b)}$ statistics applied to subtables may be very useful to identify the source of the misfit in multidimensional contingency tables where the number of categories is large.

In large and/or sparse contingency tables, M_r for small r ($r = 2, 3$) should be employed instead of X^2 as the former have more precise empirical Type I errors and may be more powerful than the latter. In the case of $K = 2$, for sequences of local alternatives involving the comparison of the two-parameter logistic IRT model versus the one-parameter (Rasch) model, Maydeu-Olivares and Joe (2005) showed that asymptotically M_2 (and typically also M_3) is more powerful than X^2. Here, we have shown that in finite samples with $K > 2$ and non-sparse data that M_2 is more powerful than X^2 for the comparison of the graded model versus a graded model with a common slope. The comparative power of M_r and X^2 for nonsparse data depends on the models and hypotheses, and we expect that X^2 will be more powerful in some cases. When the data are sparse, the power comparisons based on chi-square critical values are not meaningful as X^2 has inflated empirical Type I error rates. Also, the power comparisons are only meaningful when the source of misfit is embedded in low-order margins. When testing, using up to r-dimensional margins, M_r has no power to detect model misfit in margins higher than r.

Also, the family of statistics M_r compares favorably to the use of resampling methods for overall goodness-of-fit assessment of composite null hypothesis in multidimensional contingency tables. On the one hand, one can obtain a p-value for the overall fit of the model with considerably less computing effort than by resampling methods. On the other hand, they provide a way to detect the source of misfit of the model.

However, there are three obvious limitations to the use of the approach advocated here. First, the model must be identified from the margins. In practice, most models of interest—such as the IRT model considered here—can be identified from the bivariate or trivariate margins. The second limitation stems from the fact that by testing using only margins up to order r there is no power to detect misfit in higher-order margins. Thus, although the use of limited information may be very useful to reject a full information null hypothesis, it may not provide enough information when it fails to reject the null. In practice—as our examples illustrate—as the number of categorical variables increases, it is not an easy task to find a model that is not rejected even when testing is performed using only univariate and bivariate margins. The third limitation is computational. When some of the observed variables have a large number of categories K_i, even computing M_2 for $n > 15$ can be computationally infeasible as the dimension $s(2)$ of the matrices $\boldsymbol{\Xi}_r$ and $\boldsymbol{\Delta}_r$ gets too large.

When the categorical data are ordinal, then there exists an alternative set of limited information test statistics that are invariant to the set of permissible transformations of the ordinal data and that can be used with much larger models than those feasible using the M_r family of test statistics. This alternative approach, suitable only for ordinal variables, will be discussed in a separate report.

In closing, while we have focused on testing composite hypotheses, the common situation in applications, the general framework discussed here can also be applied to testing simple null hypotheses. This is discussed in the Appendix. Also, although the applications and simulations in this paper are focused on item response models, the theory introduced in this paper is completely general for multivariate discrete data. For example, for multivariate continuous variables with a

copula model (e.g., Joe, 1997) there is no general approach for assessing goodness of fit other than discretizing the variables. Applying the family of M_r statistics to discretized continuous variables to assess goodness of fit is another topic for future investigation.

Appendix

Goodness-of-Fit Testing of Simple Null Hypotheses

For testing the overall goodness of fit of a simple null hypotheses $H_0 : \boldsymbol{\pi}(\boldsymbol{\theta})$ for a fixed a priori vector $\boldsymbol{\theta}$ of dimension q, and as an alternative to X^2 in sparse tables, Maydeu-Olivares and Joe (2005) proposed using the family of limited information test statistics

$$L_r = N(\mathbf{p}_r - \boldsymbol{\pi}_r)' \boldsymbol{\Xi}_r^{-1} (\mathbf{p}_r - \boldsymbol{\pi}_r), \qquad r = 1, \ldots, n.$$

The choice of r depends on the sparseness of the contingency table. From (2.3), under H_0, the L_r statistics converge in distribution to a $\chi^2_{s(r)}$ distribution as $N \to \infty$. For $r = n$, $L_n = X^2$.

If the L_r test suggests significant misfit then $L_r^{(b)} = X_b^2$ for the r-dimensional subtables can be obtained to identify the source of the misfit. Under the null hypothesis, these statistics applied to subtables are asymptotically chi-square; the degrees of freedom for margin b is $[\prod_{i \in b} K_i] - 1$.

Some Computing Details

Consider a model, such as that given in (3.4)–(3.6), that has a form that is closed under margins. Then any probability in the rth-order margin and in $\boldsymbol{\Xi}_r$ can be computed directly without marginalizing the n-dimensional joint distribution. That is, for computations, one can avoid the large matrix $\mathbf{T}_r$ in Section 2.2, where it was presented for notational convenience. M_r depends on $\mathbf{C}_r$ evaluated at $\hat{\boldsymbol{\theta}}$, which depends on the matrices $\boldsymbol{\Delta}_r$ and $\boldsymbol{\Xi}_r$ evaluated at $\hat{\boldsymbol{\theta}}$. The matrix of partial derivatives $\boldsymbol{\Delta}_r$ can be computed at the same time as $\boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})$. The computation of $\boldsymbol{\Xi}_r$ is a bit more involved. $\boldsymbol{\Xi}_r$ is the covariance matrix of the vector of sample proportions of margins of order r or less. A term in $\boldsymbol{\Xi}_r$ has the form $m_c(\mathbf{y}_c) - m_a(\mathbf{y}_a) m_b(\mathbf{y}_b)$, where a, b are subsets of $\{1, \ldots, n\}$ of dimension between 1 and r, $c = a \cup b$, and m_a, m_b, m_c are marginal probabilities. Efficient computation of $\boldsymbol{\Xi}_r$ relies on a systematic way of enumerating the marginal probabilities corresponding to terms in $\mathcal{A}_r$.

Next we discuss computation of the MLE for model (3.4)–(3.6). There are similar considerations for other item response models. With Gauss–Hermite (GH) quadrature for evaluation of marginal probabilities of (3.4) and its derivatives, we have coded the computation of the MLE with the Newton–Raphson method. This is an alternative to the expectation-maximization (EM) algorithm (e.g., Bock & Aitkin, 1981). It has the advantage that the inverse observed Fisher information matrix, used as the estimated covariance matrix, is computed at the same time. For the covariance matrix of $\hat{\mathbf{e}}_r$ in (2.6), the expected Fisher information matrix is needed. Computing the information matrix is much harder than computing $\boldsymbol{\Xi}_r$ because the former requires summing through the probabilities in (3.4) for all $C = K^n$ n-dimensional probabilities, and this is essentially only feasible if $K^n < 10^9$. With most efficient use of computer memory, Fisher information $\boldsymbol{\mathcal{I}}$ can be computed as $\sum_{\mathbf{y}} (\partial \pi_{\mathbf{y}} / \partial \boldsymbol{\theta})(\partial \pi_{\mathbf{y}} / \partial \boldsymbol{\theta})' / \pi_{\mathbf{y}}$.

The BCL estimator for model (3.4)–(3.6) can also be obtained numerically with Gauss–Hermite quadrature and the Newton–Raphson method. The computations require the evaluation of marginal probabilities of (3.4) and its derivatives for dimensions 2, 3, and 4. We next indicate how to evaluate $\boldsymbol{\Sigma}_2$ in (2.6), without any matrices of order C. This technique applies to any $\sqrt{N}$ -consistent estimator that can be considered as a solution to a set of estimating equations. Let $\pi^{(ij)}_{k_1 k_2}(\boldsymbol{\theta}) = \Pr(Y_i = k_1, Y_j = k_2)$ and let $p^{(ij)}_{k_1 k_2}$ be the sample counterpart. Then the BCL estimator

$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{BCL}}$ maximizes

$$L_2(\boldsymbol{\theta}) = N \sum_{i<j} \sum_{k_1} \sum_{k_2} p_{k_1k_2}^{(ij)} \log \pi_{k_1k_2}^{(ij)}(\boldsymbol{\theta}).$$

From second-order Taylor approximation,

$$\mathbf{0} = N^{-1} \frac{\partial L_2(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = N^{-1} \frac{\partial L_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + N^{-1} \frac{\partial^2 L_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + O_p(N^{-1})$$

and using the theory of estimating equations (see, e.g., Zhao & Joe, 2005),

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \approx \mathbf{W}^{-1} N^{-1} \frac{\partial L_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \qquad \text{where} \qquad \mathbf{W} = \mathbf{W}(\boldsymbol{\theta}) = -N^{-1} \mathrm{E} \left[\frac{\partial^2 L_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right].$$

Note that

$$\begin{aligned} N^{-1} \frac{\partial L_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{i<j} \sum_{k_1} \sum_{k_2} p_{k_1k_2}^{(ij)} \frac{\partial \pi_{k_1k_2}^{(ij)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big/ \pi_{k_1k_2}^{(ij)}(\boldsymbol{\theta}) \\ &= \sum_{i<j} \sum_{k_1} \sum_{k_2} \left[p_{k_1k_2}^{(ij)} - \pi_{k_1k_2}^{(ij)}(\boldsymbol{\theta}) \right] \frac{\partial \pi_{k_1k_2}^{(ij)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big/ \pi_{k_1k_2}^{(ij)}(\boldsymbol{\theta}). \end{aligned} \tag{A. 1}$$

Let $\boldsymbol{\pi}_2^*$ be a vector containing all model-based bivariate marginal probabilities (including those with 0 indices). Also, let $\mathbf{p}_2^*$ be its sample counterpart. From (A.1), there is a matrix $\mathbf{K}$ such that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \approx \mathbf{W}^{-1}\mathbf{K}\big(\mathbf{p}_2^* - \boldsymbol{\pi}_2^*\big)$. From Section 2, each element of $\boldsymbol{\pi}_2^*$ is either an element of $\boldsymbol{\pi}_2$ or a linear function of elements of $\boldsymbol{\pi}_2$. Hence $\mathbf{p}_2^* - \boldsymbol{\pi}_2^* = \mathbf{S}(\mathbf{p}_2 - \boldsymbol{\pi}_2)$ for a matrix $\mathbf{S}$. Putting everything together, $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \approx \mathbf{H}_2(\mathbf{p}_2 - \boldsymbol{\pi}_2)$ where $\mathbf{H}_2 = \mathbf{W}^{-1}\mathbf{K}\mathbf{S}$. Since only probabilities in $\boldsymbol{\pi}_2$ are involved, $\boldsymbol{\Sigma}_2$ in (2.6) can be written as

$$\boldsymbol{\Sigma}_2 = \boldsymbol{\Xi}_2 - \boldsymbol{\Delta}_2\mathbf{H}_2\boldsymbol{\Xi}_2 - \boldsymbol{\Xi}_2\mathbf{H}_2'\boldsymbol{\Delta}_2' + \boldsymbol{\Delta}_2[\mathbf{H}_2\boldsymbol{\Xi}_2\mathbf{H}_2']\boldsymbol{\Delta}_2'.$$

For computer implementation in all of the above, a systematic way is needed to convert a multi-indexed margin to a row or column index in a matrix.

When using Gauss–Hermite quadrature for ML estimation, one must be careful in the simulation of (3.4)–(3.6) for the assessment of the null distribution of M_r. For a fixed number of quadrature points n_q, the accuracy decreases as the slope parameters increase in absolute value. This is checked by comparing Romberg integration with Gauss–Hermite quadrature. Hence, the number of quadrature points needs to increase as the slope parameter increases in order to achieve a desired accurary; $n_q = 48$ is acceptable provided β values do not exceed 3 in absolute value.

The null distribution of X^2 and M_2 depends on the simulation method if the MLE (or another estimator) is obtained based on Gauss–Hermite quadrature of the model probabilities. Rather than a standard normal latent random variable Z, Gauss–Hermite calculations with n_q quadrature points implicitly assume that the latent random variable Z' is discrete with mass w_i at point x_i for $i = 1, \ldots, n_q$ (note that $\sum_i w_i = 1$). Hence, if simulating with Z and estimating and calculating M_2 and X^2 with Z', the resulting "null distribution" of M_2 and X^2 will be stochastically larger than the (asymptotic) χ^2 distribution if the sample size N is large (relative to the number of vector categories C). This is because Z is different from Z' and the goodness-of-fit statistics can discriminate these two for large N. If estimation is based on Gauss–Hermite with Z', then simulation with Z means that a nonnull model that is close to null is used, and the M_2 and X^2 statistics will tend to be a bit larger than simulation with Z'. A rough calculation shows that the distribution of M_r in this case is approximately noncentral chi-square with noncentrality parameter $N\boldsymbol{\delta}_r'\mathbf{C}_r\boldsymbol{\delta}_r$ where $\boldsymbol{\delta}_r$ is the vector is differences in marginal moments up to order r for

probabilities based on latent variables Z and Z'. This behavior was readily seen in simulation results of Z versus Z'.

References

Agresti, A. (2002). *Categorical data analysis* (2nd edn.). New York: Wiley.

Bartholomew, D.J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd edn.). London: Arnold.

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

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

Bentler, P.M. (1995). *EQS*. Encino, CA: Multivariate Software.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

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

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

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

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5–32.

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

Collins, L.M., Fidler, P.L., Wugalter, S.E., & Long, J. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, *28*, 375–389.

Diener, E., Emmons, R.A., Larsen, R.J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment*, *49*, 71–75.

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

D'Zurilla, T.J., Nezu, A.M., & Maydeu-Olivares, A. (2002). *Manual of the social problem-solving inventory-Revised*. North Tonawanda, NY: Multi-Health Systems.

Fraser, C., & McDonald, R.P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*, 267–269.

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

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

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

Godambe, V.P. (Ed.) (1991). *Estimating functions*. Oxford: Oxford University Press.

Joe, H. (1997). *Multivariate models and dependence concepts*. London: Chapman & Hall.

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

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

Jöreskog, K.G., & Sörbom, D. (2001). *LISREL 8*. Chicago: Scientific Software.

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

Kramp, U. (2006). *Effects of the number of response options on personality rating scales*. Unpublished doctoral dissertation. University of Barcelona.

Lee, S.Y., Poon, W.Y., & Bentler, P.M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, *48*, 339–358.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Maydeu-Olivares, A. (2001a). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66*, 209–228.

Maydeu-Olivares, A. (2001b). Multidimensional item response theory modeling of binary data: Large sample properties of NOHARM estimates. *Journal of Educational and Behavioral Statistics*, *26*, 49–69.

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

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

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

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

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

Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K.A. Bollen, & J.S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.
Muthén, L., & Muthén, B. (2001). *MPLUS*. Los Angeles: Muthén & Muthén.
Rao, C.R. (1973). *Linear statistical inference and its applications*. New York: Wiley.
Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika*, *61*, 509–528.
Reiser, M., & Lin, Y. (1999). A goodness-of-fit test for the latent class model when expected frequencies are small. In M. Sobel, & M. Becker (Eds.), *Sociological methodology 1999* (pp. 81–111). Boston: Blackwell.
Reiser, M., & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, *47*, 85–107.
Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.
Schott, J.R. (1997). *Matrix analysis for statistics*. New York: Wiley.
Teugels, J.L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *Journal of Multivariate Analysis*, *32*, 256–268.
Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.
Tollenaar, N., & Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology*, *56*, 271–288.
van der Linden, W.J., & Hambleton, R.K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
Zhao, Y., & Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*, *33*, 335–356.

Manuscript received 14 FEB 2005
Final version received 11 DEC 2005
Published Online Date: 11 NOV 2006