# An Overview of Limited Information Goodness-of-Fit Testing in Multidimensional Contingency Tables

**Alberto Maydeu-Olivares[1] and Harry Joe[2]**
*(1) Faculty of Psychology, University of Barcelona, P. Valle de Hebrón 171, 08035 - Barcelona, Spain*
*(2) Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2*

**Abstract**

We provide an overview of goodness-of-fit testing in categorical data analysis with applications to item response theory modeling. A promising line of research is the use of limited information statistics. These are quadratic form statistics in marginal residuals such as univariate and bivariate residuals. We describe two approaches to obtain asymptotic p-values for these statistics: (1) matching the asymptotic moments of the statistic with those of a chi-square distribution, (b) using g-inverses. Also, we discuss statistics for piecewise assessment of model fit (i.e., for items, pairs of items, etc.).

## 1. Introduction

Until recently, researchers interested in modeling multivariate categorical data faced the problem that most often no procedure existed to assess goodness-of-fit of the fitted models that yielded trustworthy p-values except for very small models. Fortunately, this situation has recently changed, and it is now possible to reliably assess the fit of multivariate categorical data models. This breakthrough is based on the principle that for goodness-of-fit assessment one should not use all the data available. Rather, by using only a handful of the information at hand (i.e., by using *limited information*) researchers can obtain goodness-of-fit statistics that yield asymptotic p-values that are accurate even in large models and small samples. Furthermore, the power of such statistics can be larger than that of full information statistics (i.e., statistics that use all the data available).

The purpose of this article is to provide an overview of the new developments in limited information goodness-of-fit assessment of categorical data models; see also Bartholomew and Tzamourani (1999), Cai et al. (2006), Mavridis et al. (2007), Maydeu-Olivares and Joe (2005, 2006), and Reiser (in press). Although the exposition focuses on psychometric models (and in particular on item response theory models), the results provided here are completely general and can be applied to any multidimensional categorical data model.

## 2. The Challenge of Testing Goodness-of-Fit in Multivariate Categorical Data Analysis

Consider modeling $N$ independent and identically distributed observations on $n$ discrete random variables whose categories have been labeled $0, 1, \ldots, K-1$. For notational ease we assume that all observed variables consists of the same number of categories $K$. This leads to a $n$-dimensional contingency table with $C = K^n$ cells. However, the theory applies also for variables with different number of categories. We assume a parametric model for $\boldsymbol{\pi}$, the $C$-dimensional vector of cell probabilities, writing $\boldsymbol{\pi}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a $q$-dimensional parameter vector to be estimated from the data. The null and alternative hypotheses are $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$ versus $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$.

pp. 253–262 

The two most commonly used statistics for testing the overall fit of the model 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 and maximum likelihood estimation is used, the two statistics are asymptotically equivalent, $G^2 \stackrel{a}{=} X^2 \stackrel{d}{\to} \chi^2_{C-q-1}$. However, in sparse tables 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 $G^2$.

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. When the cell probabilities become small the discrepancy between the empirical and asymptotic variances of $X^2$ can be large. 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. However, for $C$ and $\boldsymbol{\pi}(\boldsymbol{\theta})$ fixed the accuracy of the the asymptotic p-values for $X^2$ depends also 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 $X^2$ depend on the model and the degree of sparseness of the contingency table.

Three alternative strategies have been proposed to obtain accurate p-values:

(a) *Pooling cells.* If cells are pooled before the model is fitted and if the estimation is based on the $C'$ pooled categories and not the $C$ original categories, then the approximate null distribution of $X^2$ is $\chi^2_{C'-1-q'}$, where $q'$ denotes the number of parameters used after pooling. However, if estimation is based on the original categories, and pooling is based based on the results of the analysis, then the resulting $X^2$ is stochastically larger than $\chi^2_{C'-1-q}$, and hence using a $\chi^2_{C'-1-q}$ reference distribution could give an unduly impression of poor fit, see Joe and Maydeu-Olivares (2007) for details. Most importantly, there is a limit in the amount of pooling that can be performed without distorting the purpose of the analysis.

(b) *Resampling methods.* P-values for goodness-of-fit statistics can be obtained by generating the empirical sampling distribution of goodness-of-fit statistics using a resampling method such as the parametric bootstrap method, see Langeheine et al. (1996), Bartholomew and Tzamourani (1999), and Tollenaar and Mooijaart (2003). However, there is strong evidence that parametric bootstrap procedures do not yield accurate p-values, see Tollenaar and Mooijaart (2003) and Mavridis et al. (2007)). Furthermore, resampling methods may be very time consuming if the researcher is interested in comparing the fit of several models.

(c) *Limited information methods.* Only the information contained in suitable summary statistics of the data, typically the low order marginals of the contingency table, is used to assess the model. This amounts to pooling cells a priori, 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.

## 3. An Overview of Limited Information Methods For Goodness-of-Fit

In this section, we consider methods for testing the overall fit of the model, followed by methods for assessing the source of any misfit.

Before proceeding, notice that one observation of the $i$th variable $Y_i$ has a Multinomial$(1; \pi_{i0}, \ldots, \pi_{i,K-1})$ distribution. Hence, the joint distribution of the random variables is multivariate multinomial (MVM). In the special case where $K = 2$, $Y_i$ has a Bernoulli distribution, and the joint distribution is multivariate Bernoulli (MVB). The MVM can be represented by the $C$ vector of joint probabilities $\boldsymbol{\pi}$, or equivalently by the $C - 1$ vector of marginal probabilities $\dot{\boldsymbol{\pi}}$. The relationship between both representations is one-to-one and can be written as $\dot{\boldsymbol{\pi}} = \mathbf{T}\boldsymbol{\pi}$, where $\mathbf{T}$ is a $(K^n - 1) \times K^n$ matrix of 1s and 0s, of full row rank. $\dot{\boldsymbol{\pi}}$ can be partitioned as $\dot{\boldsymbol{\pi}}' = (\dot{\boldsymbol{\pi}}_1', \dot{\boldsymbol{\pi}}_2', \ldots, \dot{\boldsymbol{\pi}}_n')'$, where $\dot{\boldsymbol{\pi}}_r$ is the $s_i = \binom{n}{r}(K-1)^r$ vector of $r$th way marginal probabilities, such that the marginal probabilities involving category 0 are excluded. Also, we write $\boldsymbol{\pi}_r = (\dot{\boldsymbol{\pi}}_1', \ldots, \dot{\boldsymbol{\pi}}_r')'$ for the $s = \sum_{i=1}^{r} s_i$ vector of multivariate marginal probabilities up to order $r$ ($r \leq n$). Now, let $\mathbf{p}$ and $\mathbf{p}_r$ be the vector of cell proportions, and the vector of marginal proportions up to order $r$, respectively. Also, let $\mathbf{e} = \mathbf{p} - \boldsymbol{\pi}$ and $\mathbf{e}_r = \mathbf{p}_r - \boldsymbol{\pi}_r$ be respectively the vector of cell residuals and marginal residuals. Finally, we use $\hat{\mathbf{e}}$ and $\hat{\mathbf{e}}_r$ when these residual vectors depend on the estimated parameters.

To give a completely general result, we only assume that $\hat{\boldsymbol{\theta}}$ is a $\sqrt{N}$-consistent and asymptotically normal estimator. Specifically, we assume that $\hat{\boldsymbol{\theta}}$ satisfies

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

for some $q \times C$ matrix $\mathbf{H}$. This includes minimum variance or best asymptotic normal (BAN) estimators such as the maximum likelihood estimator (MLE) or the minimum chi-square estimator. It also includes the limited information estimators for IRT models: those implemented in programs such as LISREL, EQS, MPLUS, or NOHARM, and those proposed by Christoffersson (1975) and Jöreskog and Moustaki (2001).

We have the following results for the cell residuals: $\sqrt{N}\,\mathbf{e} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma})$, and $\sqrt{N}\,\hat{\mathbf{e}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Gamma} = \boldsymbol{\Delta} - \boldsymbol{\pi}\boldsymbol{\pi}'$, and $\boldsymbol{\Sigma} = (\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})\boldsymbol{\Gamma}(\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})'$. Here, $\mathbf{D} = \mathrm{diag}(\boldsymbol{\pi})$, and $\boldsymbol{\Delta} = \partial\boldsymbol{\pi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$, which is assumed to be of full rank so that the model is identified when using full information. 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.

For the marginal residuals up to order $r$,

$$\sqrt{N}\,\mathbf{e}_r \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Xi}_r) \quad \text{and} \quad \sqrt{N}\,\hat{\mathbf{e}}_r \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_r),$$

where

$$\boldsymbol{\Xi}_r = \mathbf{T}_r\boldsymbol{\Gamma}\mathbf{T}_r',$$

$$\boldsymbol{\Sigma}_r = \mathbf{T}_r\boldsymbol{\Sigma}\mathbf{T}_r' = \boldsymbol{\Xi}_r - \boldsymbol{\Delta}_r\mathbf{H}\boldsymbol{\Gamma}\mathbf{T}_r' - \mathbf{T}_r\boldsymbol{\Gamma}\mathbf{H}'\boldsymbol{\Delta}_r' + \boldsymbol{\Delta}_r[\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}']\boldsymbol{\Delta}_r'. \tag{2}$$

In equation (2), $\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}'$ 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 $s \times q$ matrix, where $s$ denotes the dimension of the vector of residuals considered. In the special case of BAN estimators such as the MLE, we have $\boldsymbol{\Sigma} = \boldsymbol{\Gamma} - \boldsymbol{\Delta}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}'$, and $\boldsymbol{\Sigma}_r = \boldsymbol{\Xi}_r - \boldsymbol{\Delta}_r\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}_r'$, respectively.

*3.1. Testing the overall fit of the model*

Two general strategies have been proposed to obtain goodness-of-fit statistics using limited information. Both are based on quadratic forms in marginal residuals. Suppose testing is to be performed using $\hat{\mathbf{e}}_r$. We write

$$T_r = N\hat{\mathbf{e}}_r'\widehat{\mathbf{W}}\hat{\mathbf{e}}_r, \tag{3}$$

where $\widehat{\mathbf{W}}$ converges in probability to an $s \times s$ weight matrix $\mathbf{W}$.

The first strategy consists in choosing $\widehat{\mathbf{W}}$ so that the quadratic form is easily computed. Two obvious choices are (a) $\widehat{\mathbf{W}} = \mathbf{I}$, leading to $U_r = N\hat{\mathbf{e}}_r'\hat{\mathbf{e}}_r$, and (b) $\widehat{\mathbf{W}} = (\text{diag}(\widehat{\boldsymbol{\Xi}}_r))^{-1} = (\text{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}')^{-1}$, leading to $D_r = N\hat{\mathbf{e}}_r'(\text{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}')^{-1}\hat{\mathbf{e}}_r$.

Quite generally, the asymptotic distribution of $T_r$ is a mixture of independent chi-square variates. P-values for $T_r$ are then obtained by matching the moments of $T_r$ with those of a central chi-square distribution. One, two, or three moments can be matched. The first three asymptotic moments (mean, variance and third central moment) of $T_r$ are: $\mu_1(T_r) = \text{tr}\,(\mathbf{W}\boldsymbol{\Sigma}_r)$, $\mu_2(T_r) = 2\text{tr}\,(\mathbf{W}\boldsymbol{\Sigma}_r)^2$, and $\mu_3(T_r) = 3\text{tr}\,(\mathbf{W}\boldsymbol{\Sigma}_r)^3$. Let $A_\nu$ be a random variable with $\chi^2_\nu$ distribution. To obtain a p-value using a two-moment adjustment, we assume that $T_r$ can be approximated by $bA_c$. Solving for the two unknown constants $b$ and $c$ using the first two asymptotic moments of $T_r$ yields $b = \mu_2(T_r)/(2\mu_1(T_r))$, $c = \mu_1(T_r)/b$. For the three-moment adjustment, we assume that $T_r$ can be approximated by $a + bA_c$. Solving for the three unknown constants $a$, $b$, and $c$ using the first three asymptotic moments of $T_r$ yields $b = \mu_3(T_r)/(4\mu_2(T_r))$, $c = \mu_2(T_r)/(2b^2)$, and $a = \mu_1(T_r) - bc$. A p-value for the two moment adjusted statistic is obtained using $\Pr(A_c > T_r/b)$, and for the three moment adjusted statistic using $\Pr(A_c > (T_r - a)/b)$. For the one-moment approximation, we assume again that $T_r$ can be approximated by $bA_t$, where $t$ is the number of degrees of freedom available for testing. Heuristically, this can be taken to be $t = s - q$. Solving for $b$, we have $b = \mu_1(T_r)/t$, and the p-value for the first moment adjusted statistic is given by $\Pr(A_t > T_r/b)$.

Many different limited information statistics can be constructed in this way depending on the choice of (a) marginal residual in the quadratic form, (b) weight matrix, and (c) number of moments used to approximate the central chi-square distribution. Regarding (a), a typical choice is $\hat{\mathbf{e}}_2$, the set of univariate and bivariate residuals that do not include category 0. This is a vector of dimension $s = n(K-1) + \binom{n}{2}(K-1)^2$. Another choice is the set of all bivariate residuals $\tilde{\mathbf{e}}_2 = \tilde{\mathbf{p}}_2 - \tilde{\boldsymbol{\pi}}_2(\hat{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\pi}}_2$ is a $\binom{n}{2}K^2$ vector with elements $\pi^{(ij)}_{k_1k_2} = \Pr(Y_i = k_1, Y_j = k_2)$ and sample counterparts $p^{(ij)}_{k_1k_2}$. An statistic based on $\tilde{\mathbf{e}}_2$ is

$$X_2 = N\tilde{\mathbf{e}}_2'\left(\text{diag}\left(\tilde{\boldsymbol{\pi}}_2\right)\right)^{-1}\tilde{\mathbf{e}}_2 = N\sum_{i<j}\sum_{k_1}\sum_{k_2}\frac{\left(p^{(ij)}_{k_1k_2} - \pi^{(ij)}_{k_1k_2}(\hat{\boldsymbol{\theta}})\right)^2}{\pi^{(ij)}_{k_1k_2}},$$

the sum of all $\binom{n}{2}$ $X^2$ bivariate statistics.

The fact that $\boldsymbol{\Sigma}_r$, the asymptotic covariance matrix of the estimated marginal residuals, needs to be estimated in this approach results in some drawbacks. First, the estimation of $\boldsymbol{\Sigma}_r$ can be computationally involved for some estimators, such as the MLE, when the model is large. Another drawback is that a different implementation is needed for each estimator under consideration, as $\boldsymbol{\Sigma}_r$ depends on $\mathbf{H}$, which depends on the estimator chosen. Thus, (a) formulae for moment adjusted statistics for testing IRT models for binary data estimated using the MLE were given by

Cai et al. (2006); see also Bartholomew and Leung (2002), (b) formulae for testing IRT models estimated sequentially using tetrachoric/polychoric correlations were given by Maydeu-Olivares (2001a) for the binary case, and Maydeu-Olivares (2006) for the polytomous case, and (c) formulae for testing IRT models for binary data estimated using the NOHARM program were given by Maydeu-Olivares (2001a). Maydeu-Olivares (2001a, 2001b, 2006) considered one- and two-moment approximations to $U_2$. Cai et al. (2006) considered one- to three-moment approximations to $D_2$, and also to the analogous statistic based only on bivariate residuals.

In any case, available evidence on the use of this approach suggests that (a) all in all this approach gives accurate p-values, except when only one moment is matched, in which case the approximation is generally poor, (b) there is little to choose from $U_2$ and $D_2$, and that (c) statistics based on univariate and bivariate residuals are slightly more powerful than statistics based only on bivariate residuals.

The second strategy consists of choosing $\widehat{\mathbf{W}}$ so that the resulting quadratic form is asymptotically chi-square. This is the strategy followed by Reiser (1996) and Maydeu-Olivares and Joe (2005, 2006). Choosing $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Sigma}}_r^-$ ensures that $T_r \xrightarrow{d} \chi_t^2$, where $t$ equals to the rank of $\boldsymbol{\Sigma}_r$. A g-inverse (or alternatively a Moore-Penrose inverse $\widehat{\boldsymbol{\Sigma}}_r^+$) needs to be employed because $\boldsymbol{\Sigma}_r$ is almost invariably of deficient rank. Reiser (1996) considered a quadratic form in $\hat{\mathbf{e}}_2$ with $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Sigma}}_2^+$ for testing models for binary data.

The use of $\widehat{\boldsymbol{\Sigma}}_r^+$ as a weight matrix has two drawbacks. The first drawback is that, as was the case with the moment-adjustment strategy, $\boldsymbol{\Sigma}_r$ needs to be estimated. The second drawback stems from the fact that almost invariably, the rank of $\boldsymbol{\Sigma}_r$ can not be determined a priori. In that case, one can determine $t$, the degrees of freedom, by inspecting the magnitude of the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_r$. However, this may be tricky, as this matrix often has some small eigenvalues, and $t$ (and the value of the statistic itself) will depend on which eigenvalues are judged to be zero.

To overcome these difficulties, Maydeu-Olivares and Joe (2005, 2006) considered using instead a weight matrix $\widehat{\mathbf{W}}$ such that $\boldsymbol{\Sigma}_r$ is a g-inverse of $\mathbf{W}$, that is, $\mathbf{W} = \mathbf{W}\boldsymbol{\Sigma}_r\mathbf{W}$. More specifically, they proposed using

$$\mathbf{W} = \boldsymbol{\Xi}_r^{-1} - \boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r\left(\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r\right)^{-1}\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1} = \boldsymbol{\Delta}_r^{(c)}\left(\boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r^{(c)}\right)^{-1}\boldsymbol{\Delta}_r^{(c)\prime}, \quad (4)$$

evaluated at $\hat{\boldsymbol{\theta}}$, as the weight matrix in equation (3). Here, $\boldsymbol{\Delta}_r^{(c)}$ is the $s \times (s-q)$ orthogonal complement of $\boldsymbol{\Delta}_r = \mathbf{T}_r\boldsymbol{\Delta}$ (i.e, it satisfies $\boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Delta}_r = \mathbf{0}$).

One advantage of using this weight matrix is that it does not require an estimate of $\boldsymbol{\Sigma}_r$, but of the more easily computable $\boldsymbol{\Xi}_r$. Another advantage is that by construction, if the model is identified from the marginal probabilities up to order $r$, the degrees of freedom $t$ can be determined a priori: $T_r \xrightarrow{d} \chi_{s-q}^2$. Yet, another advantage is that the result holds for any estimator (1), and hence, a single implementation suits all estimators. Maydeu-Olivares and Joe (2005, 2006) considered the full class of statistics with (4) (referred to as $M_r$ statistics) and they showed that Pearson's $X^2$ is a special case of the family when the MLE is used and all marginal residuals are used.

### *3.2. Assessing the source of the misfit*

Limited information methods are also useful to identify the source of misfit in poorly fitting models. The inspection of standardized 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 standardized residuals will be either very small or very large.

Yet, dividing a marginal residual by its asymptotic standard error we obtain a standardized marginal residual that is asymptotically standard normal. To identify the source of the misfit, these residuals (univariate, bivariate, or trivariate) can be inspected. 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. For polytomous data models, a more fruitful avenue is to assess how well the model fits single variables, variable pairs, etc. (i.e., subtables). Note that this is like multiple testing after a jointly significant result.

If the model for an $r$-variate subtable is identified (with $t' > 0$ degrees of freedom), Maydeu-Olivares and Joe's $M_r$ statistic can be used to assess the fit to the subtable, where $M_r \xrightarrow{d} \chi^2_{t'}$. In contrast, when applied to an identified subtable, the asymptotic distribution of $X^2$ is stochastically larger than $\chi^2_{t'}$, because the parameters in the subtable have been estimated using the full table, see Maydeu-Olivares and Joe (2006).

## 4. Numerical Examples

To illustrate the discussion we consider two numerical examples. The first one is the well-known LSAT 7 dataset, see Bock and Lieberman (1970). It consists of 1000 observations on five binary variables; thus, $C = 2^5 = 32$. A two parameter logistic IRT model is fitted to this data. The second dataset consists of 551 young women responding to the five items of the Positive Problem Orientation (PPO) scale of the Social Problem Solving Inventory-Revised, see D'Zurilla et al. (2002). These Likert-type items consist of five categories. For this analysis the two lowest and the two highest categories were merged; thus $C = 3^5 = 243$. Samejima's (1969) graded model is fitted to these data. Maximum likelihood estimation was used in both examples.

### *4.1. LSAT 7 data*

Table 1 provides the results obtained with $X^2$ and $G^2$. Because the data are not sparse, both statistics yield similar results. The model can not be rejected at the 5% significance level. We have also included in this table the results obtained with three limited information test statistics: $M_2$, $M_3$, and $D_2$. Univariate and bivariate residuals are used in $M_2$ and $D_2$. Up to trivariate residuals are used in $M_3$. Also, from Maydeu-Olivares and Joe (2005), $X^2 = M_5$ because ML estimation was used.

One-, two-, and three-moment adjustments where used to obtain p-values for $D_2$. They are labeled $D_2^{(1)}$, $D_2^{(2)}$, and $D_2^{(3)}$ in Table 1. We see in this table that, for this example, the same approximate p-values for $D_2$ are obtained regardless of the number of moments used. Even a one-moment adjustment gives good results.

We also see in Table 1 that when data are not sparse, limited information statistics yield similar p-values than full information statistics. However, they do so at the expense of fewer degrees of freedom. It is interesting to compare the statistic/df ratios for the members of the $M$ family of statistics. These ratios are 2.39, 1.77 and 1.55 for $M_2$, $M_3$ and $M_5$, respectively. Joe and Maydeu-Olivares (2007) have theory that relate larger ratios with smaller degrees of freedom to test statistics that have more power for reasonable directional alternatives.

We now consider using $R_2 = \hat{\mathbf{e}}_2' \widehat{\boldsymbol{\Sigma}}_r^{+} \hat{\mathbf{e}}_2$, as in Reiser (1996). There are 15 univariate and bivariate residuals in $\hat{\mathbf{e}}_2$. Table 2 provides the value of the 6 smallest

**Table 1** Goodness-of-Fit Results for the LSAT7 Data.

| stat | value | df | p-value | stat | value | df | p-value |
|---|---|---|---|---|---|---|---|
| $X^2$ | 32.48 | 21 | 0.052 | $D_2^{(1)}$ | 11.33 | 5 | 0.045 |
| $G^2$ | 31.70 | 21 | 0.063 | $D_2^{(2)}$ | 11.36 | 5.0 | 0.045 |
| $M_2$ | 11.94 | 5 | 0.036 | $D_2^{(3)}$ | 10.90 | 4.7 | 0.045 |
| $M_3$ | 26.48 | 15 | 0.033 | | | | |

eigenvalues of $\mathbf{\Sigma}_r$, the value of $R_2$ for $j = 1, \ldots, 6$, if the $j$th eigenvalue and those smaller are judged to be zero, and the resulting *df* and p-values. The results illustrate how the p-value is affected by how many eigenvalues are judged to be zero. Also, notice that a larger range of p-values are obtained than when moment corrections for $D_2$ are used. Nevertheless, simulation results by Mavridis et al. (2007) reveal this statistic also works adequately. Also, determining the degrees of freedom in Reiser's approach is more numerically stable when fitting models that do not require numerical integration to obtain probabilities (such as loglinear models).

**Table 2** Range of P-values Obtainable for $R_2$

| eigenvalue | $R_2$ | *df* | p-value |
|---|---|---|---|
| $2.22 \times 10^{-5}$ | 17.82 | 9 | 0.037 |
| $3.83 \times 10^{-6}$ | 18.29 | 10 | 0.050 |
| $3.95 \times 10^{-9}$ | 19.42 | 11 | 0.069 |
| $2.90 \times 10^{-11}$ | 19.78 | 12 | 0.079 |
| $1.03 \times 10^{-13}$ | 19.78 | 13 | 0.101 |
| $-1.02 \times 10^{-15}$ | 19.78 | 14 | 0.137 |

Finally, consider obtaining a better fitting model by dropping one item. The standardized cell residuals are not very helpful to this end. There are only two standardized cell residuals significant at a 5% level, those for patterns $(0, 1, 0, 0, 0)$ and $(1, 0, 0, 0, 0)$. The inspection of univariate, bivariate and trivariate residuals is more helpful. The significant standardized residuals for up to trivariate margins are for margins $(1, 3)$, $(1, 4)$, $(1.5)$, and $(2, 3)$. They indicate that item 1 is the best selection if an item is to be dropped to fit the model. This is indeed the case, as shown in Table 3.

**Table 3** $X^2$ Obtained When Dropping One Item at a Time (df = 7)

| item dropped | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $X^2$ | 5.01 | 9.52 | 8.59 | 18.68 | 9.86 |
| p-value | 0.66 | 0.22 | 0.28 | 0.01 | 0.20 |

### *4.2. PPO data*

With polytomous data, the contingency table often becomes sparse and $X^2$ and $G^2$ sometimes yield conflicting results, indicating that both p-values are incorrect. In those cases, $G^2$ gives an overly optimistic p-value (often 1), and $X^2$ generally

gives a p-value of 0. Because in this example the data are not sparse, the discrepancy between both statistics is not large, as shown in Table 4. The table also includes the results obtained with $M_2$, and the results of using $D_2$ and $X_2$ (using 1- to 3-moment adjustments to obtain their approximate p-values). In this case, the p-values for $D_2$ appear inflated, but those for $X_2$ appear reasonable.

**Table 4** Goodness-of-Fit Results for the PPO Data.

| stat | value | df | p-value |
|---|---|---|---|
| $X^2$ | 304.89 | 227 | 0.0004 |
| $G^2$ | 271.09 | 227 | 0.024 |
| $M_2$ | 55.01 | 35 | 0.017 |
| $D_2^{(1)}$ | 42.58 | 35 | 0.177 |
| $D_2^{(2)}$ | 22.21 | 18.3 | 0.236 |
| $D_2^{(3)}$ | 14.62 | 11.5 | 0.230 |
| $X_2^{(1)}$ | 52.13 | 35 | 0.031 |
| $X_2^{(2)}$ | 45.19 | 30.3 | 0.041 |
| $X_2^{(3)}$ | 34.11 | 21.6 | 0.042 |

We also applied $R_2$ to these data. For this example, $\widehat{\mathbf{\Sigma}}_2$ is of full rank (its smallest eigenvalue is $8.42 \times 10^{-5}$), yielding a statistic of 66.65 on 50 *df*, $p = 0.058$. Thus, for this example, based on the same residuals, this statistic has 15 more *df* than $M_2$ and yields a slightly larger p-value.

**Table 5** Bivariate $X^2$ Statistics (Above the Diagonal) and $M_2$ Statistics (Below the Diagonal)

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | −− | 5.92 | 4.89 | 6.97* | 3.99 |
| 2 | 0.69 | −− | 4.99 | 7.47* | 4.53 |
| 3 | 3.36 | 2.94 | −− | 0.41 | 7.89* |
| 4 | 2.70 | 2.29 | 0.15 | −− | 2.97 |
| 5 | 1.90 | 3.45 | 4.34 | 1.57 | −− |

We next consider obtaining a better fitting model by dropping one item. To do so, we shall assess how well the model fits different subtables. The degrees of freedom for testing the fitted model one item at a time is negative. Thus, item level testing is not possible for this model. There are 2 *df* for testing the model for pairs of items, and the model for the subtable is identified. Table 5 provides the $M_2$ statistics for each pair of variables, and for comparison the $X^2$ statistics. Starred statistics are significant at $\alpha = 0.05$, based on a $\chi^2_2$.

Notice that the values of the $X^2$ statistics are larger than the values of the $M_2$ statistics. Also, they yield a misleading impression of poor fit because the asymptotic distribution of $X^2$ is stochastically larger than $\chi^2_2$. The inspection of the values of $X^2$ suggests the model misfits at the bivariate level, and it suggests that fit would improve the most by dropping item 4. In contrast, the inspection of the $M_2$ values does not reveal any model misfit at the bivariate level. These statistics are not useful to locate the source of the misfit. In this case, we can inspect the value of $M_3$ for triplets of items. With 17 degrees of freedom, there are two $M_3$ statistics significant at $\alpha = 0.05$; for triplets $(2, 3, 5)$ and $(2, 4, 5)$, which

suggests that either item 2 or 5 should be dropped. Table 6 reveals that item 2 should to be dropped to get a better fit.

**Table 6** $X^2$ Obtained When Dropping One Item at a Time (df = 68)

| item dropped | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $X^2$ | 82.92 | 66.61 | 86.97 | 106.42 | 89.17 |
| p-value | 0.105 | 0.525 | 0.060 | 0.002 | 0.044 |

## 5. Discussion

Limited information statistics appear to be a promising avenue to overcome the decades long problem of assessing the goodness-of-fit in multivariate categorical data analysis. Using these statistics, it is possible to obtain p-values for extremely large models. Extant results suggest their asymptotic p-values yield accurate results even in samples of 300 observations. However, so far the behavior of these statistics has only been investigated for a handful of models containing up to 20 variables. More research is needed to investigate the behavior of the statistics in the extremely large models that are common in Social Science applications, and for alternative models.

A critical limitation of these methods is that the model must be identified from the statistics used for testing. This needs to be verified numerically application by application and for the statistics that assess the overall fit and also the statistics used for subtables. It may be that the model is identified from the statistic of interest but that in a given application it is nearly non-identified. In that case, the statistic will become numerically unstable and yield unreliable results.

Here, we have focused on limited information statistics based on low order marginal residuals. A limitation of these statistics is that if testing is performed using up to say $r$-variate information, then they have no power to detect model misfit if it is only present in $r+1$ and higher associations. In our view, this is not a serious limitation. A researcher interested in detecting a specific model misfit can construct a limited information statistic (not necessarily based on residual moments) so that it achieves good power with respect to the specific alternatives of interest (Joe and Maydeu-Olivares, 2007), and the resulting statistic is more powerful than full information statistics.

## Acknowledgments

This research has been supported by grant SEJ2006-08204/PSIC of the Spanish Ministry of Science and Technology, and an NSERC Canada grant.

## References

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.

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

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 athematical and Statistical Psychology,59*, 173–194.

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

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, Inc.

Joe, H. & Maydeu-Olivares (2007). Constructing chi-square goodness-of-fit tests for multinomial data that are more powerful than Pearson's $X^2$. Under review.

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

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

Mavridis, D., Moustaki, I., & Knott, M. (2007). Goodness-of-fit measures for latent variable models for binary data. In S.-Y. Lee (Ed.). *Handbook of Latent Variable and Related Models. (pp. 135–162*.

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. (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.

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

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

Reiser, M. (in press). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology*.

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

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.