British Journal of Mathematical and Statistical Psychology (2006), 59, 173–194

The British Psychological Society
www.bpsjournals.co.uk

Limited-information goodness-of-fit testing of item response theory models for sparse 2^P tables

Li Cai[1]*, Albert Maydeu-Olivares[2], Donna L. Coffman[1] and David Thissen[1]

[1]University of North Carolina, Chapel Hill, USA
[2]University of Barcelona and Instituto de Empresa, Spain

Bartholomew and Leung proposed a limited-information goodness-of-fit test statistic (Y) for models fitted to sparse 2^P contingency tables. The null distribution of Y was approximated using a chi-squared distribution by matching moments. The moments were derived under the assumption that the model parameters were known in advance and it was conjectured that the approximation would also be appropriate when the parameters were to be estimated. Using maximum likelihood estimation of the two-parameter logistic item response theory model, we show that the effect of parameter estimation on the distribution of Y is too large to be ignored. Consequently, we derive the asymptotic moments of Y for maximum likelihood estimation. We show using a simulation study that when the null distribution of Y is approximated using moments that take into account the effect of estimation, Y becomes a very useful statistic to assess the overall goodness of fit of models fitted to sparse 2^P tables.

1. Introduction

It is common in social science research to encounter surveys, personality inventories, or educational tests consisting of P dichotomously scored items. The responses of a sample of individuals to items can be described by a contingency table consisting of $C = 2^P$ cells. When modelling such contingency tables one assumes that the cell probabilities depend on a set of q parameters whose values are either fixed in advance (simple hypothesis) or estimated from the data (composite hypothesis). Because most often the parameters are estimated when modelling 2^P tables, in this paper we focus on goodness-of-fit testing under composite null hypotheses.

Pearson's X^2 and the likelihood ratio G^2 are arguably the two most widely used statistics in contingency table analysis. When the model parameters are estimated using maximum likelihood (ML), these two statistics are asymptotically equivalent and their

*Correspondence should be addressed to Li Cai, Department of Psychology, UNC-CH, CB #3270, Chapel Hill, NC 27599-3270, USA (e-mail: cai@unc.edu).

DOI:10.1348/000711005X66419

asymptotic distribution is chi-squared with $C - q - 1$ degrees of freedom. However, it is well known (see Cochran, 1952) that when some cell probabilities are small the type I error rates of goodness-of-fit tests using X^2 or G^2 do not match their expected rates under their reference asymptotic distribution. Since the number of items in typical psychometric applications is large, the number of cells in the resulting contingency tables is even larger (often several million) while the number of respondents is usually in the hundreds. As a result, most of the cells have very small probabilities (Bartholomew & Tzamourani, 1999) in such large tables. Also, cell proportions are very poorly estimated (most cell proportions are zero). In sum, goodness-of-fit testing in such sparse tables poses a serious challenge to psychometricians.

To overcome the limitations of X^2 and G^2, three remedies have been proposed. One proposal is to pool cells so that the cell probabilities of the resulting table are large (see Bartholomew & Tzamourani, 1999, for an excellent discussion). The second proposal is to use resampling methods such as the parametric bootstrap to obtain an empirical p-value for X^2 and G^2 (Bartholomew & Knott, 1999; Bartholomew & Tzamourani, 1999; Collins, Fidler, Wugalter, & Long, 1993; Langeheine, Pannekoek, & van de Pol, 1996; Tollenaar & Mooijaart, 2003). The third proposal is to use limited-information statistics (Bartholomew & Leung, 2002; Maydeu-Olivares, 1997, 2001a; Reiser, 1996; Reiser & Lin, 1996; Reiser & VandenBerg, 1994).

Pooling cells after the model has been fitted often results in statistics with an unknown sampling distribution, as the procedure is data-dependent. It may also lead to a gross loss of information about model misfit. The use of resampling methods, on the other hand, has become increasingly popular given the power of today's computers. However, to obtain a stable p-value for any goodness-of-fit test, several hundred resamples are needed. In addition, if the researcher is interested in comparing the fit of different models, the resampling procedure must be repeated for each model. In sum, resampling methods are not very practical computationally. Furthermore, Tollenaar and Mooijaart (2003) showed that the validity of a bootstrap-based test depends critically upon what statistic is being bootstrapped. In particular, bootstrapping X^2 or G^2 does not provide immediate Type I error rate control under sparseness.

As an alternative to statistics such as X^2 and G^2 which use all the information contained in the contingency table (i.e. full information), several researchers in psychometrics have proposed limited-information statistics based on the lower-order margins of the contingency table (see Bartholomew & Leung, 2002, and the references therein). There is evidence that when the table is large and sparseness severe, limited-information tests can be superior to tests based on the full cross-classifications (Agresti, Lipsitz, & Lang, 1992; Agresti & Yang, 1987; Maydeu-Olivares & Joe, in press).

Limited-information methods have a long tradition in psychometrics. The classical solutions to the factor-analytic model involving dichotomous indicators (Christofferson, 1975; Muthén, 1978) use limited-information methods that yield a class of consistent and asymptotically normal estimators of the model parameters using only the first- and second-order margins. These procedures yield goodness-of-fit tests that also use limited information.

In this paper, we are primarily interested in the combination of full-information estimation (Bock & Aitkin, 1981; Bock & Lieberman, 1970) and limited-information goodness-of-fit testing. This 'hybrid' approach is similar to that found in Reiser (1996).

Bartholomew and Leung (2002) proposed a limited-information goodness-of-fit statistic, called Y in their paper, with two attractive properties. First, it is computationally simpler than other statistics suggested in the literature (such as

Reiser, 1996). Second, after conducting an overall goodness-of-fit test, it is easy to 'decompose' Y into simple additive pieces to assess the contributions of individual margins to the misfit of the model. To obtain p-values for Y, Bartholomew and Leung (2002) approximated its distribution under the simple null hypothesis by matching its exact moments to the moments of a linear transformation of a central chi-squared variable. They also showed that the exact moments of their statistic could be well approximated by asymptotic moments that are much easier to compute in practice. However, we are most often interested in composite null hypotheses where the parameters are estimated from the data. Bartholomew and Leung (2002) suggested that moment adjustments based on simple null hypotheses could also be used when testing composite null hypotheses. In this paper we investigate the usefulness of the asymptotic moment adjustments for Y proposed by Bartholomew and Leung (2002) in testing composite null hypotheses when the model parameters have been estimated by ML.

The remainder of the paper is organized as follows. In Section 2 we present a characterization of the multivariate Bernoulli (MVB) distribution using its moments. This characterization is very useful for introducing limited-information goodness-of-fit statistics. In Section 3 we discuss limited- and full-information goodness-of-fit tests for simple null hypotheses within an MVB framework. Within this framework, Y is simply a quadratic form in residual bivariate MVB moments. In Section 4 we consider the use of Y for testing composite null hypotheses. We approximate the distribution of Y with the same moment-based adjustments as in Bartholomew and Leung (2002), but taking into account that the parameters have been estimated by ML. In Section 5 we employ the two-parameter logistic (2PL) item response theory (IRT) model in a set of simulations to empirically investigate the type I error rates and power of Y with various moment adjustments. We use the moments of Y computed as in Bartholomew and Leung (2002), assuming that the parameters are fixed, and we also use the moments of Y derived in this paper for the maximum likelihood estimator (MLE). We shall see that ignoring the fact that the parameters are estimated has an adverse effect on the behaviour of Y. Finally, we analyse a real data set to illustrate our discussion.

2. A multivariate Bernoulli framework

Throughout this paper we consider a test consisting of P dichotomously scored items, administered to a sample of N examinees. Without loss of generality, we may assign a score of 1 to the 'correct' response to an item, and 0 otherwise, so that each variable is Bernoulli, and their joint distribution is MVB[1] (see Teugels, 1990, for details).

2.1. The MVB distribution

Consider one of the 2^P response patterns – a random P-vector $\mathbf{U} = (U_1, \ldots, U_i, \ldots, U_P)'$ of Bernoulli random variables – and let $\mathbf{u} = (u_1, \ldots, u_i, \ldots, u_P)'$, $u_i \in \{0, 1\}$, be a realization of $\mathbf{U}$. We write the joint distribution of this MVB variable as[2]

$$\pi_{\mathbf{u}} = P(U_i = u_i), \tag{1}$$

[1] *The notation for the MVB characterization used here follows Maydeu-Olivares (1997).*
[2] *To differentiate scalars from vectors or matrices, we use bold lower-case letters to indicate a vector, and bold capital letters to indicate a matrix.*

for $i = 1, \ldots, P$. Without loss of generality, we order the elements in $\boldsymbol{\pi}$ by the number of items correct (also known as the summed-score group in IRT terminology), $\mathbf{1}'\mathbf{u} = 0,\ 1, \ldots,\ P$, where $\mathbf{1}$ is a $P \times 1$ summing vector. Within each summed-score group, the individual joint probabilities are sorted by the (inverse) lexicographical ordering of the response patterns in $\mathbf{u}$. For example, for $P = 3$, $\boldsymbol{\pi} = (\pi_{(0,0,0)},\ \pi_{(1,0,0)},\ \pi_{(0,1,0)},\ \pi_{(0,0,1)},\ \pi_{(1,1,0)},\ \pi_{(1,0,1)},\ \pi_{(0,1,1)},\ \pi_{(1,1,1)})'$.

The MVB distribution can be characterized by the set of 2^P probabilities $\boldsymbol{\pi}$. In the psychometrics literature, rather than using an MVB sampling scheme, it is customary to consider a multinomial sampling scheme for the observed frequencies of the response patterns. Specifically, let $\mathbf{N}$ represent a random 2^P-vector of cell frequencies associated with the response patterns, and $\mathbf{n}$ be a realization of $\mathbf{N}$; the distribution of $\mathbf{N}$ is then multinomial with

$$\boldsymbol{\pi}(\mathbf{N} = \mathbf{n}) = N! \prod_{\mathbf{u}} \frac{[\pi_{\mathbf{u}}]^{n_{\mathbf{u}}}}{n_{\mathbf{u}}!}, \tag{2}$$

where the subscript $\mathbf{u}$ indicates that the product is taken over all of the 2^P patterns, and $n_{\mathbf{u}}$ refers to the corresponding element in $\mathbf{n}$.

In this paper we use an MVB framework rather than a multinomial framework as the former is more amenable to presenting limited-information methods. This is because the MVB distribution can be equivalently characterized by its joint moments. This characterization is discussed next.

2.2. *A characterization by moments of the MVB distribution*

Consider the $(2^P - 1)$-vector $\dot{\boldsymbol{\pi}}$ of joint moments of the MVB distribution. For convenience, $\dot{\boldsymbol{\pi}}$ can be written in the partitioned form $\dot{\boldsymbol{\pi}} = (\dot{\boldsymbol{\pi}}_1', \dot{\boldsymbol{\pi}}_2', \ldots, \dot{\boldsymbol{\pi}}_r', \ldots, \dot{\boldsymbol{\pi}}_P')'$, where the dimension of the vector $\dot{\boldsymbol{\pi}}_r$ is $\binom{P}{r}$. The vector $\dot{\boldsymbol{\pi}}_1 = (\dot{\pi}_1, \ldots, \dot{\pi}_i, \ldots, \dot{\pi}_P)'$ contains P univariate (first-order marginal) moments, where $\dot{\pi}_i = E(U_i) = P(U_i = 1) = \pi_i$. The $P(P-1)/2$-dimensional vector $\dot{\boldsymbol{\pi}}_2$ contains bivariate (second-order marginal) moments, $\dot{\pi}_{ij} = E(U_i U_j) = P(U_i = 1, U_j = 1) = \pi_{ij}$, for all distinct integers i and j satisfying $1 \le i < j \le P$. The joint moments are defined in this way up to the Pth order, with the last one, $\dot{\boldsymbol{\pi}}_P = E(U_1 \cdots U_P) = P(U_1 = \cdots = U_P = 1)$, having a dimension of $\binom{P}{P} = 1$.

When $\boldsymbol{\pi}$ is sorted according to the descriptions given in Section 2.1, there exists a $(2^P - 1) \times 2^P$ upper triangular matrix $\mathbf{M}$ of full row rank such that $\dot{\boldsymbol{\pi}} = \mathbf{M}\boldsymbol{\pi}$. We show an example of this mapping for $P = 3$:

$$\begin{pmatrix} \dot{\boldsymbol{\pi}}_1 \\ \dot{\boldsymbol{\pi}}_2 \\ \dot{\boldsymbol{\pi}}_3 \end{pmatrix} = \begin{pmatrix} \dot{\pi}_1 \\ \dot{\pi}_2 \\ \dot{\pi}_3 \\ \hline \dot{\pi}_{12} \\ \dot{\pi}_{13} \\ \dot{\pi}_{23} \\ \hline \dot{\pi}_{123} \end{pmatrix} = \left(\begin{array}{cccccccc} 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \begin{pmatrix} \pi_{(0,0,0)} \\ \pi_{(1,0,0)} \\ \pi_{(0,1,0)} \\ \pi_{(0,0,1)} \\ \pi_{(1,1,0)} \\ \pi_{(1,0,1)} \\ \pi_{(0,1,1)} \\ \pi_{(1,1,1)} \end{pmatrix}.$$

$\mathbf{M}$ resembles a design matrix, as it consists only of 0s and 1s. The first column of $\mathbf{M}$ is always a zero vector, so we can partition $\mathbf{M}$ as $(\mathbf{0}\ \dot{\mathbf{M}})$. Note that $\dot{\mathbf{M}}$ is a full-rank square matrix.

Hence, its inverse exists. Also notice that $\dot{\boldsymbol{\pi}} = \dot{\mathbf{M}}\breve{\boldsymbol{\pi}}$, with $\boldsymbol{\pi} = \begin{pmatrix} \pi_{(0\cdots0)} \\ \breve{\boldsymbol{\pi}} \end{pmatrix}$, where $\pi_{(0\cdots0)} = 1 - \mathbf{1}'\breve{\boldsymbol{\pi}}$. Therefore $\breve{\boldsymbol{\pi}} = \dot{\mathbf{M}}^{-1}\dot{\boldsymbol{\pi}}$, and there exists a one-to-one inverse mapping from the $(2^P - 1)$-vector $\dot{\boldsymbol{\pi}}$ of moments to the 2^P-vector $\boldsymbol{\pi}$ of joint probabilities,

$$\boldsymbol{\pi} = \begin{pmatrix} 1 - \mathbf{1}'\breve{\boldsymbol{\pi}} \\ \dot{\mathbf{M}}^{-1}\dot{\boldsymbol{\pi}} \end{pmatrix} = \begin{pmatrix} 1 - \mathbf{1}'\dot{\mathbf{M}}^{-1} \\ \dot{\mathbf{M}}^{-1} \end{pmatrix} \dot{\boldsymbol{\pi}}. \tag{3}$$

A more revealing way of partitioning $\mathbf{M}$ is to break it into parts according to the partitioning of $\dot{\boldsymbol{\pi}}$, i.e.

$$\begin{pmatrix} \dot{\boldsymbol{\pi}}_1 \\ \dot{\boldsymbol{\pi}}_2 \\ \vdots \\ \dot{\boldsymbol{\pi}}_r \\ \vdots \\ \dot{\boldsymbol{\pi}}_P \end{pmatrix} = \begin{pmatrix} \dot{\mathbf{M}}_1 \\ \dot{\mathbf{M}}_2 \\ \vdots \\ \dot{\mathbf{M}}_r \\ \vdots \\ \dot{\mathbf{M}}_P \end{pmatrix} \boldsymbol{\pi}, \tag{4}$$

where $\mathbf{M}_{\binom{P}{r}}$ is a $\binom{P}{r} \times 2^P$ matrix containing the appropriate rows of $\mathbf{M}$ to obtain the r-variate joint moments of the MVB distribution,

$$\dot{\boldsymbol{\pi}}_r = \mathbf{M}_{\binom{P}{r}}\boldsymbol{\pi}. \tag{5}$$

For convenience, we write the vector of joint moments of the MVB distribution *up to* order r ($r \leq P$) as $\boldsymbol{\pi}_r = (\dot{\boldsymbol{\pi}}_1', \ldots, \dot{\boldsymbol{\pi}}_r')'$. To obtain $\boldsymbol{\pi}_r$ directly from $\boldsymbol{\pi}$, we assemble a matrix, $\mathbf{M}_r$, in the following form

$$\boldsymbol{\pi}_r = \mathbf{M}_r\boldsymbol{\pi} = \begin{pmatrix} \dot{\mathbf{M}}_1 \\ \vdots \\ \dot{\mathbf{M}}_r \end{pmatrix} \boldsymbol{\pi}. \tag{6}$$

Then $\mathbf{M}_r$ is an $s \times 2^P$ matrix, where

$$s = s(r) = \sum_{i=1}^{r} \binom{P}{i}. \tag{7}$$

Note also that $\boldsymbol{\pi}_P \equiv \dot{\boldsymbol{\pi}}$ and $\mathbf{M}_P \equiv \mathbf{M}$, by definition.

Now, let $\mathbf{p}$ be a 2^P-dimensional vector of observed cell proportions for a random sample of size N. Also, let $\mathbf{e} = \mathbf{p} - \boldsymbol{\pi}$ be the corresponding vector of cell residuals.

We can write

$$\begin{aligned}\sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_r) &= \sqrt{N}\mathbf{e}_r = \mathbf{M}_r\sqrt{N}\mathbf{e},\\ \sqrt{N}(\dot{\mathbf{p}}_r - \dot{\boldsymbol{\pi}}_r) &= \sqrt{N}\dot{\mathbf{e}}_r = \dot{\mathbf{M}}_r\sqrt{N}\mathbf{e},\\ \sqrt{N}(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}) &= \sqrt{N}\dot{\mathbf{e}} \;\; = \mathbf{M}\sqrt{N}\mathbf{e}.\end{aligned} \tag{8}$$

It is well known (see Bishop, Fienberg, & Holland, 1975) that

$$\sqrt{N}\mathbf{e} \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}), \tag{9}$$

where $\boldsymbol{\Omega} = \text{ diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$, and $\xrightarrow{L}$ indicates convergence in law (also called weak convergence or convergence in distribution). This result, along with (8), implies that

$$\sqrt{N}\mathbf{e}_r \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Xi}_r), \quad \sqrt{N}\dot{\mathbf{e}}_r \xrightarrow{L} \mathcal{N}(\mathbf{0}, \dot{\boldsymbol{\Xi}}_r), \quad \sqrt{N}\dot{\mathbf{e}} \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Xi}), \tag{10}$$

where $\boldsymbol{\Xi}_r = \mathbf{M}_r\boldsymbol{\Omega}\mathbf{M}_r'$, $\dot{\boldsymbol{\Xi}}_r = \dot{\mathbf{M}}_r\boldsymbol{\Omega}\dot{\mathbf{M}}_r'$, and $\boldsymbol{\Xi} = \mathbf{M}\boldsymbol{\Omega}\mathbf{M}'$.

3. Goodness-of-fit statistics for simple null hypothesis in MVB notation

For full-information tests, we consider a simple null hypothesis, H_0: $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta}_0)$, for parameter values $\boldsymbol{\beta}_0$ determined in advance.

3.1. Pearson's X^2

Usually X^2 is written as

$$X^2 = N\mathbf{e}'(\text{diag}(\boldsymbol{\pi}))^{-1}\mathbf{e}. \tag{11}$$

Maydeu-Olivares (1997) showed that X^2 can be equivalently written as

$$X^2 = N\dot{\mathbf{e}}'\boldsymbol{\Xi}^{-1}\dot{\mathbf{e}}. \tag{12}$$

A proof is outlined in the Appendix.

Writing Pearson's X^2 as in (12) reveals some facts easily overlooked if the multinomial characterization is used. First, the null hypothesis can be equivalently stated in terms of joint moments instead of probabilities, i.e. $H_0 : \dot{\boldsymbol{\pi}} = \dot{\boldsymbol{\pi}}(\boldsymbol{\beta}_0)$. Next, the quadratic form X^2 is a weighted sum of squares in the difference between the sample moments and the expected moments of the MVB distribution, including higher-order joint moments such as the 2^Pth-order moment in $\dot{\boldsymbol{\pi}}_P$. When only a small sample is available, the higher-order sample moments become very unstable, and we should not expect X^2 to behave like a chi-squared variable. Finally, the MVB characterization using joint moments also facilitates the proof of the asymptotic chi-squared distribution of X^2 (and similar statistics). As the asymptotic covariance matrix of $\sqrt{N}\dot{\mathbf{e}}$ is $\boldsymbol{\Xi}$, the chi-squaredness of $N\dot{\mathbf{e}}'\boldsymbol{\Xi}^{-1}\dot{\mathbf{e}}$ is obvious because $\boldsymbol{\Xi}^{-1}\boldsymbol{\Xi} = \mathbf{I}_{2^P-1}$ is idempotent, and the degrees of freedom are then simply $\text{rank}(\mathbf{I}_{2^P-1}) = 2^P - 1$ (Rao, 1973).

3.2. Some limited-information statistics

Instead of using the full vector $\dot{\mathbf{e}}$ of all the joint MVB moments in the construction of a quadratic form, we may use the vector $\mathbf{e}_r$ containing only MVB moments up to order $r \leq P$ and the matrix $\boldsymbol{\Xi}_r$ to obtain a limited-information goodness-of-fit statistic proposed

by Maydeu-Olivares (1997),

$$L_r = N\mathbf{e}_r'\boldsymbol{\Xi}_r^{-1}\mathbf{e}_r.$$

For tests using L_r, the null hypothesis is H_0: $\boldsymbol{\pi}_r = \boldsymbol{\pi}_r(\boldsymbol{\beta_0})$. Alternatively, we could also consider using $\dot{\mathbf{e}}_r$ and $\dot{\boldsymbol{\Xi}}_r$ in a quadratic form,

$$\dot{L}_r = N\dot{\mathbf{e}}_r'\dot{\boldsymbol{\Xi}}_r^{-1}\dot{\mathbf{e}}_r.$$

For tests using $\dot{L}_r$, the null hypothesis is $H_0 : \dot{\boldsymbol{\pi}}_r = \dot{\boldsymbol{\pi}}_r(\boldsymbol{\beta_0})$. It follows directly from results in (10) that L_r is asymptotically distributed as a chi-squared variable with s degrees of freedom, where s is defined in (7), and $\dot{L}_r$ also has an asymptotic chi-squared distribution with $\binom{P}{r}$ degrees of freedom. When r is small, only the lower-order joint moments enter into the computation of L_r or $\dot{L}_r$, and the chi-squared approximation should be more accurate for L_r or $\dot{L}_r$ than for X^2 in small samples. We only mention these limited-information statistics to motivate the discussions about Bartholomew and Leung's (2002) Y statistic. We do not examine the properties of L_r or $\dot{L}_r$ further.

3.3. Bartholomew and Leung's (2002) Y

Bartholomew and Leung (2002) considered a test statistic, Y, using only bivariate MVB moments, for the simple hypothesis of H_0: $\dot{\boldsymbol{\pi}}_2 = \dot{\boldsymbol{\pi}}_2(\boldsymbol{\beta_0})$. It can be written using the MVB notation as

$$Y = N\dot{\mathbf{e}}_2'\dot{\mathbf{D}}_2^{-1}\dot{\mathbf{e}}_2, \tag{13}$$

where $\dot{\mathbf{D}}_2 = Dg(\dot{\boldsymbol{\Xi}}_2)$ and the operator $Dg(\cdot)$ sets the off-diagonal elements of $\dot{\boldsymbol{\Xi}}_2$ to zero. It can be verified that $\dot{\mathbf{D}}_2 = \text{diag}(\dot{\boldsymbol{\pi}}_2)(\mathbf{I} - \text{diag}(\dot{\boldsymbol{\pi}}_2))$, and the operator diag$(\cdot)$ creates a diagonal matrix from a vector. Y is a quadratic form in bivariate residuals, just as $\dot{L}_2$ is, except that only the diagonal elements of $\dot{\boldsymbol{\Xi}}_2$ are used in the weight matrix, instead of the full covariance matrix.

The Y statistic is not asymptotically distributed as a chi-squared variable under a simple null hypothesis. Rather, it is asymptotically distributed as a mixture of one-degree-of-freedom chi-squared variates (Box, 1954). Bartholomew and Leung (2002) suggested approximating its distribution using a central chi-squared variable by matching moments. Because the vector of bivariate residuals is asymptotically normally distributed with mean zero and covariance matrix $\dot{\boldsymbol{\Xi}}_2$, the first three asymptotic moments of Y are (see Mathai & Provost, 1992, p. 53)

$$\mu_1(Y) = \text{tr}(\dot{\mathbf{D}}_2^{-1}\dot{\boldsymbol{\Xi}}_2), \quad \mu_2(Y) = 2\text{tr}(\dot{\mathbf{D}}_2^{-1}\dot{\boldsymbol{\Xi}}_2)^2, \quad \mu_3(Y) = 8\text{tr}(\dot{\mathbf{D}}_2^{-1}\dot{\boldsymbol{\Xi}}_2)^3. \tag{14}$$

These expressions are the same as those in Bartholomew and Leung (2002). Given these moments, Bartholomew and Leung (2002) equated them to those of a linear transformation of a chi-squared variable.[3] They considered two- and three-moment adjustments for Y.

To obtain a p-value using a two-moment adjustment, we assume that Y can be approximated by $b\chi_c^2$, where χ_c^2 stands for a chi-squared distribution with c degrees of freedom. Solving for the two unknown constants b and c using the first two asymptotic

[3] *In Bartholomew and Leung's (2002) original derivations, both the exact moments and the asymptotic moments were used. However, because 'the exact moments rapidly approach their limits', we only use the asymptotic moments here for ease of exposition (Bartholomew & Leung, 2002, p. 5).*

moments of Y yields

$$b = \frac{\mu_2(Y)}{2\mu_1(Y)}, \quad c = \frac{\mu_1(Y)}{b}. \tag{15}$$

For the three-moment adjustment, we assume that Y can be approximated by $a + b\chi_c^2$. Solving for the three unknown constants a, b, and c using the first three asymptotic moments of Y yields

$$b = \frac{\mu_3(Y)}{4\mu_2(Y)}, \quad c = \frac{\mu_2(Y)}{2b^2}, \quad a = \mu_1(Y) - bc. \tag{16}$$

A p-value for the two-moment adjusted statistic is obtained using

$$\Pr(\chi_c^2 > Y/b), \tag{17}$$

and for the three-moment adjusted statistic using

$$\Pr(\chi_c^2 > (Y - a)/b). \tag{18}$$

In the structural equation modelling literature, mean-adjusted test statistics are popular (Satorra & Bentler, 1994), so in addition to the two- and three-moment approximations considered above, we also consider here a one-moment approximation. Again, we assume that Y can be approximated by $b\chi_d^2$, where d is equal to the degrees of freedom available for testing. In this case, since no parameters are estimated, we conjecture that the number of degrees of freedom simply equals the number of moments used in the computation of Y, $d = P(P - 1)/2$. Solving for b, we have

$$b = \frac{\mu_1(Y)}{d}, \tag{19}$$

and the p-value for the first-moment adjusted statistic is given by

$$\Pr(\chi_d^2 > Y/b). \tag{20}$$

4. Limited-information testing of composite null hypotheses

So far our discussion has been limited to the case of simple null hypotheses, but in practice we are most often interested in composite hypotheses, where the parameters are estimated from the data. In this section, we consider the asymptotic distribution of goodness-of-fit statistics when the model parameters are estimated by ML. That is, for full-information tests, we consider H_0: $\boldsymbol{\pi} - \boldsymbol{\pi}(\boldsymbol{\beta}) = \mathbf{0}$ for some $\boldsymbol{\beta}$ vs. H_1: $\boldsymbol{\pi} - \boldsymbol{\pi}(\boldsymbol{\beta}) \neq \mathbf{0}$ for any $\boldsymbol{\beta}$. The adaptation of the null to limited-information tests is straightforward. For example, if only second-order joint moments are used, H_0 becomes $\dot{\boldsymbol{\pi}}_2 - \dot{\boldsymbol{\pi}}_2(\boldsymbol{\beta}) = \mathbf{0}$, and if moments up to the second order are used, H_0 becomes $\boldsymbol{\pi}_2 - \boldsymbol{\pi}_2(\boldsymbol{\beta}) = \mathbf{0}$ (see Reiser, 1996, p. 521, for an equivalent statement of the null). Note that the zero vectors used above have different dimensions.

Let $\hat{\boldsymbol{\beta}}$ be the MLE of the q-dimensional parameter vector $\boldsymbol{\beta}$. We assume the necessary regularity conditions on the model (Bishop *et al.*, 1975) to ensure the consistency and asymptotic normality of the MLE. In particular, we assume that the $2^P \times q$ Jacobian matrix $\mathbf{J} = \partial\boldsymbol{\pi}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$ is of full column rank so that the model is identified. Thus, $\hat{\boldsymbol{\beta}}$ is consistent, that is, $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$, where $\xrightarrow{P}$ indicates convergence in probability. The estimator is asymptotically normally distributed

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \mathcal{I}^{-1}), \tag{21}$$

where $\mathcal{I} = \mathbf{J}'\{\text{diag}[\boldsymbol{\pi}(\boldsymbol{\beta})]\}^{-1}\mathbf{J}$ is the information matrix. The cell residuals $\hat{\mathbf{e}} = \mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$ are also asymptotically normally distributed

$$\sqrt{N}\hat{\mathbf{e}} \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{22}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Omega} - \mathbf{J}\mathcal{I}^{-1}\mathbf{J}'$.

Because the marginal residuals defined in (8) are simply linear combinations of the cell residuals, their asymptotic distribution under ML estimation of the model parameters follows immediately:

$$\mathbf{M}_r\sqrt{N}\hat{\mathbf{e}} = \sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\beta}})) = \sqrt{N}\hat{\mathbf{e}}_r \xrightarrow{L} \mathcal{N}(0, \boldsymbol{\Phi}_r),$$

$$\dot{\mathbf{M}}_r\sqrt{N}\hat{\mathbf{e}} = \sqrt{N}(\dot{\mathbf{p}}_r - \dot{\boldsymbol{\pi}}_r(\hat{\boldsymbol{\beta}})) = \sqrt{N}\hat{\dot{\mathbf{e}}}_r \xrightarrow{L} \mathcal{N}(0, \dot{\boldsymbol{\Phi}}_r),$$

$$\mathbf{M}\sqrt{N}\hat{\mathbf{e}} = \sqrt{N}(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}})) = \sqrt{N}\hat{\dot{\mathbf{e}}} \xrightarrow{L} \mathcal{N}(0, \boldsymbol{\Phi}), \tag{23}$$

where

$$\boldsymbol{\Phi}_r = \mathbf{M}_r\boldsymbol{\Omega}\mathbf{M}_r' - \mathbf{M}_r\mathbf{J}\mathcal{I}^{-1}\mathbf{J}'\mathbf{M}_r' = \boldsymbol{\Xi}_r - \mathbf{J}_r\mathcal{I}^{-1}\mathbf{J}_r', \tag{24}$$

$$\dot{\boldsymbol{\Phi}}_r = \dot{\mathbf{M}}_r\boldsymbol{\Omega}\dot{\mathbf{M}}_r' - \dot{\mathbf{M}}_r\mathbf{J}\mathcal{I}^{-1}\mathbf{J}'\dot{\mathbf{M}}_r' = \dot{\boldsymbol{\Xi}}_r - \dot{\mathbf{j}}_r\mathcal{I}^{-1}\dot{\mathbf{j}}_r', \tag{25}$$

$$\boldsymbol{\Phi} = \mathbf{M}\boldsymbol{\Omega}\mathbf{M}' - \mathbf{M}\mathbf{J}\mathcal{I}^{-1}\mathbf{J}'\mathbf{M}' = \boldsymbol{\Xi} - \dot{\mathbf{j}}\mathcal{I}^{-1}\dot{\mathbf{j}}'. \tag{26}$$

Equations (23), (24), and (25) give us the machinery to study the properties of Y under ML parameter estimation.

4.1. The distribution of Y under ML estimation

Consider the Y statistic given in (13) when the model parameters are estimated using ML,

$$Y = N\hat{\dot{\mathbf{e}}}_2'\hat{\dot{\mathbf{D}}}_2^{-1}\hat{\dot{\mathbf{e}}}_2', \tag{27}$$

where $\hat{\dot{\mathbf{D}}}_2 = \text{diag}(\dot{\boldsymbol{\pi}}_2(\hat{\boldsymbol{\beta}}))(\mathbf{I} - \text{diag}(\dot{\boldsymbol{\pi}}_2(\hat{\boldsymbol{\beta}})))$. By the consistency of the MLE and the usual regularity conditions assumed for the model (Bishop *et al.*, 1975, pp. 509–511), especially those pertaining to the continuity of the mapping $\boldsymbol{\pi}(\cdot)$ and the compactness of the parameter space, we can show that $\hat{\dot{\mathbf{D}}}_2 \xrightarrow{P} \dot{\mathbf{D}}_2$.

When ML estimates are used, the Y statistic is also asymptotically distributed as a mixture of one-degree-of-freedom chi-squared variates (Box, 1954). Following Bartholomew and Leung (2002), we can approximate its distribution using a central chi-square by matching moments. Since the vector of bivariate residuals is asymptotically normally distributed with mean zero and covariance matrix $\dot{\boldsymbol{\Phi}}_2$ (see (25)), the first three asymptotic moments of Y are similar in form to those given in (14), with $\dot{\boldsymbol{\Phi}}_2$ replacing $\dot{\boldsymbol{\Xi}}_2$:

$$\mu_1(Y) = \text{tr}(\dot{\mathbf{D}}_2^{-1}\dot{\mathbf{\Phi}}_2), \quad \mu_2(Y) = 2\text{tr}(\dot{\mathbf{D}}_2^{-1}\dot{\mathbf{\Phi}}_2)^2, \quad \mu_3(Y) = 8\text{tr}(\dot{\mathbf{D}}_2^{-1}\dot{\mathbf{\Phi}}_2)^3. \tag{28}$$

Given these expressions for the asymptotic moments of the Y statistic, approximate p-values can be obtained using a two-moment adjustment via (15) and (17), when the moments in (28) are consistently estimated by evaluating the matrices involved in (25) and (28) at the ML estimates. For example, we can write the estimated first moment as

$$\hat{\mu}(Y) = \text{tr}\left(\hat{\dot{\mathbf{D}}}_2^{-1}\dot{\mathbf{\Phi}}_2\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right).$$

Similarly, we write $\hat{\mu}_2(Y)$ and $\hat{\mu}_3(Y)$ to stand for the estimated second and third moments. Approximate p-values can also be obtained using a three-moment adjustment via (16) and (18). For the one-moment adjustment we heuristically use as degrees of freedom $d = P(P-1)/2 - q$ since q parameters are estimated. That is, we use a chi-squared distribution with $d = P(P-1)/2 - q$ degrees of freedom in (19) and (20) to obtain the p-value for the first-moment adjusted statistic.

However, Bartholomew and Leung (2002) used $\dot{\mathbf{\Xi}}_2$ in (10) rather than $\dot{\mathbf{\Phi}}_2$ in (25) as the covariance matrix of the residual moments for the composite null hypothesis. For example, their method gives the first moment as

$$\tilde{\mu}_1(Y) = \text{tr}\left(\hat{\dot{\mathbf{D}}}_2^{-1}\dot{\mathbf{\Xi}}_2\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right)'$$

and we write $\tilde{\mu}_2(Y)$ and $\tilde{\mu}_3(Y)$ for the second and third moments obtained using their method. We use the symbol $\sim$ to denote the moments obtained ignoring ML parameter estimation.

Upon examining the covariance matrix in (25), we can see that $\dot{\mathbf{\Phi}}_2$ is in general not equal to $\dot{\mathbf{\Xi}}_2$, because the second term $\dot{\mathbf{J}}_2\mathcal{I}^{-1}\dot{\mathbf{J}}_2'$ is not negligible. In our experience, Bartholomew and Leung's (2002) method often leads to substantial overestimation of the moments. Bartholomew and Leung (2002) assumed that the behaviour of Y under the composite null hypothesis is similar to that under the simple null hypothesis when the number of parameters q is much smaller than the number of cells $C = 2^P$. We report simulations below showing that the distribution of Y under the composite null hypothesis is much more constrained than under the simple null hypothesis.

4.2. An extension using both univariate and bivariate moments

In its current form, Y uses only the bivariate moments. Tollenaar and Mooijaart (2003) suggested that using both univariate and bivariate moments often results in a test statistic that is more powerful, because the lowest-order margin is the best-filled. Following their recommendation, we constructed a statistic, provisionally called Y_2, based on Y but including univariate moments in the formulation. Instead of using $P(P-1)/2$ moments in Y, we are now using $P(P+1)/2$ moments in Y_2. Again, Y_2 is not chi-squared distributed under the simple or composite null hypothesis, so we follow the same moment approximation technique outlined in the previous section. With the distributional results given in the preceding sections, it is easy to find the moments of Y_2 by assembling an appropriate design matrix $\mathbf{M}_2$, as defined in (6). In the MVB notation, Y_2 is given by

$$Y_2 = N\mathbf{e}_2'\mathbf{D}_2^{-1}\mathbf{e}_2, \tag{29}$$

for the simple null hypothesis, where $\mathbf{e}_2$ is defined in (8) and $\mathbf{D}_2 = \text{diag}(\boldsymbol{\pi}_2)(\mathbf{I} - \text{diag}(\boldsymbol{\pi}_2))$. For the composite null, we simply use the ML parameter estimates $\hat{\boldsymbol{\beta}}$ to obtain

$$Y_2 = N\hat{\mathbf{e}}_2'\hat{\mathbf{D}}_2^{-1}\hat{\mathbf{e}}_2, \tag{30}$$

where $\hat{\mathbf{e}}_2$ is defined in (23) and $\hat{\mathbf{D}}_2 = \text{diag}(\boldsymbol{\pi}_2(\hat{\boldsymbol{\beta}}))(\mathbf{I} - \text{diag}(\boldsymbol{\pi}_2(\hat{\boldsymbol{\beta}})))$.

Under ML parameter estimation, the asymptotic moments of Y_2 are

$$\mu_1(Y_2) = \text{tr}(\mathbf{D}_2^{-1}\boldsymbol{\Phi}_2), \quad \mu_2(Y_2) = 2\text{tr}(\mathbf{D}_2^{-1}\boldsymbol{\Phi}_2)^2, \quad \mu_3(Y_2) = 8\text{tr}(\mathbf{D}_2^{-1}\boldsymbol{\Phi}_2)^3. \tag{31}$$

These moments can be estimated consistently by evaluating the matrices at the ML parameter estimates, e.g.

$$\hat{\mu}_1(Y_2) = \text{tr}\left(\hat{\mathbf{D}}_2^{-1}\boldsymbol{\Phi}_2\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right).$$

We also write $\hat{\mu}_2(Y_2)$ and $\hat{\mu}_3(Y_2)$ for the second and third moments. The approximate p-values can be obtained using the same methods as in (15)–(20).

5. Simulations

In this section, we describe simulations that illustrate the difference between our method of approximating the distribution of Y under ML parameter estimation, and that of Bartholomew and Leung (2002). A secondary goal is to show that Y, together with our new moment approximations, offers a goodness-of-fit test that can keep the type I error rate at the nominal significance level when the contingency table is sparse. We also investigate how many moments are needed in order to achieve an acceptable degree of approximation. Lastly, we gauge the performance of Y_2 under sparseness by pitting it against Y and the full-information statistics G^2 and X^2.

5.1. Models in the simulation

We used the 2PL IRT model (also called the logit/normit model by Bartholomew & Knott, 1999) in our simulations to assess the empirical type I error rates of the goodness-of-fit statistics, and the three-parameter logistic (3PL) model to evaluate the power and sensitivity of the test statistics against model misspecification. Using notation that is consistent with the MVB characterization, the 2PL model relates the probability of correctly responding to a test item given a continuous latent variable θ through a 2PL function,

$$\pi(U_i = 1|\boldsymbol{\beta}_i, \theta) = \frac{1}{1 + \exp\left[-D\alpha_i(\theta - \beta_i)\right]},$$

where $U_i = 1$ represents the correct response to the ith item, $\boldsymbol{\beta}_i = (\alpha_i, \beta_i)'$ is a vector of parameters for that item, and D is a constant that puts the logistic model on the normal metric, usually taken to be 1.7. It follows that the response $u_i = 1$ or 0 has probability

$$\pi(U_i = u_i|\boldsymbol{\beta}, \theta) = \pi(U_i = 1|\boldsymbol{\beta}_i, \theta)^{u_i}[1 - \pi(U_i = 1|\boldsymbol{\beta}_i, \theta)]^{1-u_i},$$

where $\boldsymbol{\beta}$ is a $2P$-dimensional vector containing all the item parameters of the model. Since the 'person parameters' θ are incidental, the estimation of the item parameters is usually carried out by assuming that the latent trait is distributed normally with mean zero and variance one, so the marginal probability for response pattern $\mathbf{u} = (u_1, \ldots, u_i, \ldots, u_P)$ is

$$\pi_{\mathbf{u}}(\boldsymbol{\beta}) = \int_{-\infty}^{\infty} \prod_{i=1}^{P} \pi(U_i = u_i | \boldsymbol{\beta}, \theta) \phi(\theta) d\theta, \tag{32}$$

where $\phi(\theta)$ is the standard normal density function. The estimator of $\boldsymbol{\beta}$ derived from the solution to the marginal likelihood equations (Bock & Aitkin, 1981; Bock & Lieberman, 1970) is consistent, asymptotically efficient, and asymptotically normal.

The 3PL model adds a 'guessing' parameter to the response function, making it

$$\pi(U_i = 1 | \boldsymbol{\beta}_i, \theta) = \gamma_i + \frac{1 - \gamma_i}{1 + \exp\left[-D\alpha_i(\theta - \beta_i)\right]},$$

where γ_i is the guessing parameter. When γ_i is set to 0, the 3PL model reduces to the 2PL model. Except for a change in the dimensions of the parameter vector, everything else remains basically the same for the 3PL model. We use the 3PL model only as the data generating model for power evaluations, so only the 2PL model is fitted to the simulated data sets.

5.2. Data generation

We simulated the null distribution of the goodness-of-fit statistics using the following configuration of 2PL item parameters:

$$\boldsymbol{\beta}' = \begin{pmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 1.0 & 1.0 & 1.0 & 1.0 & 1.5 & 1.5 & 1.5 & 1.5 \\ -1.5 & 0.0 & 1.5 & 2.0 & -1.5 & 0.0 & 1.5 & 2.0 & -1.5 & 0.0 & 1.5 & 2.0 \end{pmatrix}. \tag{33}$$

There are 12 items in this hypothetical test. We varied the sample size ($N = 250$, $N = 1000$, and $N = 4000$) to investigate the effect of different levels of sparseness on the type I error rates of the goodness-of-fit test statistics. The number of replications was set to 1000 in each sample size condition. For each sample, a 2PL model was fitted using MULTILOG (Thissen, 2003).[4]

For power evaluations, we need to simulate under the non-null condition so that the 2PL model does not fit perfectly in the population. This is done by changing the population generating model to a 3PL with a non-zero guessing parameter. The α_i and β_i are the same as those in (33), while all the γ_i are fixed to 0.25. Therefore the population model probably reflects a test with 12 multiple choice items each having 4 response options. When 2PL models are fitted to random samples generated from test items with a non-zero lower asymptote, they are clearly misspecified, so in this condition the empirical rejection rates of goodness-of-fit tests provide some rough estimates of the

[4] *For all iterative algorithms, non-convergence always poses a problem in simulations. For the larger sample sizes (N = 1000 and N = 4000) all 1000 replications converged. For the N = 250 condition, 35 out of 1000 replications did not converge after 500 E-step iterations. These non-convergent cases were removed from the analysis based on recommendations given in Paxton, Curran, Bollen, Kirby, and Chen (2001). However, due to the smaller number of replications (965 as opposed to 1000), the results for N = 250 are subject to more sampling variability than the other conditions.*

power of the test against model misspecification. We investigated power only for $N = 1000$ as an illustration.

5.3 Test statistics

Using the new results in (28) and (31), we matched one, two and three moments for both Y and Y_2. As a comparison, we also matched one, two and three moments for Y using the moments given by Bartholomew and Leung (2002). Altogether, we included the following 11 goodness-of-fit statistics in the simulations:

- two full-information statistics[5] – G^2 and X^2;
- YBL1, YBL2, YBL3 – the one-, two- and three-moment adjustments to the quadratic form in bivariate residuals Y using Bartholomew and Leung's (2002) moments;
- YC1, YC2, YC3 – the one-, two- and three-moment adjustments to Y using corrected moments that take ML estimation into account;
- Y2C1, Y2C2, Y2C3 – the one-, two- and three-moment adjustments to the quadratic form in univariate and bivariate residuals Y_2 using moments that take ML estimation into account.

The p-values for both X^2 and G^2 were obtained in each replication with reference to their limiting chi-squared distribution with $2^P - 2P - 1 = 4071$ degrees of freedom. For methods based on moment approximations, the p-values were obtained in each replication by matching the estimated moments with those of a scaled chi-square, but because the first three moments have to be estimated from sample data, the degrees of freedom in the chi-squared approximation vary from replication to replication. However, this is consistent with the standard practice in both structural equation modelling (Satorra & Bentler, 1994), and variance components (Satterthwaite, 1946), wherein the adjusted degrees of freedom are also estimated. Once the p-values are obtained, hypothesis tests are conducted for all statistics at various significance levels, and the number of rejections recorded.

6. Simulation results

The main results from the simulations are summarized in Tables 1–5. Tables 1–3 contain the observed type I error rates for tests conducted at five significance levels: .01, .05, .10, .20, and .25, at $N = 250$, 1000, and 4000. We report limited power results in Table 4 for sample size 1000. Because our main focus is on corrections to the null distribution, we leave a more thorough investigation of power to future research. In Table 5 we show the effect of ignoring parameter estimation on the quality of moment approximations. The entries in Table 5 are for $N = 1000$, but they are representative of the situations in other sample size conditions.

6.1 Type I error rates

For sample size $N = 250$, the sparseness is very severe. The usual full-information statistics become invalid.[6] First, the type I error rates of G^2 are grossly inaccurate at any

[5] *G^2 is computed as $G^2 = 2N\sum_{\boldsymbol{u}} p_{\boldsymbol{u}} \log(p_{\boldsymbol{u}}/\pi_{\boldsymbol{u}}(\hat{\boldsymbol{\beta}}))$ (see Bishop et al., 1975, p. 513).*

[6] *We define the validity of a test according to the liberal criterion suggested by Bradley (1978), e.g. if the nominal significance level is .05, a test is deemed valid if its type I error rate is in [.025, .075].*

Table 1. Type I Error rates for $N = 250$

	Significance level				
	.01	.05	.10	.20	.25
YBL1	.000	.002	.002	.002	.002
YBL2	.000	.000	.000	.002	.002
YBL3	.000	.000	.000	.002	.002
YC1	.038	.092	.135	.222	.265
YC2	.028	.077	.114	.209	.245
YC3	.026	.075	.114	.211	.251
Y2C1	.055	.107	.167	.243	.280
Y2C2	.027	.078	.114	.212	.245
Y2C3	.027	.076	.112	.214	.252
X^2	.092	.092	.093	.095	.095
G^2	.000	.000	.000	.000	.000

Note. The numbers are based on 965 fully converged replications. The test statistics are described in Section 5.3.

of the significance levels considered. It is simply too conservative. On the other hand, X^2 can be too liberal or too conservative depending on the nominal significance level, but note that its type I error rates are almost constant across all significance levels. This is an indication that the null distribution of X^2 is not well approximated by the reference chi-squared distribution. Next, we look at the tests involving moment-based approximations. It does not matter how many moments are matched for YBL1–YBL3, because as long as Bartholomew and Leung's (2002) moments are used, the null hypothesis is almost never rejected at any of the significance levels specified. In contrast, YC3 has type I error rates that are reasonably close to the nominal significance levels. This indicates a good agreement between the true null distribution of Y and the moment

Table 2. Type I error rates for $N = 1000$

	Significance level				
	.01	.05	.10	.20	.25
YBL1	.000	.000	.000	.000	.000
YBL2	.000	.000	.000	.000	.000
YBL3	.000	.000	.000	.000	.000
YC1	.017	.051	.111	.226	.269
YC2	.008	.040	.090	.205	.252
YC3	.008	.040	.089	.209	.258
Y2C1	.024	.083	.141	.247	.289
Y2C2	.008	.041	.094	.208	.256
Y2C3	.008	.040	.093	.211	.261
X^2	.102	.106	.107	.108	.110
G^2	.000	.000	.000	.000	.000

Note. The numbers are based on 1000 fully converged replications. The test statistics are described in Section 5.3.

Table 3. Type I error rates for $N = 4000$

	Significance level				
	.01	.05	.10	.20	.25
YBL1	.000	.000	.000	.000	.000
YBL2	.000	.000	.000	.000	.000
YBL3	.000	.000	.000	.000	.000
YC1	.026	.092	.152	.253	.307
YC2	.017	.069	.129	.236	.286
YC3	.014	.066	.127	.239	.293
Y2C1	.046	.129	.191	.300	.344
Y2C2	.019	.078	.134	.253	.308
Y2C3	.016	.075	.133	.254	.313
X^2	.147	.152	.154	.157	.158
G^2	.000	.000	.000	.000	.000

Note. The numbers are based on 1000 fully converged replications. The test statistics are described in Section 5.3.

approximations based on our corrected moments in (28). The same description also applies to Y2C3.

For $N = 1000$, the performance of X^2, G^2 and YBL1–YBL3 does not improve. On the other hand, the increased sample size has a clear impact on the quality of moment approximations for YC2, YC3, Y2C2 and Y2C3. All of them performed better as compared to the $N = 250$ condition, especially at the extreme tails of the distributions. For example, when sample size increased from 250 to 1000, the empirical rejection rates of YC2 and YC3 for the .01 significance level dropped from being more than twice the nominal level to a uniform .008. We observe essentially the same phenomenon for $N = 4000$, but we also notice that some of the limited-information tests seem to be

Table 4. Power for $N = 1000$

	Significance level				
	.01	.05	.10	.20	.25
YBL1	.000	.000	.000	.000	.000
YBL2	.000	.000	.000	.000	.000
YBL3	.000	.000	.000	.000	.000
YC1	.124	.276	.380	.541	.583
YC2	.116	.270	.376	.533	.582
YC3	.106	.263	.375	.536	.583
Y2C1	.188	.331	.451	.575	.615
Y2C2	.120	.273	.390	.544	.588
Y2C3	.110	.271	.386	.548	.591
X^2	.849	.875	.886	.898	.901
G^2	.000	.000	.000	.000	.000

Note. The numbers are based on 1000 fully converged replications. The test statistics are described in Section 5.3.

Table 5. A comparison of different moment estimation methods for $N = 1000$

	Descriptives			
	Min	Max	Mean	Variance
Y	4.080	23.810	9.162	4.613
$\tilde{\mu}_2(Y)$	911.834	1306.011	1116.609	3989.944
$\hat{\mu}_1(Y)$	7.866	10.779	9.050	0.181
$\hat{\mu}_2(Y)$	3.632	6.933	4.773	0.203

Note. This table presents the simulated null distribution of $\hat{Y}$ under the composite null hypothesis. The numbers are based on 1000 converged replications; Bartholomew and Leung's (2002) first moment, $\tilde{\mu}_1(Y)$, is always equal to 66; $\hat{\mu}_1(Y)$ and $\hat{\mu}_2(Y)$ are computed from (28).

somewhat liberal.[7] However, given the rather small number of replications (1000), fine distinctions about the pattern of type I error rates cannot be made. More research is certainly warranted.

6.2 Power to detect misspecification

Table 4 presents the observed rejection rates of the 11 tests when the population generating model is a 3PL model with non-zero guessing parameters. Thus the rejection rates reflect the sensitivity of the test statistics to model misspecification. Among the statistics considered, the rejection rates for X^2 under model misspecification are the largest. This is to be expected, as it simply rejects the null hypothesis too often, even when the model fits perfectly in the population (see Tables 1–3). G^2 and YBL1–YBL3, on the other hand, continue to be extremely conservative. We also see that tests using univariate moments (Y2C1–Y2C3) are slightly more powerful than tests that do not use univariate moments (YC1–YC3), but the power advantage is small.

6.3 Quality of moment approximations

Table 5 demonstrates how ignoring the effect of parameter estimation can have a dramatic impact on the quality of moment approximations. The first line contains some basic descriptive statistics for the simulated null distribution of Y. We see that it ranges from 4.08 to 23.81, with mean 9.16 and variance 4.61. Using Bartholomew and Leung's (2002) moment approximations ignoring ML estimation, we would find the mean to be 66, and the variance to be about 1116.61, when averaged over the 1000 replications. Both the mean and the variance are grossly overestimated. On the other hand, the moments we derived under ML estimation closely match the simulated null distribution. Specifically, the average of the estimated means of Y is 9.05 (compare with 9.16), and the average of the estimated variances is 4.77 (compare with 4.61). These comparisons indicate that a much better approximation to the null distribution of Y can be obtained if our corrected moments are used,

[7] *This is a purely numerical problem caused by the number of quadrature points used in the estimation. With a larger number of quadrature points, the liberal bias would be reduced.*

and indeed we find support for this claim from the type I error rates reported in Tables 1–3.

7. A worked example

We apply the 11 statistics discussed in the preceding sections to a real data set in order to demonstrate the utility of the moment approximations we proposed in practical situations, where the parameters are estimated by ML. The data set was the Social Life Feelings (SLF) survey taken from Bartholomew (1998). The SLF data are in the form of a five-question scale with dichotomous items. The number of respondents was 1490. We fitted a 2PL model by ML to these data using MULTILOG (Thissen, 2003). The results for the different test statistics considered are reported in Table 6.

Table 6. Analysis of social life feelings data

	Value	Degrees of freedom	*p*-value
YBL1	–	–	–
YBL2	1.59	3.70	0.770
YBL3	0.38	1.89	0.810
YC1	–	–	–
YC2	19.55	4.45	<0.001
YC3	19.00	3.42	<0.001
Y2C1	23.21	5.00	<0.001
Y2C2	20.05	4.47	<0.001
Y2C3	17.16	3.42	0.001
X^2	38.96	21.00	0.010
G^2	39.09	21.00	0.010

Note. $Y = 4.3$, $Y_2 = 4.41$, but YBL1 and YC1 cannot be computed because there are no degrees of freedom left for testing. The test statistics are described in Section 5.3.

As can be seen in this table, the two full-information statistics have *p*-values roughly equal to .01, suggesting a significant lack of fit. Bartholomew and Leung (2002) reported similar results for these two statistics after pooling cells with small expected probabilities. Given the large *N*-to-*C* ratio, these full-information test statistics should be trusted to have good approximations to their limiting distributions.

For limited-information testing, we first essentially replicated Bartholomew and Leung's (2002) analyses. The value of Y is 4.3. Applying Bartholomew and Leung's (2002) moment approximations, we find $\tilde{\mu}_1(Y) = 10, \tilde{\mu}_2(Y) = 54$, and $\tilde{\mu}_3(Y) = 813$. The one-moment approximation (YBL1) with our heuristic for the degrees of freedom cannot be computed in this case, because the number of bivariate moments entering into the computation of Y is 10, and a 2PL model fitted to 5 items has 10 free parameters, so there are no degrees of freedom left for testing. For the two-moment approximation (YBL2) we compute $b = \tilde{\mu}_2(Y)/[2\tilde{\mu}_1(Y)] = 2.7$, and $c = \tilde{\mu}_1(Y)/b = 3.7$ using (15). Then, the two-moment adjusted statistic (YBL2) is $Y/b = 1.59$. This statistic is referred to a chi-squared distribution with $c = 3.7$ degrees of freedom. For the three-moment approximation (YBL3) we compute $b = \tilde{\mu}_3(Y)/[4\tilde{\mu}_2(Y)] = 3.77$, $c = \tilde{\mu}_2(Y)/(2b^2) = 1.89$, and $a = \tilde{\mu}_1(Y) - bc = 2.87$ using (16). Then the three-moment adjusted statistic is $(Y - a)/b = 0.38$. This statistic is referred to a chi-squared

distribution with $c = 1.89$ degrees of freedom. As shown in Table 6, YBL2 and YBL3 yield very large p-values, 0.77 and 0.81, respectively. These large p-values would suggest that the 2PL model fits well for the SLF data, but when our moment approximations are used, the conclusion is completely reversed.

Taking the effect of ML estimation into account, we find the first three central moments of Y to be $\hat{\mu}_1(Y) = 0.972$, $\hat{\mu}_2(Y) = 0.424$, and $\hat{\mu}_3(Y) = 0.423$. Again, the one-moment approximation (YC1) cannot be computed. For YC2, again using (15), we find $b = 0.22$ and $c = 4.45$. So YC2 $= 19.55$ on 4.45 degrees of freedom, $p < .001$. For YC3, by (16), $b = 0.25$, $c = 3.42$ and $a = 0.12$, so YC3 is equal to 19 on 3.42 degrees of freedom, $p < .001$, as well. As we have pointed out in the preceding sections, such a large discrepancy in the p-values between YBL3 and YC3, as well as the size of the estimated moments, is attributable to the fact that $\dot{\mathbf{J}}_2\mathcal{I}^{-1}\dot{\mathbf{J}}_2'$ is not trivial enough to be ignored.

When both univariate and bivariate moments are included, and our new approximations (see (31)) are invoked, we draw the same conclusion: as far as the univariate and bivariate margins are concerned, the 2PL model does not fit very well. Y_2 is found to be 4.41, and its moments under ML estimation are $\hat{\mu}_1(Y_2) = 0.974$, $\hat{\mu}_2(Y_2) = 0.425$, and $\hat{\mu}_3(Y_2) = 0.423$. For the one-moment adjustment to Y_2, heuristically there are $c = 5(5 + 1)/2 - 10 = 5$ degrees of freedom left for testing. Using (19) we compute $b = \hat{\mu}_1(Y_2)/c = 0.19$. Then the one-moment adjusted statistic is Y2C1 $= Y_2/b = 23.21$, which is referred to a chi-squared distribution with 5 degrees of freedom, $p < .001$. For Y2C2, again using (15), $b = 0.22$ and $c = 4.47$. So Y2C2 is 20.05 on 4.47 degrees of freedom, $p < .001$. For Y2C3, by (16), $b = 0.25$, $c = 3.42$ and $a = 0.12$, so Y2C3 equals 17.16 on 3.42 degrees of freedom, $p = .001$.

In sum, in a non-sparse table for which the p-values of X^2 and G^2 are likely to be trustworthy, the moment adjustments to Y and Y_2 yield p-values similar to those of the full-information statistics when the effect of parameter estimation is taken into account. When it is ignored, the p-values become erroneously large. The latter finding is consistent with the simulation results presented in the previous section.

8. Discussion and conclusions

Psychological researchers often wish to model large 2^P contingency tables in which there are many empty cells. In modelling these tables, most often the cell probabilities depend on parameters that are estimated from the data. Testing the overall goodness of fit of these models is a challenge because the distributions of the usual goodness of fit statistics (X^2 and G^2 in particular) are not well approximated by their asymptotic distributions. Three alternatives have been proposed to overcome this problem: resampling methods, pooling cells, and limited-information statistics. Resampling methods are computationally intensive, whereas pooling cells does not make the best use of the data and may yield statistics with unknown sampling distributions. Limited-information statistics are not only simpler computationally, but also have tractable asymptotic distributions even under severe sparseness.

Bartholomew and Leung (2002) proposed an appealing limited-information statistic, Y, based on bivariate MVB residual moments. They also proposed obtaining p-values for Y by matching either its first two or three moments with those of a scaled chi-squared distribution. Bartholomew and Leung (2002) provided the first three moments of Y for testing models whose parameters were known in advance. They conjectured that accurate p-values could be obtained using these equations when the parameters were

estimated from the data. Here, we have investigated their conjecture for models estimated using maximum likelihood. To do so, we have provided the asymptotic moments of Y for ML estimation. We have also considered an alternative test statistic Y_2 which differs from Y simply in that both univariate and bivariate MVB residual moments are used. For Y_2, the asymptotic moments under ML estimation have also been provided. By means of a simulation study we have compared the Type I error rates of Y and Y_2 with moment adjustments of different orders (1, 2, 3) in testing a 2PL model. Our results suggest the following:

(1) When parameter estimation is taken into account, the null distribution of Y and Y_2 can be well approximated using two- or three-moment adjustments, where the three-moment match usually gives a more accurate level of approximation. The one-moment adjustment based on a heuristic argument, borrowed from the structural equation modelling literature, does not perform nearly as well. Thus, it seems that the two-moment approximation suffices to obtain accurate *p*-values for these statistics.
(2) When parameter estimation is not taken into account (as in Bartholomew & Leung, 2002) an extremely conservative test is obtained.
(3) Finally, of the two statistics considered, Y_2 is slightly more powerful than Y. Thus, the use of Y_2 is recommended.

In summary, we have shown that the test statistic proposed by Bartholomew and Leung (2002) can be used to obtain accurate *p*-values for MVB models such as the IRT models considered here, provided that the effects of parameter estimation are taken into account. For increased power, we suggest that the test statistic based on both univariate and bivariate moments, Y_2, be used instead of Y. Two obvious drawbacks of Y and Y_2 are that they have no power to distinguish among models with the same expected lower-order marginal moments but different higher-order moments. Also, they cannot be employed with models that fit the marginals perfectly (such as log-linear models with all bivariate terms).

It should be noted that the equations provided here for the asymptotic moments of the statistics considered are only valid for models estimated using ML or other asymptotically minimum variance estimators such as minimum chi-squared that are asymptotically equivalent to ML. For other estimators, different expressions for the asymptotic moments of the statistics are needed. It is interesting to point out that Maydeu-Olivares (2001a) proposed a statistic similar to Y_2 for testing the two-parameter normal ogive model (and related models). His statistic is simply

$$T = N\hat{\mathbf{e}}_2'\hat{\mathbf{e}}_2.$$

To obtain *p*-values for his statistic he suggested using one- and two-moment approximations to a central chi-squared distribution. Maydeu-Olivares (2001a) provided the asymptotic mean and variance of T when the model parameters were estimated using the three-stage limited-information estimator implemented in LISREL (Jöreskog & Sörbom, 2001) and MPLUS (Muthén & Muthén, 2001), and in Maydeu-Olivares (2001b) for the two-stage limited-information estimator implemented in NOHARM (Fraser & McDonald, 1988). Remarkably, Maydeu-Olivares (2001a) reports a simulation study with a 2^{21} contingency table where $N = 100$ observations sufficed to obtain accurate *p*-values when the two-moment adjustment was employed.

Further research is needed to compare the empirical behaviour of the unweighted test statistic T versus the weighted test statistic Y_2. However, preliminary research suggests that the differences may be minimal. Also, for ML estimates the computation of the asymptotic moments of Y and Y_2 becomes very intensive for large P. Further research is needed to manage the computations within available computer memory for large models. Also, the MVB framework and moment-based approximations for Y and Y_2 may be readily extended to sparse multidimensional tables in which the categorical variables take more than two values.

In conclusion, we believe that the limited-information framework presented here may be a fruitful avenue for evaluating goodness of fit of ML estimated models in large and sparse binary contingency tables. On the one hand, the Y and Y_2 family of statistics enable researchers to determine the overall adequacy of their hypothesized models. On the other hand, the individual marginal residuals may be used to identify the source of the misfit for poorly fitting models. The use of these marginal residuals may be much more informative than the use of cell residuals (see Reiser, 1996).

Acknowledgements

Li Cai wishes to thank Dr Robert MacCallum for helpful comments on an earlier draft of the paper.

References

Agresti, A., Lipsitz, S., & Lang, J. B. (1992). Comparing marginal distributions of large sparse contingency tables. *Computational Statistics and Data Analysis*, *14*, 55–73.

Agresti, A., & Yang, M. C. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis*, *5*, 9–21.

Bartholomew, D. J. (1998). Scaling unobservable constructs in social science. *Applied Statistics*, *47*, 1–13.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). 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. *Sociological Methods & Research*, *27*, 525–546.

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

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

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *25*, 290–302.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*, 290–302.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152.

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

Cochran, W. G. (1952). The χ^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.

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

Jöreskog, K., & Sörbom, D. (2001). *LISREL user's guide*. Chicago: SSI International.

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

Mathai, A. M., & Provost, S. B. (1992). *Quadratic forms in random variables: Theory and applications*. New York: Marcel Dekker.

Maydeu-Olivares, A. (1997). Structural equation modeling of binary preference data (Doctoral dissertation, University of Illinois). *Dissertation Abstracts International: Section B*, *58*, 5694.

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., & Joe, H. (2005). Limited and full information estimation and 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 dichotomized variables. *Psychometrika*, *43*, 551–560.

Muthén, L., & Muthén, B. (2001). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

Paxton, P., Curran, P. J., Bollen, K., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*, 287–312.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

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

Reiser, M., & Lin, Y. (1996). Goodness-of-fit test for the latent class model when expected frequencies are small. *Sociological Methodology*, *29*, 81–111.

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.

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

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrical Bulletin*, *2*, 110–114.

Teugels, J. L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *Journal of Multivariate Analysis*, *32*, 256–268.

Thissen, D. (2003). *MULTILOG user's guide*. Chicago: SSI International.

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.

Received 15 July 2004; revised version received 23 May 2005

Appendix

We claimed in (12) that Pearson's X^2 can be equivalently written as a quadratic form in residual MVB moments. To show the equivalence, we first consider the inverse of $\boldsymbol{\Xi}$. From (10) and the partitioning of $\mathbf{M}$ and $\boldsymbol{\pi}$ given earlier in (3), it can be shown that

$$\ddot{\boldsymbol{\Xi}} = \mathbf{M}(\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{M}' = \dot{\mathbf{M}}(\text{diag}(\breve{\boldsymbol{\pi}}) - \breve{\boldsymbol{\pi}}\breve{\boldsymbol{\pi}}')\dot{\mathbf{M}}.$$

Using a result in Rao (1973), $\ddot{\boldsymbol{\Xi}}^{-1}$ is given by

$$\ddot{\boldsymbol{\Xi}}^{-1} = (\dot{\mathbf{M}}')^{-1}((\text{diag}(\breve{\boldsymbol{\pi}}))^{-1} + \pi_{(0\cdots0)}^{-1}\mathbf{1}\mathbf{1}')\dot{\mathbf{M}}^{-1}. \tag{34}$$

Next, recall that $\mathbf{e} = \mathbf{p} - \boldsymbol{\pi}$, $\dot{\mathbf{e}} = \dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}$, $\breve{\mathbf{e}} = \breve{\mathbf{p}} - \breve{\boldsymbol{\pi}}$, and $\mathbf{e}$ can be partitioned into $\mathbf{e} = (e_{(0\cdots0)}, \breve{\mathbf{e}}')'$, where $e_{(0\cdots0)} = -\mathbf{1}'\breve{\mathbf{e}}$. It can be verified that $\dot{\mathbf{e}} = \dot{\mathbf{M}}\breve{\mathbf{e}}$, and (12) can be written as

$$\begin{aligned}
N\dot{\mathbf{e}}'\ddot{\boldsymbol{\Xi}}^{-1}\dot{\mathbf{e}} &= N\breve{\mathbf{e}}'\dot{\mathbf{M}}'(\dot{\mathbf{M}}')^{-1}((\text{diag}(\breve{\boldsymbol{\pi}}))^{-1} + \pi_{(0\cdots0)}^{-1}\mathbf{1}\mathbf{1}')\dot{\mathbf{M}}^{-1}\dot{\mathbf{M}}\breve{\mathbf{e}} \\
&= N(\breve{\mathbf{e}}'(\text{diag}(\breve{\boldsymbol{\pi}}))^{-1}\breve{\mathbf{e}} + \pi_{(0\cdots0)}^{-1}\breve{\mathbf{e}}'\mathbf{1}\mathbf{1}'\breve{\mathbf{e}}) \\
&= N(\breve{\mathbf{e}}'(\text{diag}(\breve{\boldsymbol{\pi}}))^{-1}\breve{\mathbf{e}} + \pi_{(0\cdots0)}^{-1}e_{(0\cdots0)}^2) \\
&= N\mathbf{e}'(\text{diag}(\boldsymbol{\pi}))^{-1}\mathbf{e},
\end{aligned}$$

which is the same as (11).