PSYCHOMETRIKA—VOL. 71, NO. 3, 587–592
SEPTEMBER 2006
DOI: 10.1007/S11336-005-1284-Z

# ON THE ASYMPTOTIC DISTRIBUTION OF PEARSON'S $X^2$ IN CROSS-VALIDATION SAMPLES

HARRY JOE

UNIVERSITY OF BRITISH COLUMBIA

ALBERT MAYDEU-OLIVARES

UNIVERSITY OF BARCELONA AND INSTITUTO DE EMPRESA

In categorical data analysis, two-sample cross-validation is used not only for model selection but also to obtain a realistic impression of the overall predictive effectiveness of the model. The latter is of particular importance in the case of highly parametrized models capable of capturing every idiosyncracy of the calibrating sample. We show that for maximum likelihood estimators or other asymptotically efficient estimators Pearson's $X^2$ is not asymptotically chi-square in the two-sample cross-validation framework due to extra variability induced by using different samples for estimation and goodness-of-fit testing. We propose an alternative test statistic, $X^2_{xval}$, obtained as a modification of $X^2$ which is asymptotically chi-square with $C - 1$ degrees of freedom in cross-validation samples. Stochastically, $X^2_{xval} \leq X^2$. Furthermore, the use of $X^2$ instead of $X^2_{xval}$ with a $\chi^2_{C-1}$ reference distribution may provide an unduly poor impression of fit of the model in the cross-validation sample.

Key words: contingency tables, item response theory modeling, latent class analysis, quadratic form statistics, goodness-of-fit.

## 1. Introduction

Cross-validation is used not only in many areas of statistics for model selection but also to obtain a realistic impression of the overall predictive effectiveness of the model. For an up-to-date account of cross-validation procedures, see Browne (2000). The simplest form of cross-validation is the two-sample procedure. In this procedure, cross-validation is performed by drawing two samples from the population of interest. The first sample (denoted as a calibration sample) is used to estimate the model parameters. Then, the goodness-of-fit of the model is computed using the second sample (denoted as the cross-validation or holdout sample) with parameters evaluated at the values estimated in the first sample.

Cross-validation is not as frequently used in categorical data analysis as in other areas of statistics. Yet, Collins, Graham, Long, & Hansen (1994) have strongly advocated its use for model selection in latent class analysis. Also, Drasgow, Levine, Tsien, Williams, & Mead (1995) have advocated its use for model selection in item response theory (IRT) modeling. Other applications of cross-validation in IRT include Chernyshenko, Stark, Chan, Drasgow, & Williams (2001) and Maydeu-Olivares (2005). Drasgow et al. (1995) promoted the use of cross-validation also to obtain a fair assessment of the overall goodness-of-fit of IRT models. As Drasgow et al. (1995) point out, this is of particular importance in the case of highly parametrized models, such as the nonparametric IRT Multilinear Formula Score (MFS) model of Levine (1994), since these models are capable of capturing every idiosyncracy of the calibration data.

The two most commonly used statistics for assessing the overall goodness-of-fit of a categorical data model are Pearson's $X^2$ and the likelihood ratio test $G^2$. When testing a composite null hypothesis (i.e., when parameters are estimated from the data) it is well known that $X^2$ is

This paper is dedicated to the memory of Michael V. Levine.

Requests for reprints should be sent to Albert Maydeu-Olivares, Faculty of Psychology, University of Barcelona, P. Valle de Hebrón, 171, 0835 Barcelona, Spain. E-mail: amaydeu@ub.edu.

preferable to $G^2$ as $X^2$ is less adversely affected by the sparseness of the contingency table than $G^2$ (Koehler & Larntz, 1980). Under multinomial sampling the asymptotic distribution of $X^2$ for testing composite null hypothesis is chi-square with $C - q - 1$ degrees of freedom, where $C$ denotes the number of cells in the contingency table and $q$ denotes the number of estimated parameters (Agresti, 2002). Also, when testing simple null hypothesis (parameters fixed at a priori values), the asymptotic distribution of $X^2$ is chi-square with $C - 1$ degrees of freedom. The two-sample cross-validation procedure, however, does not fall into any of these two frameworks.

The purpose of this short paper is to show that for maximum likelihood estimators (MLEs)or other asymptotically efficient estimators, Pearson's $X^2$ is not asymptotically chi-square in the two-sample cross-validation framework due to extra variability induced by using different samples for estimation and goodness-of-fit testing. Consequently, we propose an alternative test statistic, $X^2_{xval}$, obtained as a modification of $X^2$ which is asymptotically chi-square with $C - 1$ degrees of freedom in cross-validation samples. Furthermore, we show that stochastically $X^2_{xval} \leq X^2$. A small simulation study is included to gauge the small sample behavior of $X^2_{xval}$ and $X^2$ in the two-sample cross-validation framework. The paper concludes with an empirical example where the well-known LSAT 6 data of Bock and Lieberman (1970) is used to show that using $X^2$ instead of $X^2_{xval}$ with a $\chi^2_{C-1}$ reference distribution may provide an unduly poor impression of fit of the model in the cross-validation sample.

## 2. Asymptotic Distribution of $X^2$ in Cross-Validation Samples

For $X^2$ with the MLE or another asymptotically efficient estimator, the distribution of $X^2$ for the cross-validation sample is not distributed as a chi-square under the null hypothesis. Furthermore, the asymptotic mean of $X^2$ for the cross-validation sample is actually larger than $C - 1$.

The analysis comes from looking at the asymptotic distribution of the residual vector. Let $N_0$ be the sample size of the calibration sample, and let $N_1$ be the sample size of the cross-validation sample. Let $\mathbf{p_1}$ be the vector of sample proportions for the cross-validation sample, and let $\mathbf{p_0}$ and $\hat{\boldsymbol{\theta}}_0$ be the vectors of sample proportions and MLEs for the calibration sample.

From Bishop, Fienberg, & Holland (1975),

$$\sqrt{N_0}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{B}\sqrt{N_0}\,(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1), \qquad \mathbf{B} = \mathcal{I}^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1},$$

where $\sqrt{N_0}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, \mathcal{I}^{-1})$ as $N_0 \to \infty$. Here, $\mathbf{D} = diag(\boldsymbol{\pi}(\boldsymbol{\theta}))$, $\boldsymbol{\Delta} = \partial\boldsymbol{\pi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$, and $\mathcal{I} = \mathcal{I}(\boldsymbol{\theta}) = \boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta}$ is the Fisher information matrix. Then, the residual vector is

$$\begin{aligned}\sqrt{N_1}\,[\mathbf{p}_1 - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_0)] &\approx \sqrt{N_1}\,[\mathbf{p}_1 - \boldsymbol{\pi}(\boldsymbol{\theta}) - \boldsymbol{\Delta}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta})] \\ &\approx \sqrt{N_1}\,[\mathbf{p}_1 - \boldsymbol{\pi}(\boldsymbol{\theta})] - (N_1/N_0)^{1/2}\boldsymbol{\Delta}\mathbf{B}\sqrt{N_0}\,[\mathbf{p}_0 - \boldsymbol{\pi}(\boldsymbol{\theta})]\end{aligned}$$

with asymptotic covariance matrix

$$\boldsymbol{\Omega} = \boldsymbol{\Gamma} + (N_1/N_0)\boldsymbol{\Delta}\mathcal{I}^{-1}\boldsymbol{\Delta}',$$

as $N_0, N_1 \to \infty$, where $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\boldsymbol{\theta}) = \mathbf{D} - \boldsymbol{\pi}(\boldsymbol{\theta})\boldsymbol{\pi}'(\boldsymbol{\theta})$.

$\mathbf{D}^{-1}$ is not a generalized inverse of $\boldsymbol{\Omega}$. Also, asymptotically, the mean of $X^2 = N_1(\mathbf{p}_1 - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_0))'\mathbf{D}^{-1}(\mathbf{p}_1 - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_0))$ is

$$\mathrm{tr}(\mathbf{D}^{-1}\boldsymbol{\Omega}) = \mathrm{tr}(\mathbf{D}^{-1}\boldsymbol{\Gamma}) + (N_1/N_0) \quad \mathrm{tr}(\mathbf{D}^{-1}\boldsymbol{\Delta}\mathcal{I}^{-1}\boldsymbol{\Delta}') = C - 1 + (N_1/N_0)q. \tag{1}$$

One can get an asymptotic chi-square quadratic form by using a generalized inverse of $\boldsymbol{\Omega}$. One simple choice of $\boldsymbol{\Omega}$ is given in the next theorem.

**Theorem.** Let $\mathbf{P} = \boldsymbol{\Delta}\mathcal{I}^{-1}\boldsymbol{\Delta}'$. *A generalized inverse of* $\boldsymbol{\Omega}$ *is*

$$\boldsymbol{\Omega}^{-} = \mathbf{D}^{-1} - \epsilon\mathbf{D}^{-1}\mathbf{P}\mathbf{D}^{-1}, \quad \textit{with} \quad \epsilon = N_1/(N_0 + N_1).$$

*Proof.* Let $\gamma = N_1/N_0$. Consider a potential generalized inverse of the form

$$\mathbf{\Omega}^- = \mathbf{D}^{-1} - \mathbf{A}, \quad \mathbf{A} = \epsilon \mathbf{D}^{-1}\mathbf{P}\mathbf{D}^{-1}.$$

The appropriate value of $\epsilon$ will be solved. Since $\mathbf{\Gamma}\mathbf{D}^{-1} = \mathbf{I} - \boldsymbol{\pi}\mathbf{1}'$, $\mathbf{D}^{-1}\mathbf{\Gamma} = \mathbf{I} - \mathbf{1}\boldsymbol{\pi}'$, $\mathbf{P}\mathbf{D}^{-1}\mathbf{P} = \mathbf{P}$, and $1'\mathbf{\Delta} = \mathbf{0}'$, then

$$\begin{aligned}
\mathbf{P}\mathbf{A} &= \epsilon \mathbf{\Delta}\mathcal{I}^{-1}\mathbf{\Delta}'\mathbf{D}^{-1} = \epsilon \mathbf{P}\mathbf{D}^{-1},\\
\mathbf{A}\mathbf{P} &= \epsilon \mathbf{D}^{-1}\mathbf{\Delta}\mathcal{I}^{-1}\mathbf{\Delta}' = \epsilon \mathbf{D}^{-1}\mathbf{P},\\
\mathbf{\Gamma}\mathbf{A} &= \epsilon \mathbf{\Gamma}\mathbf{D}^{-1}\mathbf{P}\mathbf{D}^{-1} = \epsilon(\mathbf{I} - \boldsymbol{\pi}\mathbf{1}')\mathbf{P}\mathbf{D}^{-1} = \epsilon \mathbf{P}\mathbf{D}^{-1}.
\end{aligned}$$

Next, with $\mathbf{\Omega} = \mathbf{\Gamma} + \gamma\mathbf{P}$,

$$\mathbf{\Omega}(\mathbf{D}^{-1} - \mathbf{A}) = \mathbf{\Gamma}\mathbf{D}^{-1} + \gamma\mathbf{P}\mathbf{D}^{-1} - \mathbf{\Gamma}\mathbf{A} - \gamma\mathbf{P}\mathbf{A} = \mathbf{I} - \boldsymbol{\pi}\mathbf{1}' + \gamma\mathbf{P}\mathbf{D}^{-1} - \epsilon\mathbf{P}\mathbf{D}^{-1} - \gamma\epsilon\mathbf{P}\mathbf{D}^{-1} \quad (2)$$

because

$$\mathbf{\Gamma}\mathbf{A} = \epsilon\mathbf{\Gamma}\mathbf{D}^{-1}\mathbf{P}\mathbf{D}^{-1} = \epsilon(\mathbf{I} - \boldsymbol{\pi}\mathbf{1}')\mathbf{P}\mathbf{D}^{-1} = \epsilon\mathbf{P}\mathbf{D}^{-1} - \epsilon\boldsymbol{\pi}\mathbf{1}'\mathbf{\Delta}\mathcal{I}^{-1}\mathbf{\Delta}'\mathbf{D}^{-1} = \epsilon\mathbf{P}\mathbf{D}^{-1}.$$

The right-hand side of (2) is equal to $\mathbf{I} - \boldsymbol{\pi}\mathbf{1}'$ if $\epsilon = \gamma/(1+\gamma) = N_1/(N_0 + N_1)$. For this choice of $\epsilon$, since $\mathbf{1}'\mathbf{\Gamma} = \mathbf{1}'\mathbf{P} = \mathbf{0}'$,

$$\mathbf{\Omega}(\mathbf{D}^{-1} - \mathbf{A})\mathbf{\Omega} = (\mathbf{I} - \boldsymbol{\pi}\mathbf{1}')(\mathbf{\Gamma} + \gamma\mathbf{P}) = \mathbf{\Gamma} + \gamma\mathbf{P} = \mathbf{\Omega},$$

so that $\mathbf{\Omega}^-$ is in fact a generalized inverse of $\mathbf{\Omega}$. □

The above results imply that $\mathbf{\Omega}\mathbf{\Omega}^- = \mathbf{I} - \boldsymbol{\pi}\mathbf{1}'$ is idempotent, and its rank is equal to its trace: $\mathrm{tr}(\mathbf{\Omega}\mathbf{\Omega}^-) = \mathrm{tr}(\mathbf{I} - \boldsymbol{\pi}\mathbf{1}') = C - 1$. We define the cross-validation goodness-of-fit statistic as

$$\begin{aligned}
X^2_{xval} &= N_1[\mathbf{p}_1 - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_0)]'\mathbf{\Omega}^-(\hat{\boldsymbol{\theta}}_0)[\mathbf{p}_1 - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_0)] = X^2 - N_1[\mathbf{p}_1 - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_0)]'\mathbf{A}(\hat{\boldsymbol{\theta}}_0)[\mathbf{p}_1 - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_0)]\\
&= X^2 - N_1^2(N_0 + N_1)^{-1}[\mathbf{p}_1 - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_0)]'\mathbf{D}^{-1}\mathbf{\Delta}\mathcal{I}^{-1}\mathbf{\Delta}'\mathbf{D}^{-1}[\mathbf{p}_1 - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_0)]. \quad (3)
\end{aligned}$$

The matrix of the quadratic form in (3) is evaluated at $\hat{\boldsymbol{\theta}}_0$. Under the null hypothesis that $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$, $X^2_{xval}$ is asymptotically $\chi^2_{C-1}$ as $N_0, N_1 \to \infty$. Note that as $N_0 \to \infty$ with $N_1$ fixed, then $X^2_{xval} - X^2 \to 0$.

## 3. Simulation Study

In this section we report on a small simulation study to illustrate the finite sample performance of $X^2_{xval}$ against $X^2$. Multivariate binary data were generated using the following model

$$\Pr(Y_1 = y_1, \cdots, Y_n = y_n) = \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^{n} [\Pr(Y_i = 1 \mid \theta = t)]^{y_i} [1 - \Pr(Y_i = 1 \mid \theta = t)]^{(1-y_i)} \right\} \phi(\theta)\, d\theta, \quad (4)$$

where $\phi(\theta)$ is the standard normal density, and

$$\Pr(Y_i = 1 \mid \theta = t) = \left[1 + e^{-(\alpha_i + \beta_i t)}\right]^{-1}.$$

This is the two-parameter logistic model with a normally distributed latent trait (e.g., Lord & Novick, 1968).

To get simulated contingency tables that are not sparse, so that the empirical Type I error rates match their expected rates under their asymptotic null distribution for the calibration sample, we used small values of $n$ such as 5 or 6. The pattern was similar over various choices of the parameters $\alpha_i$, $\beta_i$, so below we provide representative results for just one parameter vector with $n = 5$ and $C = 2^5 = 32$.

TABLE 1.
Small sample distribution for $X^2$ in calibration samples under a two-parameter logistic model: Mean, variance, and exceedances of asymptotic upper 0.2, 0.1, 0.05, 0.01 quantiles. The asymptotic mean and variance for $X^2$ are 21 and 42.

| $N$ | Statistic | Mean | Var. | $\alpha = .2$ | $\alpha = .1$ | $\alpha = .05$ | $\alpha = .01$ |
|---|---|---|---|---|---|---|---|
| 500 | $X^2$ | 21.0 | 40.8 | .19 | .09 | .05 | .01 |
| 1000 | $X^2$ | 21.0 | 41.5 | .20 | .10 | .05 | .01 |
| 5000 | $X^2$ | 21.0 | 43.5 | .20 | .10 | .05 | .01 |

*Note*: $10^4$ replications. $(\alpha; \beta) = (-1, -.5, 0, .5, 1; 1, 1.2, 1.4, 1.6, 1.8)$.

Samples with $N = N_0 = N_1 = 500$, 1000, and 5000 were simulated, that is, the calibration and validation sample sizes were equal. We had also tried $N_0/N_1$ in a ratio of $1/2$ to 2, and the pattern was similar. We used the first calibration sample for parameter estimation using (marginal) maximum likelihood. For the overall goodness-of-fit of the model we computed $X^2$ and $X^2_{xval}$ in the second cross-validation sample, evaluating the matrix of the quadratic form at the parameter estimates obtained in the first sample.

Table 1 has the mean, variance, and empirical rejection rates at $\alpha = 0.20, 0.10, 0.05, 0.01$ for $X^2$ in the calibration samples, whereas the results for $X^2$ and $X^2_{xval}$ in the cross-validation samples are shown in Table 2. From Table 1, $X^2$ has a small sample distribution close to its asymptotic one in the calibration samples. From Table 2, based on a reference distribution of $\chi^2_{C-1}$, $X^2$ rejects the model far too often in the cross-validation samples (for all the values of $N$); note that the mean of $X^2$ is as expected from (1). The empirical distribution of $X^2_{xval}$ is closer to $\chi^2_{C-1}$. For the model in (4), there is a tendency for the rejection rate, based on $X^2_{xval}$, to be a little too high for $N = 500$ or 1000. The slowness of the asymptotic convergence rate may be due to the estimation in the matrix of the quadratic form, and the rate is likely to be model dependent. For another simpler model with fewer cells and parameters, the convergence rate was faster.

TABLE 2.
Corresponding small sample distribution for $X^2$ and $X^2_{xval}$ in cross-validation samples under a two-parameter logistic model: Mean, variance, and exceedances of asymptotic upper 0.2, 0.1, 0.05, 0.01 quantiles. The asymptotic mean and variance for $X^2_{xval}$ are 31 and 62. A reference $\chi^2_{31}$ distribution was used for $X^2$.

| $N$ | Statistic | Mean | Var. | $\alpha = .2$ | $\alpha = .1$ | $\alpha = .05$ | $\alpha = .01$ |
|---|---|---|---|---|---|---|---|
| 500 | $X^2$ | 42.2 | 154 | .61 | .47 | .36 | .19 |
| | $X^2_{xval}$ | 31.9 | 81 | .24 | .14 | .08 | .03 |
| 1000 | $X^2$ | 41.6 | 139 | .60 | .46 | .35 | .17 |
| | $X^2_{xval}$ | 31.5 | 71 | .22 | .12 | .07 | .02 |
| 5000 | $X^2$ | 41.2 | 126 | .60 | .45 | .33 | .16 |
| | $X^2_{xval}$ | 31.2 | 64 | .21 | .11 | .05 | .01 |

## 4. Data Example

In this section we provide a numerical example to illustrate our results. In this example we use the well-known LSAT 6 data set from Bock and Lieberman (1970). This data set consists of $N = 1000$ observations on $n = 5$ binary variables. The MULTILOG manual (see Du Toit, 2003) reports the application of the one-, two-, and three-parameter logistic models to these data. In the one-parameter logistic (1PL) model the slopes of the two-parameter logistic(2PL) model are constrained to be equal to each other. In the three-parameter logistic (3PL) model

$$\Pr(Y_i = 1 \mid \theta = t) = c_i + (1 - c_i)\left[1 + e^{-(\alpha_i + \beta_i \theta)}\right]^{-1}.$$

Here, we fitted by (marginal) maximum likelihood the 1PL, 2PL, and 3PL models using MULTILOG 7 (Thissen, Chen, & Bock, 2003) as in the example scripts except that the number of quadrature points was increased to 51 and the number of EM cycles to 50. We obtained $X^2 = 17.98$ on 16 degrees of freedom ($p = 0.33$) for the 3PL model, $X^2 = 18.14$ on 21 degrees of freedom ($p = 0.64$) for the 2PL model, and $X^2 = 18.34$ on 25 degrees of freedom ($p = 0.83$) for the 1PL model. Thus, the more parsimonious 1PL model is clearly to be preferred.

For this example, the LSAT 6 data were randomly split into two samples of 500 observations each. We used the first sample to estimate the parameters of all three models using MLEs. Then, we assessed the overall goodness-of-fit of the models by computing $X^2$ and $X^2_{xval}$ in the second sample evaluating the parameters at the parameter estimates obtained in the first sample. In Table 3, we provide the values of the $X^2$ statistics in the calibration sample, as well as the values of the $X^2$ and $X^2_{xval}$ statistics in the cross-validation sample. The $\chi^2_{C-1}$ reference distribution for $X^2$ is included for comparison only, as the results in sections 2 and 3 indicate that it is incorrect.

TABLE 3.
GOF statistics for calibration and cross-validation subsamples for LSAT 6 data for three models.

| | Calibration sample | | | | | |
|---|---|---|---|---|---|---|
| | $X^2$ | $df$ | $p$ | | | |
| 3PL | 22.50 | 16 | .13 | | | |
| 2PL | 23.30 | 21 | .33 | | | |
| 1PL | 23.95 | 25 | .52 | | | |
| | Cross-validation | | | | | |
| | $X^2$ | $df$ | $p$ | $X^2_{xval}$ | $df$ | $p$ |
| 3PL | 55.68 | 31 | $<$ .01 | 33.37 | 31 | .35 |
| 2PL | 53.95 | 31 | .01 | 34.91 | 31 | .29 |
| 1PL | 43.00 | 31 | .07 | 32.54 | 31 | .39 |

As can be seen in this table, the effect of using $X^2$ versus $X^2_{xval}$ in the cross-validation sample is rather dramatic. When $X^2$ is used in the cross-validation sample with a $\chi^2_{C-1}$ reference distribution the 3PL and 2PL models are rejected and the 1PL model is barely accepted at the $\alpha = 0.05$ significance level. On the other hand, when $X^2_{xval}$ is used none of the three models can be rejected. Furthermore, in both samples, the 1PL model is to be preferred to the 2PL and 3PL models.

## 5. Discussion and Conclusions

Often researchers successively modify their model to improve its fit in the calibration sample. Such data snooping results in selection bias (Zucchini, 2000) leading to overly optimistic goodness-of-fit statistics in the calibration sample. In this situation, model fit assessment in a fresh, holdout sample, using the two-sample cross-validation procedure provides a more accurate impression of the overall goodness-of-fit of the model (Browne, 2000). Also, Drasgow et al. (1995) have advocated the use of two-sample cross-validation to obtain a fair assessment of the overall predictive effectiveness of highly parametrized IRT models. Here, we have shown that for MLEs or other asymptotically efficient estimators Pearson's $X^2$ is not asymptotically chi-square in the cross-validation sample. Also, we have proposed a modification $X^2_{xval}$ of $X^2$, which is asymptotically chi-square with $C-1$ degrees of freedom in cross-validation samples. Stochastically, $X^2_{xval} \leq X^2$, and, as our small simulation study and numerical example suggest, using $X^2$ instead of $X^2_{xval}$ may result in an unduly poor impression of fit of the models considered.

## Acknowledgments

Harry Joe was supported by an NSERC Canada grant. Albert Maydeu-Olivares was supported by the Deptartment. of Universities, Research and Information Society (DURSI) of the Catalan Government, and by grants BSO2000-0661 and BSO2003-08507 from the Spanish Ministry of Science and Technology.

## References

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

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

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

Browne, M.W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, ***44***, 108–132.

Chernyshenko, O.S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, ***36***, 523–562.

Collins, L.M., Graham, J.W., Long, J.D., & Hansen, W.B. (1994). Crossvalidation of latent class models of early substance use onset. *Multivariate Behavioral Research*, ***29***, 165–183.

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

Du Toit, M. (ed.) (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International.

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

Levine, M.V. (1984). An introduction to multilinear formula score theory. *Measurement series 84-4*. Champaign, IL: Model Based Measurement Laboratory.

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

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

Thissen, D., Chen, W.-H., & Bock, R.D. (2003). *Multilog (version 7) [Computer software]*. Lincolnwood, IL: Scientific Software International.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, ***44***, 41–61.

*Manuscript received 20 JAN 2005*
*Final version received 30 MAY 2005*
*Published Online Date: 25 AUG 2006*