Educational and Psychological Measurement
73(2) 254–274

DOI: 10.1177/0013164412453841
epm.sagepub.com
SAGE

Local Dependence Diagnostics in IRT Modeling of Binary Data

Yang Liu[1] and Alberto Maydeu-Olivares[2]

Abstract

Local dependence (LD) for binary IRT models can be diagnosed using Chen and Thissen's bivariate X^2 statistic and the score test statistics proposed by Glas and Suárez-Falcón, and Liu and Thissen. Alternatively, LD can be assessed using general purpose statistics such as bivariate residuals or Maydeu-Olivares and Joe's M_r statistic. The authors introduce a new general statistic for assessing the source of model misfit, R_2, and compare its performance to the above statistics using a simulation study. Results suggest that the bivariate and trivariate X^2 statistics have unacceptable Type I error rates. As for the remaining statistics, if their computation involves the information matrix (bivariate residuals and score tests), they show good power; if not (M_r and R_2), they lack power. Of course, the performance of the bivariate residuals and score tests depends on how the information matrix is approximated.

Keywords

item response model, local dependence, limited information goodness-of-fit statistics

Introduction

Item response theory (IRT) refers to a set of models for discrete data that are widely used in educational, psychological, and medical assessment applications (Thissen & Steinberg, 2009). In particular, these models posit that the observed responses to a set of discrete items can be accounted for by a small number of latent traits (i.e.,

[1]The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[2]University of Barcelona, Barcelona, Spain

Corresponding Author:
Yang Liu, Department of Psychology, The University of North Carolina at Chapel Hill, 352 Davie Hall, Chapel Hill, NC 27599, USA.
Email: liuy0811@live.unc.edu

unobserved continuous underlying variables). In this article, we shall restrict ourselves to IRT models for binary data involving a single underlying latent trait.

One of the defining assumptions of IRT models is the conditional independence assumption (also known as local independence; see McDonald, 1982). This assumption states that the conditional probability of observing a response pattern given a particular latent trait value equals the product of the items' conditional probabilities.

The violation of this assumption is generally referred to as local dependence (LD) and cannot be tested directly because the latent trait is not observed. Furthermore, violations of the local independence assumption cannot be isolated from the other model assumptions. Rather, all the model assumptions (local independence, dimensionality, specification of the item characteristic curve, and specification of the latent trait density) are tested simultaneously when using an overall goodness-of-fit statistic for multivariate discrete data such as Pearson's X^2, or the likelihood ratio test statistic. Due to data sparseness, these statistics can only be used with very small models (i.e., consisting of a few items; Bishop, Fienberg, & Holland, 1975). Fortunately, the overall limited information test statistics M_r proposed recently (Maydeu-Olivares & Joe, 2005, 2006; usually $r = 2$ or 3) are able to overcome the problem of data sparseness and can be used in realistic size applications.

When a particular model shows misfit using an overall goodness-of-fit test, and LD is the suspected culprit, researchers are interested in locating the source of misfit to take remedial action, for instance, by removing certain items or by modifying the model. A number of statistics have been proposed to provide information about local dependencies among item subsets (usually pairs or triplets; e.g., Chen & Thissen, 1997; Glas & Suárez-Falcón, 2003; Liu & Thissen, 2012; Yen, 1984). Alternatively, one can use general purpose statistics to assess the model misfit to subsets of items (e.g., Maydeu-Olivares & Joe, 2006; Reiser, 1996).

In this article, the performance of some of these statistics in detecting LD in IRT models for binary data is evaluated under a variety of simulated conditions. More specifically, the statistics considered are the following: (a) Pearson's X^2 (Chen & Thissen, 1997) statistic for pairs of items; (b) Glas and Suárez-Falcón's (2003) score test statistic for pairs of items; (c) Liu and Thissen's (2012) score test statistic for pairs of items; (d) Maydeu-Olivares and Joe's (2006) M_3 statistic for triplets of items; (e) Pearson's X^2 statistic for triplets of items; (f) standardized bivariate residuals (Maydeu-Olivares & Joe, 2005; Reiser, 1996); and (g) a new sum-score-based statistic for pairs of items, R_2, inspired in previous work by Glas (1988) and Thissen and Orlando (2000).

The remaining part of this article is organized as follows: First, we describe each of these statistics. Next, we examine their behavior under the null hypothesis (i.e., when the fitted model holds in the population). This is the most important condition, as we do not want to remove well-fitting items since developing items is in general expensive. Next, we examine the behavior of these statistics under two different alternatives that involve violations of local independence: (a) a bifactor structure and (b) independent clusters multidimensional structure. For the ease of exposition, the fitted

model is in all cases the 2-parameter logistic (2PL) model. Also, note that we restrict ourselves to parametric methods. There are nonparametric procedures for detecting LD such as DIMTEST (see Stout, 1987) that we do not consider here.

A Review of Existing Parametric LD Diagnostics

Consider a set of J binary items $\mathbf{Y} = (Y_1, \ldots, Y_J)$ of which the test is composed; $Y_j = y_j = \{0, 1\}$ for all j. For any unidimensional (i.e., single latent trait) IRT model, the probability of observing one of the possible 2^J response patterns $\mathbf{y} = (y_1, \ldots, y_J)$ is

$$\pi_{\mathbf{y}} = \Pr(Y_1 = y_1, \ldots, Y_J = y_J) = \int_{-\infty}^{\infty} \prod_{j=1}^{J} \Pr(y_j|\theta)\phi(\theta)d\theta. \quad (1)$$

In Equation 1, θ denotes the latent trait, and $\phi(\theta)$ its density, which is often assumed to be standard normal. The conditional probability of endorsing the item given a particular latent trait value, $\Pr(y_j = 1|\theta)$, is generally referred to as the item characteristic curve (ICC), and it is generally assumed to be monotonically increasing. One commonly used ICC is the 2PL model (Lord & Novick, 1968),

$$\Pr(y_j = 1|\theta) = \frac{1}{1 + \exp(-\alpha_j - \beta_j\theta)}. \quad (2)$$

In Equation 2, the intercept α_j and slope β_j are related to the item's difficulty and discrimination, respectively.

The conditional independence assumption is embedded in Equation 1. It is given by

$$\Pr(y_1, \ldots, y_J|\theta) = \prod_{j=1}^{J} \Pr(y_j|\theta). \quad (3)$$

One of the earliest test statistics specifically proposed to diagnose LD (i.e., the violation of Equation 3) is Yen's (1984) Q_3 statistic. Let N denote sample size. Also, let $r_{ij} = y_{ij} - \Pr(y_{ij} = 1|\hat{\theta}_i)$ be the residual of subject i ($i = 1, \ldots, N$) on item j conditional on some estimated latent score, where y_{ij} is the actual response, and $\hat{\theta}_i$ is the estimated latent score for subject i. Then Yen's Q_3 statistic is the correlation between these residuals for items j and k. That is,

$$Q_{3,jk} = \frac{\sum_{i=1}^{N} (r_{ij} - \bar{r}_j)(r_{ik} - \bar{r}_k)}{\sqrt{\sum_{i=1}^{N} (r_{ij} - \bar{r}_j)^2}\sqrt{\sum_{i=1}^{N} (r_{ik} - \bar{r}_k)^2}}, \quad (4)$$

in which $\bar{r}_j$ and $\bar{r}_k$ denote the average of r_{ij} and r_{ik} across subjects.

Yen (1984) suggested treating the residuals r_{ij} as normally distributed and apply Fisher's r-to-z transformation to obtain an approximate p-value for Q_3. However, a simulation study performed by Chen and Thissen (1997) revealed that the suggested reference distribution does not work well.

As an alternative, Chen and Thissen (1997) proposed applying Pearson's X^2 statistic to each pair of items j and k:

$$X_{jk}^2 = N \sum_{y_j=0}^{1} \sum_{y_k=0}^{1} \frac{(p_{y_j y_k} - \hat{\pi}_{y_j y_k})^2}{\hat{\pi}_{y_j y_k}}. \tag{5}$$

In Equation 5, $\pi_{y_j y_k} = \Pr(Y_j = y_j, Y_k = y_k)$ denotes a bivariate cell probability under the model, and $p_{y_j y_k}$ denotes its corresponding sample proportion. In the case of unidimensional IRT models,

$$\pi_{y_j y_k} = \int_{-\infty}^{\infty} \Pr(Y_j = y_j|\theta)\Pr(Y_k = y_k|\theta)\phi(\theta)d\theta. \tag{6}$$

As reference distribution for this statistic, Chen and Thissen suggested a χ^2 distribution with the degrees of freedom corresponding to an independence test. For binary data, this amounts to using a χ_1^2 reference distribution. However, in a simulation study using 2PL model, Liu and Thissen (2012) had shown that the empirical distribution of the bivariate X^2 is stochastically smaller than a χ_1^2 distribution; in other words, the use of Chen and Thissen's suggested reference distribution led to underrejecting the model when it is correctly specified.

Pearson's X^2 statistic is closely related to the Wald test (Bishop et al., 1975). As an alternative to the Wald test, one can use Rao's score test (e.g., Lehmann, 1999)—also known as the Lagrange multiplier test. Score tests require the specification of an alternative model; for detecting LD, Glas and Suárez-Falcón (2003) proposed the use of a threshold shift model. Suppose we suspect that LD might exist with the pair of items (*j, k*), the ICC for one item (*k* as shown here, usually the one that appears later in the test) can be written as

$$\Pr(y_k = 1|\theta) = \frac{1}{1 + \exp(-\alpha_k - \beta_k\theta - \delta_{jk}y_j)}, \tag{7}$$

while all other items, including item j, still have 2PL ICC. It is not hard to see that the threshold shift model only has one more parameter than the 2PL model, the shift parameter δ_{jk}. Indeed, the model reduces to the locally independent 2PL model when $\delta_{jk} = 0$.

Assume again we have an N-observation sample of binary item responses $\{\mathbf{y}_i\}_{i=1}^N$, and let $\boldsymbol{\eta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta_{jk})$, where $\boldsymbol{\alpha}$ is the vector of all intercepts and $\boldsymbol{\beta}$ the vector of all slopes. The likelihood function for the threshold shift model can be written as

$$L(\boldsymbol{\eta}; \mathbf{y}_1, \ldots, \mathbf{y}_N) = \prod_{i=1}^{N} \pi_{\mathbf{y}_i}, \tag{8}$$

Then the null hypothesis $H_0 : \delta_{jk} = 0$, corresponding to local independence, can be tested via the score test statistic

$$S = \frac{1}{N} \mathbf{g}(\hat{\boldsymbol{\eta}}_0)' \mathcal{I}(\hat{\boldsymbol{\eta}}_0)^{-1} \mathbf{g}(\hat{\boldsymbol{\eta}}_0). \tag{9}$$

In Equation 9, $\mathbf{g}(\hat{\boldsymbol{\eta}}_0)$ and $\mathcal{I}(\hat{\boldsymbol{\eta}}_0)$ denote the score vector and Fisher's information matrix of all item parameters under the alternative model (i.e., threshold shift model), but evaluated at $\hat{\boldsymbol{\eta}}_0$—the item parameters estimated under the 2PL model. Therefore, the score test statistic with the aforementioned threshold shift alternative has an asymptotic χ_1^2 distribution when the locally independent 2PL model is true. In practice, the Fisher information matrix for the full set of item parameters is often replaced by its consistent estimator—the cross-product approximation (see Kendall & Stuart, 1961) evaluated at the estimated 2PL parameters (Liu & Thissen, 2012).

Liu and Thissen (2012) proposed to use another alternative for the bivariate score test—a bifactor LD model. For candidate pair (*j, k*), their model is specified as following:

$$\Pr(y_j = 1 | \theta, \xi) = \frac{1}{1 + \exp(-\alpha_j - \beta_j \theta - \beta_{jk} \xi)}, \tag{10}$$

$$\Pr(y_k = 1 | \theta, \xi) = \frac{1}{1 + \exp(-\alpha_k - \beta_k \theta \pm \beta_{jk} \xi)}. \tag{11}$$

In these equations, ξ is a secondary latent variable only related to items *j* and *k*, which leads to conditional dependence between the responses, and β_{jk} is the secondary slope parameter, which is constrained to be equal in the two ICCs for identification purpose. Again, we assume that the 2PL holds for all remaining items. The corresponding score test statistic also has approximately χ_1^2 distribution under local independence, since the bifactor LD model reduces to a 2PL when $\beta_{jk} = 0$ and they only differ by one parameter for each pair of items.

In the sequel, to distinguish the two score test statistics, we denote by S_t the score statistic that uses the threshold shift model as alternative, and by S_b the score statistic that uses the bifactor model as alternative.

Recently, Maydeu-Olivares and Joe (2005, 2006) have introduced a general framework for goodness-of-fit testing in multivariate discrete data. For assessing the source of misfit in poorly fitting models, they proposed two methods: (a) using M_r statistics for marginal subtables (single items, pairs of items, triplets of items); (b) following Reiser (1996), using standardized bivariate residuals. One question that remains to be addressed is whether statistics specifically designed to assess LD are needed, or whether general all-purpose test statistics such as those proposed in Maydeu-Olivares

and Joe (2005, 2006) suffice. Using the simulation results and a numerical example, the current article aims at providing a preliminary answer to this question.

Using All-Purpose Goodness-of-Fit Statistics to Diagnose LD

First, we review some of the relevant theory. Throughout this section, a notation similar to that of Maydeu-Olivares and Joe (2005, 2006) is adopted. Let $\boldsymbol{\pi}$ be the column vector containing the probability of all 2^J possible response patterns, $\mathbf{p}$ be the vector of observed cell proportions, and N be the sample size. We assume the response vector $\mathbf{Y} \sim \text{Multinomial}(1, \boldsymbol{\pi})$. By the central limit theorem (Bishop et al., 1975), $\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = \mathbf{D} - \boldsymbol{\pi}\boldsymbol{\pi}'$, with $\mathbf{D} = \text{Diag}(\boldsymbol{\pi})$.

Now, we assume 2PL model holds, and let q denote the number of parameters to be estimated (i.e., slopes and intercepts). The overall null hypothesis that the 2PL model holds can be written as $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ versus $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Under the null hypothesis, and if maximum likelihood estimation is used, the residual vector $\hat{\mathbf{e}} = \mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ has the following asymptotic distribution:

$$\sqrt{N}\hat{\mathbf{e}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma} - \boldsymbol{\Delta}\mathcal{I}^{-1}\boldsymbol{\Delta}'), \tag{12}$$

where $\mathcal{I} = \boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta}$ is the Fisher information matrix for the 2PL model parameters, and $\boldsymbol{\Delta} = (\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\alpha}'}, \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\beta}'})$ is the $2^J \times q$ Jacobian matrix of derivatives of the cell probabilities with respect to the item parameters.

Equation 12 describes the asymptotic distribution of all the cell residuals under maximum likelihood estimation. Now, consider a $t \times 2^J$ matrix of constants $\mathbf{T}_\kappa$ ($t \leq 2^J - 1$). $\mathbf{T}_\kappa$ can be for instance a $2^2 \times 2^J$ matrix of 1s and 0s that yields the bivariate marginal probability vector for item pair (*j, k*), say $\mathbf{T}_{(jk)}$; or it can be $2^3 \times 2^J$ matrix yielding the trivariate marginal probability vector for item triplet (*j, k, l*), say $\mathbf{T}_{(jkl)}$. Then, $\sqrt{N}(\mathbf{p}_\kappa - \boldsymbol{\pi}_\kappa) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Xi}_\kappa)$ is the asymptotic distribution of the sample statistics where $\mathbf{p}_\kappa = \mathbf{T}_\kappa\mathbf{p}$, $\boldsymbol{\pi}_\kappa = \mathbf{T}_\kappa\boldsymbol{\pi}$, and $\boldsymbol{\Xi}_\kappa = \mathbf{T}_\kappa\boldsymbol{\Gamma}\mathbf{T}_\kappa'$. Similarly,

$$\sqrt{N}\,\hat{\mathbf{e}}_\kappa \overset{\text{def}}{=} \sqrt{N}(\mathbf{p}_\kappa - \boldsymbol{\pi}_\kappa(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_\kappa), \tag{13}$$

where

$$\boldsymbol{\Sigma}_\kappa = \boldsymbol{\Xi}_\kappa - \boldsymbol{\Delta}_\kappa\mathcal{I}^{-1}\boldsymbol{\Delta}_\kappa' \tag{14}$$

is the asymptotic distribution of the residuals with $\boldsymbol{\Delta}_\kappa = \mathbf{T}_\kappa\boldsymbol{\Delta}$ being the $t \times q$ matrix of derivatives of $\boldsymbol{\pi}_\kappa$ with respect to the item parameters.

Notice that terms in Equation 14 depend on the true values of parameters that we do not know in most cases. To get an estimate of $\boldsymbol{\Sigma}_\kappa$, one needs to replace the true parameter values by their consistent estimates (e.g., MLE); the convergence results remains unchanged by Slutsky's theorem. For the rest of this article, we denote the elements computed with MLE of item parameters by a hat caret (e.g., $\hat{\boldsymbol{\Delta}}_\kappa = \boldsymbol{\Delta}_\kappa(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$).

Standardized Bivariate Residuals

Maydeu-Olivares and Joe (2006) considered a special case of this general framework in which $\mathbf{T}_\kappa$ maps the vector of cell probabilities into the set of all moments up to order r. There are J univariate moments $\dot{\pi}_j = \Pr(Y_j = 1),\ j = 1, \ldots, J;\ \binom{J}{2} = \frac{n(n-1)}{2}$ bivariate moments $\dot{\pi}_{jk} = \Pr(Y_j = 1, Y_k = 1),\ j, k = 1, \ldots, J;$ and $\binom{J}{3} = \frac{n(n-1)(n-2)}{6}$ trivariate moments $\dot{\pi}_{jkl} = \Pr(Y_j = 1, Y_k = 1, Y_l = 1),$ and so forth.

Consider the set of all univariate and bivariate residual moments (i.e., $r = 2$). There are $t = \frac{n(n+1)}{2}$ such moments. Let $\mathbf{T}_2$ be the $\frac{n(n+1)}{2} \times 2^J$ transformation matrix that maps the cell residuals (Equation 12) onto the univariate and bivariate residual moments, which we shall denote by $\hat{\dot{\mathbf{e}}}_2 = \dot{\mathbf{p}}_2 - \dot{\boldsymbol{\pi}}_2(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$. By Equation 13, the asymptotic distribution of the vector of univariate and bivariate residual moments is $\sqrt{N}\hat{\dot{\mathbf{e}}}_2 \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_2)$, with $\boldsymbol{\Sigma}_2$ being their asymptotic covariance matrix, which is of the type given by Equation 14.

In particular, consider a single bivariate residual moment $\hat{\dot{e}}_{jk} = \dot{p}_{jk} - \hat{\dot{\pi}}_{jk}$ belonging to $\hat{\dot{\mathbf{e}}}_2$. Its asymptotic variance is the corresponding term in the diagonal of $\boldsymbol{\Sigma}_2$, which we can denote by $\dot{\sigma}^2_{jk}$. We can use these bivariate residuals in a fashion analogous to bivariate residual covariances in factor analysis. More specifically, the standardized bivariate residual

$$Z_{jk} = \frac{\sqrt{N}\hat{\dot{e}}_{jk}}{\hat{\dot{\sigma}}_{jk}} \tag{15}$$

is asymptotically distributed as a $N(0, 1)$, or equivalently $Z^2_{jk} \xrightarrow{d} \chi^2_1$, under the null hypothesis. Large absolute values of the Z_{jk} statistic indicate model misfit, which could be due to violations of local independence.

The General Family of Test Statistics M_κ and Its Special Case, the Overall Goodness-of-Fit Statistic M_2

For the general linear transformation of cell residuals $\hat{\mathbf{e}}_\kappa = \mathbf{T}_\kappa \hat{\mathbf{e}}$, Joe and Maydeu-Olivares (2010) showed that if (a) $\mathbf{T}_\kappa$ has full row rank t and the one vector $\mathbf{1}_{2^J}$ is not in its row span (condition T) and (b) $\boldsymbol{\Delta}_\kappa = \mathbf{T}_\kappa \boldsymbol{\Delta}$ has full column rank $q < t$ (condition D), the statistic M_κ

$$M_\kappa = N\, \hat{\mathbf{e}}'_\kappa\, \hat{\mathbf{U}}_\kappa\, \hat{\mathbf{e}}_\kappa \tag{16}$$

follows an asymptotic chi-squared distribution with $t - q$ degrees of freedom under H_0, where

$$\hat{\mathbf{U}}_\kappa = \hat{\boldsymbol{\Xi}}_\kappa^{-1} - \hat{\boldsymbol{\Xi}}_\kappa^{-1}\hat{\boldsymbol{\Delta}}_\kappa(\hat{\boldsymbol{\Delta}}'_\kappa \hat{\boldsymbol{\Xi}}_\kappa^{-1}\hat{\boldsymbol{\Delta}}_\kappa)^{-1}\hat{\boldsymbol{\Delta}}'_\kappa \hat{\boldsymbol{\Xi}}_\kappa^{-1}. \tag{17}$$

A special case of this family of statistics is M_2 statistic implemented in IRTPRO (Cai, Thissen, & du Toit, 2011) to assess the overall goodness-of-fit of IRT models. The M_2 is the statistic within this family with the $\mathbf{T}_\kappa = \mathbf{T}_2$ transformation matrix described earlier. For binary data, this statistic

$$M_2 = N\,\hat{\mathbf{e}}'_2\,\hat{\mathbf{U}}_2\,\hat{\mathbf{e}}_2 \xrightarrow{d} \chi^2_{n(n+1)/2-q} \tag{18}$$

under H_0, where $\hat{\mathbf{U}}_2$ is of the form (17).

Assessing the Source of Misfit Using M_3 Statistic on Trivariate Subtables

The general setup outlined above can be readily employed to assess the source of the misfit in marginal subtables. Let $\mathbf{T}_{jkl}$ be the 7×2^J matrix (i.e., $t = 7$) that maps the 2^J-dimensional vector of multinomial probabilities into the 7-dimensional vector of three univariate, three bivariate and one trivariate residual moments involving variables Y_j, Y_k, and Y_l:

$$\dot{\boldsymbol{\pi}}_{jkl} = (\,\dot{\pi}_j\,,\ \dot{\pi}_k\,,\ \dot{\pi}_l\,,\ \dot{\pi}_{jk}\,,\ \dot{\pi}_{jl}\,,\ \dot{\pi}_{kl}\,,\ \dot{\pi}_{jkl}\,)'. \tag{19}$$

Furthermore, let $\hat{\mathbf{e}}_{jkl} = \mathbf{T}_{jkl}\hat{\mathbf{e}}$, and $\hat{\mathbf{U}}_{jkl}$ is of the form given by Equation 17. Then, as a special case of the general theory presented above,

$$M_{jkl} = N\hat{\mathbf{e}}'_{jkl}\hat{\mathbf{U}}_{jkl}\hat{\mathbf{e}}_{jkl} \xrightarrow{d} \chi^2_1 \tag{20}$$

under the 2PL model. This is because for three items there are three intercepts and three slopes under 2PL model (i.e., $q = 6$). Hence, degrees of freedom of M_{jkl} are $t - q = 7 - 6 = 1$. Values of M_{jkl} larger than the critical value of its reference distribution provides evidence for the misfit of locally independent 2PL model in trivariate subtables. It should be pointed out that M_{jkl} is simply the M_3 statistic proposed by Maydeu-Olivares and Joe (2005) applied to a subset of three variables after the model parameters have been estimated using all J variables.

There is a one-to-one relationship between the seven moments involved in three binary variables $\dot{\boldsymbol{\pi}}_{jkl}$ and the 2^3 cell probabilities in the marginal subtable of three binary variables, which we shall denote by $\boldsymbol{\pi}_{(jkl)} = \mathbf{T}_{(jkl)}\boldsymbol{\pi}$. As a result, the trivariate M_{jkl} statistic can be alternatively written as a function of the cell probabilities as follows:

$$M_{jkl} = N\hat{\mathbf{e}}'_{(jkl)}\hat{\mathbf{U}}_{(jkl)}\hat{\mathbf{e}}_{(jkl)}, \tag{21}$$

where $\hat{\mathbf{e}}_{(jkl)} = \mathbf{T}_{(jkl)}\hat{\mathbf{e}}$ is the vector of trivariate cell residuals, and

$$\hat{\mathbf{U}}_{(jkl)} = \hat{\mathbf{D}}^{-1}_{(jkl)} - \hat{\mathbf{D}}^{-1}_{(jkl)}\hat{\boldsymbol{\Delta}}_{(jkl)}(\hat{\boldsymbol{\Delta}}'_{(jkl)}\hat{\mathbf{D}}^{-1}_{(jkl)}\hat{\boldsymbol{\Delta}}_{(jkl)})^{-1}\hat{\boldsymbol{\Delta}}'_{(jkl)}\hat{\mathbf{D}}^{-1}_{(jkl)}, \tag{22}$$

where $\hat{\mathbf{D}}_{(jkl)} = \mathrm{Diag}(\hat{\boldsymbol{\pi}}_{(jkl)})$, and $\hat{\boldsymbol{\Delta}}_{(jkl)}$ is the matrix of derivatives of the trivariate marginal cell probabilities with respect to slope and intercept parameters for items *j, k*, and *l*.

Notice that Equation 21 is equivalent to

$$M_{jkl} = X^2_{jkl} - N\hat{\mathbf{e}}'_{(jkl)}\hat{\mathbf{D}}^{-1}_{(jkl)}\hat{\boldsymbol{\Delta}}_{(jkl)}(\hat{\boldsymbol{\Delta}}'_{(jkl)}\hat{\mathbf{D}}^{-1}_{(jkl)}\hat{\boldsymbol{\Delta}}_{(jkl)})^{-1}\hat{\boldsymbol{\Delta}}'_{(jkl)}\hat{\mathbf{D}}^{-1}_{(jkl)}\hat{\mathbf{e}}_{(jkl)}, \tag{23}$$

where X^2_{jkl} simply denotes Pearson's X^2 statistic applied to triplet (j, k, l). Equation 23 implies that X^2_{jkl} rejects more than M_{jkl} (the difference term is nonnegative), and because the latter follows asymptotically a chi-squared distribution with 1 degree of freedom when 2PL is the true model, the use of X^2_{jkl} will then lead to incorrectly rejecting well fitting items (Maydeu-Olivares & Joe, 2006).

An obvious drawback of the M_{jkl} statistic is that triplets of variables need to be used in order to get enough degrees of freedom. This makes more difficult to draw conclusions as to which items do not fit the model, than if fit could be assessed two items at a time. To overcome this difficulty, in the next subsection we propose a new bivariate statistic $R_{2,jk}$.

Sum-Score-Based Bivariate R_2 Statistic

Since the 2PL model is not identified in bivariate marginal tables (i.e., $q = 4$ while $t - 1 = 3$), additional information should be incorporated into the test statistic to enable inferences about the pairwise fit of the model. One convenient solution is to consider the joint probability of univariate and bivariate margins with sum-score levels. Thissen and Orlando (2000) applied the same idea and derived an item-fit statistic; however, they only provided a conjecture of its reference distribution under the null hypothesis. To avoid this problem, in the present article we construct a statistic inspired by Thissen and Orlando, but that belongs to the M_κ family. This enables us to establish the asymptotic null distribution of our statistic.

The statistic we proposed, called $R_{2,jk}$, is based on a transformation matrix $\mathbf{T}_{R_{2,jk}}$ such that

$$\boldsymbol{\pi}_{R_{2,jk}} = \mathbf{T}_{R_{2,jk}}\boldsymbol{\pi} = \begin{bmatrix} \Pr(Y_j = 1, S = 1) \\ \vdots \\ \Pr(Y_j = 1, S = J - 1) \\ \Pr(Y_k = 1, S = 1) \\ \vdots \\ \Pr(Y_k = 1, S = J - 1) \\ \Pr(Y_j = Y_k = 1, S = 2) \\ \vdots \\ \Pr(Y_j = Y_k = 1, S = J - 1) \\ \Pr(S = J) \end{bmatrix}, \tag{24}$$

where S is the sum score. See Appendix A for an example of this matrix involving $J = 4$. The resulting statistic, like the Thissen and Orlandos (2000) statistic and Glas's (1988) R_2 statistic, conditions the bivariate marginal subtables on the observed

summed score, thereby removing the undesirable dependence on latent trait estimates in Yen's Q_3 statistic.

It can be checked that the transformation matrix $\mathbf{T}_{R_{2,jk}}$ satisfies conditions T and D of Joe and Maydeu-Olivares (2010) and, therefore, the bivariate $R_{2,jk}$ can be computed as an M_κ statistic by Equations 16 and 17 using this transformation matrix. For a pair of binary items, the number of statistics involved in the computation of $R_{2,jk}$ is $t = 2(J-1) + (J-2) + 1 = 3(J-1)$. Thus, $R_{2,jk}$ follows asymptotically a chi-squared distribution with $3(J-1) - q$ degrees of freedom, where q is the number of all item parameters; in particular, for the 2PL, $q = 2J$. This is because the joint probabilities of univariate and bivariate margins with the sum-score patterns depend on all item parameters. As one of the reviewers pointed out, when the number of items is large and the computation of all response patterns is infeasible, some iterative algorithm (e.g., Lord & Wingersky, 1984) should be used to compute the model-implied probability vector $\boldsymbol{\pi}_{R_{2,jk}}$.

The statistic $R_{2,jk}$ is also similar to the R_2 statistic proposed by Glas's (1988) to assess the overall goodness-of-fit of the Rasch model. The differences between his statistic and ours are that our statistic only considers a pair of items and that it excludes the term $\Pr(S = 0)$ from the computations. The former renders it a pairwise diagnostic statistic; the latter ensures that the resulting statistic satisfies the conditions to be an M_κ statistic, which further enables us to establish its asymptotic distribution.

There is one potential problem with this new sum-score-based statistic: when the number of items is large, the summary statistics given by Equation 24 might still suffer from sparseness, which may result in flawed asymptotic behavior (Joe & Maydeu-Olivares, 2010). When this occurs, one should use sum score ranges instead of levels (Glas, 1988). This reduces the number of summary statistics and as a result each of them has larger expected probabilities. However, the number of sum score groups is to be chosen such that the number of statistics is greater than the number of all item parameters. Otherwise, $R_{2,jk}$ will have negative degrees of freedom.

Simulation Study

The empirical distribution of a number of LD statistics was investigated first under correct model specification and also under model violations that lead to local dependencies. In all cases the fitted model was the 2PL model. The first simulation study involved fitting a 2PL to data generated using this model. Ten items were used with two conditions of sample size (300 and 1,000 respondents). One thousand replications per condition were used. The true parameter values were

$$\boldsymbol{\beta} = (1.28, 1.67, 2.27, 1.67, 1.28, 1.28, 1.67, 2.27, 1.67, 1.28)', \tag{25}$$

$$\boldsymbol{\alpha} = (0, 1.19, 2.84, -1.19, 0, -2.13, 0, 1.42, 0, 2.13)'. \tag{26}$$

These parameters correspond to the following factor loading ($\boldsymbol{\lambda}$) and threshold ($\boldsymbol{\tau}$) values on a normal ogive scale (Forero & Maydeu-Olivares, 2009; Wirth & Edwards, 2007):

$$\boldsymbol{\lambda} = (0.6, 0.7, 0.8, 0.6, 0.7, 0.8, 0.6, 0.7, 0.8, 0.6)', \tag{27}$$

$$\boldsymbol{\tau} = (0, 0.5, 1, -0.5, 0, -1, 0, 0.5, 0, 1)'. \tag{28}$$

The second simulation study involved generating data under two different alternative models: (a) a bifactor LD triplet model and (b) a 2-factor independent cluster model (with factor intercorrelation $\rho = 0.3$). Only 1,000-observation condition was used for power runs. For both models, the same intercepts as in equation (26) were used. The true slope values for the two models were, respectively,

$$\boldsymbol{\beta}_1 = \begin{bmatrix} 1.28 & 1.67 & 2.27 & 1.67 & 1.28 & 1.28 & 1.67 & 2.27 & 1.67 & 1.28 \\ 0.98 & -0.98 & 0 & 0.98 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{29}$$

$$\boldsymbol{\beta}_2 = \begin{bmatrix} 1.28 & 1.67 & 2.27 & 1.67 & 1.28 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.28 & 1.67 & 2.27 & 1.67 & 1.28 \end{bmatrix}. \tag{30}$$

In all cases, parameter estimation was performed using IRTPRO (Cai et al., 2011), and the statistics were computed using R (R Development Core Team, 2010) from the IRTPRO output.

The statistics compared were: (a) for pairs: Chen-Thissen X^2, standardized bivariate residuals Z (i.e., Z_{jk}) using expected information, two score test statistics S_t and S_b using observed information, and R_2 (i.e., $R_{2,jk}$); (b) for triplets: M_3 (i.e., M_{jkl}), and X^2 (i.e., X^2_{jkl}).

The Chen-Thissen X^2 and X^2_{jkl} are the same statistic, Pearson's X^2 statistic, except that the former is applied to pairs of variables, whereas the latter is applied to triplets of variables. Also, the reference distribution, χ^2_1, employed in both cases is the same but based on different rationales: for bivariate X^2, χ^2_1 is the reference for an independence model; for trivariate X^2, χ^2_1 is the reference for M_3, which can be regarded as an adjustment to the trivariate X^2 (see Equation 23), and hence its inclusion enables us to gauge the need for such adjustment. The bivariate X^2 with an independence reference distribution (i.e., Chen and Thissen's proposal) is probably the most widely used statistic to assess LD.

As mentioned, M_{jkl} is simply Maydeu-Olivares and Joe's (2005) M_3 statistic applied to a triplet of variables, whereas $R_{2,jk}$ is only inspired, but should not be confused with Glas's (1988) R_2 statistic. Finally, S_t and S_b are both score test statistic (i.e., Equation 9), but computed under different alternative hypotheses.

All statistics considered but X^2 have known asymptotic chi-squared distributions under the null hypothesis. Degrees of freedom is 1 for the score tests S_t and S_b, the bivariate residuals Z, and the M_3 statistic; degrees of freedom for the R_2 statistic are 7.

Table 1. Simulation Results Under a 2PL Model: $N = 300$.

Statistic	Subtable	Mean	Variance	Reference	Rejection Rate			
					0.01	0.05	0.1	0.25
Z	$(1,2)$	-0.032	0.922	$N(0,1)$	0.010	0.040	0.090	0.225
	$(3,4)$	-0.011	1.000		0.010	0.057	0.106	0.253
	$(3,6)$	0.017	1.046		0.011	0.047	0.098	0.273
R_2	$(1,2)$	7.059	15.881	χ^2_7	0.016	0.063	0.120	0.245
	$(3,4)$	6.558	43.885		0.032	0.060	0.094	0.200
	$(3,6)$	6.395	32.431		0.038	0.076	0.114	0.202
S_b	$(1,2)$	1.033	2.195	χ^2_1	0.013	0.056	0.103	0.248
	$(3,4)$	6.908	1081.550		0.062	0.123	0.175	0.321
	$(3,6)$	14.903	1460.157		0.176	0.205	0.244	0.384
S_t	$(1,2)$	1.032	2.251	χ^2_1	0.015	0.055	0.103	0.250
	$(3,4)$	4.619	418.205		0.052	0.115	0.164	0.314
	$(3,6)$	11.929	934.045		0.178	0.205	0.248	0.369
X^2	$(1,2)$	0.493	0.477	χ^2_1	0.000	0.008	0.016	0.107
	$(3,4)$	0.667	0.810		0.001	0.007	0.047	0.159
	$(3,6)$	0.706	1.146		0.003	0.020	0.048	0.194
M_3	$(1,2,3)$	0.925	1.618	χ^2_1	0.007	0.037	0.078	0.244
	$(3,4,5)$	0.951	1.922		0.011	0.038	0.086	0.248
	$(3,6,7)$	1.001	2.231		0.013	0.054	0.105	0.239
X^2	$(1,2,3)$	2.61	3.748	χ^2_1	0.046	0.217	0.376	0.740
	$(3,4,5)$	2.787	4.35		0.038	0.249	0.436	0.744
	$(3,6,7)$	2.981	5.701		0.073	0.239	0.403	0.776

Results: Correctly Specified 2PL

Some descriptive statistics and empirical rejection rates based on a sample of size 300 are shown in Table 1. For succinctness, only results for item pairs $(1,2)$, $(3,4)$, and $(3,6)$, and triplets $(1,2,3)$, $(3,4,5)$, and $(3,6,7)$ are shown.

We see in this table that the empirical distributions of M_3 and of the standardized residual Z closely match their reference distribution even at this small sample size. In contrast, we see that the empirical distribution of R_2 statistics deviates from its reference χ^2_7 for most pairs indicating inadequate small sample performance. This is probably due to small counts (sparseness) in the summary statistics employed when the sample size is not large enough. Also, we see in this table that the score test statistics S_b and S_t tend to reject more often than they should for some pairs but not for others, depending on the true intercept values. We conjecture that the differences found between the performance of the standardized residual and the score statistics are due to how the information matrix is approximated: For residuals we used the expected information matrix, for score statistics the observed, which may not behave well in small samples. To assess our conjecture, in Appendix B we provide additional simulation results for S_t using both the observed and expected Fisher information under the same small sample condition.

Table 2. Simulation Results Under a 2PL Model: N = 1,000.

Statistic	Subtable	Mean	Variance	Reference	Rejection Rate			
					0.01	0.05	0.1	0.25
Z	(1, 2)	−0.006	0.993	$N(0, 1)$	0.007	0.050	0.096	0.261
	(3, 4)	−0.012	1.044		0.011	0.061	0.109	0.250
	(3, 6)	0.010	1.013		0.013	0.051	0.095	0.256
R_2	(1, 2)	6.980	18.474	χ^2_7	0.021	0.056	0.106	0.238
	(3, 4)	6.987	19.121		0.023	0.068	0.109	0.229
	(3, 6)	6.997	14.373		0.009	0.060	0.105	0.248
S_b	(1, 2)	1.036	1.946	χ^2_1	0.012	0.046	0.107	0.279
	(3, 4)	1.193	4.010		0.028	0.071	0.118	0.272
	(3, 6)	1.434	11.438		0.035	0.089	0.140	0.304
S_t	(1, 2)	1.027	1.909	χ^2_1	0.009	0.050	0.102	0.264
	(3, 4)	1.201	4.136		0.026	0.065	0.119	0.272
	(3, 6)	1.442	13.211		0.031	0.082	0.136	0.289
X^2	(1, 2)	0.531	0.506	χ^2_1	0.000	0.006	0.027	0.115
	(3, 4)	0.691	0.956		0.002	0.019	0.051	0.158
	(3, 6)	0.778	1.222		0.005	0.028	0.058	0.190
M_3	(1, 2, 3)	0.991	2.016	χ^2_1	0.010	0.054	0.103	0.239
	(3, 4, 5)	1.033	1.924		0.012	0.054	0.098	0.273
	(3, 6, 7)	0.989	1.954		0.010	0.050	0.104	0.245
X^2	(1, 2, 3)	2.753	4.358	χ^2_1	0.055	0.238	0.402	0.730
	(3, 4, 5)	2.927	4.457		0.069	0.256	0.444	0.770
	(3, 6, 7)	2.937	4.947		0.070	0.258	0.431	0.763

As for the X^2 statistics, we see in Table 1 that the bivariate (Chen-Thissen) X^2 rejects less often than it should, whereas the trivariate X^2 rejects more often than it should. The former is consistent with simulation result of Chen and Thissen (1997), whereas the latter is consistent with the asymptotic theory of Maydeu-Olivares and Joe (2006).

Table 2 presents the simulation results for samples of size 1,000. As expected, with a large enough sample size, the score test statistics and R_2 perform better (i.e., their reference distributions more closely approximate their empirical ones). In contrast, no improvement is apparent for either the bivariate or trivariate X^2 statistics. Because the empirical distribution of the X^2 statistics is not well approximated by the reference distributions employed, their power will not be investigated.

Results: Bifactor Alternative

The power (i.e., empirical rejection rate) of various statistics evaluated using their corresponding reference distribution at commonly used nominal level (i.e., $\alpha = 0.01$, 0.05, 0.1) is tabulated in Table 3. Only the results for three pairs and three triplets are reported: negative LD pair $(1, 2)$, positive LD pair $(1, 4)$, and locally independent pair $(3, 6)$; triplet $(1, 2, 3)$, $(1, 2, 4)$, and $(3, 6, 7)$.

Table 3. Simulation Results Under a Bifactor Model: N = 1,000.

Statistic	Subtable	Mean	Variance	Reference	Rejection Rate		
					0.01	0.05	0.1
Z	(1, 2)	−3.657	0.884	$N(0, 1)$	0.880	0.963	0.982
	(1, 4)	4.390	0.993		0.962	0.992	0.998
	(3, 6)	−0.008	1.003		0.014	0.056	0.097
R_2	(1, 2)	7.036	23.208	χ^2_7	0.028	0.075	0.128
	(1, 4)	7.164	17.895		0.020	0.056	0.111
	(3, 6)	7.902	17.676		0.022	0.087	0.155
S_b	(1, 2)	16.017	68.908	χ^2_1	0.897	0.966	0.982
	(1, 4)	18.949	64.612		0.963	0.992	0.998
	(3, 6)	2.374	313.123		0.039	0.085	0.137
S_t	(1, 2)	16.462	71.282	χ^2_1	0.909	0.968	0.984
	(1, 4)	19.001	66.856		0.961	0.993	0.998
	(3, 6)	2.028	141.169		0.042	0.083	0.124
M_3	(1, 2, 3)	1.074	2.442	χ^2_1	0.011	0.058	0.113
	(1, 2, 4)	1.069	2.191		0.011	0.052	0.110
	(3, 6, 7)	1.07	2.225		0.015	0.055	0.110

We see in Table 3 that both M_κ family statistics R_2 and M_3 do not have much power to detect that the 2PL model is misspecified when the true model is the bifactor triplet model as described in Equation 29; however, R_2 has slightly higher power than M_3 for detecting negative LD pairs (i.e., Items 1 and 2). The score test statistics and the bivariate residual show higher power than R_2 and M_3 for detecting both positive (e.g., Items 1 and 4) and negative (e.g., Items 1 and 2) LD pairs within the bifactor triplet. None of the statistics rejects much more than the nominal level for the pair (3, 6). The reason is that for this pair the 2PL model is correctly specified.

Results: Independent Cluster Alternative

For the final condition, data were simulated from the independent cluster two-dimensional model of Equation 30. The empirical rejection proportions for some pairs and triplets are presented in Table 4 with a sample of size 1,000.

We see in this table that when data are generated from a two-factor independent cluster model, both score test statistics and the bivariate residual have the highest level of power for all pairs. In contrast, R_2 has relatively high power when two items are from different factors, but its power is still uniformly lower than for the first tier. The M_3 statistic, again, has very little power.

The results of the simulation study can be summarized as follows. In the null case where data are generated from 2PL model: (a) Pearson's X^2 computed from either bivariate (i.e., Chen and Thissen) or trivariate (i.e., an unadjusted M_3 in the sense of Equation 23) subtables cannot be approximated well by a chi-squared distribution. (b) The empirical distribution of both score test statistics S_b and S_t is well approximated by its reference asymptotic distribution in large samples, but might be liberal

Table 4. Simulation Results Under a Two-Dimensional Model: N = 1,000.

Statistic	Subtable	Mean	Variance	Reference	Rejection Rate		
					0.01	0.05	0.1
Z	(1, 2)	3.988	4.354	$N(0, 1)$	0.705	0.809	0.859
	(3, 4)	3.645	2.997		0.704	0.793	0.850
	(3, 6)	−1.795	1.182		0.245	0.441	0.562
R_2	(1, 2)	4.632	11.367	χ^2_7	0.005	0.021	0.038
	(3, 4)	6.521	14.952		0.010	0.045	0.088
	(3, 6)	9.844	32.109		0.080	0.186	0.282
S_b	(1, 2)	20.225	291.632	χ^2_1	0.707	0.815	0.863
	(3, 4)	32.403	1125.321		0.760	0.844	0.888
	(3, 6)	5.011	17.969		0.289	0.520	0.643
S_t	(1, 2)	20.096	286.479	χ^2_1	0.702	0.818	0.864
	(3, 4)	32.121	1125.113		0.749	0.835	0.880
	(3, 6)	4.212	14.513		0.214	0.441	0.567
M_3	(1, 2, 3)	1.581	5.117	χ^2_1	0.048	0.117	0.186
	(3, 4, 6)	0.907	1.388		0.003	0.033	0.075
	(3, 6, 7)	1.44	4.151		0.039	0.105	0.165

for certain combinations of parameter values that are likely to produce zero cell counts in small samples. Nevertheless, the results of Appendix B reveal that this deficiency could be improved by using a better approximation of the information matrix. (c) The empirical distribution of R_2 is well approximated by its reference distribution for a sample of size 1,000, but not as well in a sample of size 300 for some pairs. (d) Bivariate residual and M_3 have the best small sample behavior. They have very accurate empirical Type I error rates.

As for power, R_2 and M_3 have low power under both alternative models. Nevertheless, the power of R_2 is barely acceptable when computed for negative LD pairs under independent cluster alternative, which makes it useful in separating items belonging to different factors. Meanwhile, the score test statistics and bivariate residual are very sensitive to both underfit and overfit of the covariances between item responses. Therefore, they are recommended for the purpose of identifying LD.

In closing, the simulation study reported here has only investigated two LD data generating mechanisms. Also, by some accounts the form of LD considered here can be described as weak (for instance, the correlation between the traits in the independent clusters condition is only 0.3). Further research is necessary to investigate the performance of the statistics under alternative data generating models leading to LD, including situations leading to stronger local dependencies than those considered here.

Numerical Example: LSAT-7 Data

We fit a 2PL model to the well-known LSAT-7 data (Bock & Lieberman, 1970). These data consist of 1,000 responses to $J = 5$ binary variables. Because the data are not sparse, the overall X^2 and M_2 statistics agree (as shown in Table 5).

Table 5. Overall Goodness-of-Fit Statistics: LSAT-7 Data.

	Statistic	*df*	*p*
X^2	32.48	21	0.05
M_2	11.94	5	0.04

Table 6. LD Diagnostics for Pairs: LSAT-7 Data.

	X^2 (*df* = 1)		S_b (*df* = 1)		S_t (*df* = 1)		Z^2 (*df* = 1)		R_2 (*df* = 2)	
Pair	Statistic	*p*	Statistic	*p*	Statistic	*p*	Statistic	*p*	Statistic	*p*
(1, 2)	0.45	.50	1.15	.28	1.18	.28	1.19	.27	**8.95**	**.01**
(1, 3)	0.86	.35	**4.43**	**.04**	**4.58**	**.03**	**4.16**	**.04**	0.53	.77
(1, 4)	2.58	.11	**4.28**	**.04**	**4.30**	**.04**	**4.77**	**.03**	3.16	.21
(1, 5)	2.39	.12	3.62	.06	3.58	.06	3.90	.05	2.77	.25
(2, 3)	1.06	.30	**8.12**	**.00**	**7.67**	**.01**	**8.38**	**.00**	3.96	.14
(2, 4)	0.27	.61	0.77	.38	0.77	.38	0.70	.40	**8.63**	**.01**
(2, 5)	1.38	.24	2.89	.09	2.91	.09	2.83	.09	**8.39**	**.02**
(3, 4)	0.15	.69	0.95	.33	0.96	.33	0.67	.41	1.36	.51
(3, 5)	0.00	.96	0.01	.94	0.00	.95	0.01	.93	1.58	.45
(4, 5)	0.00	1.00	0.00	.98	0.00	.99	0.00	1.00	3.24	.20

Note. Values in boldface indicate $p < .05$.

Both statistics suggest a barely acceptable fit, and to investigate whether we can identify how to improve the fit of the model to these data, we compute the score tests S_b and S_t, Chen-Thissen's X^2, the bivariate residual Z^2, and R_2 for all pairs of items. We also compute the X^2 and M_3 statistics for all triplets of items. The results are presented separately for pairs and triplets in Tables 6 and 7. We see in these tables that the statistics fail to agree.

For pairwise diagnostics: (a) The bivariate X^2 statistic suggests that the model fits for all pairs, but we know that Chen and Thissen's proposal leads to underrejection; (b) S_b, S_t, Z consistently identify the pairs $(1, 3)$, $(1, 4)$, and $(2, 3)$ as fitting poorly; (c) in addition, the bivariate residual Z suggests that pair $(1, 5)$ does not fit well; (d) the problematic pairs suggested by the R_2 statistic are $(1, 2)$, $(2, 4)$, and $(2, 5)$. Provided one is willing to remove one item to improve model fit, then Items 1 and 3 might be the top choices as suggested by most of the statistic; however, R_2 suggests that Item 2 might also be problematic.

For triplet-wise diagnostics: (a) Pearson's X^2 suggests more problematic triplets than it should, as we know it is liberal; (b) in contrast, M_3 flags only the item triplet $(1, 2, 5)$. If one wishes to delete one item, triplet-wise diagnostics alone might not be very informative.

Table 7. LD Diagnostics for Triplets: LSAT-7 Data.

	X^2 (df = 1)		M_3 (df = 1)	
Triplet	Statistic	p	Statistic	p
(1, 2, 3)	3.27	.07	0.73	.39
(1, 2, 4)	**4.27**	**.04**	0.48	.49
(1, 2, 5)	**11.36**	**.00**	**6.12**	**.01**
(1, 3, 4)	**5.79**	**.02**	1.27	.26
(1, 3, 5)	**5.41**	**.02**	1.63	.20
(1, 4, 5)	**5.89**	**.02**	1.37	.24
(2, 3, 4)	2.05	.15	0.43	.51
(2, 3, 5)	**4.48**	**.03**	1.60	.21
(2, 4, 5)	1.79	.18	0.20	.65
(3, 4, 5)	0.25	.62	0.09	.77

Note. Values in boldface indicate $p < .05$.

Table 8. Overall X^2 and M_2 Statistics After Deleting One Item: LSAT-7 Data.

	X^2 (df = 7)		M_2 (df = 2)	
Item Omitted	Statistic	p	Statistic	p
1	**5.01**	**.66**	1.26	.53
2	9.52	.22	1.90	.29
3	8.59	.29	**1.01**	**.60**
4	18.68	.01	7.05	.03
5	9.86	.20	6.58	.04

Note. Omitting Item 1 produces the smallest X^2. Omitting Item 3 produces the smallest M_2.

Next, we fit the 2PL model to the data omitting one item at a time. The results (Table 8) reveal that deleting Item 1 will produce the smallest X^2, and Item 3 the smallest M_2. These are consistent with the diagnoses drawn from the score tests and the standardized residuals.

Concluding Remarks

The statistic that fared the worst is Pearson's X^2 applied to trivariate subtables heuristically using the same reference distribution as for M_3 (i.e., number of cells minus number of parameters involved minus one, i.e., χ^2_1). We should avoid using it because it rejects well fitting items more often than it should.

Chen and Thissen (1997), also heuristically, suggested using as reference distribution for X^2 the reference distribution corresponding to an independence

model. If this reference distribution is used, then one can apply X^2 to pairs of items. Unfortunately, we have seen that this resulted in failure to reject poorly fitting items.

The remaining statistics we considered have known asymptotic distributions, which guarantees the rejection rate being adequate when the fitted model is correct, provided that sample size is large enough.

M_3 statistic for trivariate subtables can be seen as a correction of X^2. This correction is necessary for the statistic to be asymptotically distributed as chi-squared, whenever a model is estimated using J items while the testing only involves a subset of them (e.g., pair or triplet). Triplets of variables are needed to assess the goodness of fit of the 2PL. M_3 was found to be very well approximated by its reference distribution in small samples when the fitted model was correct. Hence, it will not reject well-fitting items. However, it was also found to have low power to detect dependencies arising from a bifactor or independent cluster multidimensional models. In addition, it is generally hard to draw conclusions from tests involving triplets of items as compared to pairs.

To be able to test pairwise LD, we have proposed a new bivariate statistic in this article, termed R_2, by drawing on information from the sum score. It is similar in spirit to Glas's (1988) statistic for testing the overall goodness-of-fit of the Rasch model and also to Thissen and Orlando's (2000) item-fit statistic. Drawing on theory from Joe and Maydeu-Olivares (2010), we have been able to derive the asymptotic distribution of the R_2 statistic for pairs of variables. Larger samples are needed for the statistic to be well approximated by its asymptotic distribution than for M_3. However, we found that R_2 also lacked power to identify the two parametric forms of LD used in our simulation study.

The score test statistics and standardized bivariate residuals had the highest power in our simulation study. However, both of them require the computation of the information matrix (or the covariance matrix of all estimated item parameters). The expected information cannot be computed for long tests, while the cross-product estimation does not work well with small samples. This might limit the use of these statistics in practice.

To summarize, there exist statistics that will not reject well fitting items—namely, M_3, R_2, score statistics, standardized residuals. Among them, however, the most powerful statistics (i.e., score test statistics, standardized bivariate residuals) depend on the computation of the information matrix. We have seen that using the expected information leads to a much better performance of the statistics. However, the expected information matrix can only be computed with tests that are not too long (e.g., no more than 30 binary items). One can compute M_3 and R_2 for larger models but they may have little power to detect poorly fitting items. Future research should investigate better estimates of Fisher information that are computable for large models and have adequate small sample performance.

Appendix A

Example for $\mathbf{T}_{R_2,jk}$ When J = 4

When $J = 4$, there are in total $2^4 = 16$ possible response patterns:

$$\mathbf{Y} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}', \tag{A.1}$$

Using these ordering of the patterns, the statistics shown in Equation 24 for Items 1 and 2 can be obtained from the cell residuals by multiplication of the following 9×16 transformation matrix:

$$\mathbf{T}_{R_2,12} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{A.2}$$

This matrix is only shown for exposition purposes. In practice, the statistics are to be computed directly.

Appendix B

Simulation Results for S_t Using Expected and Observed Information When the Model Is Correctly Specified

Using the same true parameter values as in Equations 25 and 26 with sample size $N = 300$, we repeated the simulation under H_0 for the score test statistic S_t using both the expected (i.e., exact) and the observed (i.e., cross-product approximated) information. The results are tabulated in Table B.1.

For all three pairs displayed in Table B.1, the statistic computed using exact information has rejection rates much closer to the nominal α level than the one computed using the cross-product approximation. This reveals that the small sample performance of the score test is contingent on the estimates of the Fisher information matrix.

Table B.1. Simulation Results for S_t under H_0 Using Expected Versus Observed Information: $N = 300$.

Subtable	Statistic	Mean	Variance	Rejection Rate			
				0.01	0.05	0.1	0.25
(1, 2)	S_t (Expected)	1.043	1.922	0.008	0.050	0.111	0.272
	S_t (Observed)	1.146	2.395	0.015	0.066	0.126	0.303
(3, 4)	S_t (Expected)	0.980	1.501	0.006	0.031	0.091	0.254
	S_t (Observed)	4.467	361.716	0.061	0.114	0.177	0.311
(3, 6)	S_t (Expected)	0.983	2.149	0.013	0.036	0.088	0.262
	S_t (Observed)	12.825	974.230	0.185	0.223	0.261	0.368

Note. The reference distribution is χ^2_1.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Alberto Maydeu-Olivares was supported by an ICREA-Academia Award and Grant SGR 2009 74 from the Catalan Government and Grants PSI2009-07726 and PR2010-0252 from the Spanish Ministry of Education.

References

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

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

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows*. Lincolnwood, IL: Scientific Software International.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14*, 275-299.

Glas, C. (1988). The derivation of some tests for the rasch model from the multinomial distribution. *Psychometrika, 53*, 525-546.

Glas, C. A., & Suárez-Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87-106.

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika, 75*, 393-419.

Kendall, M., & Stuart, A. (1961). *The advanced theory of statistics* (Vol. II). London, England: Griffin.

Lehmann, E. (1999). *Elements of large-sample theory*. New York, NY: Springer.

Liu, Y., & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement*. Manuscript submitted for publication. In press.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score ''equatings.'' *Applied Psychological Measurement, 8*, 453-461.

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

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

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Thissen, D., & Orlando, M. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 148-177). London, England: Sage.

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58-79.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.