PSYCHOMETRIKA—VOL. 78, NO. 1, 116–133
JANUARY 2013
DOI: 10.1007/S11336-012-9293-1

HOW SHOULD WE ASSESS THE FIT OF RASCH-TYPE MODELS? APPROXIMATING THE POWER OF GOODNESS-OF-FIT STATISTICS IN CATEGORICAL DATA ANALYSIS

ALBERTO MAYDEU-OLIVARES

FACULTY OF PSYCHOLOGY, UNIVERSITY OF BARCELONA

ROSA MONTAÑO

UNIVERSIDAD DE SANTIAGO DE CHILE

We investigate the performance of three statistics, R_1, R_2 (Glas in Psychometrika 53:525–546, 1988), and M_2 (Maydeu-Olivares & Joe in J. Am. Stat. Assoc. 100:1009–1020, 2005, Psychometrika 71:713–732, 2006) to assess the overall fit of a one-parameter logistic model (1PL) estimated by (marginal) maximum likelihood (ML). R_1 and R_2 were specifically designed to target specific assumptions of Rasch models, whereas M_2 is a general purpose test statistic. We report asymptotic power rates under some interesting violations of model assumptions (different item discrimination, presence of guessing, and multidimensionality) as well as empirical rejection rates for correctly specified models and some misspecified models. All three statistics were found to be more powerful than Pearson's X^2 against two- and three-parameter logistic alternatives (2PL and 3PL), and against multidimensional 1PL models. The results suggest that there is no clear advantage in using goodness-of-fit statistics specifically designed for Rasch-type models to test these models when marginal ML estimation is used.

Key words: discrete data, power, IRT, maximum likelihood.

1. Introduction

Broadly speaking, item response theory (IRT) refers to the class of latent trait models for discrete multivariate data obtained by coding the responses to a set of questionnaire items, such as those found in educational tests, personality inventories, etc. Rasch-type models are a subset of IRT models, so named after the pioneering work of Rasch (1960). Rasch-type models are characterized by two properties (McDonald, 1999): (a) the sum score is a sufficient statistic for the latent traits, and (b) comparisons of subpopulations are made independently of the item or items used for the comparison (the so-called specific objectivity property). Although only highly restrictive IRT models can satisfy these properties, their mathematical potential has led some researchers to prefer them to all other IRT models. Thus, we may distinguish between two traditions in IRT modeling: a model-based tradition and a data-based tradition. In the model-based tradition, a model with appealing mathematical properties is selected first (a Rasch-type model) and tests are designed to fit the model. By contrast, in a data-based tradition, different models within the IRT family are explored to find the best fitting model for the available data.

Because of the availability of sufficient statistics for the latent traits that do not depend on item parameters, estimation methods (conditional maximum likelihood, or CML) and goodness-of-fit testing procedures have been developed specifically for Rasch-type models (for an overview

This research was supported by an ICREA-Academia Award and Grant SGR 2009 74 from the Catalan Government, and by Grants PSI2009-07726 and PR2010-0252 from the Spanish Ministry of Education awarded to the first author, and by a Dissertation Research Award of the Society of Multivariate Experimental Psychology awarded to the second author. The authors are indebted to the reviewers and to David Thissen for comments that improved the manuscript.

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

of Rasch modeling, see Fischer & Molenaar, 1995). However, for Rasch-type models, it is not clear whether there is any advantage in using the procedures specifically designed for these tests or if, on the other hand, those general procedures applicable to other IRT models that do not possess sufficient statistics can actually yield more accurate results.

For item parameter estimation techniques and after Thissen's (1982) pioneering work (1982), there seems to be a consensus (Pfanzagel, 1993; see also De Leeuw & Verhelst, 1986) that the more general marginal maximum likelihood estimation procedure (MML or simply ML) generally implemented via the EM algorithm (see Bock & Aitkin, 1981) is preferable to the CML procedure originally favored by Rasch modelers. However, many Rasch modelers still prefer CML because no distribution needs to be assumed for the latent traits. Finally, the reader should also note that MML estimation is sometimes referred to in the literature as full information maximum likelihood (FIML) (e.g., Jöreskog & Moustaki, 2001).

However, no such consensus has emerged regarding the use of goodness-of-fit testing procedures. Indeed, there is a large number of goodness-of-fit statistics specifically proposed for Rasch-type models (see Andersen, 1973; van den Wollenberg, 1982; Suárez-Falcon and Glas, 2003; and the excellent review by Glas & Verhelst, 1995), in addition to the general procedures proposed available for testing multivariate discrete data models, and particularly IRT models; see the reviews by Mavridis, Moustaki, and Knott (2007), Swaminathan, Hambleton, and Rogers (2007), and Maydeu-Olivares and Joe (2008). The purpose of this article is to compare the performance of certain goodness-of-fit statistics to test Rasch-type models. To do so, we concentrate on models for binary data. More specifically, we use the one-parameter logistic model, that is, the random effects version of Rasch's 1960 model. The statistics being compared are R_1, R_2, and M_2. The statistics R_1 and R_2 were proposed by Glas (1988) to assess the fit of the one-parameter logistic model, and M_2 was proposed by Maydeu-Olivares and Joe (2005, 2006) for testing general composite null hypotheses in multivariate discrete data

The remaining sections of this article are divided as follows. Sections 2 and 3 provide theoretical background. The R_1, R_2, and M_2 test statistics are described and the details of their asymptotic distribution are provided for composite nulls, and also under a sequence of local alternatives. Section 3 describes a procedure first used by Reiser (2008) to approximate asymptotically the power of the statistics for specific alternatives without having to use simulations. In many ways, this procedure is the categorical data counterpart to the procedure proposed by Satorra and Saris in structural equation modeling (1985). Section 4 compares the performance of the statistics. We report asymptotic power rates under some interesting violations of model assumptions (different item discrimination, presence of guessing, and multidimensionality) as well as empirical rejection rates for correctly specified models and some misspecified models. Finally, Section 5 provides two numerical examples using real data.

2. Rasch-Type Models for Binary Data

Consider n binary items Y_i, whose categories have been coded as 0 or 1. Rasch (1960) proposed the following model:

$$P(Y_i = 1|\xi) = \frac{\xi}{\xi + \delta_i}, \tag{1}$$

where ξ denotes the latent trait, and δ_i denotes the item difficulty parameter. The model can be reparameterized using $\xi = \exp(\eta)$ and $\delta_i = \exp(b_i)$ to yield

$$P(Y_i = 1|\eta) = \frac{\exp(\eta)}{\exp(\eta) + \exp(b_i)} = \frac{1}{1 + \exp[-(\eta - b_i)]}, \tag{2}$$

which is a special case of the three parameter logistic (3PL) model,

$$P(Y_i = 1|\eta) = c_i + (1 - c_i)\frac{1}{1 + \exp[-(a_i(\eta - b_i))]}, \tag{3}$$

introduced in Lord and Novick (1968). In (2), η and b_i denote the (reparameterized) latent trait and item difficulty parameter, respectively. In turn, the item parameters a_i and c_i in (3) may be interpreted as discrimination and guessing parameters, respectively.

Rasch (1960) treated the latent trait parameters ξ (or equivalently η) as fixed effects, so that the distribution of ξ (or equivalently η) is not specified. He proposed identifying the model by introducing the constraint

$$\sum_{i=1}^{n} b_i = 0 \tag{4}$$

and estimating the mean and variance of the latent trait. The model can be equivalently identified by fixing the mean and variance of the latent trait to some constants, say 0 and 1, estimating the b_i without the constraint (4), and rewriting (2) as

$$P(Y_i = 1|\eta) = \frac{1}{1 + \exp[-a(\eta - b_i)]}. \tag{5}$$

Note that the discrimination parameter a in (5) is common to all items.

In recent times, the latent trait η in (5) has most often been treated as a random effect, generally by specifying a standard normal distribution for η. To distinguish between the two variants of the model, we shall refer to the fixed-effects version of the model (i.e., Equation (2) with the constraint (4) and latent trait mean and variance to be estimated) as the *Rasch model*; and we shall refer to the random-effects version (given by Equation (5) with mean zero and unit variance for η), as the *one-parameter logistic model* or the *1PL*. A parametric latent trait distribution need not be assumed for the 1PL, as the latent trait density can be estimated nonparametrically. However, in this article, we shall assume that the latent trait follows a standard normal distribution for all the models considered, meaning the 1PL, the 3PL model (3), and in the special case where all c_i parameters are equal to zero, the two-parameter logistic model, 2PL. Readers should also note that regardless of whether we assume the Rasch model or the 1PL (i.e., regardless of whether the latent trait is treated as fixed- or random-effect), the specific objectivity property holds. In other words, for any item the log-odds value of two subpopulations is equal to the difference in their trait values; see Irtel (1995) for a less restrictive definition of specific objectivity that applies to the 2PL.

3. Goodness-of-Fit Assessment in Binary IRT Models

Consider modeling N observations on n binary random variables. The observed responses can then be gathered in an n-dimensional contingency table with $C = 2^n$ cells. Let π_c be the probability of one such cell, $c = 1, \ldots, C$, and let p_c be the observed proportion. Also, let $\boldsymbol{\pi}(\boldsymbol{\theta})$ be the C-dimensional vector of model probabilities expressed as a function of q model parameters $\boldsymbol{\theta}$ to be estimated from the data. Then the (composite) null hypothesis to be tested is $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ against $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$.

The two standard goodness-of-fit statistics for discrete data are Pearson's statistic, $X^2 = N\sum_{c=1}^{C}(p_c - \hat{\pi}_c)^2/\hat{\pi}_c$, and the likelihood ratio statistic $G^2 = 2N\sum_{c=1}^{C} p_c \ln(p_c/\hat{\pi}_c)$, where $\hat{\pi}_c = \pi_c(\hat{\boldsymbol{\theta}})$ denotes the probability of cell c under the model. Asymptotic p-values for both statistics can be obtained using a chi-square distribution with $C - q - 1$ degrees of freedom when maximum likelihood estimation is used. However, these asymptotic p-values are only reliable

when all expected frequencies are large (>5 is the usual rule of thumb). Unfortunately, as the number of cells in the table increases, the expected frequencies must be small (Bartholomew & Tzamourani, 1999) because the sum of all C probabilities must be equal to 1. As a result, in multivariate discrete data analysis the asymptotic p-values for these statistics can hardly ever be used.

To overcome these limitations, a number of authors (e.g., Christoffersson, 1975; Reiser, 1996, 2008; Bartholomew & Leung, 2002; Maydeu-Olivares & Joe, 2005, 2006; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006) have proposed testing using limited information, that is, pooling cells of the contingency table a priori so that the resulting statistics have a known asymptotic null distribution.

3.1. M_2 and the M_r Family of Test Statistics

Maydeu-Olivares and Joe (2005) proposed testing using a quadratic form in residual moments of the multivariate Bernoulli distribution (Teugels, 1990) up to the smallest order at which the model is identified. The family of statistics they proposed is

$$M_r = N\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})\big)'\hat{\mathbf{C}}_r\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})\big), \tag{6}$$

$$\mathbf{C}_r = \boldsymbol{\Xi}_r^{-1} - \boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r\big(\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r\big)^{-1}\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1} = \boldsymbol{\Delta}_r^{(c)}\big(\boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r^{(c)}\big)^{-1}\boldsymbol{\Delta}_r^{(c)\prime}, \tag{7}$$

where $\hat{\mathbf{C}}_r$ denotes $\mathbf{C}_r$ evaluated at $\hat{\boldsymbol{\theta}}$. In (6), $\boldsymbol{\pi}_r$ denotes the $s = \sum_{i=1}^{r}\binom{n}{i}$ vector of moments of the multivariate Bernoulli distribution up to order r, and $\mathbf{p}_r$ denotes its sample counterpart. In (7), $\boldsymbol{\Delta}_r = \frac{\partial\boldsymbol{\pi}_r(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}$, and $\boldsymbol{\Xi}_r$ denotes the asymptotic covariance matrix of $\sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_r)$. Also, $\boldsymbol{\Delta}_r^{(c)}$ is the $s \times (s-q)$ orthogonal complement of $\boldsymbol{\Delta}_r$ (i.e., it satisfies $\boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Delta}_r^{(c)} = \mathbf{0}$).

M_r is a family of statistics comprising $M_1, M_2, \ldots, M_n$. In M_1 only univariate information is used, in M_2 only univariate and bivariate information is used, and so forth up to M_n, a full information statistic that is algebraically equal to Pearson's X^2 when ML estimation is used. For the chi-square approximation to the distribution of M_r to be accurate, only the expected frequencies of the moments of order min $(2r, n)$ need to be large. Thus, the smaller the r used for testing, the more accurate the asymptotic approximation in small samples. Because most IRT models are identified from univariate and bivariate information, M_2 is the statistic of choice within this family for testing IRT models. Note that only expected frequencies for sets of four variables are involved in the computation of M_2.

Actually, the moments of the multivariate Bernoulli distribution are simply marginal probabilities obtained by a linear transformation of the cell probabilities, $\boldsymbol{\pi}_n = \mathbf{T}\boldsymbol{\pi}$, or

$$\begin{pmatrix}\dot{\boldsymbol{\pi}}_1\\ \dot{\boldsymbol{\pi}}_2\\ \vdots\\ \dot{\boldsymbol{\pi}}_n\end{pmatrix} = \begin{pmatrix}\dot{\mathbf{T}}_1\\ \dot{\mathbf{T}}_2\\ \vdots\\ \dot{\mathbf{T}}_n\end{pmatrix}\boldsymbol{\pi},$$

where $\dot{\boldsymbol{\pi}}_r$ is the $\binom{n}{r}$-dimensional vector of rth order moments. This transformation is illustrated here for $n = 3$ variables:

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

Note that $\boldsymbol{\Delta}_r = \mathbf{T}_r\boldsymbol{\Delta} = \mathbf{T}_r\frac{\partial\boldsymbol{\pi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}$, where $\mathbf{T}_r = (\dot{\mathbf{T}}'_1 \ \ldots \ \dot{\mathbf{T}}'_r)'$ is an $s \times C$ matrix, and $\boldsymbol{\Xi}_r = \mathbf{T}_r\boldsymbol{\Gamma}\mathbf{T}'_r$, where $\boldsymbol{\Gamma}$ is the asymptotic covariance matrix of $\sqrt{N}(\mathbf{p}-\boldsymbol{\pi})$, $\boldsymbol{\Gamma} = \mathbf{D} - \boldsymbol{\pi}\boldsymbol{\pi}'$, with $\mathbf{D} = \operatorname{diag}(\boldsymbol{\pi})$. Thus, M_r may be written as

$$M_r = N\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big)'\mathbf{T}'_r\hat{\mathbf{C}}_r\mathbf{T}_r\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big). \tag{8}$$

Maydeu-Olivares and Joe (2005) showed that if the model is identified from $\boldsymbol{\pi}_r$, and if $\hat{\boldsymbol{\theta}}$ is an $\sqrt{N}$-consistent and asymptotically normal estimator (not just for the ML estimator), M_r is asymptotically distributed as a chi-square with $s-q$ degrees of freedom under the null hypothesis. This follows from the asymptotic normality of the vector of cell residuals and by noting, using (7), that

$$\boldsymbol{\Sigma}_r = \mathbf{C}_r\boldsymbol{\Sigma}_r\mathbf{C}_r. \tag{9}$$

Also, note that M_2 is simply a quadratic form in residual means and cross-products, since the elements of $\boldsymbol{\pi}_2 = (\dot{\boldsymbol{\pi}}'_1 \ \dot{\boldsymbol{\pi}}'_2)'$ are of the type $\dot{\pi}_i = \Pr(Y_i = 1)$, and $\dot{\pi}_{ij} = \Pr(Y_i = 1, Y_j = 1)$. The degrees of freedom available for testing when using M_2 are $n(n+1)/2 - q$.

3.2. R_1, R_2, and the Family of Generalized Pearson Statistics

The statistic R_1 proposed by Glas (1988) has a similar form to M_r as given in (8),

$$R_1 = N\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big)'\mathbf{T}'_{R1}\hat{\mathbf{C}}_{R1}\mathbf{T}_{R1}\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big), \qquad \mathbf{C}_{R1} = \big(\mathbf{T}_{R1}\mathbf{D}\mathbf{T}'_{R1}\big)^{-1}, \tag{10}$$

with $\hat{\mathbf{C}}_{R1}$ denoting $\mathbf{C}_{R1}$ evaluated at $\hat{\boldsymbol{\theta}}$, but $\mathbf{T}_{R1}$ is an $(n(n-1)+2) \times C$ block diagonal matrix, $\mathbf{T}_{R1} = \operatorname{diag}(\mathbf{T}^{(0)}_{R1}, \ldots, \mathbf{T}^{(n)}_{R1})$, and $\mathbf{C}_{R1}$ has a much simpler form than $\mathbf{C}_r$. Furthermore,

$$\boldsymbol{\Sigma}_{R1} = \boldsymbol{\Sigma}_{R1}\mathbf{C}_{R1}\boldsymbol{\Sigma}_{R1}, \tag{11}$$

and $\boldsymbol{\Sigma}_{R1} = \mathbf{T}_{R1}\boldsymbol{\Sigma}\mathbf{T}_{R1}$. The relationship $\boldsymbol{\pi}_{R1} = \mathrm{T}_{R1}\boldsymbol{\pi}$ is illustrated below for $n = 3$ items

$$\boldsymbol{\pi}_{R1} = \mathbf{T}_{R1}\boldsymbol{\pi} = \begin{pmatrix} \pi_0 \\ \boldsymbol{\pi}_{1|x=1} \\ \boldsymbol{\pi}_{1|x=2} \\ \pi_3 \end{pmatrix} = \left(\begin{array}{c|c|c|c|c|c|c|c} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right) \begin{pmatrix} \pi_{(0,0,0)} \\ \pi_{(1,0,0)} \\ \pi_{(0,1,0)} \\ \pi_{(0,0,1)} \\ \pi_{(1,1,0)} \\ \pi_{(1,0,1)} \\ \pi_{(0,1,1)} \\ \pi_{(1,1,1)} \end{pmatrix}, \tag{12}$$

with $\mathbf{T}'_{R1} = (\mathbf{T}^{(0)\prime}_{R1}, \mathbf{T}^{(1)\prime}_{R1}, \ldots, \mathbf{T}^{(n)\prime}_{R1})$ being block diagonal.

In (12), $\boldsymbol{\pi}_{1|x=k}$ is used to denote the n-dimensional vector of probabilities of endorsing each of the n items, given that the sum score $X = Y_1 + \cdots + Y_n$ is k, that is, $\boldsymbol{\pi}_{1|x=k} = (\Pr(Y_1 = 1, X = k), \ldots, \Pr(Y_n = 1, X = k))'$. Also, π_0 is the probability of obtaining a sum score of zero, and π_n of obtaining a sum score of n.

Let $\hat{\mathbf{e}} = (\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}))$. Because $\mathbf{T}_R$ is block diagonal, the R_1 statistic may be written as

$$\begin{aligned} R_1 &= N\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big)'\mathbf{T}_{R1}\big(\mathbf{T}_{R1}\hat{\mathbf{D}}\mathbf{T}'_{R1}\big)^{-1}\mathbf{T}_{R1}\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big) \\ &= N\left\{\frac{(p_0 - \hat{\pi}_0)^2}{\hat{\pi}_0} + \sum_{k=1}^{n-1}\hat{\mathbf{e}}'\mathbf{T}^{(k)\prime}_{R1}\big(\mathbf{T}^{(k)}_{R1}\hat{\mathbf{D}}_k\mathbf{T}^{(k)\prime}_{R1}\big)^{-1}\mathbf{T}^{(k)}_{R1}\hat{\mathbf{e}} + \frac{(p_n - \hat{\pi}_n)^2}{\hat{\pi}_n}\right\}. \end{aligned} \tag{13}$$

Glas (1988) showed that for the 1PL the asymptotic distribution of R_1 under the null hypothesis for the ML estimator is chi-square with $n(n-2)$ degrees of freedom. This follows from the asymptotic normality of $\mathbf{T}_{R1}(\mathbf{p} - \hat{\boldsymbol{\pi}})$ and that (11) is satisfied for this model.

R_1 was developed to assess the assumptions of monotone increasing and parallel item response functions of the 1PL model when the item parameters are estimated by marginal maximum likelihood. The assumption of monotone increasing and parallel item response functions is not the only assumption underlying the 1PL. Another is unidimensionality of the latent trait. van den Wollenberg (1982) showed that tests based on the univariate moments conditional on sum score such as R_1 are relatively insensitive to multidimensionality. Therefore, Glas (1988) introduced the R_2 statistic, which like M_2 is based on bivariate moments. More specifically, the R_2 statistic is based on π_0, $\Pr(Y_i = 1, X = 1)$ for $i = 1, \ldots, n$, $\Pr(Y_i = 1, Y_j = 1, 2 \leq X \leq n-1)$ for $i = 1, \ldots, n-1$ and $j = i+1, \ldots, n$, and π_n.

Thus, for R_2 $\mathbf{T}_{R2} = \text{diag}(\mathbf{T}_{R2}^{(0)}, \ldots, \mathbf{T}_{R2}^{(n)})$ is an $(n(n+1)/2+2) \times C$ 4-block diagonal matrix. The relationship $\boldsymbol{\pi}_{R2} = \mathbf{T}_{R2}\boldsymbol{\pi}$ is illustrated in the appendix for $n = 4$ items; for $n = 3$ items $\mathbf{T}_{R2}$ is an identity matrix. The statistic can be written as (13) except that in this case, since $\mathbf{T}_{R2}$ only consists of 4 blocks regardless of the number of items, we write

$$\begin{aligned} R_2 &= N\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big)'\mathbf{T}_{R2}\big(\mathbf{T}_{R2}\hat{\mathbf{D}}\mathbf{T}_{R2}'\big)^{-1}\mathbf{T}_{R2}\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big) \\ &= N\left\{\frac{(p_0 - \hat{p}_0)^2}{\hat{p}_0} + \hat{\mathbf{e}}'\mathbf{T}_{R2}^{(1)\prime}\big(\mathbf{T}_{R2}^{(1)}\hat{\mathbf{D}}\mathbf{T}_{R2}^{(1)\prime}\big)^{-1}\mathbf{T}_{R2}^{(1)}\hat{\mathbf{e}} + \hat{\mathbf{e}}'\mathbf{T}_{R2}^{(2)\prime}\big(\mathbf{T}_{R2}^{(2)}\hat{\mathbf{D}}\mathbf{T}_{R2}^{(2)\prime}\big)^{-1}\mathbf{T}_{R2}^{(2)}\hat{\mathbf{e}} \right. \\ &\quad \left. + \frac{(p_n - \hat{p}_n)^2}{\hat{p}_n}\right\}. \end{aligned} \tag{14}$$

When testing the fit of the 1PL model, the statistic follows asymptotically a chi-square distribution with $(n(n-2)+2)/2$ degrees of freedom (Glas, 1988).

For the 1PL, R_1 belongs to the family of generalized Pearson statistics introduced by Glas and Verhelst (1989) and R_2 has the same form as statistics within this family. The family of generalized Pearson statistics is defined as

$$Q = N\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big)'\mathbf{U}'\hat{\mathbf{C}}_U\mathbf{U}\big(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})\big), \qquad \mathbf{C}_U = \big(\mathbf{U}\mathbf{D}\mathbf{U}'\big)^{-}, \tag{15}$$

with $\hat{\mathbf{C}}_U$ denoting $\mathbf{C}_U$ evaluated at $\hat{\boldsymbol{\theta}}$. $\mathbf{U}$ denotes a $g \times C$ matrix of constants so as to choose g linear combinations of the cell residuals such that (a) they show specific model violations, and (b) their expected probabilities are sufficiently large for applying asymptotic theory. They show that if (a) the columns of $\mathbf{D}^{-\frac{1}{2}}\boldsymbol{\Delta}$ belong to the linear manifold of the columns of $\mathbf{D}^{-\frac{1}{2}}\mathbf{U}'$, and (b) there exists a vector of constants $\mathbf{c}$ such that $\mathbf{U}'\mathbf{c} = \mathbf{1}$, then Q is asymptotically distributed as a chi-square with degrees of freedom equal to $\text{rank}(\mathbf{U}\mathbf{D}\mathbf{U}') - q - 1$. Note that in (15) a generalized inverse is used, allowing the rows of $\mathbf{U}$ to be linearly dependent. Condition (a) is verified if $\text{rank}(\mathbf{D}^{-\frac{1}{2}}\mathbf{U}') = \text{rank}(\mathbf{D}^{-\frac{1}{2}}\mathbf{U}' \vdots \mathbf{D}^{-\frac{1}{2}}\boldsymbol{\Delta})$; condition (b) is verified if $\text{rank}(\mathbf{U}') = \text{rank}\left(\begin{smallmatrix}\mathbf{1}' \\ \mathbf{U}'\end{smallmatrix}\right)$.

4. Estimating the Power of the Statistics

The asymptotic distribution of R_1, R_2, and M_r under a sequence of local alternatives can be derived from the asymptotic distribution of the cell residuals. Consider a sequence of local alternatives

$$\boldsymbol{\pi}_N = \boldsymbol{\pi}_0 + \frac{\boldsymbol{\delta}}{\sqrt{N}}, \tag{16}$$

where $\boldsymbol{\pi}_0 = \boldsymbol{\pi}(\boldsymbol{\theta}_0)$ denotes the probability vector specified under the null hypothesis. Assuming (16), and provided $\boldsymbol{\delta}$ is not too large, the asymptotic distribution of the cell residuals for the ML estimator is

$$\sqrt{N}(\mathbf{p} - \hat{\boldsymbol{\pi}}_0) \xrightarrow{d} N(\boldsymbol{\delta}, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \boldsymbol{\Gamma}_0 - \boldsymbol{\Delta}_0\big(\boldsymbol{\Delta}_0'\mathbf{D}_0^{-1}\boldsymbol{\Delta}_0\big)^{-1}\boldsymbol{\Delta}_0'. \tag{17}$$

Equation (17) follows from assuming that $\sqrt{N}(\mathbf{p}-\boldsymbol{\pi}_N) \xrightarrow{d} N(\mathbf{0},\boldsymbol{\Gamma}_0)$, or equivalently that $\sqrt{N}(\mathbf{p}-\boldsymbol{\pi}_0) \xrightarrow{d} N(\boldsymbol{\delta},\boldsymbol{\Gamma}_0)$, where $\boldsymbol{\Gamma}_0 = \mathbf{D}_0 - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'$. For the ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$, we have $\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0 \overset{a}{=} (\boldsymbol{\Delta}_0'\mathbf{D}_0^{-1}\boldsymbol{\Delta}_0)^{-1}\boldsymbol{\Delta}_0'\mathbf{D}_0^{-1}(\mathbf{p}-\boldsymbol{\pi}_N) = \mathbf{B}_0(\mathbf{p}-\boldsymbol{\pi}_N)$, where $\overset{a}{=}$ denotes asymptotic equality and $\boldsymbol{\Delta}_0 = \frac{\partial\boldsymbol{\pi}_0}{\partial\boldsymbol{\theta}'}$. Now, $\hat{\boldsymbol{\pi}}_0 \overset{a}{=} \boldsymbol{\pi}_0 + \boldsymbol{\Delta}_0(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \overset{a}{=} \boldsymbol{\pi}_0 + \boldsymbol{\Delta}_0\mathbf{B}_0(\mathbf{p}-\boldsymbol{\pi}_N)$. Equation (17) follows from $(\mathbf{p}-\hat{\boldsymbol{\pi}}_0) \overset{a}{=} (\mathbf{I}-\boldsymbol{\Delta}_0\mathbf{B}_0)(\mathbf{p}-\boldsymbol{\pi}_N) + \frac{\boldsymbol{\delta}}{\sqrt{N}}$, where we have used (16).

Thus, under the sequence of local alternatives (16), Pearson's X^2 is asymptotically distributed as a noncentral chi-square with $df = C-q-1$ and noncentrality parameter

$$\lambda = \boldsymbol{\delta}'\mathbf{D}_0^{-1}\boldsymbol{\delta} = N(\boldsymbol{\pi}_N-\boldsymbol{\pi}_0)'\mathbf{D}_0^{-1}(\boldsymbol{\pi}_N-\boldsymbol{\pi}_0), \tag{18}$$

since (16) implies $\boldsymbol{\delta} = \sqrt{N}(\boldsymbol{\pi}_N-\boldsymbol{\pi}_0)$.

Similarly, assuming (16) and (17), M_r is asymptotically noncentral chi-square with $df = s-q$ and noncentrality parameter

$$\lambda_r = \boldsymbol{\delta}'\mathbf{T}_r'\mathbf{C}_r\mathbf{T}_r\boldsymbol{\delta}. \tag{19}$$

This result follows from (9) and standard results in quadratic forms of normal random variables (e.g., Mathai & Provost, 1992) since $\mathbf{T}_r\sqrt{N}(\mathbf{p}-\hat{\boldsymbol{\pi}}_0) \xrightarrow{d} N(\mathbf{T}_r\boldsymbol{\delta},\boldsymbol{\Sigma}_r)$.

For the 1PL, using similar arguments and (11), R_1 and R_2 are asymptotically noncentral chi-square with $df = n(n-2)$ and $(n(n-2)+2)/2$, respectively, and non-centrality parameters

$$\lambda_R = \boldsymbol{\delta}'\mathbf{T}_R'\mathbf{C}_R\mathbf{T}_R\boldsymbol{\delta}. \tag{20}$$

Now, given a vector $\boldsymbol{\pi}_N$ and a null model $\boldsymbol{\pi}_0 = \boldsymbol{\pi}(\boldsymbol{\theta}_0)$, we estimate the non-centrality parameters λ in (18), λ_r in (19), and λ_R in (20) as follows: $\hat{\boldsymbol{\theta}}_0$ is obtained by minimizing the Kullback–Leibler (1951) discrepancy function

$$D_{KL}\big(\boldsymbol{\pi}_N,\boldsymbol{\pi}(\boldsymbol{\theta}_0)\big) = \boldsymbol{\pi}_N'\ln\big(\boldsymbol{\pi}_N/\boldsymbol{\pi}(\boldsymbol{\theta}_0)\big) = \boldsymbol{\pi}_N'\big[\ln(\boldsymbol{\pi}_N) - \ln\big(\boldsymbol{\pi}(\boldsymbol{\theta}_0)\big)\big] \tag{21}$$

and the noncentrality parameters are estimated by evaluating $\boldsymbol{\delta}$ and $\mathbf{D}_0$, $\boldsymbol{\delta}$ and $\mathbf{C}_r$, and $\boldsymbol{\delta}$ and $\mathbf{C}_R$ at $\hat{\boldsymbol{\pi}}_0 = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_0)$ for X^2, M_r and R_1, R_2, respectively. Note that the minimizer of (21) is the same as the maximizer of the maximum likelihood function between a "true" model $\boldsymbol{\pi}_N$ and a null model $\boldsymbol{\pi}(\boldsymbol{\theta}_0)$ (e.g., Jöreskog, 1994). For categorical data analysis, this procedure to estimate the noncentrality parameter was first used by Reiser (2008), and it is analogous to the procedure used by Satorra and Saris (1985) in structural equation modeling (SEM), except that in Satorra and Saris the function minimum (multiplied by sample size) yields an estimate of the noncentrality parameter. By contrast, here the noncentrality parameter needs to be computed given $\hat{\boldsymbol{\theta}}_0$.

In the next section, we shall assess the accuracy of these asymptotic approximations to the distribution of the statistics R_1, R_2 and M_2 under the null, and also under sequences of local alternatives.

5. An Empirical Comparison of R_1, R_2 and M_2

5.1. Accuracy of the Asymptotic p-Values Under Correct Model Specification

Table 1 reports the results of a simulation comparing empirical Type I error rates. Data were generated using the 1PL model (5) with a standard normal latent trait. Parameter estimation was performed using marginal maximum likelihood. The true value of the discrimination parameter was set to 1 and the true values of the difficulty parameters were $\mathbf{b}' = (-2.7,-2.1,-1.5,-0.9,-0.3,0.3,0.9,1.5,2.1,2.7)$. Three sample sizes were considered: $N = 300$, 500, and 1,000. A total of 1,000 replications per condition were used. All replications converged. Table 1 gives the empirical mean, variance, and rejection rates for R_1, R_2,

TABLE 1.
Empirical mean, variance, and rejection rates of R_1, R_2, and M_2 for correct model specification.

Stat	df	N	Mean	Var	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
R_1	80	300	78.6	432.6	0.07	0.11	0.14
		500	80.2	349.7	0.06	0.11	0.14
		1000	79.3	216.1	0.02	0.07	0.11
R_1^*	20	300	23.4	62.72	0.04	0.14	0.23
		500	23.7	54.04	0.04	0.13	0.24
		1000	23.5	47.93	0.04	0.12	0.22
R_2	45	300	46.6	248.1	0.05	0.10	0.15
		500	46.6	336.1	0.04	0.10	0.17
		1000	46.5	129.6	0.03	0.08	0.13
M_2	44	300	44.1	86.5	0.01	0.05	0.11
		500	44.0	79.2	0.01	0.05	0.09
		1000	44.2	86.5	0.01	0.05	0.10

Notes: $n = 10$; The true parameters were $a = 1$ and $\mathbf{b}' = (-2.7, -2.1, -1.5, -0.9, -0.3, 0.3, 0.9, 1.5, 2.1, 2.7)$. The sum score levels used in R_1 are $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ (no grouping of sum scores); in R_1^* they are $\{0, 1–3, 4–6, 7–9, 10\}$.

and M_2 at selected nominal rates ($\alpha = 0.01$, 0.05, and 0.10). The sample means of the statistics should be close to the degrees of freedom, and the sample variance close to twice the degrees of freedom. As we can see in this table, M_2 provides accurate empirical rates, whereas R_1 and R_2 tend to reject slightly more often than they should.

More extensive simulations comparing the empirical Type I rejection rates of R_1 and M_2 are reported in Montaño (2009), including model sizes from $n = 10$ to 20 and three levels of the true discrimination parameter a (0.5, 1, 1.5). M_2 was found to be very accurate across all conditions and more accurate than R_1. The accuracy of the p-values for R_1 was found to improve with decreasing model size and increasing discrimination. The highest empirical rejection rate for R_1 across all conditions was 17 % at the 5 % level, so the discrepancies between the empirical and asymptotic rates were not large.

The discrepancies between the empirical and asymptotic rejection rates for R_1 and R_2 occur because their empirical variances are much larger than those expected under their asymptotic distribution. For the asymptotic p-values of R_1 to be accurate, the expected frequencies of $N \Pr(Y_i = 1, X = k)$, should be large, say larger than 20 (Glas, 2009). This can be accomplished by grouping the sum scores in triplets of scores, quads, etc. That is, for the case of $n = 10$ items instead of using $X = 1, 2, \ldots, 9$, one can use the sum score ranges $X = 1–3, 4–6, 7–9$. This is the statistic R_1^* shown in Table 1. Thus, the sum scores are grouped in blocks of three scores. Glas (1988) points out that the asymptotic theory also applies if score ranges are used and that only the degrees of freedom need to be adjusted. They are now n times the number of score levels $-1 - q$. Unfortunately, when grouping sum scores in uniform blocks the discrepancy between observed and expected rejection rates reported in Table 1 appears larger than for the ungrouped R_1 statistic because the empirical mean of R_1^* is larger than the expected mean.

We considered alternative groupings of score levels and the results are provided in Table 2. For this configuration of item parameters, it appears that grouping the sum scores in uniform blocks of three scores as we did in Table 1 is the worst choice. In all the cases shown in Table 2, grouping the scores reduces the number of degrees of freedom; and, as a result, the discrepancy between the empirical and asymptotic variance of the statistic is also reduced, particularly in small samples. Also, grouping of sum scores can be performed using an iterative procedure after

TABLE 2.
Empirical mean, variance, and rejection rates of R_1^* with alternative sum score groupings.

Sum score grouping	df	N	Mean	Var	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
{0, 1–3, 4–5, 6–7, 8–10}	30	300	30.24	71.70	0.02	0.06	0.12
		500	30.42	63.64	0.02	0.06	0.11
		1000	30.25	59.51	0.01	0.06	0.11
{0, 1–2, 3–4, 5, 6–7, 8–9, 10}	40	300	40.53	101.39	0.02	0.07	0.12
		500	40.67	91.44	0.02	0.07	0.13
		1000	40.86	88.11	0.01	0.06	0.12
{0, 1–3, 4, 5, 6, 7–9, 10}	40	300	40.00	75.57	0.01	0.04	0.09
		500	40.33	76.95	0.01	0.05	0.10
		1000	40.49	84.44	0.02	0.05	0.11

Notes: The number of items and true parameters are the same as in Table 1.

observing the parameter estimates (for details, see Suárez-Falcon and Glas, 2003) leading to more accurate empirical Type I errors. Thus, when appropriate score ranges are used, the empirical rejection rates of R_1 should closely match the theoretical rejection rates. A similar grouping of the sum scores for the univariate moments in R_2 could be used to reduce the discrepancy between empirical and theoretical rejection rates, but this was not attempted here.

5.2. *Power to Reject a 2PL*

In this subsection, we use the asymptotic approximation to the power of the test statistics described in the previous section without having to use simulations for each condition of interest. For some of the conditions investigated, we provide simulation results in the next subsection to gauge the performance of the asymptotic approximation to the power of the statistics.

Power was approximated using asymptotic methods for three levels of average slopes (0.5, 1, and 1.5), two model sizes ($n = 10$ and 15), and three levels of sample size ($N = 300$, 500, and 1,000). The actual $\mathbf{a}$ values used were for $n = 10$ and $\bar{\mathbf{a}} = 0.5$, $\mathbf{a}' = (0.35, 0.25, 0.8, 0.5, 0.6, 0.6, 0.5, 0.8, 0.25, 0.35)$; for $\bar{\mathbf{a}} = 1$, $\mathbf{a}' = (1.05, 0.85, 1.5, 0.6, 1.0, 1.0, 0.6, 1.5, 0.85, 1.05)$; and for $\bar{\mathbf{a}} = 1.5$, $\mathbf{a}' = (2.0, 1.35, 1.65, 1.0, 1.5, 1.5, 1.0, 1.65, 1.35, 2.0)$. For $n = 15$ and $\bar{\mathbf{a}} = 0.5$, $\mathbf{a}' = (0.6, 0.35, 0.25, 0.8, 0.5, 0.6, 0.4, 0.5, 0.4, 0.6, 0.5, 0.8, 0.25, 0.35, 0.6)$; for $\bar{\mathbf{a}} = 1$, $\mathbf{a}' = (1.1, 1.05, 0.85, 1.5, 0.6, 1.0, 0.9, 1.0, 0.9, 1.0, 0.6, 1.5, 0.85, 1.05, 1.1)$; and for $\bar{\mathbf{a}} = 1.5$, $\mathbf{a}' = (1.6, 2.0, 1.35, 1.65, 1.0, 1.5, 1.4, 1.5, 1.4, 1.5, 1.0, 1.65, 1.35, 2.0, 1.6)$. Results are given in Table 3 for R_1, R_1^*, R_2, and M_2. In R_1^* the sum scores are grouped {0, 1–3, 4–6, 7–9, 10} for $n = 10$; and {0, 1–3, 4–6, 7–8, 9–11, 12–14, 15} for $n = 15$. We see in Table 3 that the expected rejection rates for R_1, R_2, and M_2 increase with increasing discrimination and decrease with increasing model size. Obviously, they also increase with increasing sample size. None of these statistics has power to distinguish a 1PL from a 2PL when the average slope is 0.5 even at $N = 1{,}000$. When $\bar{\mathbf{a}} \geq 1$, all statistics have acceptable power (>50 %) only when $N = 1{,}000$; and $N = 500$ if $\bar{\mathbf{a}} = 1.5$ and $n = 10$. We also see in this table that power is uniformly higher for M_2 than for R_1, R_1^*, and R_2.

For the purposes of comparison, Table 3 also includes results for Pearson's X^2. We see in this table that R_1, R_1^*, R_2, and M_2 are considerably more powerful than X^2 in distinguishing a 1PL from a 2PL. For X^2, expected power often increases only slightly with increasing sample size and it did not reach 30 % for any of the conditions investigated.

TABLE 3.
Asymptotic power rates of R_1, M_2, and Pearson's X^2 at $\alpha = 0.05$ when the true model is a two-parameter logistic model.

$\bar{a}_i$	n	N	R_1	R_1^*	R_2	M_2	X^2
0.5	10	300	0.08	0.09	0.08	0.10	0.06
		500	0.11	0.12	0.10	0.14	0.06
		1000	0.19	0.22	0.17	0.27	0.08
	15	300	0.08	0.08	0.07	0.09	0.05
		500	0.10	0.10	0.09	0.13	0.05
		1000	0.17	0.17	0.15	0.24	0.06
1.0	10	300	0.24	0.26	0.20	0.34	0.09
		500	0.44	0.45	0.36	0.61	0.12
		1000	0.85	0.82	0.74	0.95	0.24
	15	300	0.18	0.18	0.16	0.26	0.06
		500	0.32	0.32	0.28	0.47	0.06
		1000	0.71	0.69	0.63	0.89	0.07
1.5	10	300	0.29	0.28	0.26	0.40	0.09
		500	0.53	0.50	0.46	0.69	0.14
		1000	0.93	0.87	0.86	0.98	0.29
	15	300	0.23	0.22	0.20	0.30	0.06
		500	0.44	0.40	0.36	0.56	0.06
		1000	0.87	0.80	0.77	0.94	0.08

Notes: No grouping of sum score levels is used in R_1. In R_1^*, the sum scores are grouped {0, 1–3, 4–6, 7–9, 10} for $n = 10$; and {0, 1–3, 4–6, 7–8, 9–11, 12–14, 15} for $n = 15$.

5.3. *Power to Reject a 2PL: Accuracy of the Asymptotic p-Values Under Model Misspecification*

We performed a simulation study to investigate how well the asymptotic procedure used in the previous subsection approximates the empirical rejection rates under model misspecification. Empirical rejection rates for R_1, R_1^*, R_2, and M_2, $n = 10$ and $\bar{\mathbf{a}} = 1$ are shown in Table 4. They are not adjusted for differential empirical Type I errors. A comparison of empirical and asymptotic rejection rates across a much larger number of conditions, but only for R_1 and M_2, is reported in Montaño (2009) where on average and across conditions the absolute deviation at $\alpha = 0.05$ is 0.06 for R_1 and 0.04 for M_2.

A comparison of the results of Tables 3 and 4 shows that the proposed asymptotic procedure approximates fairly well the empirical rejection rates of R_1, R_2, and M_2, but not of R_1^*. Results are not shown for X^2 in this table because it is well known that asymptotic rates of X^2 are only accurate in very small models (Koehler & Larntz, 1980; Agresti & Yang, 1987). For models that are not so small, p-values based on asymptotic methods almost invariably lead X^2 to reject the model, even for correctly specified models.

5.4. *Power to Reject a 3PL*

In this subsection, we use asymptotic methods to approximate the power of R_1, R_1^*, R_2, M_2 and X^2 to reject a 3PL model. The setup was identical to the setup in the previous section except that now $c = 0.1$ or 0.25, whereas in the previous subsection, $c = 0$. Table 5 lists the asymptotic power at $\alpha = 0.05$ for these statistics when fitting a 1PL to these 3PL models. Table 3 lists the asymptotic power when $c = 0$ (the 2PL model) at the same parameter values. We see in

TABLE 4.
Empirical power rates of R_1, R_2, and M_2 when the true model is a two-parameter logistic model.

Stat	df	N	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
R_1	80	300	0.15	0.26	0.35
		500	0.26	0.44	0.55
		1000	0.68	0.84	0.90
R_1^*	20	300	0.16	0.40	0.52
		500	0.36	0.62	0.73
		1000	0.76	0.92	0.96
R_2	45	300	0.12	0.25	0.33
		500	0.22	0.39	0.51
		1000	0.53	0.76	0.85
M_2	44	300	0.16	0.32	0.47
		500	0.38	0.61	0.73
		1000	0.86	0.95	0.98

Notes: $n = 10$; The true parameters are $\mathbf{a}' = (1.05, 0.85, 1.5, 0.6, 1.0, 1.0, 0.6, 1.5, 0.85, 1.05)$ and $\mathbf{b}' = (-2.7, -2.1, -1.5, -0.9, -0.3, 0.3, 0.9, 1.5, 2.1, 2.7)$.

TABLE 5.
Asymptotic power rates of R_1, M_2, and Pearson's X^2 at $\alpha = 0.05$ when the true model is a three-parameter logistic model.

$\bar{a}_i$	n	N	$c = 0.1$					$c = 0.25$				
			R_1	R_1^*	R_2	M_2	X^2	R_1	R_1^*	R_2	M_2	X^2
0.5	10	300	0.08	0.09	0.08	0.10	0.06	0.07	0.08	0.07	0.09	0.06
		500	0.11	0.11	0.10	0.14	0.06	0.09	0.10	0.08	0.12	0.06
		1000	0.20	0.21	0.17	0.29	0.08	0.15	0.16	0.13	0.23	0.07
	15	300	0.08	0.07	0.07	0.09	0.05	0.07	0.07	0.06	0.09	0.05
		500	0.10	0.09	0.08	0.12	0.05	0.09	0.08	0.07	0.12	0.05
		1000	0.16	0.15	0.13	0.23	0.06	0.14	0.12	0.10	0.22	0.06
1.0	10	300	0.23	0.24	0.14	0.29	0.09	0.21	0.22	0.10	0.29	0.09
		500	0.41	0.42	0.22	0.52	0.12	0.37	0.37	0.15	0.52	0.12
		1000	0.82	0.80	0.48	0.91	0.23	0.78	0.73	0.32	0.90	0.23
	15	300	0.24	0.24	0.13	0.29	0.06	0.27	0.24	0.11	0.38	0.06
		500	0.44	0.44	0.21	0.54	0.07	0.50	0.43	0.18	0.68	0.07
		1000	0.87	0.85	0.46	0.93	0.09	0.92	0.84	0.38	0.98	0.10
1.5	10	300	0.39	0.44	0.18	0.35	0.12	0.21	0.22	0.10	0.29	0.09
		500	0.69	0.72	0.31	0.62	0.19	0.37	0.37	0.15	0.52	0.11
		1000	0.98	0.98	0.66	0.96	0.44	0.78	0.73	0.32	0.90	0.23
	15	300	0.55	0.63	0.15	0.35	0.07	0.68	0.66	0.15	0.56	0.08
		500	0.87	0.91	0.26	0.64	0.09	0.95	0.93	0.26	0.87	0.11
		1000	1.00	1.00	0.59	0.97	0.15	1.00	1.00	0.59	1.00	0.22

Notes: No grouping of sum score levels is used in R_1. In R_1^* the sum scores are grouped {0, 1–3, 4–6, 7–9, 10} for $n = 10$; and {0, 1–3, 4–6, 7–8, 9–11, 12–14, 15} for $n = 15$.

TABLE 6.
Asymptotic power rates of R_1, M_2, and Pearson's X^2 at $\alpha = 0.05$ when the true model is a two-dimensional one-parameter logistic model, $n = 10$.

a	N	$\rho = 0$					$\rho = 0.5$					$\rho = 0.9$				
		R_1	R_1^*	R_2	M_2	X^2	R_1	R_1^*	R_2	M_2	X^2	R_1	R_1^*	R_2	M_2	X^2
0.5	300	0.06	0.07	0.36	0.22	0.07	0.05	0.08	0.07	0.08	0.06	0.05	0.05	0.05	0.05	0.05
	500	0.06	0.08	0.45	0.39	0.09	0.05	0.09	0.10	0.11	0.06	0.05	0.05	0.05	0.05	0.05
	1000	0.07	0.12	0.85	0.78	0.16	0.06	0.12	0.20	0.19	0.07	0.05	0.05	0.05	0.05	0.05
1.0	300	0.30	0.45	0.95	0.99	0.65	0.10	0.52	0.68	0.43	0.10	0.05	0.06	0.05	0.06	0.05
	500	0.54	0.65	0.98	1.00	0.95	0.15	0.82	0.93	0.72	0.15	0.05	0.06	0.05	0.06	0.05
	1000	0.93	0.95	1.00	1.00	1.00	0.30	0.99	1.00	0.99	0.34	0.06	0.08	0.05	0.08	0.06
1.5	300	0.95	0.95	0.98	1.00	1.00	0.32	0.62	0.55	0.87	0.34	0.06	0.08	0.08	0.07	0.05
	500	1.00	1.00	1.00	1.00	1.00	0.59	0.90	0.85	0.99	0.64	0.06	0.11	0.11	0.08	0.06
	1000	1.00	1.00	1.00	1.00	1.00	0.95	1.00	0.99	1.00	0.99	0.07	0.19	0.19	0.12	0.06

Notes: No grouping of sum score levels is used in R_1. In R_1^* the sum scores are grouped {0, 1–3, 4–6, 7–9, 10}.

these tables that the main determinant of power, in addition to sample size, is the average item discrimination. The larger the average discrimination, the higher the power. The value of the guessing parameter only influences the power of these tests through its interaction with model size. Also, usually there is more power to distinguish a 3PL from a 1PL than a 2PL from a 1PL. Furthermore, power need not be higher at $c = 0.25$ than at $c = 0.1$. Again, the power of X^2 is low and does not reach 30 % for any of the conditions investigated. Also, power for R_2 is lower than for R_1, R_1^*, and M_2. When distinguishing a 1PL from a 3PL, the power of M_2 is slightly higher than for R_1 and R_1^* except for most conditions involving $\bar{\mathbf{a}} = 1.5$, where the reverse is true.

All in all, neither R_1 nor M_2 show a high power to detect the presence of guessing. Thus, power is unacceptable (<50 %) for $\bar{\mathbf{a}} = 0.5$, even at $N = 1{,}000$. Even for $\bar{\mathbf{a}} \geq 1$ a sample of size 500 is needed to ensure the detection of the presence of guessing.

5.5. *Power to Reject a Multidimensional Model*

In this subsection, we used asymptotic methods to compute the power of R_1, R_1^*, R_2, M_2, and X^2 to reject a two-dimensional 1PL model for $n = 10$. More specifically, we let the density of the latent traits be standard normal with correlation coefficient $\rho = 0$, 0.5, or 0.9. Also, now

$$P(Y_i = 1|\eta_1, \eta_2) = \frac{1}{1 + \exp[-a_{1i}\eta_1 - a_{2i}\eta_2 + d_i]} \tag{22}$$

with $\mathbf{a}_1' = a \odot (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$, $\mathbf{a}_2' = a \odot (0, 0, 0, 0, 0, 1, 1, 1, 1, 1)$, $a = 0.5$, 1, or 1.5, $\mathbf{d} = a \odot \mathbf{b}$, and $\odot$ denotes a Hadamard (elementwise) product. The same $\mathbf{b}$ parameters used throughout were also used here. Table 6 provides the asymptotic power at $\alpha = 0.05$ for the three statistics when fitting a one-dimensional 1PL (i.e., assuming that $\rho = 1$) but the true $\rho = 0$, 0.5, or 9. As can be seen in this table, for all the statistics considered power is roughly equal to the significance level when the correlation between the traits is 0.9; but it is at least 95 % when the correlation between the traits is zero and $\bar{\mathbf{a}} = 1.5$. For the remaining conditions, power for R_2 and M_2 is higher than for R_1, R_1^*, and X^2.

TABLE 7.
Results of fitting a 1PL model and a 2PL model to the LSAT 7 data and asymptotic power at $\alpha = 0.05$ for distinguishing between a 1PL model and a 2PL model.

	1PL				Power at $\alpha = 0.05$		
	Value	df	p	Value/df	Non-centrality	df	Power
R_1	31.95	15	0.01	2.13	12.92	15	0.59
R_2	34.65	10	<0.01	3.47	24.69	6	0.98
M_2	23.17	9	0.01	2.57	13.13	9	0.71
X^2	44.15	25	0.01	1.77	13.24	25	0.49

6. Numerical Examples

In this section, we provide two numerical examples. In both examples, a 1PL model was fitted by maximum likelihood. In the first example we fit the model to the well-known LSAT 7 data of Bock and Lieberman (1970). These data consist of $N = 1{,}000$ observations on $n = 5$ variables. Because the data are not sparse, we expect R_1, R_2, and M_2 to lead to the same conclusions as X^2. We also fit a 2PL model to these data and give details on how to approximate the power of R_1, R_2, M_2, and X^2 to distinguish a 1PL from a 2PL.

In the second example, we fit a 1PL to $n = 15$ variables. In this sparse situation, we expect the likelihood ratio statistic to yield a p-value of 1 and Pearson's X^2 a p-value of 0, suggesting that the asymptotic approximations to the distribution of these statistics are inappropriate. By contrast, the simulation results reported in Montaño (2009) suggest that the p-values of M_2 should be right on target and that the p-values of R_1 should be slightly inflated. In sparse settings like these, the accuracy of the p-values of R_1 can be improved by judicious grouping of the sum scores. We use R_1^* to refer to R_1 based on grouped sum scores. An extrapolation of the results presented in Tables 1 and 2 would suggest that the p-values of R_2 should be also slightly inflated in sparse settings, although in this case a grouping of the sum scores could also be used to obtain more accurate p-values.

6.1. LSAT 7 Data

Table 7 lists the values of the statistics R_1, R_2, M_2, and X^2, their degrees of freedom, and ratios of the test statistics to their degrees of freedom. Also provided in the table is the power of the statistics to reject a 1PL at the $\alpha = 0.05$ nominal level if the true model is a 2PL.

The power of the statistics is easily computed using the asymptotic approximation described previously. All that is needed is a program for ML estimation that computes R_1, M_2, and X^2 and that allows the input of response patterns, sample proportions, and sample size. First, the alternative model is estimated (in this case a 2PL). The estimated cell probabilities under the alternative model are then used as if they were sample proportions for estimating the null model (in this case a 1PL model) by ML with sample size equal to the actual sample size (in this case 1,000). This amounts to minimizing the Kullback–Leibler divergence (21). The R_1, M_2, and X^2 statistics estimated in this fashion are the estimated noncentrality parameters (18), (19), and (20). Then power for statistic $t = R_1$, R_2, M_2, or X^2 is computed as $1 - F(k_t; \nu_t, \lambda_t)$, where $F(\bullet; \nu_t, \lambda_t)$ is the noncentral chi-square distribution function with ν_t degrees of freedom and non-centrality parameter λ_t, and k_t is the upper α quantile of a chi-square distribution with ν_t degrees of freedom.

Because these data are not sparse, the asymptotic p-values and power values of X^2 are accurate in this case. Table 7 reveals an extremely high agreement among the p-values for all the statistics. A 1PL should be rejected at the $\alpha = 0.05$ significance level. However, there are

TABLE 8.
Results of fitting a 1PL to the Chilean mathematical achievement data: Estimated difficulty parameters and goodness-of-fit tests.

Item	Value	SE
1	4.08	0.10
2	3.31	0.08
3	2.98	0.08
4	2.19	0.07
5	1.67	0.06
6	0.82	0.06
7	0.44	0.06
8	−0.35	0.06
9	−1.47	0.06
10	−1.85	0.07
11	−2.50	0.07
12	−3.37	0.09
13	−3.91	0.10
14	−4.32	0.11
15	−4.84	0.13

Stat	Value	df	p-value	Value/df
R_1	227.69	195	0.054	1.17
R_1^*	70.41	60	0.168	1.17
R_2	149.27	105	0.003	1.42
M_2	130.66	104	0.040	1.26
X^2	7,324,094.30	32,751	0	223.63
G^2	1,420.90	32,751	1	0.04

Notes: $N = 2{,}810$, $\hat{a} = 1.81$, SE $= 0.03$.

differences in power in the statistics considered, with R_2 being more powerful than M_2, which in turn is more powerful than R_1, and X^2 being the least powerful statistic to distinguish between a 1PL and a 2PL.

Notice the agreement in ordering the models between the value/df ratios for the 1PL with the power to distinguish a 1PL from a 2PL. Joe and Maydeu-Olivares (2010) introduced a very general family of test statistics that includes R_1, M_2 and X^2 as members. It remains to be seen whether R_2 is a member of this family. For their family of statistics, Joe and Maydeu-Olivares provide theory that explains why statistics with a higher value/df ratio are generally more powerful to distinguish the fitted model from alternatives of interest.

6.2. *Chilean Mathematical Proficiency Data*

We fitted a 1PL model to the responses of a 15-item test aimed at measuring mathematical proficiency in Chilean adults. The sample size was 2,810, ages ranged from 17 to 77 years, and 56.6 % of the respondents were women. Table 8 lists the estimated parameters and their standard errors, as well as goodness-of-fit results.

These data are very sparse and, as expected, the G^2 statistic yields a p-value of 1 and Pearson's X^2 a p-value of 0. The asymptotic approximation to the distribution of these statistics is completely useless in this case. In contrast, the simulation results presented earlier reveal that the asymptotic p-values for M_2 are very accurate and that the p-values for R_1, R_1^* and R_2 are somewhat optimistic. As this example reveals, however, the p-values for R_1 may be quite close

to those of M_2. Indeed, in this example R_1 and M_2 yield quite similar p-values suggesting that the fit of the model is reasonable given the restrictiveness of the model, the number of items involved, and sample size. In contrast, R_1^* suggests a better fit than R_1 and M_2, whereas R_2 suggests that the model's fit is poorer.

Note that the value/df ratio for R_2, M_2, and R_1 suggest that R_2 is the most powerful of the three to detect misspecifications from the 1PL model. Also note that the value/df ratio for X^2 in this case (over 200) is much larger. This should not be taken to mean that X^2 is the most powerful statistic in this case because X^2 always rejects a model of this size, regardless of whether it is correct. Comparing the power of statistics is only meaningful when they have the same null empirical rejection rates.

7. Discussion and Concluding Remarks

We have compared two overall test statistics specifically designed to target specific assumptions of Rasch-type models, namely R_1 and R_2 (Glas, 1988), with a general purpose test statistic, M_2 (Maydeu-Olivares & Joe, 2005). All three are quadratic form limited information statistics, meaning they do not use all the information available for testing. Rather, they concentrate the information available in the contingency table cells in some summaries with large expected counts so that the resulting statistic is better approximated by asymptotic methods in small samples: M_2 uses as summaries univariate and bivariate moments, i.e., means and cross-products. For n variables, there is a one-to-one map between the set of means and cross-products used in M_2 and the set of all bivariate probabilities (Maydeu-Olivares & Liu, 2012). Thus, M_2 can alternatively be described as a quadratic form goodness-of-fit statistic in all bivariate tables. By contrast, R_1 uses as summaries univariate moments for each sum score. Finally, R_2 uses as summaries the univariate moments corresponding to a sum score of one and bivariate moments excluding the perfect score.

The distribution of M_2 is more closely approximated using asymptotic methods than the distributions of R_1 and R_2. The reason for this is that as the number of items grows, some of the univariate probabilities involved in R_1 and R_2 will become small, hindering the asymptotic approximation. For R_1, this problem can be overcome (Glas, 1988) by breaking down the univariate moments by sum score ranges, so that the resulting summaries have larger expected probabilities. It may be difficult to determine how to best group the sum scores into ranges before the data is seen, and grouping the sum score ranges in equidistant ranges does not seem to be a good choice; but, however the sum score ranges are grouped, our simulation results suggest that the discrepancies between the empirical distributions of R_1 and R_2 and their reference asymptotic distributions are not large and are in any case very much smaller than for full information statistics such as X^2. Then the choice between R_1, R_2, and M_2 could be based largely on the power of each statistic to detect model misspecification. Examining the power of test statistics using simulations is time-consuming and in this paper we have used a procedure that enables researchers to approximate the power of test statistics for categorical data analysis using asymptotic methods. Our simulation results suggest that the procedure yields a good approximation to the distribution of the statistics under model misspecification. That is, if the distribution of the statistic is well approximated by asymptotic methods under the null, it is likely that it is also well approximated by our procedure under model misspecification.

Armed with this asymptotic approximation to the power of the statistics, when a 1PL model is fitted we have examined the power of R_1, R_2, and M_2 to detect (a) unequal slopes (the true model is a 2PL), (b) the presence of guessing and unequal slopes (the true model is a 3PL), and (c) multidimensionality (the true model is a multidimensional 1PL). When fitting a 1PL model,

R_1 was designed to target response function misspecification and R_2 to target multidimensionality. By contrast, M_2 was designed as an omnibus test statistic, with neither a null model nor an alternative model in mind. We found that M_2 is more powerful than R_1 and R_2 to distinguish a 1PL from a 2PL, and also to distinguish a 1PL from a 3PL except when the slopes are high ($\bar{a} = 1.5$), in which case R_2 is most powerful. M_2 is also the most powerful statistic to distinguish a 1PL from a multidimensional 1PL except when the slopes are low ($\bar{a} = 0.5$), in which case R_2 is most powerful. All three statistics are asymptotically more powerful than X^2 to distinguish a 1PL from a 2PL, a 3PL and a multidimensional 1PL. This should not be taken to imply that limited information statistics are always more powerful than X^2 as Reiser (2008) has shown that statistics based on univariate and bivariate information are less powerful than X^2 if the true model contains three-way associations. In practical applications involving the fit of a 1PL, the true model is unknown. Fortunately, Joe and Maydeu-Olivares (2010) have shown that for the members of the family of statistics they describe (which includes R_1, M_2 and X^2) the statistic with the highest value/df ratio will be most powerful against most alternatives.

On a final note, the reader should bear in mind that we have compared just four of the test statistics that can be used to assess the goodness-of fit of Rasch-type models: R_1, R_2, M_2, and X^2. Many others could be used, some specifically proposed for Rasch-type models (see Glas & Verhelst, 1995), others being omnibus tests (see, for example, Reiser, 1996, 2008; Cai et al., 2006); and the asymptotic approximation described enables researchers to assess their power. However, the limited results presented here suggest that there is no clear advantage in using goodness-of-fit statistics specifically designed for Rasch-type models to test these models when marginal ML estimation is used. However, if Rasch models are estimated using conditional ML estimation, then different statistics are needed and there are counterparts of the R_1 and R_2 test statistics which can be used in this case but which do not exist in the case of the M_2 statistic.

Appendix

Relationship $\boldsymbol{\pi}_{R2} = \mathbf{T}_{R2}\boldsymbol{\pi}$ for $n = 4$ items

$$\boldsymbol{\pi}_{R2} = \mathbf{T}_{R2}\boldsymbol{\pi} = \begin{pmatrix} \pi_0 \\ \boldsymbol{\pi}_{1|x=1} \\ \boldsymbol{\pi}_{2|2\leq x\leq 4} \\ \pi_4 \end{pmatrix}$$

$$= \left(\begin{array}{c:ccc:cccccc:ccc:c} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ \hdashline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \begin{pmatrix} \pi_{(0,0,0,0)} \\ \pi_{(1,0,0,0)} \\ \vdots \\ \pi_{(0,0,0,1)} \\ \pi_{(1,1,0,0)} \\ \pi_{(1,0,1,0)} \\ \vdots \\ \pi_{(0,0,1,1)} \\ \pi_{(1,1,1,0)} \\ \vdots \\ \pi_{(0,1,1,1)} \\ \pi_{(1,1,1,1)} \end{pmatrix}.$$

References

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

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.

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

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

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*, *35*, 179–197.

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

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

De Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational and Behavioral Statistics*, *11*, 183–196.

Fischer, G.H. & Molenaar, I.W. (Eds.) (1995). *Rasch models: foundations, recent developments and applications*. New York: Springer.

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

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

Glas, C.A.W., & Verhelst, N.D. (1995). Testing the Rasch model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: foundations, recent developments and applications* (pp. 69–96). New York: Springer.

Glas, C.A.W. (2009). Personal communication.

Irtel, H. (1995). An extension of the concept of specific objectivity. *Psychometrika*, *60*, 115–118.

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

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

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

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

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

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

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

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

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

Maydeu-Olivares, A., & Joe, H. (2008). An overview of limited information goodness-of-fit testing in multidimensional contingency tables. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 253–262). Tokyo: Universal Academy Press.

Maydeu-Olivares, A., & Liu, Y. (2012). Item diagnostics in multivariate discrete data. Manuscript under review.

Mavridis, D., Moustaki, I., & Knott, M. (2007). Goodness-of-fit measures for latent variable models for binary data. In S.-Y. Lee (Ed.), *Handbook of latent variables and related models* (pp. 135–162). Amsterdam: Elsevier.

McDonald, R.P. (1999). *Test theory: a unified treatment*. Mahwah: Lawrence Erlbaum.

Montaño, R. (2009). *Una comparación de las estadísticas de bondad de ajuste R_1 y M_2 para modelos de la Teoría de Respuesta al Ítem [Comparing the R_1 and M_2 statistics for goodness of fit assessment in IRT models]*. Unpublished Ph.D. dissertation, University of Barcelona.

Pfanzagel, J. (1993). A case of asymptotic equivalence between conditional and marginal maximum likelihood estimators. *Journal of Statistical Planning and Inference*, *35*, 301–307.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.

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

Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical & Statistical Psychology*, *61*, 331–360.

Satorra, A., & Saris, W.E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83–90.

Suárez-Falcon, J.C., & Glas, C.A.W. (2003). Evaluation of global testing procedure for item fit to the Rasch model. *British Journal of Mathematical & Statistical Psychology*, *56*, 127–143.

Swaminathan, H., Hambleton, R.K., & Rogers, H.J. (2007). Assessing the fit of item response models. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 683–718). Amsterdam: Elsevier.

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

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic models. *Psychometrika*, *47*, 175–186.
van den Wollenberg, A.L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123–139.

Manuscript Received: 12 JAN 2009
Final Version Received: 26 FEB 2012
Published Online Date: 20 OCT 2012