CRESST REPORT 839

A NEW STATISTIC FOR EVALUATING ITEM RESPONSE THEORY MODELS FOR ORDINAL DATA

MARCH, 2014

Li Cai

Scott Monroe

National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

A New Statistic for Evaluating Item Response Theory Models for Ordinal Data

CRESST Report 839

Li Cai and Scott Monroe
CRESST/University of California, Los Angeles

March 2014

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Institute of Education Sciences statistical methodology grant (R305D100039). Li Cai's research is also supported by grants from the Institute of Education Sciences (R305B080016) and the National Institute on Drug Abuse (R01DA026943 and R01DA030466).

The findings and opinions expressed here do not necessarily reflect the positions or policies of the Institute of Education Sciences, or the National Institute on Drug Abuse.

To cite from this report, please use the following as your APA reference: Cai, L. & Monroe, S. (2014). *A New Statistic for Evaluating Item Response Theory Models for Ordinal Data* (CRESST Report 839). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

A NEW STATISTIC FOR EVALUATING ITEM RESPONSE THEORY MODELS FOR ORDINAL DATA

Li Cai and Scott Monroe
CRESST/ University of California, Los Angeles

Abstract

We propose a new limited-information goodness of fit test statistic C_2 for ordinal IRT models. The construction of the new statistic lies formally between the M_2 statistic of Maydeu-Olivares and Joe (2006), which utilizes first and second order marginal probabilities, and the M_2^* statistic of Cai and Hansen (2013), which collapses the marginal probabilities into means and product moments. Unlike M_2^*, C_2 may be computed even when the number of items is small and the number of categories is large. It is as well calibrated as the alternatives and can be more powerful than M_2. When all items are dichotomous, C_2 becomes equivalent to M_2^*, which is also equivalent to M_2. We analyze empirical data from a patient-reported outcomes measurement development project to illustrate the potential differences in substantive conclusions that one may draw from the use of different statistics for model fit assessment.

Keywords: item response theory, goodness of fit, limited-information

Introduction

Recent years have witnessed an increased interest in the formal evaluation of item response theory (IRT) models (Maydeu-Olivares, 2013). In particular, great technical strides have been made in the area of limited-information fit statistics (Bartholomew & Leung, 2002; Maydeu-Olivares & Joe, 2006; Joe & Maydeu-Olivares, 2010). In contrast to the classical full-information statistics such as Pearson's X^2 statistic or the likelihood ratio statistic G^2, which utilize full response pattern frequencies and residuals, these limited-information statistics are based on observed and model-implied lower-order margins, e.g., first- and second order marginal frequencies. As Thissen and Steinberg (1997) noted, for a handful of polytomous items, the contingency table upon which the item response model is defined becomes extremely sparse. The full-information test statistics do not approach the asymptotic chi-square reference distributions under such sparseness (see e.g., Bartholomew & Tzamourani, 1999) and consequently have limited utility for evaluating IRT models, particularly for ordinal data. On the other hand, because the lower-order margins tend to be better filled than the full item response cross-classifications, limited-information statistics not only retain better calibration than full-information statistics under the null, but are also more powerful under the alternative (Joe & Maydeu-Olivares, 2010). Maydeu-Olivares and Joe's (2006) M_2 and Cai and Hansen's (2013)

M_2^* statistics are examples that have found their way into widely distributed software (Cai, Thissen, & du Toit, 2011; Cai, 2013) and have begun to demonstrate their usefulness in empirical measurement research that requires evaluating IRT model fit for ordinal data.

Technical and practical challenges remain, however, and we submit that both the original M_2 statistic, which is based on uncollapsed first- and second order marginal residuals, and the Cai-Hansen updated M_2^* statistic, which utilizes a further condensing/collapsing of the first- and second order marginal residuals into residual moments, have limitations that result in diminished practical utility for IRT models fitted to ordinal data. The original M_2 statistic suffers from a more subtle sparseness issue that Cai & Hansen (2013) discussed. For example, suppose two ordinal items each with 5 categories (perhaps on a Likert-type scale) both load strongly on the same latent variable(s). Then, the observed item responses will tend to covary. Respondents who endorse the extreme response categories for item 1 tend to have similar responses for item 2. By virtue of the shared underlying latent variable(s), certain cells in the bivariate contingency table will have very small expected frequencies, e.g., the combination of the most positive response option on item 1 and the most negative response option on item 2. The number of cells that may be sparse is exacerbated by an increase in the number of categories, eventually leading to a break-down of the asymptotic chi-square approximation. Generally the statistic will be stochastically smaller than the reference distribution, leading to lower than nominal Type I error rates under the null, and a loss of statistical power under the alternative. Cai and Hansen (2013) proposed M_2^* as a remedy because it uses conventional item scores (0-1-2-3-…) assigned to ordinal categories to compute residual moments from the first- and second order margins. Given our 2-item 5-category example from above, the original M_2 would require $2 \times (5-1) = 8$ first order expected marginal probabilities, and $(5-1) \times (5-1) = 16$ second order expected marginal probabilities.[1] For M_2^*, the 8 first order marginal expected probabilities are used to compute 2 first order moments, and the 16 second order expected marginal probabilities collapses into a single second order moment. This further collapsing guarantees that sparseness no longer affects M_2^*, and the statistic is shown to be well-calibrated and powerful in Cai and Hansen's (2013) simulations.

A major issue still remains: M_2^* appears to have collapsed the contingency table far too aggressively. For assessments made up of items with 5 ordered response categories (very popular in social and behavioral sciences research), it would take at least 10 items for M_2^* to begin to have positive degrees of freedom, even for a simple unidimensional graded response model. For 9 items, there are $9 + 9 \times (9-1)/2 = 45$ first- and second order residual moments, but a

[1] The number of first order marginal probabilities is equal to 4 because the 5 probabilities must sum to 1.0 and there are only 4 independent probabilities. The same argument applies to the bivariate table.

unidimensional graded response model for 5 categories also has 45 parameters, leaving zero degree of freedom for model fit testing with M_2^*. In this case, the IRT model is not locally identified from the set of marginal residual moments. We note that, for instance, in the patient-reported outcomes measurement context (Hansen, Cai, Stucky, Tucker, Shadel, & Edelen, in press), most short forms contain fewer than 10 items and the items tend to have ordinal response formats. Thus, M_2^* cannot be used to evaluate model fit for such short form measures. This clearly limits the utility of M_2^*.

However, a closer examination of the logic of the further collapsing used in M_2^* shows that the situations for the first order margins and the second order margins are in fact very different. Typically, the first order margins are adequately filled, (mostly) as a result of standard operating procedures in routine item analysis. If a response category is endorsed by very few respondents, either the item is removed altogether, or the categories are collapsed before model-fitting commences. In other words, in practice, sparseness of the second order margins is generally not accompanied by sparseness of the first order margins.

Therefore we arrive at the dominating insight of this research: The first order margins should not be collapsed into moments, but the second order margins should. We propose a new test statistic that stands between the original M_2, which does not collapse marginal residuals, and M_2^*, which further collapses the marginal residuals into moments. The new statistic still relies on first and second order information, but only the second order margins are collapsed into moments. This new statistic remedies the weaknesses of M_2 and M_2^*. We call it C_2.

With C_2, Samejima's (1969) unidimensional graded response (GR) model or Muraki's (1992) unidimensional generalized partial credit (GPC) model is locally identified (and has positive degrees of freedom) for as few as 4 items. More generally, unlike with M_2^*, the ability to compute C_2 does not depend on the number of categories per item. We show that C_2 is as well calibrated as the competition (namely M_2 and M_2^*), and can be more powerful. Finally, for C_2, the structure of the first and second order margins has an appealing connection to the parameters of an IRT model. The uncollapsed raw first order margins are strongly related to the item location parameters, while the collapsed second order margins (i.e., moments) are essentially covariances and are directly related to the item discrimination/loading parameters.

The remainder of the paper is organized as follows. We introduce basic notation in Section 2, and discuss maximum marginal likelihood estimation in Section 3. Properties of multinomial residuals are demonstrated in Section 4 to facilitate the introduction of the proposed test statistic. In Section 5, we report results from a simulation study to examine the calibration and power of the new test statistic. In Section 6, empirical data from a patient-reported outcomes measurement

development project will be used to further illustrate the new statistic. We conclude with a discussion of possible future research directions.

Some Notation

Let there be a total of $i = 1, \ldots I$ items. For an item with K_i ordered polytomous responses, let the response categories be coded as $k = 0, \ldots K_i - 1$. Let θ denote the underlying latent variable, and $T_i(k|\theta)$ the category response function for item i and category k. Without loss of generality, let us consider a logistic version of Samejima's (1969) graded response model for the remainder of this paper, noting that the theory developed in the sequel applies equally well to other ordinal IRT models such as the GPC model. The graded model sets the cumulative response function for item i in categories k and above as

$$T_i^+(k|\theta) = \frac{1}{1 + \exp[-(\alpha_{ik} + \beta_i\theta)]}, \tag{1}$$

for $k = 1, \ldots K_i - 1$, where α_{ik} is the intercept (location) and β_i is the slope (discrimination) parameter. Let the boundary cases be $T_i^+(0|\theta) = 1$ and $T_i^+(K_i|\theta) = 0$. The category response function can be written as

$$T_i(k|\theta) = T_i^+(k|\theta) - T_i^+(k+1|\theta), \tag{2}$$

for $k = 0, \ldots K_i - 1$.

Let Y_i be a random variable whose realization y_i is a response to item i. The probability mass function of Y_i, conditional on θ, is that of a multinomial with trial size 1:

$$P(Y_i = y_i|\theta;\boldsymbol{\gamma}) = \prod_{k=0}^{K_i-1}[T_i(k|\theta)]^{1_k(y_i)}, \tag{3}$$

where $1_k(y_i)$ is an indicator function such that

$$1_k(y_i) = \begin{cases} 1, & \text{if } k = y_i \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

and $\boldsymbol{\gamma}$ collects together all item parameters. Let the dimensionality of $\boldsymbol{\gamma}$ be equal to d, which is the number of free and unconstrained parameters in the model.

Under the assumption of conditional independence (Lord, 1968), the conditional probability for the response pattern $\boldsymbol{y} = (y_1, \ldots, y_I)'$ factors into a product:

$$\pi_{\boldsymbol{y}}(\theta;\boldsymbol{\gamma}) = P\left(\bigcap_{i=1}^{I} Y_i = y_i \,\middle|\, \theta;\boldsymbol{\gamma}\right) = \prod_{i=1}^{I} P(Y_i = y_i|\theta;\boldsymbol{\gamma}). \tag{5}$$

Assuming that the latent variable distribution has density $g(\theta)$, typically standard normal in applications, the marginal probability of the response pattern is

$$\pi_{\boldsymbol{y}}(\boldsymbol{\gamma}) = \int \prod_{i=1}^{I} P(Y_i = y_i|\theta;\boldsymbol{\gamma})\, g(\theta) d\theta. \tag{6}$$

Recall that K_i is the number of categories for item i. For I items, the IRT model generates a total of $K = \prod_{i=1}^{I} K_i$ cross-classifications or possible item response patterns in the form of a contingency table. For example, with 2 dichotomous items, the 4 possible response patterns are $(0,0), (0,1), (1,0), (1,1)$, in reverse lexicographical order. Note that K may become very large for polytomous items, e.g., for ten 5-category items, K is just under 10 million.

Maximum Likelihood Estimation of IRT Models

Based on a sample of N respondents, let the observed proportion of individuals with response pattern $\boldsymbol{y}$ be denoted as $p_{\boldsymbol{y}}$. These observed proportions can be collected into a $K \times 1$ vector $\boldsymbol{p}$. Correspondingly, the K model-implied probabilities $\pi_{\boldsymbol{y}}(\boldsymbol{\gamma})$ can be collected into a $K \times 1$ vector $\boldsymbol{\pi}(\boldsymbol{\gamma})$. We recognize that it is a parametric structural model. The model-implied probability vector $\boldsymbol{\pi}(\boldsymbol{\gamma})$ imposes a specific moment structure on $K - 1$ independent probabilities (as the sum of the K probabilities must be 1) with d parameters.

Suppose there is a $K \times 1$ vector $\boldsymbol{\pi}_0$ containing the true (population) response pattern probabilities. If the IRT model is exactly correctly specified, i.e., it fits perfectly in the population, then there exists a parameter vector $\boldsymbol{\gamma}_0$ such that $\boldsymbol{\pi}(\boldsymbol{\gamma}_0) = \boldsymbol{\pi}_0$. The elements of $\boldsymbol{\gamma}_0$ may be taken as the true parameters. When this is the case, parameter estimation is straightforward.

The sampling model for this contingency table is that of a multinomial with K cells and N trials (Reiser, 1996). The log-likelihood for the item parameters $\boldsymbol{\gamma}$ is proportional to

$$\log L(\boldsymbol{\gamma}) \propto N \sum_{\boldsymbol{y}} p_{\boldsymbol{y}} \log \pi_{\boldsymbol{y}}(\boldsymbol{\gamma}), \tag{7}$$

where the summation is nominally over all K response patterns. In reality, when a particular pattern is not observed in the data, the corresponding $p_{\boldsymbol{y}}$ is zero and the term does not contribute to the log-likelihood. Maximization of the $\log L(\boldsymbol{\gamma})$, e.g., with Bock and Aitkin's (1981) EM algorithm, leads to the maximum marginal likelihood estimator $\widehat{\boldsymbol{\gamma}}$.

It is a standard result from discrete multivariate analysis (e.g., Bishop, Fienberg, & Holland, 1975) that the maximum likelihood estimator is $\sqrt{N}$-consistent, asymptotically normal and asymptotically efficient under correct model specification. In other words, we have

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{D} \mathcal{N}_d(\mathbf{0}, \mathcal{F}_0^{-1}), \tag{8}$$

where $\mathcal{F}_0 = \boldsymbol{\Delta}_0' \, [diag(\boldsymbol{\pi}_0)]^{-1} \boldsymbol{\Delta}_0$ is the $d \times d$ Fisher information matrix evaluated at the true parameter values, and the Jacobian $\boldsymbol{\Delta}_0$ is a $K \times d$ matrix of all first order partial derivatives of the response pattern probabilities with respect to the parameters, evaluated at $\boldsymbol{\gamma}_0$: $\boldsymbol{\Delta}_0 = \dfrac{\partial \boldsymbol{\pi}(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'}$.

Furthermore, let $\hat{\pi}_{\boldsymbol{y}} = \pi_{\boldsymbol{y}}(\hat{\boldsymbol{\gamma}})$ denote the model-implied probability for response pattern $\boldsymbol{y}$ under maximum likelihood estimation. The direct comparison between $\hat{\pi}_{\boldsymbol{y}}$ and $p_{\boldsymbol{y}}$ leads to classical full-information fit statistics such as the likelihood ratio $G^2$and Pearson's X^2:

$$G^2 = 2N \sum_{\boldsymbol{y}} p_{\boldsymbol{y}} \log \frac{p_{\boldsymbol{y}}}{\hat{\pi}_{\boldsymbol{y}}}, \qquad X^2 = N \sum_{\boldsymbol{y}} \frac{\left(p_{\boldsymbol{y}} - \hat{\pi}_{\boldsymbol{y}}\right)^2}{\hat{\pi}_{\boldsymbol{y}}}. \tag{9}$$

Under the null hypothesis that the IRT model fits exactly, these two statistics are asymptotically distributed as central chi-square variables with degrees of freedom equal to $K - 1 - d$, against the general multinomial alternative (Bishop et al., 1975). Unfortunately for IRT models, K is exponential in the number of items. As argued earlier, when K is large, the expected response pattern probabilities necessarily become small, resulting in an extremely sparse table. This sparseness invalidates the chi-square approximation and renders these full-information statistics unsuitable for model fit testing (Cochran, 1952).

Limited-information Goodness-of-fit Testing

Distribution of Multinomial Residuals under Maximum Likelihood Estimation

It can be shown that the asymptotic distribution of $(p - \pi_0)$ is K-variate normal:

$$\sqrt{N}(\boldsymbol{p} - \boldsymbol{\pi}_0) \xrightarrow{D} \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Xi}_0), \tag{10}$$

where $\boldsymbol{\Xi}_0 = \text{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0'$ is the population multinomial covariance matrix. Recall that $\hat{\pi}_{\boldsymbol{y}} = \pi_{\boldsymbol{y}}(\hat{\boldsymbol{\gamma}})$ is the model-implied probability for response pattern $\boldsymbol{y}$ under maximum likelihood estimation. The K model-implied probabilities may be collected into a $K \times 1$ vector $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\gamma}})$. It can also be shown that the residual vector $(\boldsymbol{p} - \hat{\boldsymbol{\pi}})$ is asymptotically K-variate normally distributed under maximum likelihood estimation, albeit with a different limiting covariance matrix to take maximum likelihood estimation of parameters into account,

$$\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{p}) \xrightarrow{D} \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Sigma}_0), \tag{11}$$

where $\boldsymbol{\Sigma}_0 = \boldsymbol{\Xi}_0 - \boldsymbol{\Delta}_0 \mathcal{F}_0^{-1} \boldsymbol{\Delta}_0'$ (Bishop et al., 1975).

For example, for a test made up of 3 items, where item 1 is dichotomous and items 2 and 3 have 3-categories each, there are 18 possible item response patterns. In reverse lexicographical order, the model-implied response pattern probabilities and observed proportions are:

$$\hat{\boldsymbol{\pi}} = \begin{pmatrix} \hat{\pi}_{(000)} \\ \hat{\pi}_{(001)} \\ \hat{\pi}_{(002)} \\ \hat{\pi}_{(010)} \\ \hat{\pi}_{(011)} \\ \hat{\pi}_{(012)} \\ \hat{\pi}_{(020)} \\ \hat{\pi}_{(021)} \\ \hat{\pi}_{(022)} \\ \hat{\pi}_{(100)} \\ \hat{\pi}_{(101)} \\ \hat{\pi}_{(102)} \\ \hat{\pi}_{(110)} \\ \hat{\pi}_{(111)} \\ \hat{\pi}_{(112)} \\ \hat{\pi}_{(120)} \\ \hat{\pi}_{(121)} \\ \hat{\pi}_{(122)} \end{pmatrix}, \quad \boldsymbol{p} = \begin{pmatrix} p_{(000)} \\ p_{(001)} \\ p_{(002)} \\ p_{(010)} \\ p_{(011)} \\ p_{(012)} \\ p_{(020)} \\ p_{(021)} \\ p_{(022)} \\ p_{(100)} \\ p_{(101)} \\ p_{(102)} \\ p_{(110)} \\ p_{(111)} \\ p_{(112)} \\ p_{(120)} \\ p_{(121)} \\ p_{(122)} \end{pmatrix}. \tag{12}$$

First Order Margins

Using this arrangement, marginal probabilities can be obtained as linear functions of $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{p}$. Consider the 3-item example from above. There are 5 independent first order marginal probabilities: 1 from item 1, which is dichotomous; 2 from items 2 and 3 each. In general, for I items, there are $q_1 = \sum_{i=1}^{I}(K_i - 1)$ independent first order marginal probabilities. Without loss of generality and by convention, we can obtain an independent set of marginal probabilities for item i by removing the marginal probability for the lowest category with code 0. These first order marginal probabilities can be obtained from the full K-dimensional probability vector using a $q_1 \times K$ reduction operator matrix (see e.g., Joe & Maydeu-Olivares, 2010). The first order reduction matrix $\dot{\mathbf{L}}$ is a fixed incidence matrix that contains zeroes and ones, where the ones serve to select and sum over those full response pattern probabilities that correspond to a particular item and a particular category code to yield the desired marginal probability. Importantly, this matrix has full row rank. An example is given below:

$$\dot{\boldsymbol{\pi}} = \begin{pmatrix} \hat{\pi}_1^{(1)} \\ \hat{\pi}_1^{(1)} \\ \hat{\pi}_1^{(2)} \\ \hat{\pi}_2^{(1)} \\ \hat{\pi}_2^{(2)} \end{pmatrix} = \dot{\mathbf{L}}\hat{\boldsymbol{\pi}} = \begin{pmatrix} 0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1 \\ 0&0&0&1&1&1&0&0&0&0&0&0&1&1&1&0&0&0 \\ 0&0&0&0&0&0&1&1&1&0&0&0&0&0&0&1&1&1 \\ 0&1&0&0&1&0&0&1&0&0&1&0&0&1&0&0&1&0 \\ 0&0&1&0&0&1&0&0&1&0&0&1&0&0&1&0&0&1 \end{pmatrix} \begin{pmatrix} \hat{\pi}_{(000)} \\ \hat{\pi}_{(001)} \\ \hat{\pi}_{(002)} \\ \hat{\pi}_{(010)} \\ \hat{\pi}_{(011)} \\ \hat{\pi}_{(012)} \\ \hat{\pi}_{(020)} \\ \hat{\pi}_{(021)} \\ \hat{\pi}_{(022)} \\ \hat{\pi}_{(100)} \\ \hat{\pi}_{(101)} \\ \hat{\pi}_{(102)} \\ \hat{\pi}_{(110)} \\ \hat{\pi}_{(111)} \\ \hat{\pi}_{(112)} \\ \hat{\pi}_{(120)} \\ \hat{\pi}_{(121)} \\ \hat{\pi}_{(122)} \end{pmatrix}, \tag{13}$$

where $\hat{\boldsymbol{\pi}}_1$ denotes the q_1-vector of all independent first order marginal probabilities, and $\hat{\pi}_i^{(k)}$ denotes the first order marginal probability for item i in category k. By analogy, $\boldsymbol{p}_1 = \dot{\mathbf{L}}\boldsymbol{p}$ is a q_1-vector of independent first order observed proportions. The elements of $\boldsymbol{p}_1$ may be denoted $p_i^{(k)}$. Comparing $\hat{\boldsymbol{\pi}}_1$ and $\boldsymbol{p}_1$ can show the extent to which the IRT model has successfully reproduced the first order proportions.

On the other hand, Cai and Hansen (2013) considered collapsing the marginal probabilities into marginal moments (see also Joe & Maydeu-Olivares, 2010). Cai and Hansen (2013) reasoned that for each ordinal item, one could use the usual category codes to compute an item mean from both the model-implied and the observed first order marginal probabilities:

$$\hat{\mu}_i = \sum_{k=0}^{K_i-1} k\hat{\pi}_i^{(k)}, \quad m_i = \sum_{k=0}^{K_i-1} kp_i^{(k)}. \tag{14}$$

This setup has the side benefit of effectively eliminating the first category for each item so that no special treatment is required to obtain independent probabilities. They also showed that the computation of item means can be computed via reduction operator matrices. In general one only has to pre-multiply $\dot{\mathbf{L}}$ by a $I \times q_1$ block-diagonal matrix $\dot{\mathbf{R}}$. The I diagonal blocks of $\dot{\mathbf{R}}$ are row vectors made up of item category codes (sans 0): $\boldsymbol{r}_i = (1, \ldots, K_i - 1)$. The reduction operator

matrix $\dot{\mathbf{L}}_* = \dot{\mathbf{R}}\dot{\mathbf{L}}$ is of order $I \times K$. Because $\dot{\mathbf{R}}$ has full row rank, $\dot{\mathbf{L}}_*$ has full row rank as well. An example of item mean computation for the 3-item test from above is shown below:

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{pmatrix} = \dot{\mathbf{R}}\hat{\boldsymbol{\pi}}_1 = \begin{pmatrix} \boldsymbol{r}_1 & & \\ & \boldsymbol{r}_2 & \\ & & \boldsymbol{r}_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} \hat{\pi}_1^{(1)} \\ \hat{\pi}_1^{(1)} \\ \hat{\pi}_1^{(2)} \\ \hat{\pi}_2^{(1)} \\ \hat{\pi}_2^{(2)} \end{pmatrix} = \left(\dot{\mathbf{R}}\dot{\mathbf{L}}\right)\hat{\boldsymbol{\pi}} = \dot{\mathbf{L}}_*\hat{\boldsymbol{\pi}}, \tag{15}$$

where $\hat{\boldsymbol{\mu}}_1$ is a vector containing the I model-implied item means. By analogy, $\boldsymbol{m}_1 = \dot{\mathbf{L}}_*\boldsymbol{p}$ is a vector of observed item means. The model-implied and observed means can be compared just as in model fit assessment for mean and covariance structure models.

Second Order Margins

Table 1

Bivariate Table of Marginal Probabilities for Item Pair $(2,3)$ from the Example

Item 3 Category Code	Item 2 Category Code			Marginal Probability for Item 3
	0	1	2	
0	$\hat{\pi}_{32}^{(00)}$	$\hat{\pi}_{32}^{(01)}$	$\hat{\pi}_{32}^{(02)}$	$\hat{\pi}_3^{(0)}$
1	$\hat{\pi}_{32}^{(10)}$	$\hat{\pi}_{32}^{(11)}$	$\hat{\pi}_{32}^{(12)}$	$\hat{\pi}_3^{(1)}$
2	$\hat{\pi}_{32}^{(20)}$	$\hat{\pi}_{32}^{(21)}$	$\hat{\pi}_{32}^{(22)}$	$\hat{\pi}_3^{(2)}$
Marginal Probability for Item 2	$\hat{\pi}_2^{(0)}$	$\hat{\pi}_2^{(1)}$	$\hat{\pi}_2^{(2)}$	1.0

Generalizing the notation from first order marginal probabilities, let $\hat{\pi}_{ij}^{(kl)}$ denote the second order marginal probability for item pair (i,j), where item i is in category k and item j is in category l. With I items, there are $I(I-1)/2$ item pairs for $1 \leq j < i \leq I$. For each pair, these second order marginal probabilities form a $K_i \times K_j$ contingency table. Table 1 presents an example using the 3-item test from above. Each cell of the table corresponds to a second order marginal probability. On the margins of the table are the first order probabilities. Locally in this two-way table, given the first order margins, there are only $(K_i - 1) \times (K_j - 1)$ independent second order probabilities. The shaded cells indicate joint probabilities that we routinely remove to obtain independent probabilities. By this point it should not become a surprise that these second order marginal probabilities can be obtained via reduction operator matrices. These

reduction matrices are also fixed incidence matrix that contain zeroes and ones only. Each row of a reduction matrix sums over the response pattern probabilities corresponding to a particular second order marginal probability. As such a reduction operator matrix has full row rank and the number of rows is equal to $q_2 = \sum_{i=1}^{I} \sum_{j=1}^{I-1} (K_i - 1) \times (K_j - 1)$. The number of columns is equal to K. An example is shown below:

$$\hat{\boldsymbol{\pi}}_2 = \begin{pmatrix} \hat{\pi}_{21}^{(11)} \\ \hat{\pi}_{21}^{(21)} \\ \hat{\pi}_{31}^{(11)} \\ \hat{\pi}_{31}^{(21)} \\ \hat{\pi}_{32}^{(11)} \\ \hat{\pi}_{32}^{(12)} \\ \hat{\pi}_{32}^{(21)} \\ \hat{\pi}_{32}^{(22)} \end{pmatrix} = \ddot{\mathbf{L}}\hat{\boldsymbol{\pi}} = \begin{pmatrix} 0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&0&0&0&1&0&0&1&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&0&0&1&0&0&1&0&0&1 \\ 0&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&1 \end{pmatrix} \hat{\boldsymbol{\pi}}, \tag{16}$$

where $\ddot{\mathbf{L}}$ is the $q_2 \times K$ (in this case 8×18) reduction matrix, and $\hat{\boldsymbol{\pi}}_2$ denotes the q_2-vector of all independent second order marginal probabilities. Again by analogy, $\boldsymbol{p}_2 = \ddot{\mathbf{L}}\boldsymbol{p}$ is a q_2-vector of independent second order observed proportions. The elements of $\boldsymbol{p}_2$ may be denoted $p_{ij}^{(kl)}$. Comparison of $\hat{\boldsymbol{\pi}}_2$ and $\boldsymbol{p}_2$ tells us how well the IRT model fits the second order proportions.

Returning to the example in Table 1, Cai and Hansen (2013) noted that if there is reason to believe that the items in a test are strongly influenced by a common latent variable, as is typically the case due to common assessment development practices in educational and psychological testing, the second order marginal probabilities for the "inconsistent" response patterns in the two-way table will necessarily become small. For example, if both items 2 and 3 provide evidence about the respondents' severity in depression symptoms, then in aggregate, endorsement of a category indicating high severity on item 2 will tend to be correlated with endorsement of a similar category on item 3. Thus the cells in Table 1 that are close to the main diagonal (i.e., the consistent response patterns) will be better filled than the cells that are far removed from the diagonal (i.e., the inconsistent response patterns). As the number of categories increases, the sparseness of the second order margins becomes increasingly severe. Some of the observed second order marginal proportions could be equal to zero, and the model-implied probabilities are similarly small. Thus a direct comparison of $\hat{\boldsymbol{\pi}}_2$ and $\boldsymbol{p}_2$ has limited utility in practical data analysis settings involving ordered polytomous IRT models.

This observation led Cai and Hansen (2013) to apply Joe and Maydeu-Olivares' (2010) general framework. Instead of examining the two-way probabilities, Cai and Hansen used the ordinal item scores to compute a raw moment statistic for each item pair:

$$\hat{\mu}_{ij} = \sum_{k=0}^{K_i-1} \sum_{l=0}^{K_j-1} kl\, \hat{\pi}_{ij}^{(kl)}, \quad m_{ij} = \sum_{k=0}^{K_i-1} \sum_{l=0}^{K_j-1} kl\, p_{ij}^{(kl)}, \tag{17}$$

where $\hat{\mu}_{ij}$ and m_{ij} are the model-implied and observed second order moments for item pair (i, j). The moment statistic further collapses the two-way contingency table into a single-number summary, thereby avoiding the sparseness issue even when the number of categories is large. The bivariate moments are effectively measuring pairwise correlations between items.

Again, the second order moments may be computed via reduction operator matrices. The reduction operator matrix for second order moments can be obtained by pre-multiplying $\ddot{\mathbf{L}}$ with a $I(I-1)/2 \times q_2$ block-diagonal matrix $\ddot{\mathbf{R}}$. The $I(I-1)/2$ diagonal blocks of $\ddot{\mathbf{R}}$ contain Kronecker products of row vectors that are made up of item category codes: $\boldsymbol{r}_j \otimes \boldsymbol{r}_i$, for $1 \leq j < i \leq I$, where $\boldsymbol{r}_i = (1, \ldots, K_i - 1)$ is as defined in Section 4.2. Note that $\ddot{\mathbf{R}}$ has full row rank. Thus $\ddot{\mathbf{L}}_* = \ddot{\mathbf{R}}\ddot{\mathbf{L}}$ has full row rank. An example for the 3-item test is shown below:

$$\hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} \hat{\mu}_{21} \\ \hat{\mu}_{31} \\ \hat{\mu}_{32} \end{pmatrix} = \ddot{\mathbf{R}}\hat{\boldsymbol{\pi}}_2 = \begin{pmatrix} \boldsymbol{r}_2 \otimes \boldsymbol{r}_1 & & \\ & \boldsymbol{r}_3 \otimes \boldsymbol{r}_1 & \\ & & \boldsymbol{r}_3 \otimes \boldsymbol{r}_2 \end{pmatrix} \hat{\boldsymbol{\pi}}_2 \tag{18}$$

$$= \begin{pmatrix} 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 2 & 4 \end{pmatrix} \begin{pmatrix} \hat{\pi}_{21}^{(11)} \\ \hat{\pi}_{21}^{(21)} \\ \hat{\pi}_{31}^{(11)} \\ \hat{\pi}_{31}^{(21)} \\ \hat{\pi}_{32}^{(11)} \\ \hat{\pi}_{32}^{(12)} \\ \hat{\pi}_{32}^{(21)} \\ \hat{\pi}_{32}^{(22)} \end{pmatrix} = \left(\ddot{\mathbf{R}}\ddot{\mathbf{L}}\right)\hat{\boldsymbol{\pi}} = \ddot{\mathbf{L}}_*\hat{\boldsymbol{\pi}},$$

where $\hat{\boldsymbol{\mu}}_2$ is a vector containing the $I(I-1)/2$ model-implied second order moments. By analogy, we define $\boldsymbol{m}_2 = \mathbf{L}_2^*\boldsymbol{p}$ as a vector of observed second order moments.

Existing Test Statistics: M_2 and M_2^*

Maydeu-Olivares and Joe (2006) proposed the M_2 statistic, which utilizes the first and second order marginal probabilities. Let $\mathbf{L}$ be a $(q_1 + q_2) \times K$ matrix that vertically concatenates $\dot{\mathbf{L}}$ and $\ddot{\mathbf{L}}$ such that its first q_1 rows come from $\dot{\mathbf{L}}$ and the remaining rows come from $\ddot{\mathbf{L}}$. What this

implies is that by pre-multiplying $\mathbf{L}$ with $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{p}$, we can obtain the $(q_1 + q_2) \times 1$ vector of first *and* second order marginal residual probabilities $(\hat{\boldsymbol{\pi}}_{12} - \boldsymbol{p}_{12})$ as a linear function of the multinomial cell residuals $(\hat{\boldsymbol{\pi}} - \boldsymbol{p})$ defined in Equation (11):

$$\hat{\boldsymbol{\pi}}_{12} - \boldsymbol{p}_{12} = \begin{pmatrix} \hat{\boldsymbol{\pi}}_1 - \boldsymbol{p}_1 \\ \hat{\boldsymbol{\pi}}_2 - \boldsymbol{p}_2 \end{pmatrix} = \begin{pmatrix} \dot{\mathbf{L}} \\ \ddot{\mathbf{L}} \end{pmatrix} (\hat{\boldsymbol{\pi}} - \boldsymbol{p}) = \mathbf{L}(\hat{\boldsymbol{\pi}} - \boldsymbol{p}). \tag{19}$$

Equation (19) implies that the asymptotic distribution of $(\hat{\boldsymbol{\pi}}_{12} - \boldsymbol{p}_{12})$ is normal:

$$\sqrt{N}(\hat{\boldsymbol{\pi}}_{12} - \boldsymbol{p}_{12}) \xrightarrow{D} \mathcal{N}_{q_1+q_2}(\mathbf{0}, \boldsymbol{\Sigma}_{12}), \tag{20}$$

where $\boldsymbol{\Sigma}_{12} = \mathbf{L}\boldsymbol{\Sigma}_0\mathbf{L}' = \mathbf{L}\boldsymbol{\Xi}_0\mathbf{L}' - \mathbf{L}\boldsymbol{\Delta}_0\mathcal{F}_0^{-1}\boldsymbol{\Delta}_0'\mathbf{L}' = \boldsymbol{\Xi}_{12} - \boldsymbol{\Delta}_{12}\mathcal{F}_0^{-1}\boldsymbol{\Delta}_{12}'$. In particular the marginal Jacobian matrix $\boldsymbol{\Delta}_{12} = \mathbf{L}\boldsymbol{\Delta}_0$ is the matrix of all partial derivatives against the first and second order marginal probabilities

$$\boldsymbol{\Delta}_{12} = \mathbf{L}\frac{\partial \boldsymbol{\pi}(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'} = \frac{\partial \mathbf{L}\boldsymbol{\pi}(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'} = \frac{\partial \boldsymbol{\pi}_{12}(\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}'}.$$

The rank of $\boldsymbol{\Delta}_{12}$ determines whether the IRT model is locally identified from the marginal probabilities. If $\boldsymbol{\Delta}_{12}$ has full column rank, the model is locally identified.

Let $\hat{\boldsymbol{\Xi}} = diag(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}'$ be the multinomial covariance matrix evaluated at the maximum likelihood solution $\hat{\boldsymbol{\gamma}}$, and let $\hat{\boldsymbol{\Xi}}_{12} = \mathbf{L}\hat{\boldsymbol{\Xi}}\mathbf{L}'$. Also evaluate the marginal Jacobian $\boldsymbol{\Delta}_{12}$ at $\hat{\boldsymbol{\gamma}}$:

$$\hat{\boldsymbol{\Delta}}_{12} = \frac{\partial \boldsymbol{\pi}_2(\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}'}.$$

By Proposition 4 in Browne (1984), the test statistic

$$M_2 = N(\hat{\boldsymbol{\pi}}_{12} - \boldsymbol{p}_{12})'\hat{\boldsymbol{\Omega}}_{12}\,(\hat{\boldsymbol{\pi}}_{12} - \boldsymbol{p}_{12}), \tag{21}$$

where $\hat{\boldsymbol{\Omega}}_{12} = \hat{\boldsymbol{\Xi}}_{12}^{-1} - \hat{\boldsymbol{\Xi}}_{12}^{-1}\hat{\boldsymbol{\Delta}}_{12}\left(\hat{\boldsymbol{\Delta}}_{12}'\hat{\boldsymbol{\Xi}}_{12}^{-1}\hat{\boldsymbol{\Delta}}_{12}\right)^{-1}\hat{\boldsymbol{\Delta}}_{12}'\hat{\boldsymbol{\Xi}}_{12}^{-1}$, is asymptotically chi-squared with $q_1 + q_2 - d$ degrees of freedom under the null hypothesis that the model fits exactly in the population.

Similarly, let $\mathbf{L}_*$ be a $I(I+1)/2 \times K$ matrix that vertically concatenates $\dot{\mathbf{L}}_*$ and $\ddot{\mathbf{L}}_*$ such that its first I rows come from $\dot{\mathbf{L}}_*$ and the remaining $I(I-1)/2$ rows come from $\ddot{\mathbf{L}}_*$. The $I(I+1)/2 \times 1$ vector of first *and* second order marginal residual moments $(\hat{\boldsymbol{\mu}}_{12} - \boldsymbol{m}_{12})$ is a linear function of the multinomial cell residuals $(\hat{\boldsymbol{\pi}} - \boldsymbol{p})$:

$$\hat{\boldsymbol{\mu}}_{12} - \boldsymbol{m}_{12} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 - \boldsymbol{m}_1 \\ \hat{\boldsymbol{\mu}}_2 - \boldsymbol{m}_2 \end{pmatrix} = \begin{pmatrix} \dot{\mathbf{L}}_* \\ \ddot{\mathbf{L}}_* \end{pmatrix} (\hat{\boldsymbol{\pi}} - \boldsymbol{p}) = \mathbf{L}(\hat{\boldsymbol{\pi}} - \boldsymbol{p}), \tag{22}$$

and the asymptotic distribution of $(\hat{\boldsymbol{\mu}}_{12} - \boldsymbol{m}_{12})$ is normal:

$$\sqrt{N}(\hat{\boldsymbol{\mu}}_{12} - \boldsymbol{m}_{12}) \xrightarrow{D} \mathcal{N}_{I(I+1)/2}(\mathbf{0}, \boldsymbol{\Sigma}_{12}^*), \tag{23}$$

where $\boldsymbol{\Sigma}_{12}^{*} = \mathbf{L}_{*}\boldsymbol{\Sigma}_0\mathbf{L}_{*}' = \mathbf{L}_{*}\boldsymbol{\Xi}_0\mathbf{L}_{*}' - \mathbf{L}_{*}\boldsymbol{\Delta}_0\mathcal{F}_0^{-1}\boldsymbol{\Delta}_0'\mathbf{L}_{*}' = \boldsymbol{\Xi}_{12}^{*} - \boldsymbol{\Delta}_{12}^{*}\mathcal{F}_0^{-1}(\boldsymbol{\Delta}_{12}^{*})'$. Let $\widehat{\boldsymbol{\Xi}}_{12}^{*} = \mathbf{L}_{*}\widehat{\boldsymbol{\Xi}}\mathbf{L}_{*}'$ and evaluate the Jacobian with respect to marginal moments $\boldsymbol{\Delta}_{12}^{*}$ at $\widehat{\boldsymbol{\gamma}}$:

$$\widehat{\boldsymbol{\Delta}}_{12}^{*} = \mathbf{L}_{*}\frac{\partial\boldsymbol{\pi}(\widehat{\boldsymbol{\gamma}})}{\partial\boldsymbol{\gamma}'} = \frac{\partial\mathbf{L}_{*}\boldsymbol{\pi}(\widehat{\boldsymbol{\gamma}})}{\partial\boldsymbol{\gamma}'} = \frac{\partial\boldsymbol{\mu}_{12}(\widehat{\boldsymbol{\gamma}})}{\partial\boldsymbol{\gamma}'}.$$

By an analogous argument, the test statistic

$$M_2^{*} = N(\widehat{\boldsymbol{\mu}}_{12} - \boldsymbol{m}_{12})'\widehat{\boldsymbol{\Omega}}_{12}^{*}(\widehat{\boldsymbol{\mu}}_{12} - \boldsymbol{m}_{12}), \tag{24}$$

where $\widehat{\boldsymbol{\Omega}}_{12}^{*} = (\widehat{\boldsymbol{\Xi}}_{12}^{*})^{-1} - (\widehat{\boldsymbol{\Xi}}_{12}^{*})^{-1}\widehat{\boldsymbol{\Delta}}_{12}^{*}\left[(\widehat{\boldsymbol{\Delta}}_{12}^{*})'(\widehat{\boldsymbol{\Xi}}_{12}^{*})^{-1}\widehat{\boldsymbol{\Delta}}_{12}^{*}\right]^{-1}(\widehat{\boldsymbol{\Delta}}_{12}^{*})'(\widehat{\boldsymbol{\Xi}}_{12}^{*})^{-1}$, is also asymptotically chi-squared, but with $I(I+1)/2 - d$ degrees of freedom under the null hypothesis that the model fits exactly in the population (Cai & Hansen, 2013). Note that when I is small, the degrees of freedom may become negative for polytomous items.

The Proposed Test Statistic

Given the foregoing development, we are ready introduce the new statistic. Let $q = q_1 + I(I-1)/2$ denote the number of first order marginal probabilities and the number of second order marginal moments. Let $\mathbf{M}$ be a $q \times K$ matrix that vertically concatenates $\dot{\mathbf{L}}$ and $\ddot{\mathbf{L}}_{*}$ such that its first q_1 rows come from $\dot{\mathbf{L}}$ and the remaining $I(I-1)/2$ rows come from $\ddot{\mathbf{L}}_{*}$. Let $\boldsymbol{\sigma}(\widehat{\boldsymbol{\gamma}}) = \widehat{\boldsymbol{\sigma}} = \mathbf{M}\widehat{\boldsymbol{\pi}} = \mathbf{M}\boldsymbol{\pi}(\widehat{\boldsymbol{\gamma}})$ be the $q \times 1$ vector of model-implied first order marginal residual probabilities and second order expected marginal moments, and $\boldsymbol{s} = \mathbf{M}\boldsymbol{p}$ be the corresponding observed proportions and sample moments, i.e.,

$$\widehat{\boldsymbol{\sigma}} = \mathbf{M}\widehat{\boldsymbol{\pi}} = \begin{pmatrix}\widehat{\boldsymbol{\pi}}_1\\ \widehat{\boldsymbol{\mu}}_2\end{pmatrix} = \begin{pmatrix}\dot{\mathbf{L}}\\ \ddot{\mathbf{L}}_{*}\end{pmatrix}\widehat{\boldsymbol{\pi}} = \mathbf{M}\widehat{\boldsymbol{\pi}}, \quad \boldsymbol{s} = \mathbf{M}\boldsymbol{p} = \begin{pmatrix}\boldsymbol{p}_1\\ \boldsymbol{m}_2\end{pmatrix} = \begin{pmatrix}\dot{\mathbf{L}}\\ \ddot{\mathbf{L}}_{*}\end{pmatrix}\boldsymbol{p} = \mathbf{M}\boldsymbol{p}. \tag{25}$$

The $q \times d$ Jacobian matrix $\mathbf{J}_0 = \mathbf{J}(\boldsymbol{\gamma}_0) = \mathbf{M}\boldsymbol{\Delta}_0 = \mathbf{M}\boldsymbol{\Delta}(\boldsymbol{\gamma}_0)$ is therefore

$$\mathbf{J}_0 = \mathbf{J}(\boldsymbol{\gamma}_0) = \mathbf{M}\frac{\partial\boldsymbol{\pi}(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}'} = \frac{\partial\mathbf{M}\boldsymbol{\pi}(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}'} = \frac{\partial\boldsymbol{\sigma}(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}'}.$$

Note that the number of independent first order marginal probabilities q_1 is generally equal to the number of location/intercept parameters in GR or GPC models. As long as the number of discrimination parameters does not equal or exceed the number of second order marginal moments, q is typically larger than d and $\mathbf{J}(\boldsymbol{\gamma})$ may have full column rank, indicating local identification of the IRT model, in contrast to the case of M_2^{*}.

It is clear then the $q \times 1$ marginal residual vector $(\widehat{\boldsymbol{\sigma}} - \boldsymbol{s})$ is still a linear function of the multinomial cell residuals $(\widehat{\boldsymbol{\pi}} - \boldsymbol{p})$ as defined in Equation (11):

$$\widehat{\boldsymbol{\sigma}} - \boldsymbol{s} = \begin{pmatrix}\widehat{\boldsymbol{\pi}}_1 - \boldsymbol{p}_1\\ \widehat{\boldsymbol{\mu}}_2 - \boldsymbol{m}_2\end{pmatrix} = \begin{pmatrix}\dot{\mathbf{L}}\\ \ddot{\mathbf{L}}_{*}\end{pmatrix}(\widehat{\boldsymbol{\pi}} - \boldsymbol{p}) = \mathbf{M}(\widehat{\boldsymbol{\pi}} - \boldsymbol{p}). \tag{26}$$

Equation (26) implies that the asymptotic distribution of $(\widehat{\boldsymbol{\sigma}} - \boldsymbol{s})$ is q-variate normal:

$$\sqrt{N}(\widehat{\boldsymbol{\sigma}} - \boldsymbol{s}) \xrightarrow{D} \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Phi}), \tag{27}$$

where $\boldsymbol{\Phi} = \mathbf{M}\boldsymbol{\Sigma}_0\mathbf{M}' = \mathbf{M}\boldsymbol{\Xi}_0\mathbf{M}' - \mathbf{M}\boldsymbol{\Delta}_0\mathcal{F}_0^{-1}\boldsymbol{\Delta}_0'\mathbf{M}' = \mathbf{M}\boldsymbol{\Xi}_0\mathbf{M}' - \mathbf{J}_0\mathcal{F}_0^{-1}\mathbf{J}_0'$.

Let $\boldsymbol{\Upsilon} = \mathbf{M}\boldsymbol{\Xi}_0\mathbf{M}'$ and let $\widehat{\boldsymbol{\Upsilon}} = \mathbf{M}\widehat{\boldsymbol{\Xi}}\mathbf{M}$ be an estimate under maximum likelihood estimation. Define a weight matrix $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Upsilon}}^{-1} - \widehat{\boldsymbol{\Upsilon}}^{-1}\hat{\mathbf{J}}\left[\hat{\mathbf{J}}'\widehat{\boldsymbol{\Upsilon}}^{-1}\hat{\mathbf{J}}\right]^{-1}\hat{\mathbf{J}}'\widehat{\boldsymbol{\Upsilon}}^{-1}$, where $\hat{\mathbf{J}}$ is simply the Jacobian $\mathbf{J}(\boldsymbol{\gamma})$ evaluated at $\widehat{\boldsymbol{\gamma}}$. We argue that the new test statistic

$$C_2 = N(\widehat{\boldsymbol{\sigma}} - \boldsymbol{s})'\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\sigma}} - \boldsymbol{s}), \tag{28}$$

is asymptotically chi-squared with $q - d$ degrees of freedom under the null hypothesis that the model fits exactly in the population. Needless to say, C_2, M_2, and M_2^* become equivalent when all items are dichotomous.

To see this is the case, assume that $\mathbf{J}(\boldsymbol{\gamma}_0)$ has full column rank and the model is locally identified. By the continuity of the matrix inverse and the consistency of the maximum likelihood estimator, the probability limit of $\widehat{\boldsymbol{\Omega}}$ is $\boldsymbol{\Omega} = \boldsymbol{\Upsilon}^{-1} - \boldsymbol{\Upsilon}^{-1}\mathbf{J}_0[\mathbf{J}_0'\boldsymbol{\Upsilon}^{-1}\mathbf{J}_0]^{-1}\mathbf{J}_0'\boldsymbol{\Upsilon}^{-1}$. Since the statistic C_2 is a quadratic form in an asymptotically normal random vector with zero means, it is sufficient to show that the product of the limiting covariance matrix and the weight matrix of the quadratic form $\boldsymbol{\Phi}\boldsymbol{\Omega} = (\boldsymbol{\Upsilon} - \mathbf{J}_0\mathcal{F}_0^{-1}\mathbf{J}_0')\boldsymbol{\Omega} = \mathbf{I} - \boldsymbol{\Upsilon}^{-1}\mathbf{J}_0[\mathbf{J}_0'\boldsymbol{\Upsilon}^{-1}\mathbf{J}_0]^{-1}\mathbf{J}_0'$ is idempotent. This is true. By Cochran's theorem and Slutsky's theorem, C_2 is asymptotically chi-squared. The degrees of freedom is equal to the trace (rank) of $\mathbf{I} - \boldsymbol{\Upsilon}^{-1}\mathbf{J}_0[\mathbf{J}_0'\boldsymbol{\Upsilon}^{-1}\mathbf{J}_0]^{-1}\mathbf{J}_0'$, which is $q - d$.

A Measure of Model Error

When the model does not fit exactly in the population, there does not exist a $\boldsymbol{\gamma}_0$ such that $\boldsymbol{\pi}(\boldsymbol{\gamma}_0) = \boldsymbol{\pi}_0$. In general, for any parameter vector $\boldsymbol{\gamma}$, $\boldsymbol{\pi}_{12}(\boldsymbol{\gamma}) \neq \mathbf{L}\boldsymbol{\pi}_0$, $\boldsymbol{\mu}_{12}(\boldsymbol{\gamma}) \neq \mathbf{L}_*\boldsymbol{\pi}_0$, and $\boldsymbol{\sigma}(\boldsymbol{\gamma}) \neq \mathbf{M}\boldsymbol{\pi}_0$, unless the misspecification only affects third-order margins or above. The limiting means of the random vectors in Equations (20), (23), and (27) are generally no longer zero, and M_2, M_2^*, and C_2 are no longer distributed as central chi-square random variables. Maydeu-Olivares (2013) suggested that we borrow from the model fit assessment literature in structural equation modeling, and utilize the quadratic forms in M_2, M_2^*, and C_2 to compute Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993; Maydeu-Olivares, 2013) type indices to characterize the per degree of freedom error of approximation in the population.

Generically, for observed discrepancy measure F, an unbiased estimate of the population discrepancy is $\hat{F} = F - df/N$ (Browne & Cudeck, 1993), where df is the degrees of freedom available for testing. The sample RMSEA estimate is defined (with truncation at 0):

$$\hat{\varepsilon} = \max\left(\sqrt{\frac{\hat{F}}{df}}, 0\right). \tag{29}$$

Confidence intervals of RMSEA may be easily computed from the noncentral chi-square distribution by following established procedures in Browne and Cudeck (1993).

Let $F_M = (\hat{\boldsymbol{\pi}}_{12} - \boldsymbol{p}_{12})'\hat{\boldsymbol{\Omega}}_{12}\,(\hat{\boldsymbol{\pi}}_{12} - \boldsymbol{p}_{12})$ be the observed discrepancy measure based on the M_2 statistic, $F_M^* = (\hat{\boldsymbol{\mu}}_{12} - \boldsymbol{m}_{12})'\hat{\boldsymbol{\Omega}}_{12}^*\,(\hat{\boldsymbol{\mu}}_{12} - \boldsymbol{m}_{12})$ the observed discrepancy measure based on the M_2^* statistic, and finally let $F_C = (\hat{\boldsymbol{\sigma}} - \boldsymbol{s})'\hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\sigma}} - \boldsymbol{s})$ be the observed discrepancy measure based on the C_2 statistic. Then we can define $\hat{\varepsilon}_M$, $\hat{\varepsilon}_M^*$, and $\hat{\varepsilon}_C$ as three different versions of the RMSEA index, each from a different underlying test statistic. To the extent that the test statistics have different behavior under the alternative hypothesis, the different RMSEAs will exhibit differences in magnitude. The variation is important to understand because the conclusions drawn from the RMSEA values may be quite different, depending on which version of the limited-information test statistic one has chosen to evaluate the fit of the IRT model.

Simulations

A small simulation study was conducted to examine the calibration and power of the proposed statistic, C_2. Along with C_2, the M_2 statistic of Maydeu-Olivares and Joe (2006) and the fully collapsed M_2^* statistic of Cai and Hansen (2013) were considered. In all conditions, a sample size of $N = 500$ was used. The data were generated using Samejima's (1969) GR model, with $K_i = 4$ ordered response categories per item.

All generating parameter values are presented in Table 2. For the null condition, the generating model was unidimensional, with $I = 4, 6,$ or 8, adding successively more items from Table 2 to the generating moel. The β_{i1} column shows the slopes of the unidimensional GR model. As mentioned above, a shortcoming of the M_2^* statistic is that it cannot be used for smaller models with relatively large K_i, due to lack of local identification and negative degrees of freedom. Such is the case here, as M_2^* cannot be computed except for the $I = 8$ condition. On the other hand, both C_2 and M_2 can be computed for all the I considered here. However, since the items are polytomous, it is possible that the distribution of M_2 will be distorted because of poorly filled second order marginal tables, as demonstrated by Cai and Hansen (2013), leading to a reduction of power. To study power, model misspecification was introduced through the presence of a second latent variable, $\theta_2 \sim \mathcal{N}(0, 1)$, uncorrelated with θ_1, but influencing a doublet of items. More specifically, for all non-null conditions, data for items 1 and 2 were

generated using a two-dimensional GR model, wherein the cumulative category probabilities are defined as

$$T_i^+(k|\theta_1,\theta_2) = \frac{1}{1+\exp[-(\alpha_{ik}+\beta_{i1}\theta_1+\beta_{i1}\theta_2)]}. \quad (30)$$

The additional slopes on θ_2 were set to 0.8, as shown in Table 2. The number of items for the non-null condition was also $I = 4, 6,$ or 8. The fitted model in all conditions was the unidimensional GR model. There were 1,000 replications under each of the 6 conditions.

Table 2

Generating Parameters for Simulation Study

Item i	α_{i1}	α_{i2}	αi_3	β_{i1}	β_{i2}
1	2.0	0.5	-1.0	1.5	0.8
2	2.0	0.5	-1.0	1.7	0.8
3	2.0	0.5	-1.0	1.9	
4	2.0	0.5	-1.0	2.1	
5	1.0	-0.5	-2.0	1.5	
6	1.0	-0.5	-2.0	1.7	
7	1.0	-0.5	-2.0	1.9	
8	1.0	-0.5	-2.0	2.1	

Type I Error Rate

Table 3 displays means, variances, and empirical rejection rates for the 3 null conditions. Also, p-values from two-sided Kolmogorov-Smirnov (KS) tests are reported to detect any departures from the reference chi-square distributions. Immediately apparent is the range of degrees of freedom for the test statistics. While M_2 has 50 degrees of freedom with just 4 items, M_2^* has barely positive degrees of freedom (df = 4) with as many as 8 items. C_2, by construction, lies somewhere in between. All statistics behave well. The observed mean and variance relationships closely track that of a central chi-square variable, with variance approximately equal to twice the mean. The non-significant KS p-values and observed rejection rates support the proposition that all of the statistics are well-calibrated under the null. Overall, the results support the theoretical development claiming that C_2 is approximately chi-square distributed. As a consequence, the power of M_2, C_2, and M_2^* can be more directly compared.

Table 3

Simulation Results: Null Conditions

		First order	Second order					Rejection Rates at α			
I	Statistics	Information	Information	d	df	Mean	Var	.010	.050	.100	KS
8	M_2	24	252	32	244	244.36	463.74	.014	.037	.093	.503
	C_2	24	28	32	20	19.86	36.19	.005	.046	.084	.671
	M_2^*	8	28	32	4	4.00	8.61	.014	.050	.105	.790
6	M_2	18	135	24	129	128.70	237.48	.008	.041	.100	.215
	C_2	18	15	24	9	8.90	16.64	.012	.035	.089	.692
	M_2^*	6	15	24	-3						
4	M_2	12	54	16	50	50.27	102.40	.017	.052	.105	.519
	C_2	12	6	16	2	2.03	4.71	.013	.053	.110	.182
	M_2^*	4	6	16	-6						

Note. For M_2 and C_2, first order information refers to the total number of independent first order marginal probabilities, and for M_2^*, first order information comes in the form of item-specific marginal means. For M_2, second order information refers to the number of independent second order marginal probabilities, whereas for C_2 and M_2^*, second order information are bivariate product moments. Note that for two conditions, M_2^* cannot be computed because of negative degrees of freedom.

Power

Empirical rejection rates for M_2, C_2, and M_2^* under the non-null conditions are presented in Table 4. Overall, the rejection rates increase as the number of items decreases from $I = 8$ to 6 to 4. This is expected, as the misspecification only affects the first two items regardless of I. Consequently, the misspecification is more severe with a smaller number of items. Comparing the three statistics, C_2 is clearly the most powerful, for all conditions. As one example, consider the rejection rates at $\alpha = .05$ for $I = 8$. The rejection rate of C_2 (.335) is nearly triple that of M_2 (.119), while the rejection rate of M_2^* (.052) barely exceeds the nominal α level.

Table 4

Simulation Results: Power and RMSEA

		First order	Second order			Rejection Rates at α				RMSEA		
I	Statistics	Information	Information	d	df	.010	.050	.100	F_0	ε_0	M	90% CI
8	M_2	24	252	32	244	.027	.119	.196	.017	.008	.008	(0, .020)
	C_2	24	28	32	20	.125	.335	.457	.016	.029	.025	(0, .048)
	M_2^*	8	28	32	4	.011	.052	.112	< .001	.001	.015	(0, .053)
6	M_2	18	135	24	129	.034	.124	.212	.015	.011	.010	(0, .024)
	C_2	18	15	24	9	.188	.386	.506	.014	.040	.035	(0, .067)
	M_2^*	6	15	24	-3							
4	M_2	12	54	16	50	.043	.146	.237	.011	.015	.013	(0, .032)
	C_2	12	6	16	2	.278	.504	.603	.010	.072	.061	(0, .123)
	M_2^*	4	6	16	-6							

Note. For M_2 and C_2, first order information refers to the total number of independent first order marginal probabilities, and for M_2^*, first order information comes in the form of item-specific marginal means. For M_2, second order information refers to the number of independent second order marginal probabilities, whereas for C_2 and M_2^*, second order information are bivariate product moments. Note that for two conditions, M_2^* cannot be computed because of negative degrees of freedom.

Table 4 also shows the sample mean and empirical 90% confidence intervals for $\hat{\varepsilon}_M$, $\hat{\varepsilon}_M^*$, and $\hat{\varepsilon}_C$. Interestingly, for a given condition, the means vary considerably depending on which statistic is used to compute the sample RMSEA. For instance, for the $I = 4$ condition, the mean of $\hat{\varepsilon}_M$ is .013, while the mean of $\hat{\varepsilon}_C$ is .061. Under commonly used guidelines, the former would indicate "excellent" fit, while the latter would indicate merely "acceptable" fit.

Some insight into this phenomenon can be gained by computing the population RMSEA values. For purposes of illustration, consider the $I = 4$ condition. Under the alternative model, the $K = 4^4 = 256$ population multinomial probabilities may be computed and collected in $\boldsymbol{\pi}_0$. Then, treating $\boldsymbol{\pi}_0$ as $\boldsymbol{p}$, that is, treating the population probabilities as the sample multinomial proportions, Equation (7) may be maximized under the null model to yield $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\gamma}}) \neq \boldsymbol{\pi}_0$. Using $\boldsymbol{\pi}_0$ and $\hat{\boldsymbol{\pi}}$, three different population discrepancy measures may be computed, based on M_2, C_2, and M_2^*. These population discrepancy measures, in turn, may be used to find corresponding population RMSEA values.

The population discrepancy measures and population RMSEA values are shown in Table 4, in the columns labeled F_0 and ε_0, respectively. Generally, for any condition and statistic, the mean of the sample RMSEA values is quite similar to its population RMSEA value, indicating consistency of the sample estimation. For instance, for the case of $I = 4$, for M_2 we see that

$\varepsilon_0 = .015$, while the sample estimate is .013. Similarly, for C_2, $\varepsilon_0 = .072$, while the sample estimate is .061. Note, however, that the two population RMSEA values may lead to different evaluations of fit, given common interpretive guidelines. This incongruity underscores the need to consider the underlying statistics in interpreting RMSEA. In other words, each test statistic provides a different measure of the same model misspecification.

Analysis of Empirical Data

The empirical data (a random sample of $N = 1000$) come from the PROMIS® Smoking Initiative (Edelen, Tucker, Shadel, Stucky, & Cai, 2012). One task of this initiative is to develop and evaluate short forms measures of cigarette smoking related psycho-bio-social constructs, which are more practical to administer in clinical and research settings. In this process, the research team considered short forms with as few as 4 items (Hansen et al., in press). As an illustration of the application of the C_2 statistic, we analyze a 4-item subset. The item stems, presented in Table 5, all pertain to perceived positive benefits of cigarette smoking. The response scale elicits respondents' degree of agreement to the statements presented in the item stems with 5 ordered categories: 0 = Not at All, 1 = A Little Bit, 2 = Somewhat, 3 = Quite a Bit, 4 = Very Much.

Table 5

Item Stems for the Four Smoking Items

S1	Smoking helps me concentrate.
S2	Smoking makes me feel better in social situations.
S3	If I'm feeling irritable, a cigarette will help me relax.
S4	Smoking a cigarette energizes me.

A unidimensional GR model was fit to all items, and several overall tests were calculated. The statistics, associated probabilities, and RMSEA estimates are shown in Table 6. M_2^* can not be computed because there are no degrees of freedom left for model fit testing. The null hypothesis of exact fit is rejected by all of the statistics except G^2 ($p = .277$). However, even with just $I = 4$ items, the number of response patterns is $K = 5^4 = 625$. And due to the covariation among the item responses, not all of the response patterns are observed in the sample data. Given the sparseness and prior research on the behavior of the full-information statistics, we should be skeptical that G^2 and X^2 actually follow their purported distributions.

Table 6

Model Fit Statistics for the Four Smoking Items

Statistic	Value	*df*	*p*	$\hat{\varepsilon}$	90% CI
G^2	624.17	604	.277	.006	(.001, .012)
X^2	933.63	604	< .001	.023	(.020, .026)
M_2	245.60	92	< .001	.041	(.035, .047)
C_2	13.05	2	.002	.074	(.040, .115)

Turning to the limited information statistics, M_2 and C_2, several interesting observations can be made. First, using guidelines developed in the context of linear factor analysis and structural equation modeling for continuous data (e.g., Browne & Cudeck, 1993), assessment of model fit depends on whether $\hat{\varepsilon}_M$ or $\hat{\varepsilon}_C$ is used. Second, the relative magnitudes of $\hat{\varepsilon}_M$ and $\hat{\varepsilon}_C$ are consistent with the simulation study results, where the means of the C_2-based RMSEA estimates were consistently greater than those of the RMSEA estimates based on M_2. Again, it is apparent that the underlying statistic matters when interpreting RMSEAs.

Table 7

Marginal Frequencies for Item Pair (1,3) from the Empirical Example

		Item 1 Category Code					Marginal Frequency
Item 3 Category Code		0	1	2	3	4	for Item 3
Observed	0	93	8	5	0	0	106
(Model-Implied)		(87.4)	(14.4)	(5.1)	(1.2)	(0.4)	(108.6)
	1	129	68	20	3	4	224
		(138.0)	(55.4)	(25.1)	(6.7)	(2.5)	(227.7)
	2	79	79	63	15	7	243
		(90.1)	(74.4)	(49.2)	(16.9)	(7.2)	(237.7)
	3	56	61	76	40	13	246
		(48.1)	(67.4)	(70.5)	(35.7)	(21.0)	(242.7)
	4	22	32	31	37	59	181
		(14.6)	(29.3)	(48.2)	(40.8)	(50.3)	(183.2)
Marginal Frequency		379	248	195	95	83	1000
for Item 1		(378.2)	(240.9)	(198.1)	(101.4)	(81.4)	(1000)

Finally, for the given data set, there is reason to suspect that the M_2 statistic may not perform well. A number of the sample second order marginal tables are poorly-filled (see e.g., Table 7), which might reduce the power of M_2 against model misspecification (Cai & Hansen, 2013). On the other-hand, there is no similar concern for C_2, since it is based on a further collapsing of the second order marginal tables.

Discussion

Motivated by Maydeu-Olivares and Joe's (2006) seminal work, the limited-information test statistic M_2 has become an important new tool in formal evaluations of IRT model fit. M_2 relies on a comparison between the observed and expected first order and second order marginal probabilities. The formalism to establish asymptotic chi-squaredness of M_2 involves reduction operator matrices. Building on Joe and Maydeu-Olivares' (2010) important insight that test statistics could be formed from linear functions of the first and second order marginal residuals, Cai and Hansen (2013) proposed a limited-information test statistic M_2^* for polytomous IRT models that utilizes a comparison between observed and expected item means and second order moments, which are further reductions of the marginal probabilities. They show that in certain conditions M_2^* can be more powerful than M_2 because some of the second order marginal probabilities can become sparse in M_2.

In this research, we propose a hybrid statistic C_2 that compares the observed and expected first order marginal probabilities in unreduced form, but further collapses the second order marginal probabilities into observed and expected moments. This new statistic circumvents a limitation of M_2^*, namely, that the number of items required to compute the statistic depends on the number of categories per item. This is because M_2^* collapses several first order marginal probabilities into a single number for each item, which can render the model not locally identified from the item means and second order moments. On the other hand, the ability to compute C_2 does not depend on the number of categories per item. Also, C_2 is potentially more powerful than M_2 because it has none of the sparseness issues associated with M_2. We demonstrate the effectiveness of C_2 with simulation studies and empirical data analysis. We also make the observation that to the extent approximate model fit evaluation is desirable via the use of RMSEA indices (e.g., as advocated by Maydeu-Olivares, 2013), it is important to keep in mind the statistical properties of the underlying test statistics. Sample discrepancy measures based on different limited-information statistics, M_2, M_2^*, C_2, or full-information statistics G^2 and X^2, may paint different pictures of the degree of model error because they "estimate" different population RMSEAs.

There are obvious future directions with this line of research. We have chosen to focus on IRT models for ordinal data, completely bypassing nominal categories models. The development presented here is also limited to unidimensional IRT models. It would be desirable to implement and study a version of C_2 for hierarchical multidimensional IRT models, but it would probably require technical devices similar to those employed in Cai and Hansen (2013). Also of interest is the extension of these statistics to the case of IRT models that do not have continuous underlying latent traits. Finally, new step-down model error diagnostics would have to be developed to locate the source of misspecification and to explain the rejection of the overall goodness of fit hypothesis. There is reason to be excited about these possibilities.

References

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

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

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

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

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62-83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Cai, L. (2013). *flexMIRT® 2.0: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*, 245-276.

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 and Statistical Psychology, 59*, 173–194.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling* [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Cochran, W. G. (1952). The X^2 test of goodness of fit. *Annals of Mathematical Statistics, 23*, 315-345.

Edelen, M. O., Tucker, J. S., Shadel, W. G., Stucky, B. D., & Cai, L. (2012). Toward a more systematic assessment of smoking: Development of a smoking module for PROMIS. *Addictive Behaviors, 37*, 1278–1284.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (in press). Methodology for developing and evaluating the PROMIS© smoking item banks. *Nicotine and Tobacco Research*.

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

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

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

Maydeu-Olivares, A. (2013). Focus article: Goodness of fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*, 71-101.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs, 17*.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51–65). New York, NY: Springer.