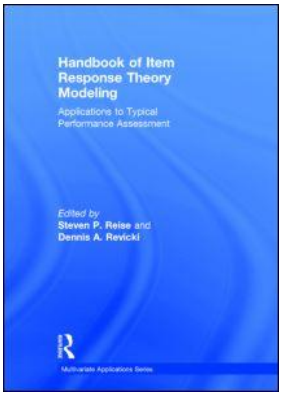

Handbook of Item Response Theory Modeling Applications to Typical Performance Assessment

Steven P. Reise, Dennis A. Revicki

Evaluating the Fit of IRT Models

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch6
Alberto Maydeu-Olivares
Published online on: 16 Dec 2014

6 Evaluating the Fit of IRT Models

Alberto Maydeu-Olivares[1]

Introduction

The goodness of fit (GOF) of a statistical model, such as an item response theory (IRT) model, describes how well the model matches a set of observations. It is useful to distinguish between goodness of fit *indices* and goodness of fit statistics. Goodness of fit indices summarize the discrepancy between the values observed in the data and the values expected under a statistical model. Goodness of fit *statistics* are GOF indices used in statistical hypothesis testing. In other words, GOF statistics are GOF indices with known sampling distributions usually obtained using asymptotic methods. Because *p*-values obtained using asymptotic methods may behave poorly in small samples, a great deal of research has been devoted to investigate using simulation studies under which conditions the asymptotic *p*-values of GOF statistics are accurate (e.g., Maydeu-Olivares & Montaño, 2013).

Assessing the absolute model fit of a model (i.e., the discrepancy between a model and the data) is critical in applications, as inferences drawn on poorly fitting models may be badly misleading. Applied researchers must examine not only the overall fit of their models, but they should also perform a piecewise assessment. It may well be that a model fits well overall but that it fits poorly some parts of the data, suggesting the use of an alternative model. Also, piecewise GOF assessment may reveal the source of misfit in poorly fitting models.

Assessing the absolute fit of a statistical model involves determining whether the model could have generated the observed data. In IRT applications, however, degrees of freedom are most often so large that no model can be expected to fit the data exactly. For example, an IRT model for 20 polytomous items, each one consisting of five response categories, involves modeling 5^{20} response patterns and it yields more than 95×10^{12} degrees of freedom. In models with so many degrees of freedom I recommend instead to assess whether the model approximately fits the data. By this we mean determining whether a goodness of fit statistic is smaller than some arbitrary nonzero value. In contrast, assessing whether the model fits exactly amounts to testing whether the value of a goodness of fit statistic equals zero.

This work is organized as follows: In this section, I review the classical statistics for assessing the overall fit of categorical data models (such as IRT models) and their limitations. In the next section, I review some new developments in this area. Thus, I describe the new limited information overall goodness of fit statistics that have been proposed in the literature as these overcome the limitations of classical statistics. I also briefly introduce

1 This research was supported by an ICREA-Academia Award and grant SGR 2009 74 from the Catalan Government and grants PSI2009–07726 and PR2010–0252 from the Spanish Ministry of Education.

methods for assessing approximate fit, as well as methods for piecewise assessment of fit. The next section includes an application to the PROMIS® depression short form (Pilkonis et al., 2011). This chapter concludes with a discussion and recommendations for applied users.

Classical Goodness of Fit Statistics

Consider the responses given by N individuals to n test items, each with K categories coded as 0,1, . . . , $K - 1$. The resulting data can be gathered in a n-dimensional contingency table with $C = K^n$ cells. Within this setting, assessing the goodness of fit of a model involves assessing the discrepancy between the observed proportions and the probabilities expected under the model across all cells of the contingency table. More formally, let π_c be the probability of one such cell (i.e., a response pattern to the n test items) and let p_c be the observed proportion, $c = 1, \ldots, C$. Also, let $\boldsymbol{\pi}(\boldsymbol{\theta})$ be the C-dimensional vector of model probabilities expressed as a function of the, say q, model parameters to be estimated from the data. Then, the 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 $X^2 = N\sum_c (p_c - \hat{\pi}_c)^2 / \hat{\pi}_c$, and the likelihood ratio statistic $G^2 = 2N\sum_c p_c \ln(p_c / \hat{\pi}_c)$, where $\hat{\pi}_c = \pi_c\left(\hat{\boldsymbol{\theta}}\right)$. 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 correct when all expected frequencies are large (>5 is the usual rule of thumb). A practical way to evaluate whether the asymptotic p-values for X^2 and G^2 are valid is to compare them. If the p-values are similar, then both are likely to be correct. If they are very different, it is most likely that both p-values are incorrect.

Unfortunately as the number of cells in the table increases, the expected frequencies must be small because the sum of all C probabilities must be equal to one (Bartholomew & Tzamourani, 1999). As a result, in IRT modeling, most often the p-values for these statistics cannot be used (Thissen & Steinberg, 1997). In fact, when the number of categories is large (say > 4) the asymptotic p-values almost invariably become inaccurate as soon as $n > 5$. To overcome the problem of the inaccuracy of the asymptotic p-values for these statistics two general methods have been proposed: resampling methods (e.g., bootstrap), and pooling cells. Unfortunately, existing evidence suggests that resampling methods do not yield accurate p-values for the X^2 and G^2 statistics (Tollenaar & Mooijart, 2003).

Pooling cells results in statistics whose asymptotic distribution may be well approximated by asymptotic methods because pooled cells must have larger expected frequencies. However, pooling must be performed before the analysis is made to obtain a statistic with the appropriate asymptotic reference distribution. A straightforward way to pool cells a priori for goodness of fit testing is to use low order margins, that is, probabilities that are univariate, bivariate, and so forth. Goodness of fit statistics based on low order margins are referred to in the literature as limited information statistics because they do not use all the information available in the data for testing the overall goodness of fit of the model. Because they are based on pooled cells, the p-values of limited information statistics are accurate in very large models even with samples as small as $N = 100$ observations. Furthermore, because they "concentrate" the information available for testing, they are most often more powerful than full information statistics such as Pearson's X^2 to detect alternatives of interest.

Research Methods

Overall Goodness of Fit Testing Using Limited Information Statistics

To understand limited information methods consider the following 2 × 3 contingency table:

	$Y_2 = 0$	$Y_2 = 1$	$Y_2 = 2$
$Y_1 = 0$	π_{00}	π_{01}	π_{02}
$Y_1 = 1$	π_{11}	π_{11}	π_{12}

This table can be characterized using the cell probabilities $\boldsymbol{\pi}' = (\pi_{00}, \cdots, \pi_{12})$. Alternatively, it can be characterized using the univariate $\dot{\boldsymbol{\pi}}_1' = \left(\pi_1^{(1)}, \pi_2^{(1)}, \pi_2^{(2)}\right)$ and bivariate $\dot{\boldsymbol{\pi}}_2' = \left(\pi_{1\ 2}^{(1)(1)}, \pi_{1\ 2}^{(1)(2)}\right)$ probabilities, where

	$Y_2 = 0$	$Y_2 = 1$	$Y_2 = 2$	
$Y_1 = 0$				
$Y_1 = 1$		$\pi_{1\ 2}^{(1)(1)}$	$\pi_{1\ 2}^{(1)(2)}$	$\pi_1^{(1)}$
		$\pi_2^{(1)}$	$\pi_2^{(2)}$	

and $\pi_2^{(2)} = \Pr(Y_2 = 2)$ and $\pi_{1\ 2}^{(1)(2)} = \Pr(Y_1 = 1, Y_2 = 2)$. Both characterizations are equivalent, and the equivalence extends to contingency tables of any dimension. In other words, one can always transform the cell probabilities into the moments $\boldsymbol{\pi}_2' = \left(\dot{\boldsymbol{\pi}}_1', \dot{\boldsymbol{\pi}}_2'\right)$ and vice versa. $\dot{\boldsymbol{\pi}}_1$ and $\dot{\boldsymbol{\pi}}_2$ are clearly univariate and bivariate moments if the variables are binary, and moments of indicator variables used to denote each category except the zero category if the variables are polytomous (Maydeu-Olivares & Joe, 2006). I use the term *moments* to distinguish $\dot{\boldsymbol{\pi}}_1$ and $\dot{\boldsymbol{\pi}}_2$ from the set of univariate and bivariate probabilities, $\hat{\boldsymbol{\pi}}_1' = \left(\pi_1^{(0)}, \pi_1^{(1)}, \pi_2^{(0)}, \pi_2^{(1)}, \pi_2^{(2)}\right)$ and $\hat{\boldsymbol{\pi}}_2 = \boldsymbol{\pi}$ (in this example). Notice that the moments of order *r* simply consist of the *r*-way marginal probabilities that do not involve the category 0.

A limited information goodness of fit statistic uses only the moments up to order $r < n$ for testing. Thus, in the example cited earlier, a statistic that only involves univariate moments would be a limited information test statistic. In contrast, full information statistics use all moments (up to order *n* or $\boldsymbol{\pi}_n$). Pearson's X^2 statistic is a full information statistic and therefore it can be written as a function of the cell probabilities:

$$X^2 = N(\mathbf{p} - \hat{\pi})' \hat{\mathbf{D}}^{-1} (\mathbf{p} - \hat{\pi}) \tag{6.1}$$

where $\mathbf{p} - \hat{\pi}$ are the cell residuals, and $\hat{\mathbf{D}} = diag\left(\pi\left(\hat{\theta}\right)\right)$ is a diagonal matrix of estimated cell probabilities, or as a function of the moments:

$$X^2 = N\left(\mathbf{p}_n - \hat{\pi}_n\right)' \hat{\Xi}_n^{-1} \left(\mathbf{p}_n - \hat{\pi}_n\right) \tag{6.2}$$

where $\mathbf{p}_n - \hat{\pi}_n$ are the residual moments, and $N\hat{\Xi}_n$ is the asymptotic covariance matrix of the sample moments up to order n, $\mathbf{p}_n$, evaluated at the parameter estimates.

For IRT applications Maydeu-Olivares and Joe (2005, 2006) suggested testing using $r = 2$, that is, using only univariate and bivariate moments because the lower the order of moments used the more accurate the p-values and (generally) the higher the power. More specifically, they suggested testing using the limited information test statistic:

$$M_2 = N\left(\mathbf{p}_2 - \hat{\pi}_2\right)' \hat{\mathbf{C}}_2 \left(\mathbf{p}_2 - \hat{\pi}_2\right), \qquad \mathbf{C}_2 = \Xi_2^{-1} - \Xi_2^{-1}\boldsymbol{\Delta}_2(\boldsymbol{\Delta}_2'\Xi_2^{-1}\boldsymbol{\Delta}_2)^{-1}\boldsymbol{\Delta}_2'\Xi_2^{-1}, \tag{6.3}$$

where $\boldsymbol{\Delta}_r$ denotes the matrix of derivatives of the univariate and bivariate moments with respect to the parameter vector $\boldsymbol{\theta}$, and $N\Xi_2$ denotes the asymptotic covariance matrix of the univariate and bivariate sample moments. These matrices are evaluated at the parameter estimates, $\hat{\boldsymbol{\theta}}$. When all items consist of the same number of categories, K, M_2 is asymptotically distributed as a chi-square with $df_2 = n(K-1) + \frac{n(n-1)}{2}(K-1)^2 - q$ degrees of freedom. M_2 is a member of the M_r class of test statistics (M_1, M_2, M_3, . . . , M_n). The members of this class of statistics are of the form (6.3) and simply differ from M_2 in the amount of information used. Thus, in M_1 only univariate moments are used. IRT models cannot be tested using only univariate information as there are no degrees of freedom available for testing. In M_3, univariate, bivariate, and trivariate moments are used, whereas in M_n all moments (up to order n) are used. For maximum likelihood estimation, M_n equals X_2 algebraically (i.e., the second term in the weight matrix equals zero).

Testing Models for Large and Sparse Ordinal Data

When the number of categories per item is large M_2 suffers from two limitations. The first limitation is that if the number of items is also large, M_2 may not be computable because of the size of the matrices that need to be stored in memory. The second limitation is that the bivariate tables may be sparse, particularly in one or both extremes of the response scale. In this case, the asymptotic p-values of M_2 may not be accurate enough (Cai & Hansen, 2013).

If the number of variables and categories is large, or if the bivariate tables are sparse, one should assess the overall goodness of fit of the model using M_{ord},

$$M_{ord} = N\left(\mathbf{k} - \hat{\kappa}\right)' \hat{\mathbf{C}}_{ord} \left(\mathbf{k} - \hat{\kappa}\right), \; \mathbf{C}_{ord} = \boldsymbol{\Xi}_{ord}^{-1} - \boldsymbol{\Xi}_{ord}^{-1}\boldsymbol{\Delta}_{ord}(\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_{ord}^{-1}\boldsymbol{\Delta}_{ord})^{-1}\boldsymbol{\Delta}_{ord}'\boldsymbol{\Xi}_{ord}^{-1}, \tag{6.4}$$

which is a statistic for ordinal data only. This statistic has the same form as M_2 but the statistics in the quadratic form are now the sample means and cross-products $\mathbf{k}$. Thus, $N\Xi_{ord}$ is their asymptotic covariance matrix, $\boldsymbol{\kappa}$ is the population counterpart of $\mathbf{k}$ (the population means and cross-products of the multinomial variables ignoring the multivariate nature of the multinomial variables), and $\boldsymbol{\Delta}_{ord}$ is the matrix of derivatives of $\boldsymbol{\kappa}$ with respect to the model parameters, $\boldsymbol{\theta}$. $\boldsymbol{\kappa}$, Ξ_{ord} and $\boldsymbol{\Delta}_{ord}$ are to be evaluated at the parameter estimates—that is, $\hat{\kappa}$ denotes $\kappa\left(\hat{\theta}\right)$. More specifically the elements of $\boldsymbol{\kappa}$ are of the form:

$$\kappa_i = E[Y_i] = 0 \times \Pr(Y_i = 0) + \ldots + K_i \times \Pr(Y_i = K_i), \tag{6.5}$$

$$\kappa_{ij} = E[Y_i Y_j] = 0 \times 0 \times \Pr(Y_i = 0, Y_j = 0) + \ldots + K_i \times K_j \times \Pr(Y_i = K_i, Y_j = K_j), \tag{6.6}$$

with sample counterparts $k_i = \bar{y}_i$ (the sample mean), and $k_{ij} = \mathbf{y}_i' \mathbf{y}_j / N$ (the sample cross-product), respectively. In particular, for our previous example, the elements of $\boldsymbol{\kappa}$ are

$$\begin{aligned}
\kappa_1 &= E[Y_1] = 1 \Pr(Y_1 = 1) = \pi_1^{(1)} \\
\kappa_2 &= E[Y_2] = 1 \Pr(Y_2 = 1) + 2 \Pr(Y_2 = 1) = \pi_2^{(1)} + 2\pi_2^{(2)} \\
\kappa_{12} &= E[Y_1 Y_2] = 1 \times 1 \Pr(Y_1 = 1, Y_2 = 1) + 1 \times 2 \Pr(Y_1 = 1, Y_2 = 2) = \pi_{1\,2}^{(1)(1)} + 2\pi_{1\,2}^{(1)(2)}.
\end{aligned} \tag{6.7}$$

Thus, for our 2 × 3 example, M_2 is a quadratic form in the sample counterparts of $\pi_2' = \left(\pi_1^{(1)}, \pi_2^{(1)}, \pi_2^{(2)}, \pi_{1\,2}^{(1)(1)}, \pi_{1\,2}^{(1)(2)}\right)$, and M_{ord} is a quadratic form in the sample counterparts of $\boldsymbol{\kappa}$ given in (6.7). Clearly, $\boldsymbol{\kappa}$ is obtained as a linear combination of $\boldsymbol{\varphi}_2$ where the weights are used as given by the coding of the categories. Thus, it only makes sense to use $\boldsymbol{\kappa}$ and their sample counterparts, and therefore M_{ord}, when the data is ordinal. When the data is binary, M_{ord} equals M_2. In general, M_{ord} is asymptotically distributed as a chi-square with $df_{ord} = \dfrac{n(n+1)}{2} - q$ degrees of freedom.

M_{ord} cannot be used if the number of categories is large and the number of items is small because of lack of degrees of freedom for testing. For instance, for a unidimensional logistic graded model (e.g., Samejima, 1969), the number of items must be larger than the number of categories plus two (i.e., $n \geq K + 2$) for the degrees of freedom M_{ord} to be positive.

To summarize this subsection, for ordinal data, if the model involves a large number of variables and categories one must resort to M_{ord} as M_2 cannot be computed. On the other hand, when the number of categories is large and the number of items is small, M_{ord} cannot be computed because of lack of degrees of freedom. In some medium-sized models for ordinal data, there is a choice between M_2 and M_{ord}. Because $\boldsymbol{\kappa}$ concentrates the information available in $\boldsymbol{\varphi}_2$, M_{ord} may be more powerful than M_2 (Joe & Maydeu-Olivares, 2010). On the other hand, if the concentration of the information is not along the alternative of interest, M_2 will be more powerful than M_{ord} along that direction.

Testing for Approximate Fit

In IRT applications to patient-reported outcomes, degrees of freedom are so large that it is unrealistic to expect that any model will fit the data. In other words, it is unrealistic to expect that the fitted IRT model is the data-generating mechanism. Hence, it is more reasonable to test for approximate fit than for exact fit. By this I simply mean testing whether some statistic is smaller than some cutoff. Drawing from work on the structural equations modeling literature by Browne and Cudeck (1993), Maydeu-Olivares and Joe (2014) have recently suggested the use of the sample bivariate root mean square error of approximation ($RMSEA_2$):

$$\hat{\varepsilon}_2 = \sqrt{\frac{M_2 - df_2}{N \times df_2}} \tag{6.8}$$

to estimate the corresponding population bivariate population RMSEA_2. They suggested that a cutoff of $\varepsilon_2 \leq 0.05$ indicates adequate fit. They show that this cutoff separates rather well mis-specified IRT models with correctly specified latent trait dimensionality from mis-specified IRT models with mis-specified latent trait dimensionality. They also show that the population RMSEA_2 is relatively unaffected by the number of variables being tested, but that it is strongly affected by the number of categories. The larger the number of categories, the smaller the value of the population RMSEA_2. They also showed that dividing the RMSEA_2 by the number of categories minus one, one obtains an RMSEA_2 relatively unaffected by the number of categories. Consequently, they suggest using $\varepsilon_2 \leq 0.05/(K-1)$ as a cutoff for good fit.

A RMSEA_{ord} can be similarly constructed around M_{ord}:

$$\hat{\varepsilon}_{ord} = \sqrt{\frac{M_{ord} - df_{ord}}{N \times df_{ord}}}. \tag{6.9}$$

However, if M_{ord} is more powerful than M_2, then RMSEA_{ord} must be larger than RMSEA_2 as the RMSEAs are simply a function of the estimated non-centrality parameters. Thus, a larger cutoff must be used for RMSEA_{ord} than for RMSEA_2. Most importantly, RMSEA_{ord} is strongly affected by the number of variables: the larger the number of variables the smaller the population RMSEA_{ord}, all other factors constant. For these reasons, to assess the approximate fit of large models for ordinal data I advocate instead the use of a Standardized Root Mean Square Residual (SRMSR) borrowed from the factor analysis literature (see for instance Hu & Bentler, 1999). For a pair of items i and j, the standardized residual is defined as the sample (product-moment or Pearson) correlation minus the expected correlation. In turn, the expected correlation simply equals the expected covariance divided by the expected standard deviations:

$$r_{ij} - \hat{\rho}_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} - \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}}\sqrt{\hat{\sigma}_{jj}}} = r_{ij} - \frac{\hat{\kappa}_{ij} - \hat{\kappa}_i\hat{\kappa}_i}{\sqrt{\hat{\kappa}_{ii} - \hat{\kappa}_i^2}\sqrt{\hat{\kappa}_{jj} - \hat{\kappa}_j^2}}. \tag{6.10}$$

where the means (κ_i and κ_j) and the cross-product κ_{ij} were given in (6.5) and (6.6), and κ_{ii} is:

$$\kappa_{ii} = E\left[Y_i^2\right] = 0^2 \times \Pr\left(Y_i = 0\right) + \ldots + K_i^2 \times \Pr\left(Y_i = K_i\right). \tag{6.11}$$

The SRMSR is simply the squared root of the average of these squared correlation residuals:

$$SRMSR = \sqrt{\sum_{i<j} \frac{\left(r_{ij} - \hat{\rho}_{ij}\right)^2}{n(n-1)/2}}. \tag{6.12}$$

An advantage of the SRMSR over the RMSEAs is that because the SRMSR is an average of standardized residuals, its interpretation is straightforward. In contrast, the RMSEAs cannot be readily interpreted. An advantage of the RMSEAs (6.8) and (6.9) over the SRMSR is that it is straightforward to compute confidence intervals and hypothesis testing for them because they are simply transformations of the M_2 and M_{ord} statistics, which are chi-square distributed when the fitted model is correctly specified (Maydeu-Olivares & Joe, 2014). In contrast, computation of confidence intervals and hypothesis testing for the SRMSR is cumbersome as the asymptotic distribution of SRMSR is a mixture of independent chi-squares when the fitted model is correctly specified. Thus, the SRMSR is best used as a goodness of fit index with SRMSR ≤ 0.05 indicating adequate fit. Of course, another

advantage of the RMSEA is that it takes model complexity into account, although this is only of interest when comparing different models fitted to a data set.

Piecewise Assessment of Fit

After examining the overall fit of a model, it is necessary to perform a piecewise goodness of fit assessment. If the overall fit is poor, a piecewise assessment of fit may suggest how to modify the model. Even if the model fits well overall, a piecewise goodness of fit assessment may reveal parts of the model that misfit. A useful starting point for our discussion of piecewise fit assessment is the bivariate Pearson's X^2 statistic. After the IRT model parameters have been estimated using the full data, a X^2 may be computed for each pair of variables:

$$X_{ij}^2 = N\left(\mathbf{p}_{ij} - \hat{\pi}_{ij}\right)' \hat{\mathbf{D}}_{ij}^{-1}\left(\mathbf{p}_{ij} - \hat{\pi}_{ij}\right). \tag{6.13}$$

This is just the standard X^2 statistic (6.1) applied to the bivariate table involving variables i and j. Thus, for a model fitted to K category items, $\mathbf{p}_{ij}$ is K^2 vector of observed bivariate proportions, $\hat{\pi}_{ij} = \pi_{ij}\left(\hat{\theta}_{ij}\right)$ is the vector of expected probabilities that depend only on the q_{ij} parameters involved in the bivariate table, $\hat{\theta}_{ij}$, and $\hat{\mathrm{D}}_{ij} = \mathrm{diag}(\hat{\boldsymbol{\pi}}_{ij})$. Suppose the model is unidimensional graded logistic, in this case, q_{ij} involves 2 slopes and $2 \times (K - 1)$ intercepts. It is tempting to refer X_{ij}^2 to a chi-square distribution degrees of freedom equal to the number of parameters in the unrestricted model $\boldsymbol{\pi}_{ij}$, $K^2 - 1$, minus the number of parameters in the restricted model $\pi_{ij}\left(\theta_{ij}\right)$, q_{ij}, so that $df_{ij} = K^2 \times q_{ij} - 1$. However, Maydeu-Olivares and Joe (2006) showed that the distribution of X_{ij}^2 is larger than this reference distribution. This means that referring X_{ij}^2 to this distribution leads to rejecting well-fitting items. They also showed that the M_2 statistic (6.3) applied to a bivariate subtable is asymptotically distributed as chi-square with df_{ij} degrees of freedom. Finally, they also showed that when applied to a single marginal subtable M_2 can be written in terms of the bivariate cell residuals as:

$$M_2^{(ij)} = X_{ij}^2 - N\left(\mathbf{p}_{ij} - \hat{\pi}_{ij}\right)' \hat{\mathbf{D}}_{ij}^{-1}\hat{\Delta}_{ij}\left(\hat{\Delta}_{ij}'\hat{\mathbf{D}}_{ij}^{-1}\hat{\Delta}_{ij}\right)^{-1}\hat{\Delta}_{ij}'\hat{\mathbf{D}}_{ij}^{-1}\left(\mathbf{p}_{ij} - \hat{\pi}_{ij}\right), \tag{6.14}$$

where $\boldsymbol{\Delta}_{ij}$ denotes the matrix of derivatives of the bivariate probabilities $\boldsymbol{\pi}_{ij}$ with respect to the parameters involved in the bivariate table, $\hat{\boldsymbol{\theta}}_{ij}$. Unfortunately, Maydeu-Olivares and Liu (in press); see also Liu & Maydeu-Olivares, 2012) have recently shown that $M_2^{(ij)}$ does not have much power against certain alternatives. $M_2^{(ij)}$ is simply a correction to X_{ij}^2. Alternatively, X_{ij}^2 can be corrected by its asymptotic mean and variance:

$$\mathrm{MV}\left(X_{ij}^2\right) = X_{ij}^2\sqrt{\frac{df_{ij}}{tr_2}} + df_{ij} - \sqrt{\frac{df_{ij}tr_1^2}{tr_2}}, \tag{6.15}$$

which is referred to a chi-square distribution with $df_{ij} = K^2 \times q_{ij} - 1$ degrees of freedom. In (6.15):

$$\begin{aligned} tr_1 &= \mathrm{tr}\left(\mathbf{D}_{ij}^{-1}\Omega_{ij}\right) \\ tr_2 &= \mathrm{tr}\left(\mathbf{D}_{ij}^{-1}\Omega_{ij}\mathbf{D}_{ij}^{-1}\Omega_{ij}\right) \end{aligned} \tag{6.16}$$

and:

$$N\hat{\Omega}_{ij} = \mathbf{D}_{ij} - \pi_{ij}\pi_{ij}' - \Delta_{ij}\left(\mathcal{I}^{-1}\right)_{ij}\Delta_{ij}' \tag{6.17}$$

is the asymptotic covariance matrix of the cell residuals for the pair of variables i and j when the model parameters have been estimated by maximum likelihood using the full table. In Equation (6.17), $\mathcal{I}^{-1}$ denotes the covariance matrix of the full set of item parameters and $\left(\mathcal{I}^{-1}\right)_{ij}$ denotes the rows and columns of this matrix corresponding to the item parameters involved in the subtable for variables i and j. The covariance matrix $\mathcal{I}^{-1}$ is generally estimated using the cross-products information matrix (e.g., Bock & Lieberman, 1970):

$$\mathcal{I}_O = \Delta_O' \mathbf{D}_O \Delta_O,\ \mathbf{D}_O = \mathrm{diag}\left(\mathbf{p}_O / \pi_O^2\right), \tag{6.18}$$

where $\mathbf{p}_O$ and $\boldsymbol{\pi}_O$ denote the proportions and probabilities of the C_O observed patterns, and $\boldsymbol{\Delta}_O$ is a $C_O \times q$ matrix of derivatives of the observed patterns with respect to the full set of q item parameters.

A drawback of using quadratic form statistics such as $\mathrm{MV}\left(X_{ij}^2\right)$ and $M_2^{(ij)}$ is that they do not convey information about the direction of misfit, as these statistics are necessarily positive. Thus, they are best combined with the residual correlations of Equation (6.10) as these indicate the direction of the misfit: A positive residual correlation implies a model expected correlation larger than the observed correlation, whereas a negative residual correlation implies an observed correlation larger than the model expected correlation.

Another drawback of quadratic form statistics such as $\mathrm{MV}\left(X_{ij}^2\right)$ and $M_2^{(ij)}$ is that they cannot be employed with binary data because of lack of degrees of freedom. For the same reason, they cannot be applied to assess the misfit of single items. Z-statistics for univariate and bivariate residual moments can be used instead to diagnose the fit of models for binary data (Maydeu-Olivares & Joe, 2005; Reiser, 1996). These z-statistics are simply:

$$z_i = \frac{p_i - \hat{\pi}_i}{\mathrm{SE}\left(p_i - \hat{\pi}_i\right)} = \frac{p_i - \hat{\pi}_i}{\sqrt{\hat{\omega}_{ii} / N}} \quad z_{ij} = \frac{p_{ij} - \hat{\pi}_{ij}}{\mathrm{SE}\left(p_{ij} - \hat{\pi}_{ij}\right)} = \frac{p_{ij} - \hat{\pi}_{ij}}{\sqrt{\hat{\omega}_{ij,ij} / N}}. \tag{6.19}$$

Here, $\pi_i = \Pr\left(Y_i = 1\right)$, $\pi_{ij} = \Pr\left(Y_i = 1, Y_j = 1\right)$, p_i and p_{ij} are their corresponding proportions and $\hat{\omega}_{ii}$ and $\hat{\omega}_{ij,ij}$ are the corresponding diagonal elements of $\hat{\Omega}_{ij}$ in (6.17).

In polytomous data, a z-statistic can also be easily computed for the bivariate residual cross-product:

$$z_{ord} = \frac{k_{ij} - \hat{\kappa}_{ij}}{\mathrm{SE}\left(k_{ij} - \hat{\kappa}_{ij}\right)} = \frac{k_{ij} - \hat{\kappa}_{ij}}{\sqrt{\hat{\omega}_{ord} / N}} \tag{6.20}$$

where κ_{ij} is given in (6.6) and:

$$N\hat{\omega}_{ord} = N\mathbf{v}'\hat{\Omega}_{ij}\mathbf{v} = \mathbf{v}'\left(\mathbf{D}_{ij} - \pi_{ij}\pi_{ij}' - \Delta_{ij}\left(\mathcal{I}^{-1}\right)_{ij}\Delta_{ij}'\right)\mathbf{v}, \tag{6.21}$$

with $\mathbf{v}' = (0,1,\ldots,K-1)$. However, in some instances (6.21) becomes negative when the cross-products information (6.18) is used and z_{ord} cannot be computed. Univariate counterparts of (6.20) cannot be computed.

Application

To illustrate the described procedures I will use the $n = 8$ item PROMIS® depression short form (Pilkonis et al., 2011). Respondents are asked to report the frequency with which they experienced certain feelings in the past seven days using a $K = 5$ point rating scale ranging from "never" to "always." The responses were coded from 0 to 4 for the analyses. I used the $N = 768$ complete responses to these data kindly provided by the editors. A unidimensional logistic graded response model (Samejima, 1969) with a normally distributed latent trait was estimated by maximum likelihood using flexMIRT (Cai, 2012); 100 rectangular quadrature points between −8 and 8 were used and standard errors were computed using the cross-products information matrix (6.18). The item stems and the estimated intercepts and slopes are reported in Table 6.1 (in logistic metric). There are $q = 5 \times 8 = 40$ estimated parameters.

The model does not fit the data exactly as the value of the statistic M_2 in (6.3) is 767.58 on 440 degrees of freedom, $p < 0.01$. The bivariate RMSEA (6.8) estimate is $\text{RMSEA}_2 = 0.03$.[2] We can compute a 90 percent confidence interval around its true parameter (Maydeu-Olivares & Joe, 2014) obtaining (0.03; 0.03).[3] Thus, the fit of the model is adequate ($\text{RMSEA}_2 \leq 0.05$) but falls short of our criterion for excellent fit, $\text{RMSEA}_2 \leq (0.05 / (K - 1) = 0.0125$.

The residual correlations (6.10) provide us with an easy-to-interpret assessment of the magnitude of the misfit (the effect size of the misfit). The standardized squared root mean squared residual is low, SRMSR = 0.02, indicating that the average size of the misfit is very small. Examining the individual residual correlations shown in Table 6.2, we see that all of them are small. In fact, only three of them are larger than 5 percent in absolute value: those corresponding to the item pairs (5,1), (7,4), and (8,3). I have also included in this table the average of the absolute values of the residual correlations involving each item. Interestingly, the average residual correlation is similar for all items.

Table 6.1 PROMIS® Depression Short Form: Estimated Item Parameters and Standard Errors for a Logistic Graded Model

Item	*Stem*	*Slope*	*Intercept 1*	*Intercept 2*	*Intercept 3*	*Intercept 4*
1	I felt worthless	4.05(0.29)	−2.24(0.24)	−4.62(0.35)	−7.15(0.50)	−9.70(0.71)
2	I felt like a failure	3.35(0.22)	−1.40(0.19)	−3.26(0.25)	−5.73(0.39)	−8.60(0.72)
3	I felt depressed	3.66(0.28)	−1.70(0.21)	−3.92(0.30)	−6.74(0.44)	−9.51(0.72)
4	I felt hopeless	3.40(0.23)	1.36(0.18)	−1.60(0.19)	−4.88(0.33)	−8.39(0.58)
5	I felt that I had nothing to look forward to	3.71(0.27)	−1.11(0.20)	−3.11(0.26)	−6.50(0.45)	−8.48(0.61)
6	I felt helpless	3.63(0.24)	0.11(0.18)	−2.61(0.23)	−5.31(0.34)	−8.69(0.56)
7	I felt unhappy	4.08(0.29)	1.28(0.21)	−2.09(0.24)	−5.74(0.39)	−9.16(0.61)
8	I felt sad	4.65(0.39)	−2.25(0.27)	−5.15(0.45)	−8.01(0.65)	−11.70(1.18)

Notes: $N = 768$; maximum likelihood estimation was used; standard errors in parentheses.

2 FlexMIRT provides both M_2 and M_{ord} (as their associated RMSEAs) whereas IRTPRO currently only provides M_2.

3 The extremes of the confidence interval are equal to two decimal digits.

Table 6.2 Residual Correlations After Fitting a Graded Model

Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
1		0.03	0.02	<–0.01	**0.06**	–0.01	–0.01	0.02	0.02
2	0.03		<–0.01	–0.01	0.03	–0.01	0.02	0.02	0.02
3	0.02	<–0.01		<–0.01	0.01	0.01	–0.02	**0.08**	0.02
4	<–0.01	–0.01	<–0.01		–0.03	0.04	**0.06**	<–0.01	0.02
5	**0.06**	0.03	0.01	–0.03		<–0.01	–0.01	0.02	0.02
6	–0.01	–0.01	0.01	0.04	<–0.01		0.04	<–0.01	0.02
7	–0.01	0.02	–0.02	**0.06**	–0.01	0.04		–0.01	0.02
8	0.02	0.02	**0.08**	<–0.01	0.02	<–0.01	–0.01		0.02

Notes: *df* = 14; I have marked in bold the correlations larger than |0.05|.

Table 6.3 Mean and Variance Adjusted Bivariate X^2_{ij} Statistics

Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
1		24.36	22.52	**37.20**	23.58	26.10	**37.70**	19.30	27.25
2	24.36		17.30	23.16	**38.48**	22.16	30.45	23.55	25.64
3	22.52	17.30		**35.55**	33.61	19.35	17.76	27.13	24.75
4	**37.20**	23.16	**35.55**		32.83	33.67	33.14	**38.83**	33.48
5	23.58	**38.48**	33.61	32.83		21.16	**56.01**	23.77	32.78
6	26.10	22.16	19.35	33.67	21.16		23.27	17.55	23.32
7	**37.70**	30.45	17.76	33.14	**56.01**	23.27		25.68	32.00
8	19.30	23.55	27.13	**38.83**	23.77	17.55	25.68		25.12

Notes: *df* = 14; I have marked in bold the statistics statistically significant at the 5 percent significance level with a Bonferroni adjustment.

The mean and variance corrected X^2 statistics for each pair of variables (6.15) reported in Table 6.3 provide us with an alternative way to locate the source of the misfit. Because the tests are not independent, and to control for the multiple testing, I use a Bonferroni adjusted *p*-value. Because there are (8 × 7) / 2 = 28 statistics the cut-off *p*-value used is 0.05 / 28 = 0.002. The critical value for a chi-square distribution with $5^2 - 2 \times 5 - 1 = 14$ degrees of freedom yielding this *p*-value is 34.43. I have boldfaced all the values above this critical value in Table 6.3. Next to the statistic's value I could have indicated whether the corresponding residual is positive (+) or negative (–); this is not necessary, as I provide the residual correlations in the previous table. As we can see in this table, there are six statistically significant residuals, corresponding to the item pairs (4,1), (4,3), (5,2), (7,1), (7,5), and (8,4). Table 6.3 also reports the average of the values of these statistics for each item. An inspection of these averages suggests that the worst-fitting item in this short form is item 4.

Alternative Statistics for Piecewise Model Fit Assessment

At this time, I believe that the two statistics that show greatest promise for detecting the source of misfit in IRT models are the standardized residual correlations and the mean

Table 6.4 Bivariate X_{ij}^2 Statistics

Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
1		24.63	23.33	**38.95**	24.02	27.33	**39.79**	18.99	28.15
2	24.63		17.30	23.50	**40.21**	22.54	31.58	24.48	26.32
3	23.33	17.30		**37.15**	**35.16**	20.32	18.51	28.05	25.69
4	**38.95**	23.50	**37.15**		34.09	**35.20**	**34.87**	**40.58**	34.91
5	24.02	**40.21**	**35.16**	34.09		21.83	**59.54**	23.85	34.10
6	27.33	22.54	20.32	**35.20**	21.83		24.03	17.57	24.12
7	**39.79**	31.58	18.51	**34.87**	**59.54**	24.03		26.18	33.50
8	18.99	24.48	28.05	**40.58**	23.85	17.57	26.18		25.67

Notes: I have marked in bold the statistics larger than 34.43, the critical point for a chi-square distribution with 14 *df* and 5 percent significance level with a Bonferroni adjustment.

Table 6.5 Bivariate $M_2^{(ij)}$ Statistics

Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
1		22.26	20.83	30.99	13.50	19.18	23.16	16.77	20.96
2	22.26		15.33	22.30	**35.69**	16.42	25.54	22.27	22.83
3	20.83	15.33		**34.88**	33.23	16.07	9.42	12.89	20.38
4	30.99	22.30	**34.88**		17.13	30.72	20.83	32.29	27.02
5	13.50	**35.69**	33.23	17.13		20.09	**34.49**	20.01	24.88
6	19.18	16.42	16.07	30.72	20.09		14.27	16.15	18.99
7	23.16	25.54	9.42	20.83	**34.49**	14.27		14.90	20.37
8	16.77	22.27	12.89	32.29	20.01	16.15	14.90		19.33

Notes: $df = 14$; I have marked in bold the statistics statistically significant at the 5 percent significance level with a Bonferroni adjustment.

and variance-adjusted X_{ij}^2 statistics. In this subsection, I present the results obtained using the other statistics discussed in this chapter, X_{ij}^2, $M_2^{(ij)}$, and z_{ord}, given in Equations (6.13), (6.14), and (6.20), respectively.[4] X_{ij}^2 must reject more often than expected under a reference a chi-square distribution with df_{ij} degrees of freedom as its distribution is larger than this reference distribution. Results for X_{ij}^2 are shown in Table 6.4. As we can see in this table, it incorrectly suggests that nine pairs of items show misfit even after applying a Bonferroni correction. In contrast, the use of $M_2^{(ij)}$ suggests that only three of these nine pairs are statistically significant using the same reference distribution (see Table 6.5). $M_2^{(ij)}$ may not be powerful enough to detect some misspecifications because the mean and variance-corrected X_{ij}^2 suggests that six out of the nine pairs items flagged by X_{ij}^2 are statistically significant.

4 All these statistics can be computed using *R* code provided in Liu and Maydeu-Olivares (2014).

Table 6.6 Chen and Thissen's Standardized LD X2 Statistics (CT X^2_{ij})

Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
1		1.6	1.4	**4.1**	1.5	2.0	**4.3**	0.6	2.21
2	1.6		0.3	1.4	**4.4**	1.2	2.8	1.8	1.93
3	1.4	0.3		**3.8**	**3.4**	0.8	0.5	2.3	1.79
4	**4.1**	1.4	**3.8**		**3.2**	**3.4**	**3.5**	**4.4**	3.40
5	1.5	**4.4**	**3.4**	**3.2**		1.1	**7.8**	1.5	3.27
6	2.0	1.2	0.8	**3.4**	1.1		1.5	0.3	1.47
7	**4.3**	2.8	0.5	**3.5**	**7.8**	1.5		1.8	3.17
8	0.6	1.8	2.3	**4.4**	1.5	0.3	1.8		1.81

Notes: I have marked in bold the statistics statistically significant at the 5 percent significance level with a Bonferroni adjustment ($z > |2.91|$).

The X^2_{ij}, $M_2^{(ij)}$, and MV(X^2_{ij}) statistics are closely related to the standardized LD X^2 statistic introduced by Chen and Thissen (1997). When all items consist of the same number of categories, the LD X^2 statistic is:

$$\text{LD } X^2 = \frac{X^2_{ij} - (K-1)^2}{\sqrt{2(K-1)^2}}. \tag{6.22}$$

This statistic is conveniently printed as an option by the software used in this application, flexMIRT, and the results obtained using this statistic are shown in Table 6.6.[5] Given the lack of an appropriate reference distribution for X^2_{ij}, Chen and Thissen (1997) observed empirically that its distribution could be approximated when fitting a two-parameter logistic model for binary data using a chi-square degrees of freedom corresponding to an independence model. If a chi-square distribution with independence degrees of freedom closely matches the distribution of X^2_{ij}, then the statistic (6.22) should be approximately standard normal. Yet we see in Table 6.6 that standardized LD X^2 is even more liberal than X^2_{ij}: it rejects too often.

I now turn to the z_{ord} statistic given in Equation (6.20). Results for this statistic are presented in Table 6.7. It is simply the z-statistic for the cross-product of two item scores. I mentioned that this statistic may not be available for some pairs, particularly in small samples. We see in Table 6.9 that this is the case for three item pairs, none of which was flagged as misfitting by any of the previous procedures. Yet the z_{ord} statistics suggest that there is only one misfitting pair, that of items (5,4). This was only flagged by the LD X^2 statistic and it was not the pair with highest LD value.

In Table 6.8 I conveniently summarize the application of all these statistics for piecewise model fit to the Depression Short scale by listing the item pairs flagged by each procedure. This table shows a large degree of agreement between X^2_{ij}, $M_2^{(ij)}$, MV(X^2_{ij}), and Chen and Thissen's LD X^2. This is not surprising as all of them are based on the same statistic, Pearson's X^2 applied to a pair of variables, which I have denoted

5 In Cai (2012) *p* next to a bivariate local dependence (LD) statistic indicates positive LD—a negative residual correlation, whereas *n* indicates a negative LD—a positive residual correlation.

Table 6.7 Bivariate Standardized Cross-Products (z_{ord} statistics)

Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1		0.59	0.76	–0.41	1.59	–0.49	–0.49	0.88
2	0.59		–0.25	–1.01	0.22	–1.26	0.28	0.31
3	0.76	–0.25		–0.32	0.32	0.33	*n.a.*	2.63
4	–0.41	–1.01	–0.32		**–4.23**	1.22	2.62	–0.17
5	1.59	0.22	0.32	**–4.23**		–0.63	–1.18	0.71
6	–0.49	–1.26	0.33	1.22	–0.63		*n.a.*	–0.06
7	–0.49	0.28	*n.a.*	2.62	–1.18	*n.a.*		*n.a.*
8	0.88	0.31	2.63	–0.17	0.71	–0.06	*n.a.*	

Notes: I have marked in bold the statistics statistically significant at the 5 percent significance level with a Bonferroni adjustment ($z > |2.91|$); *n.a.* = not available because the estimated variance of the residual cross-product is negative.

Table 6.8 Item Pairs That Show Misfit Using Different Procedures for Piecewise Model Fit Assessment

Item pair	X_{ij}^2	$MV(X_{ij}^2)$	$M_2^{(ij)}$	$CT\ X_{ij}^2$	z_{ord}	$r_{ij} - \hat{\rho}_{ij}$
4,1	√	√		√		
4,3	√	√	√	√		
5,1						√
5,2	√	√	√	√		
5,3	√			√		
5,4				√	√	
6,4	√			√		
7,1	√	√		√		
7,4	√			√		√
7,5	√	√	√	√		
8,3						√
8,4	√	√		√		

Notes: $MV(X_{ij}^2)$ = mean and variance adjusted X_{ij}^2, $r_{ij} - \hat{\rho}_{ij}$ = residual correlation, $CT\ X_{ij}^2$ = Chen and Thissen's standardized LD X_{ij}^2.

by X_{ij}^2. The statistics differ in the extent to which they reject pairs. We know that X_{ij}^2 is too liberal (it rejects too often), $M_2^{(ij)}$ may be too conservative, and we see in this table that the mean and variance-corrected X_{ij}^2 lies in between. We also see in Table 6.8 that the results obtained using these statistics do not agree with results obtained using *z*-scores for residual cross-products, or with residual correlations. The latter do not agree with each other. More work is needed to gauge the performance of all these competing statistics.

Table 6.9 PROMIS® Depression Short Form: Estimated Item Parameters and Standard Errors for a Logistic Graded Model with Correlated Errors

Item	*Slope 1*	*Slope 2*	*Slope 3*	*Slope 4*	*Intercept 1*	*Intercept 2*	*Intercept 3*	*Intercept 4*
1	4.54(0.34)	1	0	0	–2.54(0.28)	–5.27(0.41)	–8.13(0.57)	–10.99(0.80)
2	3.43(0.24)	0	0	0	–1.43(0.20)	–3.33(0.26)	–5.86(0.41)	–8.78(0.75)
3	4.39(0.35)	0	1	0	–2.08(0.27)	–4.80(0.38)	–8.19(0.59)	–11.57(0.95)
4	3.97(0.26)	0	0	1	1.62(0.21)	–1.88(0.23)	–5.80(0.42)	–9.92(0.74)
5	4.26(0.34)	1	0	0	–1.28(0.23)	–3.60(0.33)	–7.54(0.59)	–9.82(0.79)
6	3.71(0.26)	0	0	0	0.12(0.18)	–2.65(0.24)	–5.42(0.36)	–8.84(0.59)
7	4.87(0.36)	0	0	1	1.58(0.25)	–2.46(0.29)	–6.93(0.52)	–11.03(0.81)
8	5.61(0.49)	0	1	0	–2.75(0.34)	–6.30(0.57)	–9.75(0.81)	–14.27(1.44)

Notes: Standard errors in parentheses. The latent dimensions are uncorrelated. The variances of the latent dimensions are 1, 1.20 (0.48), 2.11 (0.60), 1.82 (0.47).

Improving the Fit of the Model

The fitted model makes three assumptions: 1) depression as measured by these items is unidimensional, 2) the latent trait representing depression is normally distributed, and 3) the item response functions follow Samejima's graded response model. The fact that the model does not fit perfectly may be due to the violation of any combination of these three assumptions. Assumption 2) may be relaxed by estimating the latent trait nonparametrically (Mislevy, 1984). Assumption 1) may also be relaxed by using a nonparametric IRT model (Chernyshenko et al., 2001; Maydeu-Olivares, 2005) or an alternative parametric IRT model.[6]

In my opinion, in this example unidimensionality is the least-likely culprit, as an inspection of the item stems does not suggest any multidimensionality. However, violations of the unidimensionality assumption are most easily remedied. Within an ordinal factor analysis framework, all that is needed is adding a correlated residual parameter for each pair of items with an outstanding bivariate residual. Unfortunately, existing software for IRT estimation using full information maximum likelihood (e.g., Bock & Aitkin, 1981) do not have currently have capabilities for specifying correlated residual parameters. However, they can be tricked as follows: Specify an additional latent dimension for each correlated residual parameter to be added to the model. This latent dimension is defined to be uncorrelated with all existing "substantive" dimensions and it consists of only two nonzero slope parameters, one for each item in the outstanding residual pair. The slopes are to be fixed to (1,1) if the residual is positive and to (1,–1) if the residual is negative. The variance of the latent dimension is estimated. The estimated variance for this additional dimension is simply a reparameterization of the covariance of the residuals.

I used this method to add the three correlated residual parameters—corresponding to the item pairs (5,1), (7,4), and (8,3)—suggested by the residual correlation analysis. I obtained $M_2 = 697.86$ on 437 degrees of freedom, $p < 0.01$, $\text{RMSEA}_2 = 0.02$. Thus, the model appears to fit better. The estimated parameters are shown in Table 6.9. We see in this table that all three additional parameters added to the model to account for the correlated

6 However, there is evidence that the graded model is the best-fitting model for rating data among existing parametric models with monotonically increasing category trace lines (Maydeu-Olivares, 2005).

residuals are statistically significant. Also, slope estimates are reduced when these correlated errors are introduced in the model. However, the fit improvement obtained by relaxing the unidimensionality assumption is modest. I used the same method to add instead the six correlated residual parameters suggested by the mean and variance corrected X^2 statistics; those corresponding to the item pairs (4,1), (4,3), (5,2), (7,1), (7,5), and (8,4). The model did not converge, suggesting that it is an inappropriate model for these data. I conclude that, as expected, the unidimensionality assumption is the least likely culprit for the misfit of the graded model to these data. If a better fit is desired, an alternative model should be sought.

Summary

Researchers should always assess the overall GOF of their models using a GOF statistic to assess the magnitude of the discrepancy between the data and the model taking into account sampling variability. In IRT applications to patient-reported outcomes, degrees of freedom will generally be so large that no model will fit exactly. In other words, we should always expect to reject the null hypothesis that the fitted model is the data-generating model.

Having rejected the null hypothesis, we need to judge the magnitude of the misfit. If the data are ordinal the standardized residual correlations provide the most convenient way to gauge the effect size of the misfit. In particular, the Standardized Root Mean Square Residual (SRMSR) correlation provides the average effect size of the misfit. Of course, the average misfit may be small but some parts of the model may show a large misfit. Hence, it is necessary to inspect all standardized residual correlations, not just the SRMSR.

The standardized residual correlations are just one way to assess piecewise model fit. One alternative that I have offered here is the use of X^2 statistics applied to pairs of items adjusting them by their mean and variance so that their distribution is asymptotically chi-square. The use of unadjusted X_{ij}^2 is incorrect. The statistic is too liberal and it rejects well-fitting items. An alternative is the $M_2^{(ij)}$ statistic. $M_2^{(ij)}$ works well when the model fits. However, when the model is mis-specified, it may lack power, rejecting too few items. The behavior of the mean and variance-adjusted X_{ij}^2 lies between that of X_{ij}^2 and $M_2^{(ij)}$: it is not as liberal as X_{ij}^2 nor as conservative as $M_2^{(ij)}$.

The SRMSR and standardized residual correlations are probably best used as goodness-of-fit indices with some arbitrary cutoff, say 5 percent, as computing confidence intervals for them is cumbersome. Confidence intervals can be computed for RMSEA statistics based on M_2, or if the model is large and data is ordinal, based on M_{ord}. However, as our example illustrates, for IRT applications to patient-reported outcomes, confidence intervals for RMSEA statistics are generally very narrow. Also, because RMSEA statistics take not only model fit into account but also model parsimony, they will decrease (all other factors held constant) as the number of categories increases. An adjustment can be made to the cutoff for RMSEA_2 to accommodate this fact, but not to RMSEA_{ord}. Hence, when M_{ord} is used I recommend the use of the SRMSR instead of RMSEA_{ord}.

Because the RMSEA statistics take not only model fit into account but also model parsimony, their use is particularly indicated when several alternative models are being considered. However, model selection necessarily involves subjective judgment and it is wise to examine all the implications of the competing models under examination (latent trait estimates, measurement errors, etc.) in addition to goodness of fit statistics such as the RMSEA when selecting among competing models.

Future Directions

IRT modeling involves identifying a plausible process that individuals may have used to respond to items. From this point of view, it is important to assess how well the fitted IRT model reproduces the data at hand. However, IRT models are also fitted to serve some purpose (scoring, linking, etc.) and therefore it is important also to assess how well the model meets these purposes. IRT users should not be unnecessarily obsessed with the goodness of fit of their models. Rather, they need to take the necessary time and effort to evaluate whether their IRT model serves its intended purpose. By routinely reporting the fit of their fitted models, together with an assessment of how well the model serves its intended purpose, we may learn "how bad is this fit for this purpose" and establish reasonable fit criteria. Different fit criteria may be needed for different purposes. A model that shows a substantial degree of misfit may still prove useful for purpose A. But a model with the same degree of misfit may prove useless for purpose B. What degrees of misfit are acceptable for different purposes is what we ought to determine. Further research is needed to link the goodness-of-fit statistics described in this chapter to specific research questions.

References

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

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. (2012). *flexMIRT: A numerical engine for multilevel item factor analysis and test scoring* [Computer software]. Seattle, WA: Vector Psychometric Group.

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.

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

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*, 523–562. doi:10.1207/S15327906MBR3604_03

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

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

Liu, Y., & Maydeu-Olivares, A. (2012). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*, 254–274.

Liu, Y., & Maydeu-Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research*, *49*(4), 354–371. doi:10.1080/00273171.2014.910744

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

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

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

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. doi:10.1080/00273171.2014.911075

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

Maydeu-Olivares, A., & Montaño, R. (2013). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika*, *78*, 116–133.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment*, *18*, 263–283.

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

Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 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. 51–66). New York: Springer Verlag.

Tollenaar, N., & Mooijart, A. (2003). Type I errors and power of the parametric goodness-of-fit test. Full and limited information. *British Journal of Mathematical and Statistical Psychology*, *56*, 271–288.