Published in final edited form as:
Psychol Methods. 2007 March ; 12(1): 58–79. doi:10.1037/1082-989X.12.1.58.

Item Factor Analysis: Current Approaches and Future Directions

R. J. Wirth and
L. L. Thurstone Psychometric Laboratory, Department of Psychology, University of North Carolina at Chapel Hill

Michael C. Edwards
Department of Psychology, The Ohio State University

Abstract

The rationale underlying factor analysis applies to continuous and categorical variables alike; however, the models and estimation methods for continuous (i.e., interval or ratio scale) data are not appropriate for item-level data that are categorical in nature. The authors provide a targeted review and synthesis of the item factor analysis (IFA) estimation literature for ordered-categorical data (e.g., Likert-type response scales) with specific attention paid to the problems of estimating models with many items and many factors. Popular IFA models and estimation methods found in the structural equation modeling and item response theory literatures are presented. Following this presentation, recent developments in the estimation of IFA parameters (e.g., Markov chain Monte Carlo) are discussed. The authors conclude with considerations for future research on IFA, simulated examples, and advice for applied researchers.

Keywords

item factor analysis; parameter estimation; categorical data; confirmatory factor analysis; item response theory

Item-level data within the social and behavioral sciences are often categorical in nature. Dichotomous (e.g., *disagree* vs. *agree*) or polytomous (e.g., *strongly disagree, disagree, neither, agree*, and *strongly agree*) item response formats may fail to maintain the scale and distributional properties assumed by models such as ordinary least squares regression or common linear factor analysis. Traditional regression techniques describe the outcome variable as an optimal linear function of observed predictors. The proper implementation of these techniques requires assumptions such as independent and normally distributed residuals, a continuous conditionally normal outcome, and that the model is correctly specified (i.e., a linear relationship exists between the outcome and predictors). The common linear factor model, which describes the covariances among observed variables as a function of a smaller number of latent factors, makes many of the same assumptions. It is assumed that the unique factors (those that affect only one measured variable) are normally distributed, the outcomes are continuous and conditionally normally distributed, and a linear relationship exists between the observed and latent variables. Although this list of assumptions is not exhaustive, it does represent assumptions that are easily violated with itemlevel ordered-categorical data. Attempting to estimate model parameters, for example,

Correspondence concerning this article should be addressed to R. J. Wirth, L. L. Thurstone Psychometric Laboratory, Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3270. rjwirth@unc.edu.

Supplemental materials: http://dx.doi.org/10.1037/1082-989X.12.1.58.supp

with dichotomous outcomes within the standard confirmatory factor model (as described by Jöreskog, 1969) results in parameter estimates that are biased and impossible to interpret accurately (Di-Stefano, 2002). However, all is not lost; just as logistic and ordinal regression techniques offer appropriate alternatives to linear regression when modeling dichotomous or polytomous (e.g., ordinal, Likert-type scales) outcomes, item factor analysis (IFA) offers an appropriate alternative to the common linear factor model when modeling categorical item responses (Mislevy, 1986).

This article is intended to offer a targeted review of IFA within the structural equation modeling (SEM) and item response theory (IRT) frameworks. Specific attention is paid to the subset of models most relevant for psychological research. In doing so, we consider model parameterizations for ordered-categorical data, outline the analytic relationships between these parameterizations, and offer an introduction to standard, as well as recently developed, estimation measurement models within psychology, a consideration of when IFA models are an appropriate choice for modeling psychological constructs, and an overview of the work that lies ahead. Another fundamental goal of this article is to offer advice and recommendations for the use of these models in substantive research. With the use of simulated data, we offer guidelines for choosing an appropriate parameterization, choosing an appropriate estimation method, and gaining confidence in results.

Item Factor Analysis Models

There are a wide range of IFA models available in both the SEM and IRT literatures including models for items with ordered responses (Samejima, 1969), models for items with unordered responses (Bock, 1972), models that allow for partial credit (Masters, 1982; Muraki, 1992), and models that allow for guessing (Birnbaum, 1968, pp. 404–405). All of these models are considered variants of a general item factor analytic framework (see Thissen & Steinberg, 1986, for a taxonomy of these models). We do not cover all of these models; instead, we focus on IFA models appropriate for full-credit ordered responses with no guessing, that is, for the type of data most often encountered in psychological research. This discussion includes the categorical confirmatory factor analysis (CCFA) model as well as the two-parameter logistic (2PL) and graded response models.

Latent Response Distributions and CCFA

CCFA assumes that ordered-categorical item responses are discrete representations of *continuous latent responses*. That is, it is assumed that individuals possess a latent score, denoted x^*_{ij} for individual *i* on item *j*, that reflects their level on a continuous, normally distributed[1] latent construct. An observed item response, x_{ij}, such as choosing *disagree* for the item “I generally feel sad” is the categorical manifestation of the latent continuous response. Thus, the distribution of categorical responses to a particular item is the manifestation of the x^*_{ij} distribution corresponding to that item.

The proportion of individuals who endorse each categorical response option provides information about the latent response distribution by way of threshold (τ) parameters. Analytically, τ can be defined as

$$x_{ij}=c,\ \text{if}\ \tau_{jc}<x^*_{ij}<\tau_{jc+1}, \tag{1}$$

[1] Although the continuous latent response distribution is generally assumed to be normally distributed, it is important to note that this assumption is required only in the presence of correlated unique factors; otherwise, the assumption is one of statistical convenience (i.e., it reduces complexities associated with estimation; see Pearson, 1900). However, we are unaware of any software that is currently capable of accommodating nonnormal latent response distributions.

where τ_{jc} defines C ordered-categorical responses ($\tau_{j0} = -\infty$, $\tau_{jC} = \infty$, and c equals, e.g., 0, 1, …, $C - 1$) with respect to the continuous latent response distribution of item j.

More specifically, τ parameters denote the point on the continuous latent response scale that separates one manifest discrete response (e.g., a response option on a Likert-type scale) from the next. Assuming a normally distributed latent response distribution, as in Figure 1, if the observed proportion for the *disagree* response of a dichotomous item (i.e., $x_{ij} = 0$) is .1587, τ_{11} equals −1.0 (i.e., the corresponding z score). The value τ_{11} in this case suggests that individuals with x^{*}_{ij}s (i.e., the individual's latent score) less than one standard deviation below the mean of the distribution, $\overline{x}^{*}_{ij}$, will choose the *disagree* option, whereas individuals with $\overline{x}^{*}_{ij}$s greater than −1.0 will endorse the *agree* option.

Suppose individuals responded to two items. There are now two univariate latent response distributions as well as their bivariate latent response distribution. As can be seen in Figure 2, obtaining the threshold values for two items provides information about their joint distribution, which is assumed to follow a bivariate normal distribution. With two dichotomous items, the thresholds denote the points through the bivariate normally distributed latent response distribution that gave rise to the corresponding 2 × 2 table of observed proportions. However, these proportions are also a function of the correlation between the latent response variables, called tetrachoric and polychoric correlations for dichotomous and polytomous items, respectively. In Figure 2, the ellipses represent the top view of the underlying latent response distribution (a three-dimensional representation of this distribution can be found in Figure 3), with the inner ellipses representing higher sections of the distribution. If the correlation between two response variables was zero, the ellipses would appear as circles. As the correlation between two response variables increases, the distribution becomes more narrow, resulting in the ellipses becoming more narrow and a larger proportion of individual responses falling on the diagonal of four quadrants in Figure 2. More specifically, as the correlation increases, the subsequent expected proportions in the lower left and upper right quadrants in the 2 × 2 table of observed responses also increase.

Following the earlier work of Christoffersson (1975) and B. O. Muthén (1978), Olsson (1979) introduced a maximum-likelihood method for finding the correlations between two or more latent response variables using the proportion of responses in the observed response contingency table (for two variables, the contingency table would comprise the four quadrants presented in Figure 2). Although simultaneously estimating all of the thresholds and correlations for a particular set of items is ideal, as the number of items increases there is a corresponding increase in the complexity of the estimation process. Finding the correlation among latent response variables requires integration (e.g., computing the area under the bivariate latent response distribution in Figure 3). As the number of items increases, so does the number of dimensions requiring integration. For example, to obtain simultaneous estimates of a tetrachoric/polychoric correlation matrix with 10 items, 10-dimensional integration is required, as opposed to the two-dimensional integration required in Figure 2. Noting the analytic difficulties with simultaneously estimating thresholds and correlations, Olsson suggested a two-step approach. This approach relies on the univariate estimation of thresholds, as has been discussed above. Then, treating the thresholds as fixed, the correlations are estimated bivariately, just as in Figure 2. The difference between the two-step approach and the simultaneous approach is that in the two-step approach, the thresholds and correlations are estimated separately and each element of the correlation matrix is estimated independently of all other elements.

Although Olsson (1979) suggested that the two-step method provides comparable estimates while reducing the analytic burden of the simultaneous method, this approach has limitations. Estimating each correlation independently no longer provides true maximum-likelihood estimates of the correlation matrix and can result in a matrix that is nonpositive definite (Song & Lee, 2003) and therefore cannot be inverted.

CCFA attempts to model the correlational structure of the latent responses $\mathbf{\Sigma}^*_{xx}$ by way of a linear combination of model parameters. The unconditional CCFA model can be defined as

$$\mathbf{\Sigma}^*_{xx}=\mathbf{\Lambda}^*_x\mathbf{\Phi}^*_{\xi\xi}\mathbf{\Lambda}^{*'}_x+\mathbf{\Theta}^*_{\delta\delta}. \tag{2}$$

In Equation 2, $\mathbf{\Sigma}^*_{xx}$ denotes the population covariance matrix of the *j*-dimensional latent response distribution (i.e., the distribution of the theoretical, continuous, latent responses), $\mathbf{\Lambda}^*_x$ denotes a matrix of factor loadings, $\mathbf{\Phi}^*_{\xi\xi}$ denotes a covariance matrix of normally distributed latent factors, and $\mathbf{\Theta}^*_{\delta\delta}$ denotes a (typically) diagonal matrix of unique variances. This model is similar to the common factor model; however, the CCFA model is slightly more difficult to estimate because of the presence of unobserved variables on both sides of the equation.

Item Response Models

IFA models within the IRT framework were specifically developed for categorical responses. For example, the 2PL model was created for dichotomous data and is commonly expressed as

$$P(x_j=1|\theta)=\frac{1}{1+\exp[-Da_j(\theta-b_j)]}, \tag{3}$$

where x_j is an observed response to item *j*, θ is a latent variable being measured, a_j is a slope parameter describing the strength of the relationship between item *j* and the latent factor, b_j is a severity or difficulty parameter describing how much of the latent construct someone must possess to have a 50% probability of endorsing item *j*, and *D* is a scaling constant. It is important to note that *D* is something of a historical artifact. Early IRT development focused on the normal ogive model (Lord, 1952; Lord & Novick, 1968; Samejima, 1969), similar to the probit regression model. However, likelihood-based estimation procedures have relied more heavily on the logistic approximation to the normal ogive. The scaling constant (typically 1.7) is used to place results from the logistic model on the same scale as the normal ogive model.

Figure 4A is a graphical representation of Equation 3 generally referred to as an *item characteristic curve* or *trace line*. The trace line for a given item represents or traces the probability of endorsing an item as a function of an individual's level on the underlying construct. If this were an item assessing depression, the trace line found in Figure 4A would suggest that an individual who is roughly 1.6 standard deviations below the population average would have a 50% chance of endorsing this item. Because most of the area under the normal curve falls above −1.6 (approximately 95% of it), almost all of the population would have a greater than 50% probability of endorsing this item. This tells us that the item in question indicates a relatively low level of depression, as individuals with less-than-average depression would still have a good chance of endorsing it.

Closely related to the 2PL is the graded response model (Samejima, 1969) for polytomous data. In fact, when examining the two models side by side, it is easy to see that the 2PL is a constrained version of the more general graded response model. More specifically, this model can be defined as

$$P(x_j=c|\theta)=\frac{1}{1+\exp[-a_j(\theta-b_{jc})]}-\frac{1}{1+\exp[-a_j(\theta-b_{jc+1})]}=P^*(c)-P^*(c+1), \tag{4}$$

where all parameters are as previously defined, with one slight modification. Given the now polytomous nature of the data, the graded response model incorporates $C-1$ severity parameters, b_{jc}, representing the boundaries between the C categories. These severity parameters denote the point on the latent variable separating category c from category $c+1$. With more than two categories, the probability of endorsing a particular response option can be estimated by taking the difference between the probability of a response in category c (e.g., *disagree*) or higher and the probability of a response in category $c+1$ (e.g., *neither*) or higher.

Figure 4B contains a sample trace line plot for an item using a five-category response scale. Each response category (0–4) has its own curve. These curves represent the probability of choosing any given category as a function of the level on the underlying construct. At any given level of the latent construct, it is possible to see not only which category is the most likely to be chosen but also the probabilities attached to the endorsement of any of the five categories. As one might expect, the higher the level of the latent construct, the more likely a given individual is to choose a higher category (assuming the categories are coded such that higher categories represent higher levels of the construct of interest).

Analytic Relationship Between CCFA and Item Response Models

The IFA parameterizations presented above, although found within two different modeling frameworks, are closely related. In fact, the analytic relationship between a one-factor CCFA model and the unidimensional 2PL (and graded) IRT model was originally presented (see Lord, 1952; Lord & Novick, 1968) and demonstrated mathematically (Takane & de Leeuw, 1987) many years ago. Transforming the parameter estimates from one framework to another is straightforward. Although these transformations are rarely needed in practice, they do highlight the differences in interpretation of the parameters obtained within the SEM and IRT frameworks.

For example, parameters from the 2PL IRT model can be shown to be equivalent to parameters from the CCFA model such that

$$a_j=\left(\frac{\lambda_j^*}{\sqrt{1-\lambda_j^{*2}}}\right)D \text{ and } b_j=\frac{\tau_j}{\lambda_j^*} \tag{5}$$

and, vice versa,

$$\lambda_j^*=\frac{a_j/D}{\sqrt{1+(a_j/D)^2}} \text{ and } \tau_j=\frac{(a_j/D)b_j}{\sqrt{1+(a_j/D)^2}}, \tag{6}$$

where λ_j^* is the factor loading for item *j* and all other parameters are as previously defined. It is worth noting that the $D = 1.7$ scaling constant comes into play only when the IRT parameters come from a program that uses the logistic form. The most popular IRT software packages use this form of the two-parameter IRT model, so the scaling constant is included in the conversion equations above. If one is dealing with the normal ogive, $D = 1$ and drops out of the equations. Using the normal ogive model as opposed to the logistic model is analogous to using probit regression as opposed to logistic regression. Note that in the case of the graded response model, the only change to Equations 5 and 6 is that all occurrences of τ_j and b_j receive an additional *c* subscript to denote the corresponding threshold.

An interesting relationship can be seen in Equation 5, where a_j is shown to be a factor loading weighted by the square root of its unique factor variance. This highlights the role of unique variability on the relationship between the latent construct and the probability of endorsing an item within the IRT framework. An important point often overlooked in the applied CCFA literature is that as an element of λ_j^* approaches unity (uniqueness goes to zero), the strength of the relationship goes to infinity, thereby suggesting perfect measurement (i.e., reliability equal to one).

Items are almost never perfectly related to the underlying construct. Practical experience with IRT suggests that, when dealing with dichotomous data, slope parameters (in the normal ogive metric) much greater than three should be viewed with skepticism and values greater than four should, in most situations, be considered unreasonable.[2] These values logically translate into skepticism for $\lambda_j^*\text{s} > .95$ and objection to λ_i^* values greater than .97 (when the CCFA model is identified by standardizing the latent construct). Slopes tend to be higher for the graded response model than for the 2PL, but even in this instance, slopes greater than four are unusual. Given the experience with item factor models within the IRT literature and the relationship between these models, greater care should be taken when attempting to interpret parameters close to their statistical or applicable boundary (i.e., values approaching Heywood, 1931, cases).

Obtaining accurate estimates of IFA parameters can be difficult. The examples and recommendations offered in this article are intended to aid in the use of IFA models in psychological research.

The Challenge of Dimensionality

To appropriately address the use of IFA in substantive research, it is important to understand the challenge dimensionality continues to pose when working with IFA models. Parameter estimation for IFA models typically requires integration. As described above, in the context of the CCFA model, the dimensionality of this integration is related to the number of items. In the IRT context, as is discussed in detail below, the dimensionality of the integration is a function of the number of latent factors. Unfortunately, high dimensional integration is extremely difficult and computer intensive. Although the challenge of high dimensional integration has not been insurmountable, it has been a focus of much of the IFA research in the past 2 decades.

Parameter Estimation and the Challenge of Dimensionality

A common misconception is that IFA models within a particular modeling framework are limited to either a few items (in the SEM framework) or a few factors (in the IRT

[2] Items with slopes greater than four are essentially Guttman-type items that are perfectly discriminating. Items not specifically designed with this end in mind are rarely encountered in the social sciences.

framework). Many of these misconceptions may arise from the characteristics of currently available software. For example, most IRT software programs focus exclusively on unidimensional models (i.e., models that have only a single latent factor). However, theoretically, all of the models presented in this article can be extended to any number of factors and any number of items. The difficulty is not with the models; rather, it is with estimating the parameters of the models in question.

Standard Estimation of CCFA Parameters

There are a number of methods available for the estimation of CCFA model parameters. This section introduces three of these methods: weighted least squares for categorical data (WLS$_C$), modified weighted least squares for categorical data (MWLS$_C$), and full-information maximum likelihood (FIML). Each of these methods has its own set of advantages and disadvantages (see Table 1 for a full list of estimators discussed in this and the following sections).

We begin with WLS$_C$, a common method for CCFA parameter estimation. A useful way to understand the issues facing WLS$_C$ is to begin with weighted least squares (WLS) for *continuous* indicators. The WLS fit function can be defined as

$$F_{\mathrm{WLS}}=(\mathbf{s}-\boldsymbol{\sigma})'\mathbf{W}^{-1}(\mathbf{s}-\boldsymbol{\sigma}), \tag{7}$$

where **s** is a vector containing the unique elements of a $p \times p$ sample covariance matrix and **σ** is a vector containing the unique elements of the $p \times p$ model implied covariance matrix. The weight matrix, $\mathbf{W}^{-1}$, is the inverse of a positive definite matrix of order $u \times u$, where $u = p(p + 1)/2$, and is a consistent estimate of the asymptotic covariance matrix (Browne, 1984).

Even with continuous indicators, difficulties arise in finite samples, the greatest of which is the ability to obtain an accurate estimate of the weight matrix. A few examples will demonstrate why this can be such a difficult task. First, note that the weight matrix comprises $u(u + 1)/2$ distinct elements. This means that when $p = 3$, $u = 6$ and the weight matrix comprises 21 unique elements (see Figure 5A). In a slightly more realistic case, say, with 20 items ($p = 20$), $u = 210$, and the weight matrix comprises 22,155 distinct elements. Clearly, the number of distinct elements grows rapidly as the number of indicators increases. Minimally, the sample size (N) must be larger than u to ensure that the matrix can be inverted; however, in practice, N is typically required to be much larger than u (Browne, 1974, 1984) to ensure an accurate estimate of the matrix. Indeed, the quality of the weight matrix (i.e., the accuracy of the estimated weight matrix) can often be questioned even when the sample size is large and the weight matrix is invertable. Methods to increase confidence in the quality of the weight matrix are addressed in a later section.

The fit function in Equation 7 can be rewritten in the presence of categorical data in terms of correlations and defined as

$$F_{\mathrm{WLS}_C}=(\mathbf{r}-\boldsymbol{\rho})'\mathbf{W}^{-1}(\mathbf{r}-\boldsymbol{\rho}), \tag{8}$$

where **ρ** is a vector containing the unique elements of the $p \times p$ model implied correlation matrix (Jöreskog, 1994) and **r** is a vector containing the unique elements of a $p \times p$ sample tetrachoric or polychoric correlation matrix. Other researchers (e.g., B. O. Muthén, du Toit, & Spisic, 1997) have included the observed and model implied threshold values in the **r** and **ρ** vectors, respectively.3 Regardless of whether or not the thresholds are included in the fit function, the estimation and inversion of a suitable positive definite weight matrix are still necessary. As with WLS for continuous indicators, WLS$_C$ requires a sufficiently large sample for estimation of an accurate weight matrix (B. O. Muthén et al., 1997). For psychological applications, this is not a trivial issue as the number of items found on popular scales can be quite large. For example, the Child Behavior Checklist for Ages 6–18 (Achenbach & Edelbrock, 1983) has 118 items (24,650,731 unique elements in the weight matrix), whereas the Revised NEO Personality Inventory (Costa & McCrae, 1992) has 240 items (418,197,660 unique elements in the weight matrix).

Sample size and the ability to obtain stable, accurate estimates are not the only obstacles faced when using WLS$_C$. There is no closed form solution (i.e., a single answer that can be mathematically derived) to the asymptotic covariance matrix of categorical data as implemented in Equation 8. Furthermore, multiple approaches for estimating the weight matrix in the presence of ordered-categorical data are available (Jöreskog, 1990, 1994; Lee, Poon, & Bentler, 1990b, 1995; B. O. Muthén, 1984; B. O. Muthén et al., 1997). Although the various methods should converge asymptotically, researchers should be aware that differences may arise in finite samples (see Oranje, 2003, for a comparison of these methods). One way to overcome the limitation associated with using a full weight matrix is to reduce the analytic burden by using only the diagonal elements of the weight matrix for the estimation of model parameters.

Modified weighted least squares—Modified (or diagonally) WLS estimators for ordered-categorical indicators, MWLS$_C$s, are a variation of the methods presented above and can be defined as

$$F_{\mathrm{MWLS}_C} = (\mathbf{r} - \boldsymbol{\rho})' \mathbf{W}_D^{-1} (\mathbf{r} - \boldsymbol{\rho}), \tag{9}$$

where **r** and **ρ** are defined as before. However, unlike Equation 8, the weight matrix in Equation 9, $\mathbf{W}_D^{-1}$, contains only the diagonal elements of the full weight matrix. This modification greatly reduces the number of nonzero elements and thereby reduces the computational (and sample size) burden. Returning to the example above, when $p = 3$, $u = 6$, and the diagonal weight matrix comprises only 6 unique elements; see Figure 5B. In the slightly more realistic case, when $p = 20$, $u = 210$, and the diagonal weight matrix comprises 210 as opposed to 22,155 distinct elements.

Two notable MWLS$_C$ estimators that use this general strategy are diagonally WLS (Jöreskog & Sörbom, 2001; see also Christoffersson, 1975) and robust WLS (B. O. Muthén et al., 1997), also commonly denoted as WLSm and WLSmv for mean-adjusted and mean/variance-adjusted WLS, respectively. These methods provide accurate estimates of the model parameters given a stable weight matrix.

Because of the reduction in information, estimation using MWLS$_C$ is not statistically efficient. That is, once all the off-diagonal elements of the weight matrix have been removed, it is no longer the optimal weight matrix (B. O. Muthén et al., 1997). This reduction in efficiency leads to biased standard errors and test statistics. One way to correct these inaccuracies is with the use of the Satorra-Bentler scaled chi-square and robust standard errors (Satorra & Bentler, 1994; Yuan & Bentler, 1998; see also Oranje, 2003). These methods adjust the chi-square test statistic and standard errors of the parameters but leave the model degrees of freedom unadjusted. Another method to correct the inaccurate

[3] This difference has little influence on the estimation of CCFA parameterizations discussed here (i.e., the thresholds are saturated and thus do not add any information to the estimation of other model parameters).

test statistic and standard errors similar to the Satorra-Bentler adjustments can be found in B. O. Muthén et al. (1997). This method also corrects the chi-square test statistic and standard errors of the parameters. However, the mean-and-variance-adjusted chi-square presented by Muthén and colleagues also adjusts the model degrees of freedom. Early research exploring these modified methods, as well as their associated chi-square and standard error adjustments, suggests that they perform well in practice (Flora & Curran, 2004).

Full-information maximum likelihood—An alternative to modifying WLS is to use a FIML method (see Jöreskog & Moustaki, 2001; Lee, Poon, & Bentler, 1990a; Neale, Boker, Xie, & Maes, 2002). These methods rely on *p*-dimensional integration over a multivariate distribution. As opposed to solving for the correlational structure of the multivariate latent response distribution, these methods attempt to account for the correlational structure by way of the model parameters (i.e., $\mathbf{\Lambda}^{*}_{x}$, $\mathbf{\Phi}^{*}_{\xi\xi}$, $\mathbf{\Theta}^{*}_{\delta\delta}$). The full-information methods provide accurate standard errors and fit statistics, unlike the aforementioned MWLS$_C$ estimators. However, the problem of dimensionality remains. The multiple integration required with these methods is computationally intensive, thereby limiting their practical application with large models (i.e., many items, many factors; Jöreskog & Moustaki, 2001; Lee et al., 1990a).

Standard Estimation of IRT Parameters

Much like the FIML methods addressed above, methods for the estimation of IFA models within the IRT framework generally rely on raw data. These methods take full advantage of the information in the data and are thus referred to as full-information estimators. However, unlike the FIML methods available within the SEM framework, these methods require integration (or approximations to integration) over the latent factors—not items.

One of the most commonly used estimation methods for IRT parameters is maximum marginal likelihood (MML) with an expectation-maximization (EM) algorithm (MML/EM; Bock & Aitkin, 1981). The MML/EM algorithm is a reformulation of the Bock and Lieberman (1970) MML method that incorporates Dempster, Laird, and Rubin's (1977) EM algorithm. In this section, we detail the historical developments that have given rise to the MML/EM algorithm along with explanations of the key concepts.

One of the earliest approaches developed to estimate IRT parameters directly is joint maximum likelihood (JML). When using JML, one attempts to estimate all the parameters, person and item, simultaneously. In such instances, there are more parameters being estimated than there are observations (one parameter for each individual plus varying numbers of parameters per item, depending on the model). In such cases, some of the important properties of maximum-likelihood estimates may not hold. The most important of these lost properties is consistency. An estimator is consistent if, as sample size approaches infinity, the estimated value approaches the true value. Even if we are concerned only with the item parameters, using JML there is no guarantee that increasing sample size will yield better estimates.

Rather than attempting to estimate all item and person parameters simultaneously, the MML approach integrates over the person-specific parameters (this process is known as *marginalization*) and estimates the item parameters in the marginal distribution. The rationale to this approach is attributable to Neyman and Scott (1948), who made a distinction between *structural* and *incidental* parameters. In the context of IRT, the item parameters are the structural parameters, and the person parameters (θ_i) are the incidental parameters. By assuming that the person parameters are randomly drawn from some distribution (typically assumed to be normal), it is then possible to integrate over that distribution. In essence, the marginalization process removes the person parameters from the likelihood. Then, it is possible to find the item parameters that maximize that likelihood without having to be concerned about the person parameters. In the case of MML, the marginal likelihood is what remains once the person parameters have been removed from the likelihood. It is this likelihood that is maximized to find the item parameter estimates.

A necessary step in MML estimation involves computing the unconditional probability of subject *i* giving a particular response pattern ($\mathbf{x}_i$) as given by

$$P(\mathbf{X}=\mathbf{x}_i)=\int_{-\infty}^{\infty}P(\mathbf{X}=\mathbf{x}_i|\theta)g(\theta)d(\theta), \tag{10}$$

where $g(\theta)$ represents the continuous latent distribution (typically assumed to be a standard normal distribution, but see Woods & Thissen, 2006, for other alternatives). It is common practice for integrals of the form found in Equation 10 to be approximated using numerical integration. In IRT estimation, the numerical integration is generally conducted using Gauss-Hermite quadrature (discussed below), which works well for relatively smooth functions.

Quadrature-based integration, rather than analytically computing integrals to determine the area under the curve, uses a series of rectangles to approximate the area. The area of a rectangle is very easy to compute, and simply by adding up the areas of the different rectangles, it is possible to get a numerical approximation of the area under any given curve. By having more (and smaller) rectangles, it is possible to more closely follow the contours of the curve in question, but at a cost of greater computational burden. Computing the probability in Equation 10 using Gauss-Hermite quadrature to approximate the integral results in

$$P(\mathbf{X}=\mathbf{x}_i)=\sum_{k}^{q}P(\mathbf{X}=\mathbf{x}_i|Q_k)A(Q_k), \tag{11}$$

where the summation occurs over quadrature points, Q_k, and $A(Q_k)$ is the quadrature weight for each point Q_k. Although calculation is straightforward for a one-dimensional problem, as the number of dimensions increases, it becomes more complicated. The number of quadrature points (per dimension) corresponds to the number of rectangles used to approximate the area under the curve. Thus, more quadrature points lead to a better approximation. The Gauss-Hermite quadrature-based integration used by Bock and Aitkin (1981) is not recommended for more than five factors, which is often too few for many applications in the social sciences.

Despite MML's theoretical importance, computational difficulties limit the number of items for which parameters can be estimated using MML to no more than 12. The reformulation of the Bock and Lieberman (1970) approach found in Bock and Aitkin (1981) overcomes the computational difficulty by adopting a strategy now recognized as being equivalent to the EM algorithm. The EM algorithm is an iterative method for finding maximum-likelihood estimates in the presence of incomplete data. The term *incomplete data* is intentionally vague and covers instances from missing data to latent variables. Rubin (1991, p. 242) gave a very general description of the idea underlying the EM algorithm as "fill in missing data, estimate parameters, re-estimate missing data, etcetera until no real change." In the IRT framework, the missing data are the individual latent scores on a factor (θ_i). The expectation step (E-step) of the EM algorithm implemented by Bock and Aitkin involves using provisional estimates of the item parameters to obtain estimates for the expected number of endorsements for item *j* at a particular level of the latent variable and for the expected number of examinees at that level. The maximization step (M-step) involves obtaining new estimates of the item parameters by substituting the E-step estimates in the likelihood equations. This process continues until some convergence criterion is satisfied.

In addition to the many applications of MML/EM to unidimensional IRT models, there have been extensions to multidimensional IRT (MIRT) models. The earliest such extension was an exploratory MIRT model by Bock, Gibbons, and Muraki (1988), which they called a full-information IFA model. Subsequent research by Gibbons and Hedeker (1992) has expanded the scope of MIRT models to include some limited kinds of confirmatory models (e.g., bifactor models).

Unfortunately, the standard implementation of MML/EM is still based on MML and the use of Gauss-Hermite quadrature, which limits the number of latent constructs or factors. This limitation of the IRT-based methods is not unique to the two-parameter and graded response models discussed here, but applies to most of the IFA models found in the IRT literature (Bartholomew & Knott, 1999). This issue is especially salient in psychology, where many of the commonly used inventories measure multiple factors. For instance, the revised version of the Minnesota Multiphasic Personality Inventory (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) consists of over 30 clinical subscales (factors). Measurement scales of this magnitude pose a serious challenge to the methods of estimation that have traditionally been used in psychometrics.

Recent Advances: Meeting the Challenge

The issue of marginalization and integration, as well as at what point in the estimation process they are performed, remains one of the critical distinctions separating IRT and SEM parameter estimation. This issue, most explicitly discussed in Takane and de Leeuw (1987), also sheds light on current dimensional challenges faced in IFA. When marginalization occurs after the latent response distribution (x^*) has been categorized (i.e., is broken up via the thresholds into observed categories), integration is required over the number of factors. This is generally the problem encountered when using the standard estimation approaches for IFA models found in the IRT framework. When marginalization occurs on the multivariate latent response distribution prior to categorization, integration is required over the number of items. This is generally the problem encountered when using the standard estimation approaches for IFA models found in the SEM framework, where the analysis is performed on the estimated correlations among the latent responses.

A common theme in the literature is how best to handle this integration (regardless of where the integration occurs) while obtaining simultaneous estimates of all IFA model parameters. Here, we focus on three recent developments: the underlying bivariate normal (UBN) approach of Jöreskog and Moustaki (2001), Markov chain Monte Carlo (MCMC) estimation, and adaptive numerical integration.

Underlying Bivariate Normal

The UBN approach (Jöreskog & Moustaki, 2001) for item-level factor models was developed as an alternative to the FIML or multivariate normal approach first proposed for the CCFA model by Lee et al. (1990b). Recall that the FIML approach requires integrating over the number of latent response variables (i.e., items). As the number of items increases, the integration becomes exceedingly computer intensive and eventually becomes unfeasible (somewhere around four items; Jöreskog & Moustaki, 2001).

The UBN approach circumvents integration over many dimensions by relying solely on the univariate and bivariate marginal distributions. Similar to other CCFA estimation methods, marginalization with the UBN approach occurs across the number of items and not factors. In contrast to many of the estimation methods readily available for CCFA parameter estimation, the UBN approach simultaneously estimates all of the model parameters (including the thresholds). Moreover, the UBN procedures estimate the parameters without requiring the inversion of a weight matrix.

Traditional methods for estimating the CCFA model (e.g., WLS_C) comprise three steps. First, the thresholds are estimated; then, the correlation matrix is estimated; finally, the model parameters are estimated. In the case of WLS_C, parameter estimation involves the estimation of a large matrix that must then be inverted. In the case of other maximum-likelihood-based methods (e.g., FIML), the weight matrix may have to be reestimated for every iteration. Unlike these other methods, the UBN approach minimizes all univariate and bivariate fit functions. In doing so, the UBN method requires only one- and two-dimensional integration. The integrals required with the UBN method correspond to finding the threshold for a particular item (as shown in Figure 1) and finding the relationship between two items (as shown in Figure 2). Therefore, the integration required for the simultaneous estimation of all model parameters using the UBN approach is no more computationally intensive than the integration required to estimate the polychoric correlation matrix used with the WLS_C and MWLS_C methods.

Early research suggests that the UBN method with ordered- categorical data may be appropriate for many factors and many items (Jöreskog & Moustaki, 2001). Although more work is needed to fully understand the behavior of the UBN approach in finite samples, this method is an exciting development in the estimation of IFA model parameters. Indeed, the ability to simultaneously estimate all parameters in models with many items and many factors would be a significant improvement over current, multiple-step CCFA estimation methods.

Markov Chain Monte Carlo

Another approach that, although relatively new, has been gaining popularity in the social sciences is MCMC estimation. MCMC methods evolved from work conducted at the Los Alamos National Laboratory by Nicholas Metropolis and colleagues (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Metropolis & Ulam, 1949) in the 1940s while working on the Manhattan Project. These methods were generalized by Hastings in the early 1970s (Hastings, 1970). The resulting algorithm is commonly referred to as the Metropolis-Hastings algorithm and is a very general MCMC method. As explained by some of the originators of the method, “the essential feature of the process is that we avoid dealing with multiple integrations or multiplications of the probability matrices, but instead sample single chains of events” (Metropolis & Ulam, 1949, p. 339). Before we provide an overview4 of the specifics of MCMC, it is useful to consider the more basic idea of Monte Carlo integration.

Imagine for a moment that you want to know the area of a circle but that no formula is available to compute it directly. Using the following procedure, it is possible to obtain an estimate of the area of the circle. Start off by creating a square (with known area) that is larger than the circle and then place the circle inside. Next, consider what will happen if you repeatedly place dots at random locations in the square—some of the dots will fall in the circle, and some will fall outside. If you then calculate the proportion of dots that fall inside the circle and multiply that by the total area of the square, you will have a reasonable

[4] For a more detailed introduction to MCMC, refer to Casella and George (1992); Chib and Greenberg (1995); Gilks, Richardson, and Spiegelhalter (1996); and Gill (2002).

estimate of the area of the circle. The more dots you use, the more precise your estimate will be.

Now suppose that instead of wanting to know the area of a circle, you want to know the area of some irregular shape. In many such instances, no equation exists to directly compute the area, but the "dot method" just discussed will still be able to provide an estimate of the area. This is the essential idea of Monte Carlo integration.

In the context of IFA models, we are not dealing with circles and squares but with distributions. There are many variants of the basic Monte Carlo integration strategy (see Liu, 2001). When dealing with distributions, it is often the case that we would like to obtain estimates of quantities such as the mean and the standard deviation. Consider two different ways to obtain information about the expected value of some density function $f(x)$ that represents the distribution and ranges from $-\infty$ to ∞. One way to obtain information about the expected value is to compute it directly using

$$E(X)=\int_{-\infty}^{\infty} xf(x)\, dx. \tag{12}$$

If it is possible to generate independent random draws from $f(x)$, we can also gain information about the expected value using

$$\hat{E}(X)=\frac{1}{n}\sum_{i=1}^{n} x_i, \tag{13}$$

where the summation is over the n draws from $f(x)$. Rather than analytically solving the integral in Equation 12, we obtain an estimate of the same quantity by drawing n samples from the distribution of interest and computing their mean. In essence, Monte Carlo integration replaces what can be very complex analytic integration with very simple computations.

Unfortunately, it is often the case that the distribution in question cannot be sampled from easily. For the IFA models discussed up to this point, the distribution we would like to sample from is intractable, which means we cannot draw samples from it directly. This is where MCMC proves to be invaluable. By constructing a Markov chain with the desired distribution (also called a *target distribution*) as its stationary distribution (i.e., the distribution the chain converges to), one is able to make dependent draws from the distribution of interest.

A Markov chain consists of a series of random variables that are ordered in time. If it is the case that what happens at time $t + 1$ depends only on what happened at time t and no earlier time points, that series is considered a Markov chain. To make the idea more concrete, imagine you are trying to decide what shoes to wear tomorrow. Suppose that you have a rule that you cannot wear the same pair of shoes 2 days in a row. In this case, the shoes you choose to wear tomorrow depend on the shoes you are currently wearing but not on the shoes worn yesterday or at any earlier time. This is the idea of a Markov chain—it is a sequence of events where what happens next depends only on the present set of circumstances and not on any earlier ones.

Once a given Markov chain has converged, samples from that chain will approximate samples from the distribution of interest. *Convergence* refers to when the Markov chain moves from its initial state to its stationary distribution (the distribution from which we would like to draw samples). To further extend the shoe metaphor, imagine that you have recently moved and that all of your possessions are in unlabeled boxes. One of these boxes has all of your shoes in it, but you do not know which box it is. You decide to randomly open boxes and take an object out until the object you find is a shoe. All the objects you find that are not shoes are like what you get from the Markov chain before it converges. They are objects but not the right kind. In fact, just as you would not try to keep one of those objects and use it as a shoe, in MCMC, you discard the draws that occur before convergence (called *burn-in*) as they are not what you are looking for. Once you find the box with shoes (i.e., your search has converged), you do not have to keep searching, and every additional draw you make is what you want. In MCMC, once the Markov chain has converged, then the values it produces behave as if they had come from the desired distribution. Once a suitable number of samples are taken from a converged chain, a point estimate for a given parameter can be taken as the mean (or mode) of the draws, and the standard error can be estimated as the standard deviation of the draws.

The bulk of MCMC research in IFA has focused on the models found in the IRT literature (but see Shi & Lee, 1998), beginning with the work of Albert (1992) on the normal ogive form of the two-parameter model (Equations 5 and 6 with $D = 1$). Much of the IRT-based MCMC work has focused on the normal ogive, rather than the logistic, forms of IRT models as the benefits of using the logistic approximation are not present in the MCMC context. Albert used data-augmented Gibbs sampling (see Geman & Geman, 1984; Tanner & Wong, 1987), a widely used form of MCMC. Albert and Chib (1993) expanded the work of Albert to incorporate polytomous data. A number of studies have compared MCMC estimation with other popular estimation methods (Baker, 1998; Bradlow, Wainer, & Wang, 1999; Kim, 2001; Patz & Junker, 1999; Sahu, 2002; Wollack, Bolt, Cohen, & Lee, 2002) for some of the most commonly found unidimensional IRT models. Although there are differences in the various findings reported in these studies, the typical finding is that MCMC and more standard methods (such as MML/EM) provide estimates of similar quality.

The introduction of MCMC estimation methods for IFA models has been an exciting recent development, but these methods have their own set of complexities. Draws from a Markov chain can only be considered approximate draws from the distribution of interest if the chain has converged, or reached its stationary distribution. Although there is a wide variety of methods available for assessing convergence (see Cowles & Carlin, 1996, for an excellent review), there are as many chains as there are parameters to be estimated, so assessing each chain for convergence can be time consuming. In addition, given the computer-intensive aspect of MCMC estimation, analyses can take substantial amounts of time even on fast computers. Moreover, the available software requires a substantial quantitative knowledge to be used properly.

MCMC deals with the problems encountered using Gauss-Hermite quadrature by avoiding the use of quadrature entirely. Although this is certainly one possible solution, given the additional complexities introduced by MCMC, it is reasonable to explore other solutions that mitigate the difficulties encountered in the estimation of IFA parameters.

Adaptive Numerical Integration

One potential remedy for the shortcomings of traditional Gauss-Hermite quadrature is the use of adaptive numerical integration. As discussed by Meng and Schilling (1996), there are several different ways in which numerical integration can be made adaptive, thereby expanding the number of dimensions for which it is a feasible alternative. The Monte Carlo EM algorithm, which replaces quadrature-based integration with Monte Carlo integration, is one such approach. This approach is appealing in that it remains within the traditional MML/EM framework but uses new developments to deal with the difficulties in numerical integration. This solution has been implemented in the newest versions of TESTFACT (Bock et al., 2002) and Mplus (B. O. Muthén & Muthén, 2006) but has not yet seen widespread use.

A related adaptation of the EM algorithm is the Stochastic EM algorithm (Diebolt & Ip, 1996). The Stochastic EM algorithm replaces the E-step of the traditional EM algorithm with a stochastic step. This stochastic step uses MCMC to fill in the missing data, which include the latent factor scores. Although this particular estimation scheme has not been used for IFA models to date, it appears to be a very promising avenue for future research. The use of MCMC to aid in the integration step of the EM algorithm highlights the fact that there are a number of ways to use MCMC. One such way is to perform Bayesian computations. This is the most popular use of MCMC in the IFA literature, where MCMC is used to produce draws from posterior distributions of interest, which are the focus of Bayesian inference. The second use of MCMC, highlighted by the Stochastic EM algorithm, is in conjunction with likelihood-based methods as an integration aid. It is not necessary to be a Bayesian to benefit from MCMC.

A third version of adaptive numerical integration, called *adaptive quadrature*, has been one of the primary driving forces behind the recent emergence of the generalized linear latent and mixed models (GLLAMM) approach of Rabe-Hesketh, Skrondal, and Pickles (2004). GLLAMM is a very general framework that includes as a special case many of the IFA models considered here. Adaptive quadrature does not use fixed quadrature points, as does Gauss-Hermite quadrature. Instead, part of the numerical integration procedure involves determining the optimal location for each of the quadrature points. This adaptability allows for more efficient use of quadrature points, which allows for fewer points per dimension. Although this method can also be computer intensive, it is no more so than many of the other numerical alternatives currently available. GLLAMM is currently available within STATA (StatCorp, 2003). For an excellent overview and comparison of some of these recent developments in adaptive integration as applied to IFA models, see Schilling and Bock (2005).

Summary

Recent advances in IFA have significantly improved researchers' ability to estimate models with many items and many factors regardless of the modeling framework. Some methods (e.g., UBN) circumvent the challenge of dimensionality by limiting the information used in the estimation. This approach can provide stable parameter estimates with reduced analytic and computational burden. Other methods (e.g., MCMC) avoid quadrature-based integration by relying on a sampling-based estimation strategy. Although all of these approaches allow for the estimation of more complicated IFA models, they are also more difficult to implement in practice than the methods discussed in earlier sections. The difficulty in practical implementation is not surprising given the cutting-edge nature of this technology, but these methods are certainly worth considering when making choices regarding IFA estimation.

The Future of IFA Parameter Estimation

It is impossible to predict the future of estimation; however, MCMC appears to offer a promising area of research for the simultaneous estimation of IFA parameters. These methods were specifically developed to contend with high dimensional integration and thus are well suited for the problems facing item-level factor analysis. Several researchers have already begun applying these methods in the psychological literature with some success (see Béguin & Glas, 2001; Segall, 2002; Shi & Lee, 1998). As demonstrated by the work of Fox and Glas (2001), who incorporated a multilevel structure into an IFA model, another advantage of MCMC is the relative ease with which one can estimate more complex models. This is not meant to imply that MCMC is easy, but rather that there are situations where MCMC will be easier to implement than likelihood-based approaches.

There are still many aspects of MCMC that need to be explored with respect to item factor models. For example, little work has been done to examine which variant of MCMC (e.g., data-augmented Gibbs sampling vs. Metropolis-Hastings within Gibbs sampling) is best suited for the estimation of IFA parameters. Different variants have strengths and weaknesses that may be more or less beneficial depending on the complexities of the model. Work on such issues has begun (e.g., Edwards, 2004), but it is too early to draw definitive conclusions. Although we believe that MCMC holds great promise for the future of IFA parameter estimation, we do not believe that other estimation methods will become obsolete.

In the presence of multiple factors and a small number of items per factor, the WLS-based methods can provide stable solutions. These methods are less computationally intensive than many of the more recently developed methods, and it seems reasonable to continue using WLS-based methods when feasible. However, given that many commonly used methods are based on tetrachoric or polychoric correlations (e.g., WLS_C and MWLS_C), more research is needed to better understand the role of low cell counts, non-centrally located thresholds, and missing data on the estimation of these correlations and their associated asymptotic covariances. Such work could lead to greater flexibility in research design as well as greater confidence in parameter estimates. Aspects of this work have already begun (e.g., see Neale et al., 2002, and Song & Lee, 2003, for tetrachoric or polychoric correlations with missing data), whereas other researchers are deriving new estimation methods that avoid the need for such correlations entirely (Jöreskog & Moustaki, 2001). However, much remains to be done before these methods can be easily applied to the wide variety of data encountered in psychological research. In fact, a logical next step in the advancement of IFA would be a study examining all of the estimation methods addressed in this article.

IFA in Applied Research

The choice of an IFA model and estimation method often depends on the research question of interest. For instance, if a researcher is interested in individual item characteristics or obtaining scores for individual participants, IRT-based IFA may be more practical. The IRT literature offers numerous models with parameters that apply directly to the items and are intended to explain the interaction between people and items. The history of IRT has largely been focused on these item-level properties and has seen extensive research on scale development, scoring, and other aspects of assessment. On the other hand, if the research question focuses on the structural makeup of a scale (e.g., number of factors, cross-loadings, correlated errors, higher order factors), SEM-based IFA may provide a more natural framework for addressing such questions. The SEM literature offers the CCFA model with parameters that are intended to explain the relationship between constructs (or factors) and latent response distributions. The history of SEM has largely been focused on the latent factors and has seen extensive research on multiple factor analysis, higher order factor analysis, and measurement models within larger structural models such as latent growth models. Regardless of the IFA model implemented, there are a number of estimation options available (see Table 1). We apply a subset of these methods to simulated data below. Although each example comprises only a single random sample, the results of these examples highlight the potential pros and cons of the methods.

Examples

Examples 1 and 2—Examples 1 and 2 rely on the same factor model with 10 dichotomous indicators. This model is intended to be relatively simple in terms of producing parameter estimates. The number of items is low (j = 10), the items range from moderate (i.e., λ = .6) to strong (i.e., λ = .8) in their relationship to a single latent factor, and the thresholds vary around the center of the latent response distributions (see Table 2 for the generating parameter values). With a sufficient sample size, say, an N of 300, any of the estimation methods previously addressed should accurately recover the population parameters. These examples are intended to highlight the similarities between the modeling frameworks as well as the advantages and disadvantages of a subset of estimation methods (i.e., WLS$_C$, MWLS$_C$, MML/EM, and MCMC).

There is no single best method of estimation for all data types, sample sizes, or model parameterizations. The first two examples include WLS$_C$ and MML/EM, which are two of the more common methods found in the categorical SEM and IRT literatures, respectively. Two other methods, MWLS$_C$ and MCMC, reflect estimation methods that are increasing in popularity as well as methods that we believe will play significant roles in the future of IFA estimation. These methods are very adept at handling large numbers of factors and items. The parameter estimates obtained using WLS$_C$ and MWLS$_C$ (specifically, robust WLS) were done using Mplus Version 2.14 (L. K. Muthén & Muthén, 2001)5; for MML/EM estimates, MULTILOG version 7.0.3 (Thissen, Chen, & Bock, 2003) was used, whereas MCMC estimates for Examples 1 and 2 were obtained with code written for R (R Development Core Team, 2005). The MCMC code used for Examples 1 and 2 was derived from the work of Albert (1992). The MCMC estimates for Example 3 were obtained using C ++ code developed by Michael C. Edwards (see Edwards, 2005, 2006).6 All MCMC results presented here use a slope–intercept parameterization of the two-parameter normal ogive (2PNO) model that changes how the resulting intercept is converted to a CCFA threshold. To convert the MCMC estimated intercept to a CCFA τ, use

$$\tau_j=\frac{\gamma_j}{\sqrt{1+a_j^2}}, \tag{14}$$

where γ_j is the 2PNO intercept term for item j. The scaling constant (D) is omitted as the model used in the MCMC estimation is already in the normal ogive metric. The conversion for the slope remains unchanged from Equation 6.

Results for Example 1—Example 1 consists of the aforementioned one-factor model estimated with a sample size of 300. The results, found in Table 2, highlight two important points. First, similar parameter estimates have been obtained from the SEM and IRT parameterizations. Even with a small sample and the IRT estimates converted to SEM parameters (see Equations 6 and 14), the MWLS$_C$, MML/EM, and MCMC results are strikingly similar. Second, although WLS$_C$ estimates have been obtained, the estimates are consistently more discrepant than those produced by the other estimation techniques. Recall that WLS$_C$ relies on a large weight matrix during estimation (see Equation 8). This weight matrix contains $u(u + 1)/2$ unique elements. In the current example, p = 10, so u = 55, and the weight matrix has 1,540 unique elements. A sample size of 300 allows for the estimation and inversion of the asymptotic variance/covariance matrix, thus allowing estimates to be obtained. However, it appears that a sample of 300 does not provide enough information for an accurate estimate of the weight matrix, thereby leading to more discrepant estimates of the parameters. Obtaining a solution, regardless of the estimation method, does not ensure an accurate solution (Gagné & Hancock, 2006).

[5] Equivalent estimates are obtained using more recent versions (3 and 4) of Mplus.

[6] Data and code for Mplus, MULTILOG, and R can be found online at the supplemental material Web site identified at the beginning of the article. MCMC code for Example 3 will be made available once the software has been released.

Results for Example 2—The model was reestimated with a sample size of 1,000 for Example 2. Much like Example 1, the MWLS$_C$, MML/EM, and MCMC methods provide very similar results, both to one another and to the generating values (see Table 3). However, unlike Example 1, WLS$_C$ also provides similar results. The necessary sample size is often a function of model complexity (e.g., number of variables, number of parameters, scale reliability, etc.; see Gagné & Hancock, 2006, for a discussion and sample size recommendations based on various measurement characteristics). Simply put—there is no good rule of thumb. However, relying solely on a software's ability to return results could potentially lead to accepting poor estimates. We strongly recommend, when possible, exploring a model with at least two of the estimators discussed in this article. With a single-factor model, estimates can easily be obtained using both SEM and IRT software. Therefore, one may estimate the model using, for example, MWLS$_C$ and MML/EM methods. For multiple-factor models, one may have to rely solely on a single parameterization. In such cases, one may want to estimate a model using WLS$_C$ and MWLS$_C$ methods (or MCMC and UBN methods as they become more readily available). The level of agreement among these different estimation methods can serve as a means for gauging confidence in the parameter estimates.

Example 3—Example 3 examines four correlated factors with 10 items on each factor. The first factor comprises indicators with five categories (representing Likert-type item responses), whereas the other three factors have dichotomous indicators (see Tables 4, 5, 6, and 7 for generating parameter values). Given the number of correlated factors, the SEM parameterization seems most appropriate. This example is intended to highlight that models with many items and many factors require close scrutiny even when using a relatively large sample.

Results for Example 3—Using a sample size of 1,000, parameter estimates obtained using WLS$_C$ are quite different from the MWLS$_C$ and MCMC methods (see Table 4). For example, for the first three items, WLS$_C$ converges to λ values of .69, .79, and .86, whereas MWLS$_C$ converges to λ values of .55, .67, and .78, respectively. MCMC provides estimates (λ_1 = .57, λ_2 = .68, and λ_3 = .80) that are slightly closer to the population values than the MWLS$_C$ estimates. Further exploration of the WLS$_C$ parameter values finds that λ39 has a value of .98. Values this high (i.e., >.97) may suggest an inappropriate solution. Given the differences between the WLS$_C$, MWLS$_C$, and MCMC results, as well as the λ_{39} value obtained with WLS$_C$, the results suggest that the sample size is too small to obtain a stable WLS$_C$ solution. An alternative estimator should be used: Here, those alternatives are MWLS$_C$ and MCMC. However, one may find that other methods, such as UBN or a maximum-likelihood approach with adaptive quadrature, work just as well. A comparison of the MWLS$_C$ and MCMC estimates with the population values found in Tables 4–7 shows that the methods generally obtained estimates within just a few hundredths of a point to the population values. Again, careful examination and comparison of parameter estimates obtained using multiple methods can lead to much greater confidence in a researcher's final conclusions.

Considerations and Recommendations

There are a number of issues researchers should consider when using IFA. For example, the purpose of the research can be used to decide which IFA methods are most appropriate for a particular question or whether IFA is required at all. Moreover, sample sizes are generally required to be larger with categorical data than with continuous data, and additional steps may need to be taken to gain confidence in the model results.

The methods available for IFA are complicated (e.g., nonlinear, highly parameterized, etc.) and make strong demands on the data that often necessitates the collection of large samples to achieve stable, accurate solutions. Careful planning prior to data collection can help to alleviate the difficulties associated with estimating IFA models. Although it is recommended that researchers compare results from a number of the methods outlined in this article, this is hard to accomplish without first collecting the data. A number of steps can be taken prior to data collection to help ensure that the models of interest can be adequately evaluated. The first step is to decide on the need for IFA.

At times, IFA may not be required, and with proper planning, it may be avoided. If the purpose of the research is to explore structural relationships among constructs, the researcher may circumvent the use of IFA by relying on established scales. In this case, it is important to find scales with well-validated scoring algorithms. These algorithms would preferably be based on previous IFA studies and would take individual item characteristics into consideration (i.e., differentially weight the items). In doing so, a researcher may be able to obtain continuous scale scores to be used as indicators, thereby allowing traditional methods to be used for parameter estimation.

Similarly, some research has suggested that if a sufficient number of item response options are used (e.g., five or more) per item, traditional maximum-likelihood methods with Satorra-Bentler or Yuan-Bentler adjustments may be appropriate (DiStefano, 2002; Dolan, 1994). However, it is important to note that relying solely on traditional maximum-likelihood methods with Pearson product–moment correlations ignores the categorical nature of the data and implicitly introduces a misspecification into the series of equations. These methods were developed for continuous data. More recent research has shown that using standard estimation methods with categorical data, even on raw data, and a single-moment adjustment (e.g., Satorra-Bentler or Yuan-Bentler adjustments) can fail to accurately capture the true fit of the model to the data (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006). Thus, when considering use of traditional factor analytic techniques, it remains important to compare parameter and standard error estimates using various estimation methods. This comparison should include estimation methods appropriate for categorical data and should focus not on the best fitting model, but on the triangulation of a stable solution.

If the purpose of a research project is to examine the measurement characteristics of a scale and the number of response options per item is small (e.g., five or fewer), IFA is required. The same holds true when using a scale that has not been well validated in the population of interest. In such cases, it is imperative that the items are carefully scrutinized and studied prior to data collection. Substantive theory should be able to help predict characteristics of the item distributions (and subsequent item parameters). If theory suggests that the item responses will be uniformly distributed among the various response options, sample sizes similar to those found above may be more than sufficient. If, however, substantive theory predicts skewed item responses (e.g., many individuals endorse *agree* and very few endorse *disagree* on a dichotomous item), then the sample size should be increased. Predicting the frequency of individual item responses can be difficult. Researchers such as Krosnick and Fabrigar (in press) have written extensively on this topic, and we encourage researchers to explore this literature prior to data collection.

The number of categories can also influence the model and estimation method used. Much research exploring the robustness of normal theory estimators has been done; see Hoogland and Boomsma (1998) and Boomsma and Hoogland (2001) for reviews of this literature. Standard estimation techniques such as maximum likelihood (Lawley & Maxwell, 1963), asymptotic distribution free (Browne, 1984), or ordinary least squares can, at times, provide accurate parameter estimates for item-level data. The number of response categories used by the participants can often act as a gauge for the likelihood that normal theory estimation techniques will accurately recover the population parameters. A general rule of thumb is to use categorical estimation techniques like those outlined in this article with fewer than five response categories. However, simply maintaining five response categories will not ensure that the normal theory estimators obtain equivalent or superior estimates compared with the methods discussed here. In fact, research has shown that even with five categories, relying on standard estimation methods may result in biased estimates of the parameters and standard errors (DiStefano, 2002; Dolan, 1994). When considering relying on a normal theory estimator, a comparison of the parameter estimates and standard errors with those obtained by one (or more) of the methods presented above can increase confidence in the results. If the number of categories is large and the results mimic those of the more analytically intensive methods, reporting the results from the normal theory method seems appropriate.

Sample size issues take on a different character in IFA models than in models for continuous data. Although we are hesitant to make any definitive recommendations, in our experience, sample sizes less than 200 need to be treated with particular care when conducting IFA. In addition to concerns about the overall sample size, there must be consideration of the number of responses in each observed category. Any further divisions of the data (e.g., for multiple group analyses) can exacerbate these issues. With any of the estimation approaches discussed here, item-level frequency distributions (and graphical representations) play a crucial role in the process of better understanding one's data. In addition, when using any of the procedures that rely on tetrachoric/polychoric correlations, it is important to consider item-by-item contingency tables. Sparseness in these tables can prove problematic when estimating the correlations. Moreover, missing data can significantly reduce a researcher's effective sample size when using tetrachoric/polychoric correlation approaches (e.g., WLS_C or MWLS_C). Many of the statistical software programs default to listwise deletion for the estimation of correlation matrices (see, e.g., Jöreskog & Sörbom, 2001; L. K. Muthén & Muthén, 2001). Thus, in the presence of substantial missing data, researchers may find it beneficial to consider estimation approaches such as FIML, MML/EM, or MCMC.

Having a sufficient sample size is only the first step in estimating IFA parameter values. Researchers will also want to carefully examine the parameter estimates not just for common issues such as Heywood cases but also for parameter estimates approaching their boundaries. Another useful way to gain confidence in the parameter estimates is to estimate the parameters using a number of different methods. Inconsistencies in the results suggest that at least one of the methods is providing poor estimates. Although using various methods of estimation may seem complicated, many software programs offer a choice of estimation methods that can usually be defined with a single command.

Software

At the time this article was written, CCFA parameters can be estimated using WLS_C or MWLS_C estimators in commercially available software packages such as EQS (Bentler, 2005), LISREL (Jöreskog & Sörbom, 2004), and Mplus (B. O. Muthén & Muthén, 2006), as well as in Mx (Neale et al., 2002), a freely available package. Each of these programs estimates tetrachoric/polychoric correlations, and although many of them rely on listwise deletion, Mx offers the estimation of these correlations in the presence of data missing completely at random. Many of these programs also offer FIML estimation methods. Although not specifically addressed in this article, EQS, Mplus, Mx, and the freely available CEFA program (Browne, Cudeck, Tateneni, & Mels, 2004) can estimate exploratory factor analysis models with ordered-categorical data.

There are a number of software programs for estimating the IFA models found in the IRT literature; many of these programs rely on the MML/EM estimator. Currently, unidimensional model parameters for dichotomous data can be estimated using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), MULTILOG (Thissen et al., 2003), PARSCALE (Muraki & Bock, 2003), and the ltm package (Rizo-poulos, 2006) for R. In addition to unidimensional models for dichotomous data, TESTFACT (Bock et al., 2002) can also estimate several kinds of multidimensional models, although the majority are of an exploratory nature. Unidimensional model parameters for polytomous data can be estimated using MULTILOG, PARSCALE, and the ltm package. Some multidimensional 2PL and graded response model parameterizations can currently be obtained using Mplus.

To date, there are very few software programs specifically designed to estimate IFA models using the methods outlined in the recent advances section above. Although the UBN approach is not currently available in any commercial software, it will be soon be available in LISREL. WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003), a freely available MCMC software program, can perform estimation for a subset of IFA models. However, it is important to note that WinBUGS is a general MCMC program and not specifically designed for IFA. Of course, many of the methods discussed in this article can be programmed in general statistical software programs such as R, GAUSS (Aptech Systems, 2003), or MATLAB (MathWorks, 2003).

Concluding Remarks

The issue of how to model measures with many items and many factors has been a motivating question for methodologists involved with categorical measurement model research for some time now. It seems realistic to expect that in the next decade or two continued progress in this area will render estimation of IFA models with many items and many factors commonplace. As these models become better able to adapt to the size and complexity of psychological assessment, they will undoubtedly play a more central role in the study of human behavior.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by National Institute on Drug Abuse Grants F31DA017546 awarded to R. J. Wirth and R01DA015398 awarded to Andrea Hussong. We also thank Li Cai, Robert MacCallum, Darcy McMullin, and Kris Preacher for their comments on this and earlier drafts.

References

Achenbach, T.; Edelbrock, C. Manual for the Child Behavior Checklist and Revised Child Behavior Profile. Burlington, VT: University Associates in Psychiatry; 1983.

Albert JH. Bayesian estimation of normal ogive item response curves using Gibbs sampling. Journal of Educational Statistics. 1992; 17:251–269.

Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association. 1993; 88:669–679.

Aptech Systems. GAUSS systems (Version 7.0). Black Diamond, WA: Author; 2003.

Baker FB. An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. Applied Psychological Measurement. 1998; 22:153–169.

Bartholomew, DJ.; Knott, M. Latent variable models and factor analysis. 2nd ed.. London: Arnold; 1999.

Béguin AA, Glas CAW. MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika. 2001; 66:541–561.

Bentler, P. EQS 6.1. Encino, CA: Multivariate Software; 2005.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In: Lord, FM.; Novick, MR., editors. Statistical theories of mental test scores. 1968. p. 395-479.

Bock RD. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika. 1972; 37:29–51.

Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. Psychometrika. 1981; 46:443–459.

Bock RD, Gibbons R, Muraki E. Full-information item factor analysis. Applied Psychological Measurement. 1988; 12:261–280.

Bock, RD.; Gibbons, R.; Schilling, SG.; Muraki, E.; Wilson, DT.; Wood, R. TESTFACT 4. Chicago: Scientific Software International; 2002.

Bock RD, Lieberman M. Fitting a response model for *n* dichotomously scored items. Psychometrika. 1970; 35:179–197.

Boomsma, A.; Hoogland, JJ. The robustness of LISREL modeling revisited. In: Cudeck, R.; du Toit, SHC.; Sörbom, D., editors. Structural equation modeling: Present and future. Lincolnwood, IL: Scientific Software International; 2001. p. 139-168.

Bradlow ET, Wainer H, Wang X. A Bayesian random effects model for testlets. Psychometrika. 1999; 64:153–168.

Browne MW. Generalized least-squares estimators in the analysis of covariance structures. South African Statistical Journal. 1974; 8:1–24.

Browne MW. Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology. 1984; 37:62–83. [PubMed: 6733054]

Browne MW, Cudeck R, Tateneni K, Mels G. CEFA: Comprehensive exploratory factor analysis, Version 2. 2004 from http://faculty.psy.ohio-state.edu/browne/software.php.

Butcher, JN.; Dahlstrom, WG.; Graham, JR.; Tellegen, A.; Kaemmer, B. Minnesota Multiphasic Personality Inventory (MMPI-2): Manual for administration and scoring. Minneapolis: University of Minnesota Press; 1989.

Cai L, Maydeu-Olivares A, Coffman DL, Thissen D. Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. British Journal of Mathematical and Statistical Psychology. 2006; 59:173–194. [PubMed: 16709285]

Casella G, George EI. Explaining the Gibbs sampler. American Statistician. 1992; 46:167–174.

Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. American Statistician. 1995; 49:327–335.

Christoffersson A. Factor analysis of dichotomized variables. Psychometrika. 1975; 40:5–32.

Costa, PT.; McCrae, RR. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Odessa, FL: Psychological Assessment Resources; 1992.

Cowles MK, Carlin B. Markov chain Monte Carlo convergence diagnostics: A comparative review. Journal of the American Statistical Association. 1996; 91:883–904.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1977; 39:1–38.

Diebolt, J.; Ip, EHS. Stochastic EM: Method and application. In: Gilks, WR.; Richardson, S.; Spiegelhalter, DJ., editors. Markov chain Monte Carlo in practice. London: Chapman & Hall; 1996. p. 259-273.

DiStefano C. The impact of categorization with confirmatory factor analysis. Structural Equation Modeling. 2002; 9:327–346.

Dolan CV. Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. British Journal of Mathematical and Statistical Psychology. 1994; 47:309–326.

Edwards, MC. A Markov chain Monte Carlo approach to item factor analysis; Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology; Naples, FL. 2004 October.

Edwards, MC. Unpublished doctoral dissertation. University of North Carolina at Chapel Hill; 2005. A Markov chain Monte carlo approach to confirmatory item factor analysis.

Edwards, MC. Invited presentation for the dissertation award at the annual meeting of the Psychometric Society. Montreal, Quebec, Canada: 2006 June. A Markov chain Monte Carlo approach to confirmatory item factor analysis.

Flora DB, Curran PJ. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychologival Methods. 2004; 9:466–491.

Fox J, Glas CAW. Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika. 2001; 66:269–286.

Gagné P, Hancock GR. Measurement model quality, sample size, and solution propriety in confirmatory factor models. Multivariate Behavioral Research. 2006; 41:65–83.

Geman S, Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984; 6:721–741.

Gibbons RD, Hedeker DR. Full-information item bi-factor analysis. Psychometrika. 1992; 57:423–436.

Gilks, WR.; Richardson, S.; Spiegelhalter, DJ. Introducing Markov chain Monte Carlo. In: Gilks, WR.; Richardson, S.; Spiegelhalter, DJ., editors. Markov chain Monte Carlo in practice. New York: Chapman & Hall; 1996. p. 1-19.

Gill, J. Bayesian methods: A social and behavioral sciences approach. New York: Chapman & Hall/CRC Press; 2002.

Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970; 57:97–109.

Heywood HB. On finite sequences of real numbers. Proceedings of the Royal Society, Series A. 1931; 134:486–501.

Hoogland JJ, Boomsma A. Robustness studies in covariance structure modeling: An overview and a meta-analysis. Sociological Methods and Research. 1998; 26:329–367.

Jöreskog KG. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika. 1969; 32:183–202.

Jöreskog KG. New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. Quality and Quantity. 1990; 24:387–404.

Jöreskog KG. On the estimation of polychoric correlations and their asymptotic covariance matrix. Psychometrika. 1994; 59:381–389.

Jöreskog KG, Moustaki I. Factor analysis of ordinal variables: A comparison of three approaches. Mulitvariate Behavioral Research. 2001; 36:347–387.

Jöreskog, KG.; Sörbom, D. LISREL user's guide. Chicago: Scientific Software International; 2001.

Jöreskog, KG.; Sörbom, D. LISREL (Version 8.71). Chicago: Scientific Software International; 2004.

Kim SH. An evaluation of a Markov chain Monte Carlo method for the Rasch model. Applied Psychological Measurement. 2001; 25:163–176.

Krosnick, JA.; Fabrigar, LR. The handbook of questionnaire design. New York: Oxford University Press; in press

Lawley, DN.; Maxwell, AE. Factor analysis as a statistical method. London: Butterworth; 1963.

Lee S-Y, Poon W-Y, Bentler PM. Full maximum likelihood analysis of structural equation models with polytomous variables. Statistics and Probability Letters. 1990a; 9:91–97.

Lee S-Y, Poon W-Y, Bentler PM. A three-stage estimation procedure for structural equation models with polytomous variables. Psychometrika. 1990b; 55:45–51.

Lee S-Y, Poon W-Y, Bentler PM. A two-stage estimation of structural equation models with continuous and polytomous variables. British Journal of Mathematical and Statistical Psychology. 1995; 48:339–358. [PubMed: 8527346]

Liu, JS. Monte Carlo strategies in scientific computing. New York: Springer-Verlag; 2001.

Lord, FM. A theory of test scores. New York: Psychometric Society; 1952.

Lord, FM.; Novick, MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.

Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982; 47:149–174.

MathWorks. MATLAB (Version 6.5.1). Natick, MA: Author; 2003.

Meng XL, Schilling S. Fitting full-information item factor models and an empirical investigation of bridge sampling. Journal of the American Statistical Association. 1996; 91:1254–1267.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. Journal of Chemical Physics. 1953; 21:1087–1092.

Metropolis N, Ulam S. The Monte Carlo method. Journal of the American Statistical Association. 1949; 44:335–341. [PubMed: 18139350]

Mislevy RJ. Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics. 1986; 11:3–31.

Muraki E. A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement. 1992; 16:159–176.

Muraki, E.; Bock, RD. PARSCALE 4. Chicago: Scientific Software International; 2003..

Muthén BO. Contributions to factor analysis of dichotomous variables. Psychometrika. 1978; 43:551–560.

Muthén BO. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984; 49:115–132.

Muthén BO, du Toit SHC, Spisic D. Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. 1997 Unpublished manuscript.

Muthén, BO.; Muthén, LK. Mplus: Statistical analysis with latent variables, Version 4.1. Los Angeles: Muthén & Muthén; 2006.

Muthén, LK.; Muthén, BO. Mplus user's guide. 2nd ed.. Los Angeles: Muthén & Muthén; 2001.

Neale, MC.; Boker, SM.; Xie, G.; Maes, HH. Mx: Statistical modeling. Richmond: Virginia Commonwealth University, Department of Psychiatry; 2002.

Neyman J, Scott EL. Consistent estimates based on partially consistent observations. Econometrica. 1948; 16:1–32.

Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika. 1979; 44:443–460.

Oranje, A. Comparison of estimation methods in factor analysis with categorical variables: Applications to NAEP data; Paper presented at the annual meeting of the American Educational Research Association; Chicago, IL. 2003 April.

Patz RJ, Junker BW. A straightforward approach to Markov chain Monte Carlo methods for item response models. Journal of Educational and Behavioral Statistics. 1999; 24:146–178.

Pearson K. Mathematical contributions to the theory of evolution: VII. On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society, Series A. 1900; 195:1–47.

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: Author; 2005.

Rabe-Hesketh S, Skrondal A, Pickles A. Generalized multilevel structural equation modeling. Psychometrika. 2004; 69:167–190.

Rizopoulos D. ltm: Latent trait models under IRT, Version 0.5-1. 2006 from http://cran.r-project.org/src/contrib/Descriptions/ltm.html.

Rubin D. EM and beyond. Psychometrika. 1991; 56:241–254.

Sahu SK. Bayesian estimation and model choice in item response models. Journal of Statistical Computer Simulations. 2002; 72:217–232.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. New York: Psychometric Society; 1969.

Satorra, A.; Bentler, PM. Corrections to test statistics and standard errors in covariance structure analysis. In: von Eye, A.; Clogg, CC., editors. Latent variable analysis: Applications to developmental research. Thousand Oaks, CA: Sage; 1994. p. 399-419.

Schilling S, Bock RD. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. Psychometrika. 2005; 70:533–555.

Segall, DO. Confirmatory item factor analysis using Markov chain Monte Carlo estimation with applications to online calibration in CAT; Paper presented at the annual meeting of the National Council on Measurement in Education; New Orleans, LA. 2002 April.

Shi J-Q, Lee S-Y. Bayesian sampling-based approach for factor analysis models with continuous and polytomous data. British Journal of Mathematical and Statistical Psychology. 1998; 51:233–252.

Song X-Y, Lee S-Y. Full maximum likelihood estimation of polychoric and polyserial correlations with missing data. Multivariate Behavioral Reserach. 2003; 38:57–79.

Spiegelhalter, DJ.; Thomas, A.; Best, NG.; Lunn, D. WinBUGS (Version 1.4). Cambridge, England: University of Cambridge, Institute of Public Health, Medical Research Council Biostatistics Unit; 2003.

StatCorp. STATA Statistical Software: Release 8.0. College Station, TX: Author; 2003.

Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika. 1987; 52:393–408.

Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association. 1987; 82:528–550.

Thissen, D.; Chen, W-H.; Bock, RD. MULTILOG 7. Chicago: Scientific Software International; 2003.

Thissen D, Steinberg L. A taxonomy of item response models. Psychometrika. 1986; 51:567–577.

Wollack JA, Bolt DM, Cohen AS, Lee YS. Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. Applied Psychological Measurement. 2002; 26:339–352.

Woods CM, Thissen D. Item response theory with estimation of the latent population distribution using splinebased densities. Psychometrika. 2006; 71:281–301.

Yuan K, Bentler PM. Normal theory based test statistics in structural equation modeling. British Journal of Mathematical and Statistical Psychology. 1998; 51:289–309. [PubMed: 9854947]

Zimowski, M.; Muraki, E.; Mislevy, R.; Bock, RD. BILOG-MG 3. Chicago: Scientific Software International; 2003.

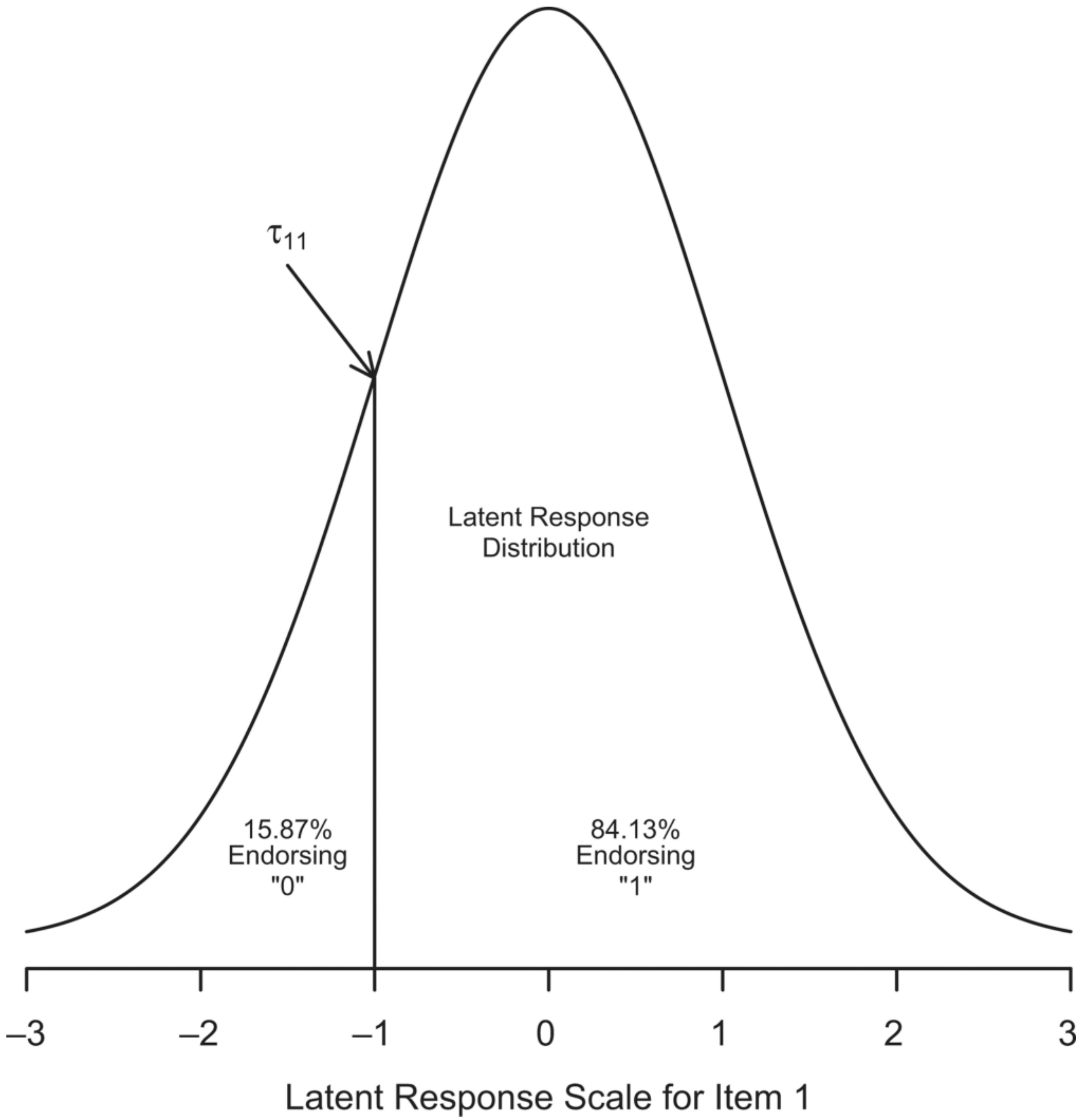

Figure 1.

Latent response distribution for a single dichotomous item representing the latent distribution of interest. τ_{11} marks the latent cut-point between observed responses.

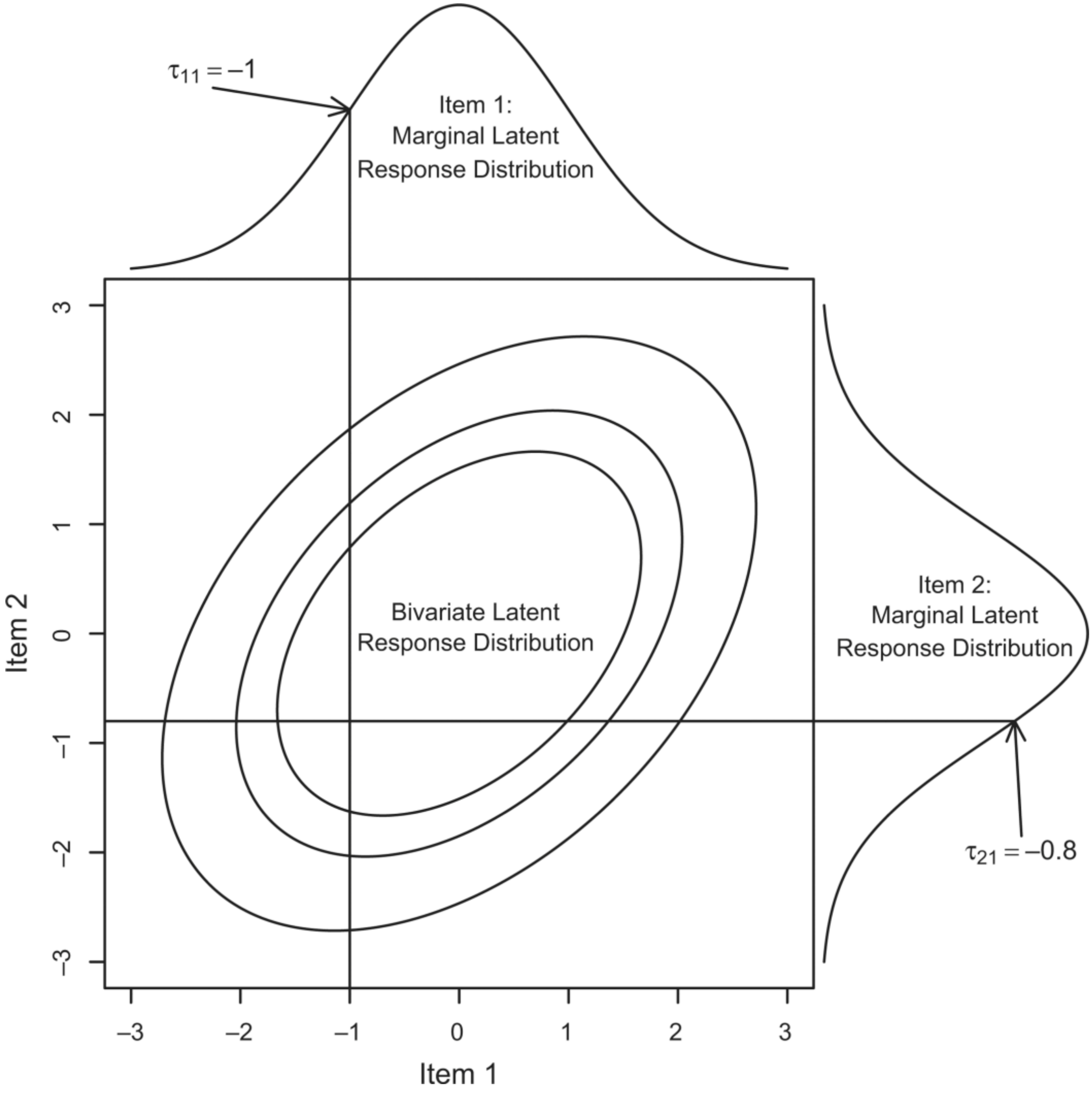

Figure 2.
Bivariate and marginal latent response distributions for two dichotomous items. The bivariate latent response distribution, with a correlation of .70, represents the distribution of interest. The ellipses represent the .01, .05, and .10 regions. The threshold parameters τ_{11} and τ_{21} denote the cut-points for Items 1 and 2, respectively.

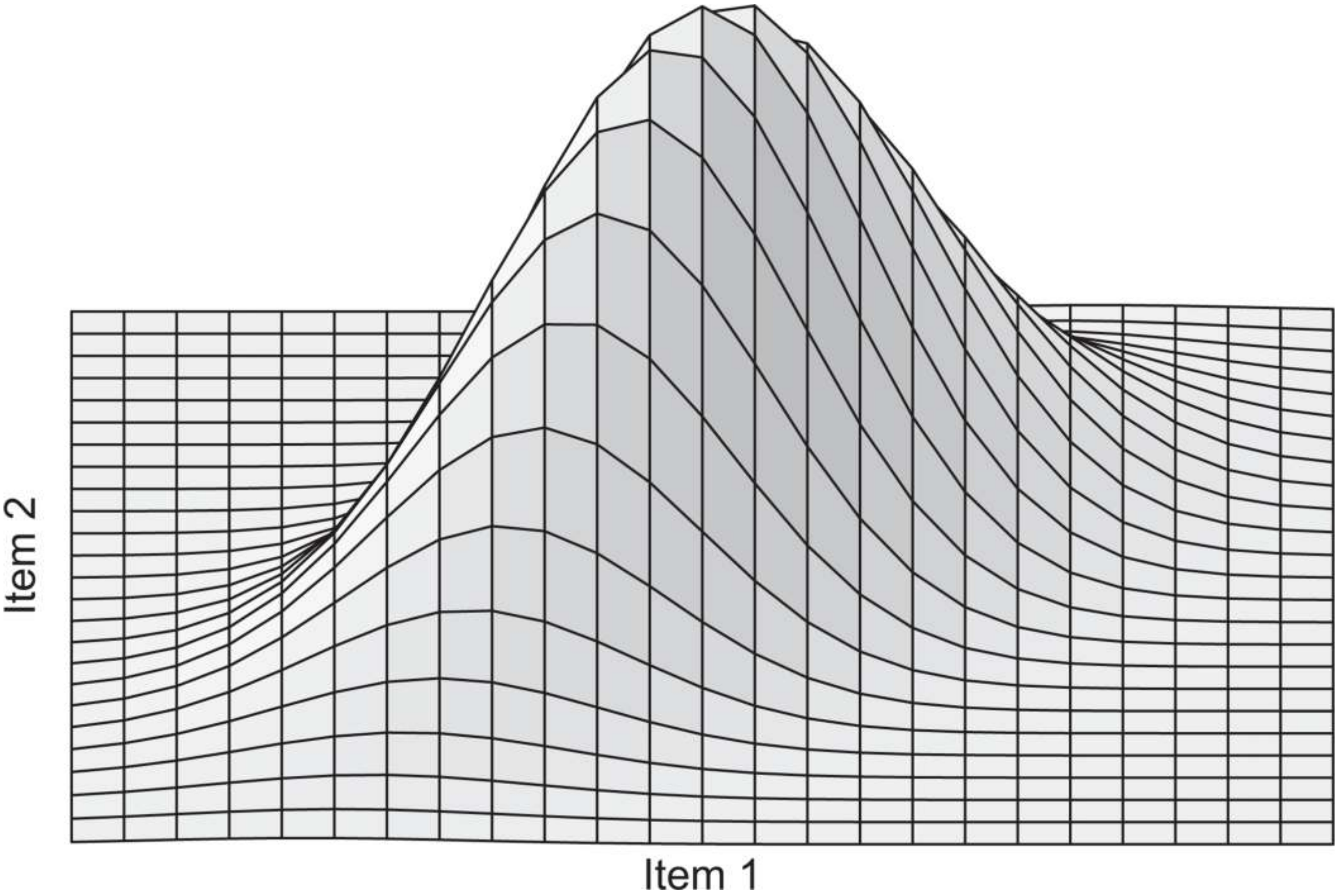

Figure 3.
Three-dimensional bivariate latent response distribution for two items with a correlation of .70.

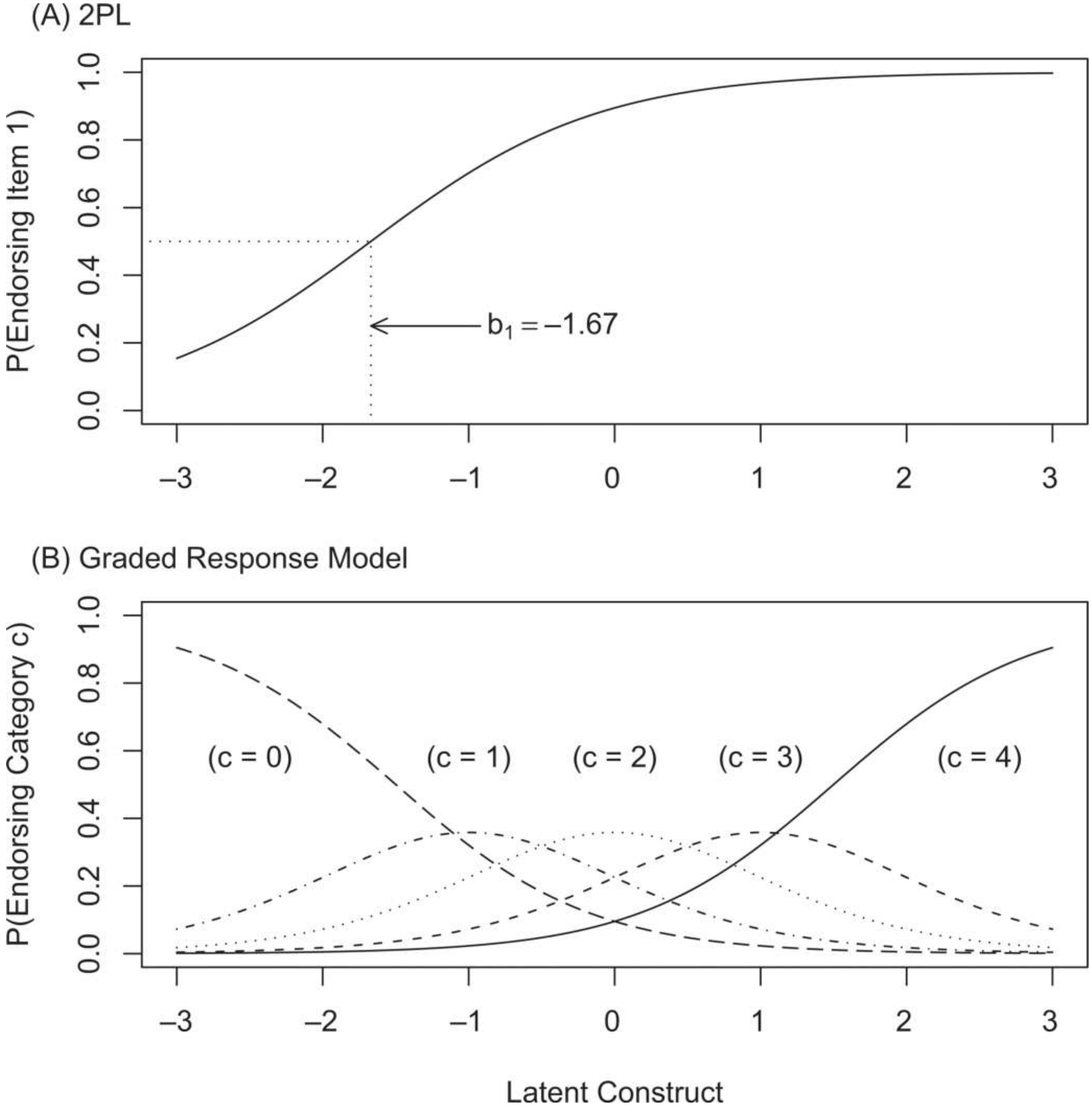

Figure 4.

A: Two-parameter logistic (2PL) model trace line for a single dichotomous item with a difficulty (i.e., *b*) equal to −1.67. The slope of the trace line ($a = 1.28$) describes the strength of the relationship between the underlying latent construct (i.e., θ) and the probability of an individual endorsing the item. B: Graded response model trace lines for an item with $C = 5$ response categories. Each line represents the corresponding probability of endorsing the c^{th} category given θ. Category 0 is denoted with a long-dashed line (– – –), Category 1 is denoted with dash–dot–dash line (– · –), Category 2 is denoted with a dotted line (…), Category 3 is denoted with a short-dashed line (- - -), and Category 4 is denoted with a solid line (—).

(A)

$$\begin{array}{cc} & \begin{array}{cccccc} \sigma_{x1x1} & \sigma_{x1x2} & \sigma_{x1x3} & \sigma_{x2x2} & \sigma_{x2x3} & \sigma_{x3x3} \end{array} \\ \begin{array}{c} \sigma_{x1x1} \\ \sigma_{x1x2} \\ \sigma_{x1x3} \\ \sigma_{x2x2} \\ \sigma_{x2x3} \\ \sigma_{x3x3} \end{array} & \left[\begin{array}{cccccc} \sigma_{x1x1x1x1} & \sigma_{x1x1x1x2} & \sigma_{x1x1x1x3} & \sigma_{x1x1x2x2} & \sigma_{x1x1x2x3} & \sigma_{x1x1x3x3} \\ \sigma_{x1x2x1x1} & \sigma_{x1x2x1x2} & \sigma_{x1x2x1x3} & \sigma_{x1x2x2x2} & \sigma_{x1x2x2x3} & \sigma_{x1x2x3x3} \\ \sigma_{x1x3x1x1} & \sigma_{x1x3x1x2} & \sigma_{x1x3x1x3} & \sigma_{x1x3x2x2} & \sigma_{x1x3x2x3} & \sigma_{x1x3x3x3} \\ \sigma_{x2x2x1x1} & \sigma_{x2x2x1x2} & \sigma_{x2x2x1x3} & \sigma_{x2x2x2x2} & \sigma_{x2x2x2x3} & \sigma_{x2x2x3x3} \\ \sigma_{x2x3x1x1} & \sigma_{x2x3x1x2} & \sigma_{x2x3x1x3} & \sigma_{x2x3x2x2} & \sigma_{x2x3x2x3} & \sigma_{x2x3x3x3} \\ \sigma_{x3x3x1x1} & \sigma_{x3x3x1x2} & \sigma_{x3x3x1x3} & \sigma_{x3x3x2x2} & \sigma_{x3x3x2x3} & \sigma_{x3x3x3x3} \end{array}\right]^{-1} \end{array}$$

(B)

$$\begin{array}{cc} & \begin{array}{cccccc} \rho_{x1x1} & \rho_{x1x2} & \rho_{x1x3} & \rho_{x2x2} & \rho_{x2x3} & \rho_{x3x3} \end{array} \\ \begin{array}{c} \rho_{x1x1} \\ \rho_{x1x2} \\ \rho_{x1x3} \\ \rho_{x2x2} \\ \rho_{x2x3} \\ \rho_{x3x3} \end{array} & \left[\begin{array}{cccccc} \rho_{x1x1x1x1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho_{x1x2x1x2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \rho_{x1x3x1x3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \rho_{x2x2x2x2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \rho_{x2x3x2x3} & 0 \\ 0 & 0 & 0 & 0 & 0 & \rho_{x3x3x3x3} \end{array}\right]^{-1} \end{array}$$

Figure 5.

A: Weighted least squares weight matrix where $p = 3$ and $u = 6$. B: Modified weighted least squares for categorical data weight matrix where $p = 3$ and $u = 6$.

Table 1

Estimators' Feasibility With Many Items or Many Factors, as Well as Some Pros and Cons

		Many			
Estimator	**Abbreviation**	**Items**	**Factors**	**Pro**	**Con**
Weighted least squares[a]	WLS_C		✓	Easily handles multiple factors and is asymptotically efficient	Requires a very large sample
Modified weighted least squares[a]	MWLS_C	✓	✓	Works well with smaller samples	Requires χ^2 and standard error adjustments
Full-information maximum likelihood[b]	FIML	✓	✓	Can handle many items or many factors	Can be computer intensive
Maximum marginal likelihood—EM[a]	MML/EM	✓		Can handle many items and is widely available	As typically implemented, limited to one factor
Underlying bivariate normal	UBN	✓	✓	Allows simultaneous estimation of CCFA parameters	Not widely available
Underlying multivariate normal	UMN		✓	Allows simultaneous estimation of CCFA parameters	Computer intensive and not widely available
Markov chain Monte Carlo[a]	MCMC	✓	✓	Extremely flexible	Computer intensive and not widely available
Monte Carlo EM[c]	MML/MCEM	✓	✓	Can handle many items and many factors	Computer intensive and not widely available
Adaptive quadrature[c]	MML/AQ	✓	✓	Can handle many items and many factors	Potentially computer intensive and not widely available

Note. CCFA = categorical confirmatory factor analysis; EM = expectation maximization.

[a] Denotes estimators used in the examples.

[b] Technically, FIML can be used in the structural equation modeling and item response theory frameworks. However, given categorical data, it is generally limited to either many factors or many items depending on whether one is working in the structural equation modeling or item response theory framework, respectively.

[c] Monte Carlo EM and Adaptive quadrature are both used as part of a maximum marginal likelihood estimation procedure.

Table 2

Item Factor Analysis of a One-Factor Model With 10 Dichotomous Indicators (N = 300) Using WLS$_C$, MWLS$_C$, MML/EM, and MCMC Estimation Methods

			SEM				IRT[a]			
	Population		WLS$_C$		MWLS$_C$[b]		MML/EM[c]		MCMC[d]	
Item	λ	τ	λ	τ	λ	τ	λ*	τ*	λ*	τ*
1	.6	−1.00	.78	−0.99	.69	−1.07	.71	−1.08	.70	−1.06
2	.7	−0.80	.73	−0.85	.62	−0.92	.61	−0.92	.61	−0.91
3	.8	−0.60	.86	−0.51	.84	−0.53	.85	−0.54	.86	−0.53
4	.6	−0.40	.73	−0.39	.58	−0.47	.57	−0.46	.59	−0.46
5	.7	−0.20	.83	−0.24	.73	−0.25	.71	−0.24	.72	−0.23
6	.8	0.00	.86	−0.13	.79	−0.06	.78	−0.06	.80	−0.05
7	.6	0.40	.68	0.41	.64	0.40	.63	0.41	.64	0.40
8	.7	0.60	.78	0.48	.78	0.55	.79	0.56	.79	0.55
9	.8	0.80	.85	0.65	.77	0.73	.78	0.74	.78	0.73
10	.6	1.10	.72	1.17	.65	1.14	.69	1.16	.67	1.14
χ^2			78.75		47.23					
df			35		25					
p value			<.001		.005					

Note. SEM = structural equation modeling; IRT = item response theory; WLS$_C$ = weighted least squares for categorical data; MWLS$_C$ = modified weighted least squares for categorical data; MML/EM = maximum marginal likelihood—expectation maximization; MCMC = Markov chain Monte Carlo.

[a] IRT estimates have been converted to SEM parameters (denoted as λ* and τ*) using Equations 6 and 14.

[b] Robust weighted least squares (WLSmv [mv = mean/variance adjusted] option in Mplus) were used for all MWLS$_C$ estimates. Note that the χ^2 and *df* are estimated (see B. O. Muthén, du Toit, & Spisic, 1997).

[c] An omnibus IRT test statistic, generally denoted G^2, is available for IRT estimates but requires many more people than possible response patterns.

[d] There is currently no well-developed chi-square equivalent statistic for MCMC estimation.

Table 3

Item Factor Analysis of a One-Factor Model With 10 Dichotomous Indicators (N = 1,000) Using WLS$_C$, MWLS$_C$, MML/EM, and MCMC Estimation Methods

			SEM				IRT[a]			
	Population		WLS$_C$		MWLS$_C$[b]		MML/EM[c]		MCMC[d]	
Item	λ	τ	λ	τ	λ	τ	λ	τ*	λ*	τ*
1	.6	−1.00	.66	−0.94	.65	−0.94	.67	−0.95	.65	−0.94
2	.7	−0.80	.74	−0.80	.73	−0.81	.74	−0.81	.73	−0.81
3	.8	−0.60	.82	−0.61	.81	−0.62	.82	−0.62	.82	−0.62
4	.6	−0.40	.55	−0.42	.53	−0.41	.53	−0.40	.54	−0.41
5	.7	−0.20	.77	−0.20	.76	−0.20	.76	−0.20	.76	−0.20
6	.8	0.00	.84	−0.01	.83	−0.00	.83	−0.00	.83	−0.01
7	.6	0.40	.62	0.37	.60	0.38	.59	0.38	.60	0.37
8	.7	0.60	.73	0.57	.72	0.58	.71	0.58	.71	0.57
9	.8	0.80	.80	0.72	.80	0.73	.81	0.74	.79	0.72
10	.6	1.10	.67	1.07	.66	1.09	.69	1.10	.65	1.08
χ^2			45.59		33.91					
df			35		32					
p value			.109		.376					

Note. SEM = structural equation modeling; IRT = item response theory; WLS$_C$ = weighted least squares for categorical data; MWLS$_C$ modified weighted least squares for categorical data; MML/EM = maximum marginal likelihood—expectation maximization; MCMC = Markov chain Monte Carlo.

[a] IRT estimates have been converted to SEM parameters (denoted as λ* and τ*) using Equations 6 and 14.

[b] Robust weighted least squares (WLSmv [mv = mean/variance adjusted] option in Mplus) were used for all MWLS$_C$ estimates. Note that the χ^2 and *df* are estimated (see B. O. Muthén, du Toit, & Spisic, 1997).

[c] An omnibus IRT test statistic, generally denoted G^2, is available for IRT estimates but requires many more people than possible response patterns.

[d] There is currently no well-developed chi-square equivalent statistic for MCMC estimation.

Table 4

Item Factor Analysis of a Four-Factor Model (N = 1,000) Using WLS$_C$, MWLS$_C$, and MCMC Methods

						SEM					SEM									
	Population					WLS$_C$					MWLS$_C$^{b}					IRT^a: MCMC^c				
Item	λ	τ_1	τ_2	τ_3	τ_4	λ	τ_1	τ_2	τ_3	τ_4	λ	τ_1	τ_2	τ_3	τ_4	λ*	τ_1^*	τ_2^*	τ_3^*	τ_4^*
1	.6	−1.0	−0.5	0.0	0.5	.69	−1.11	−0.54	0.00	0.39	.55	−0.99	−0.50	0.03	0.48	.57	−0.99	−0.50	0.03	0.48
2	.7	−1.0	−0.5	0.0	0.5	.79	−0.88	−0.42	−0.01	0.33	.67	−1.03	−0.48	0.01	0.47	.68	−1.02	−0.48	0.00	0.46
3	.8	−1.0	−0.5	0.0	0.5	.86	−1.04	−0.54	−0.05	0.38	.78	−1.03	−0.49	−0.03	0.52	.80	−1.02	−0.48	−0.02	0.51
4	.6	−1.0	−0.5	0.0	0.5	.74	−0.98	−0.45	−0.02	0.35	.61	−0.97	−0.46	0.05	0.52	.62	−0.96	−0.45	0.05	0.52
5	.7	−1.0	−0.5	0.0	0.5	.77	−0.90	−0.45	0.03	0.57	.69	−0.95	−0.46	0.03	0.53	.69	−0.94	−0.46	0.02	0.53
6	.8	−0.5	0.0	0.5	1.0	.85	−0.59	−0.02	0.39	0.93	.80	−0.50	−0.01	0.52	1.02	.80	−0.50	0.00	0.52	1.01
7	.6	−0.5	0.0	0.5	1.0	.71	−0.48	−0.05	0.46	0.80	.61	−0.46	0.03	0.56	0.98	.60	−0.46	0.03	0.56	0.98
8	.7	−0.5	0.0	0.5	1.0	.74	−0.44	0.09	0.44	0.85	.69	−0.49	0.04	0.52	1.05	.70	−0.48	0.03	0.52	1.05
9	.8	−0.5	0.0	0.5	1.0	.89	−0.44	0.03	0.41	0.81	.83	−0.45	0.01	0.53	1.02	.84	−0.45	0.00	0.52	1.01
10	.6	−0.5	0.0	0.5	1.0	.72	−0.41	0.02	0.60	1.06	.63	−0.48	0.04	0.56	1.10	.62	−0.47	0.04	0.56	1.09
11	.7	−1.0				.80	−1.05				.67	−1.03				.68	−1.02			
12	.8	−0.8				.78	−0.87				.82	−0.85				.83	−0.85			
13	.6	−0.6				.68	−0.58				.55	−0.56				.55	−0.55			
14	.7	−0.4				.79	−0.45				.71	−0.46				.72	−0.46			
15	.8	−0.2				.85	−0.19				.82	−0.22				.79	−0.21			
16	.6	0.0				.71	0.00				.55	0.03				.58	0.03			
17	.7	0.4				.85	0.37				.73	0.40				.74	0.40			
18	.8	0.6				.86	0.45				.80	0.58				.78	0.58			
19	.6	0.8				.67	0.78				.69	0.81				.67	0.81			
20	.7	1.1				.72	0.97				.63	1.05				.66	1.04			
21	.8	−1.1				.81	−1.09				.80	−1.14				.82	−1.14			
22	.6	−0.8				.55	−0.73				.58	−0.85				.60	−0.85			
23	.7	−0.6				.80	−0.64				.66	−0.66				.68	−0.66			
24	.8	−0.4				.89	−0.36				.82	−0.39				.81	−0.39			

						SEM					SEM									
	Population					WLS$_C$					MWLS$_C$[b]					IRT[a]: MCMC[c]				
Item	λ	τ_1	τ_2	τ_3	τ_4	λ	τ_1	τ_2	τ_3	τ_4	λ	τ_1	τ_2	τ_3	τ_4	λ*	τ^*_1	τ^*_2	τ^*_3	τ^*_4
25	.6	0.0				.70	−0.08				.66	−0.10				.68	−0.02			
26	.7	0.2				.85	0.08				.77	0.19				.73	0.19			
27	.8	0.4				.88	0.34				.79	0.35				.79	0.34			
28	.6	0.6				.71	0.56				.58	0.63				.61	0.63			
29	.7	0.8				.87	0.68				.73	0.78				.73	0.77			
30	.8	1.0				.91	0.79				.81	0.95				.82	0.94			
31	.6	−1.0				.63	−1.14				.56	−1.00				.56	−1.00			
32	.7	−0.8				.73	−0.78				.71	−0.87				.72	−0.86			
33	.8	−0.6				.88	−0.62				.83	−0.61				.84	−0.60			
34	.6	−0.4				.67	−0.47				.62	−0.41				.61	−0.41			
35	.7	−0.2				.75	−0.30				.66	−0.17				.68	−0.17			
36	.8	0.0				.83	−0.16				.76	−0.05				.79	−0.06			
37	.6	0.4				.81	0.32				.61	0.40				.61	0.40			
38	.7	0.6				.77	0.37				.68	0.62				.70	0.61			
39	.8	0.8				.98	0.53				.88	0.85				.87	0.83			
40	.6	1.1				.72	0.87				.64	1.03				.61	1.02			
χ^2								3,249.19					294.49							
df								734					287							
p value								<.001					.369							

Note. SEM = structural equation modeling; IRT = item response theory; WLS$_C$ = weighted least squares for categorical data; MWLS$_C$ = modified weighted least squares for categorical data; MCMC = Markov chain Monte Carlo.

[a] IRT estimates have been converted to SEM parameters (denoted as λ* and τ*) using Equations 6 and 14.

[b] Robust weighted least squares (WLSmv [mean/variance adjusted] option in Mplus) were used for all MWLS$_C$ estimates. Note that the χ^2 and *df* are estimated (see B. O. Muthén, du Toit, & Spisic, 1997).

[c] There is currently no well-developed chi-square equivalent statistic for MCMC estimation.

Table 5

Correlation Matrix of Latent Factors From Example 3 (N = 1,000) Using the WLS_C Estimation Method

Factor	1	2	3	4
1	1.00	0.30	0.59	0.65
2	0.30	1.00	0.35	0.36
3	0.40	0.30	1.00	0.66
4	0.50	0.40	0.50	1.00

Note. Population (lower triangle) and WLS_C (upper triangle). WLS_C = weighted least squares for categorical data.

Table 6

Correlation Matrix of Latent Factors From Example 3 (N = 1,000) Using the $\mathrm{MWLS_C}$ Estimation Method

Factor	1	2	3	4
1	1.00	0.27	0.42	0.49
2	0.30	1.00	0.28	0.39
3	0.40	0.30	1.00	0.50
4	0.50	0.40	0.50	1.00

Note. Population (lower triangle) and MWLS_C (upper triangle). MWLS_C = modified weighted least squares for categorical data.

Table 7

Correlation Matrix of Latent Factors From Example 3 (N = 1,000) Using the MCMC Estimation Method

Factor	1	2	3	4
1	1.00	0.28	0.41	0.49
2	0.30	1.00	0.28	0.38
3	0.40	0.30	1.00	0.50
4	0.50	0.40	0.50	1.00

Note. Population (lower triangle) and MCMC (upper triangle). MCMC = Markov chain Monte Carlo.